Sizes, conformational fluctuations, and SAXS profiles for Intrinsically Disordered Proteins

Mauro L. Mugnai[1,2], Debayan Chakraborty[1,3], Abhinaw Kumar[1], Hung T. Nguyen[1], Wade Zeno[4], Jeanne C. Stachowiak[5], John E. Straub[6], and D. Thirumalai[1,7]

[1]Department of Chemistry, The University of Texas at Austin, 105 E 24th Street, Stop A5300, Austin, TX 78712, USA.
[2]Present Address: Institute of Soft Matter Synthesis and Metrology, Georgetown University, Washington DC 20057, USA.
[3]Present Address: The Institute of Mathematical Sciences, Chennai, India
[4]Mork Family Department of Chemical Engineering and Materials Science, University of Southern California, 3650 McClintock Ave, Los Angeles, CA 90089, USA.
[5]Department of Biomedical Engineering, The University of Texas at Austin, 105 E 24th Street, Stop A5300, Austin, TX 78712, USA.
[6]Department of Chemistry, Boston University, MA 02215, USA.
[7]Department of Physics, The University of Texas at Austin, 2515 Speedway, Austin, TX 78712, USA.
Email: maurolm83@gmail.com, dave.thirumalai@gmail.com

Abstract

The preponderance of Intrinsically Disordered Proteins (IDPs) in the eukaryotic proteome, and their ability to interact with each other, folded proteins, RNA, and DNA for functional purposes, have made it important to quantitatively characterize their biophysical properties. Toward this end, we developed the transferable Self-Organized Polymer (SOP-IDP) model to calculate the properties of several IDPs. The values of the radius of gyration (R_g) obtained from SOP-IDP simulations are in excellent agreement (correlation coefficient of 0.96) with those estimated from SAXS experiments. For AP180 and Epsin, the predicted values of the hydrodynamic radii (R_hs) are in quantitative agreement with those from Fluorescence Correlation Spectroscopy (FCS) experiments. Strikingly, the calculated SAXS spectra for thirty-six IDPs are also nearly superimposable on the experimental profiles. The dependence of R_g and the mean end-to-end distance (R_{ee}) on chain length, N, follows Flory's scaling law, $R_\alpha \approx a_\alpha N^{0.588}$ ($\alpha = g$, and e), suggesting that globally IDPs behave as synthetic polymers in a good solvent. The values of a_g, and a_e are 0.20 nm and 0.48 nm respectively. Surprisingly, finite size corrections to scaling, expected on theoretical grounds, are negligible for R_g and R_{ee}. In contrast, only by accounting for the finite sizes of the IDPs, the dependence of experimentally measurable R_h on N can be quantitatively explained using $\nu = 0.588$. Although Flory scaling law captures the estimates for R_g, R_{ee}, and R_h accurately, the spread of the simulated data around the theoretical curve is suggestive of of sequence-specific features that emerge through a fine-grained analysis of the conformational ensembles using hierarchical clustering. Typically, the ensemble of conformations partitiones into three distinct clusters, having different equilibrium populations and structural properties. Without any further readjustments to the parameters of the SOP-IDP model, we also obtained excellent agreement with paramagnetic relaxation enhancement (PRE) measurements for α-synuclein. The transferable SOP-IDP model sets the stage for several applications, including the study of phase separation in IDPs and interactions with nucleic acids.

1 Introduction

There are several compelling reasons to quantitatively characterize the biophysical properties of Intrinsically Disordered Proteins (IDPs), which unlike globular proteins, do not have a unique three-dimensional fold, but like homopolymers explore an ensemble of conformations. (1) Despite variations in the precise estimate of the number [1], it is suspected that about half of the human proteins have either a large fraction ($> 30\%$) of disordered residues, or feature long (> 30 or > 40 residues, depending on the location along the chain) disordered regions. [2] (2) IDPs play a central role in a variety of cellular functions because they are not only capable of self-association, but also interacting with multiple partners, including nucleic acids. [3] (3) They organize themselves to form condensates, which are suspected to offer advantages in cellular functions [4–7] such as transcription [8], and enzyme reactions [9, 10]. (4) Although insights into the formation of these biomolecular condensates, in which a high-density droplet coexists with a low-density dispersed phase, may be obtained by borrowing concepts from polymer physics,[11–13] sequence effects must and do play an important role. For these and other reasons, there is great interest in understanding the physico-chemical properties of IDPs using experiments, simulations, and theories that faithfully take the exact sequence into account.[14, 15]

Despite considerable progress in computer simulations of IDPs [16–24], the need to develop transferable models, with as few parameters as possible, exists. The accuracies of computational models can only be assessed by direct and unbiased comparisons with different experiments. Small-Angle X-ray Scattering (SAXS) profiles, available for a number of IDPs, could be used as the basis for quantifying the accuracy of computational models. There are potentially a large number of energy functions that one could create to describe the shapes and conformational fluctuations of IDPs. Ideally, these models should be transferable, so that reasonably accurate predictions may be obtained for a given sequence and specified external conditions without having to tweak the parameters repeatedly. All-atom simulations, have had some success in describing the sequence-dependent properties of IDPs,[25–28] but their level of accuracy is often dictated by the particular choice of the force-field/water model,[29–33] or even the sampling protocol.[34, 35] In this regard, coarse-grained simulations have emerged as a viable alternative for predicting sequence-dependent properties on large scale. Several coarse-grained models, with different resolutions, have been proposed [19, 22, 36–42]. Recently, we introduced the Self-Organized Polymer (SOP) model for IDPs [43] (abbreviated as SOP-IDP) in which three adjustable parameters in the energy function (see below) were calibrated to reproduce the experimentally determined radii of gyration (R_gs) for three IDPs. We then showed that the calculated scattering curves for 24 other IDPs were in good agreement with the experimental SAXS profiles ($I(q)$s with q being scattering wave vector) without any further fine-tuning of the parameters. However, for some IDPs (Histatin-5 for example), the calculated $I(q)$ at $q > 1.7\,\mathrm{nm}^{-1}$ ($qR_g \approx 2.29$) deviated from experiment (see Fig. S1A in the Supporting Information). Although the calculated and experimental R_gs were in good agreement (Fig. S1B in the Supporting Information) with each other, in some instances the predicted R_gs were smaller than the experimental values by about 10%. The initial formulation [43] did not include side-chain for alanine, whose inclusion is important, at least for some IDPs. When the alanine side-chain was added to our model, without altering the functional form of the SOP-IDP energy function, we observed significant deviations from experimentally measured $I(q)$ for α-synuclein. Therefore, we found it necessary to recalibrate the parameters in the SOP-IDP energy function to reduce the deviations from experiments.

Here, we employ a robust optimization procedure to recalibrate the three parameters in the SOP-IDP model [43] using the experimentally determined R_gs for five test proteins. Second,

using the resulting energy function, we establish using simulations that the $I(q)$ profiles for 36 IDP sequences are in *quantitative* agreement with SAXS measurements. Third, we show that on average, IDPs behave like random coils, obeying the famous Flory's scaling law for polymers in good solvent, $R_g \sim N^{\nu}$ with $\nu \approx 0.588$. Strikingly, the scaling corrections due to the finite sizes of the IDPs are small for R_g, and the mean end-to-end distance, R_{ee}. For the hydrodynamic radius, R_h, the apparent Flory exponent obtained from simulations is below the value expected for self-avoiding polymers. However, we resolve the incongruity by including finite-size corrections for R_h. Sequence effects manifest themselves in the spread of the data around the expected, theoretical curve. Their importance is clearly highlighted if the fine structural details are probed using a hierarchical clustering scheme. In particular, the fractions of compact, semi-compact, and extended conformations vary depending on the sequence and external conditions. Finally, the proposed SOP-IDP model also reproduces the residue-specific paramagnetic relaxation enhancement (PRE) for α-synuclein nearly quantitatively. As testable predictions, we report the $I(q)$s of AP180 and FUS.

2 Computational Methodology

Energy Function: Following an earlier study [44], and subsequent reports [43, 45–47], each amino acid is represented by two beads. One bead represents the C_α atom, and the other is located at the center of mass of the side-chain (SC) of the amino acid residues. The only exception is glycine, which is modeled using only one bead located at the backbone. The SOP-IDP interaction potential is,

$$U_{SOP-IDP} = U_{FENE} + U_R + U_{SC-SC} + U_{C_\alpha - C_\alpha} + U_{C_\alpha - SC} + U_{ELE}. \tag{1}$$

The Finitely Extensible Nonlinear Elastic (FENE) potential (U_{FENE}) accounts for chain connectivity of the backbone C_α atoms and the chemical bonds between the C_α and the SC beads. The U_R term is a repulsive potential that prevents unphysical overlap between the beads, acting only between beads that are separated by another bead (1-3 interaction). The terms $U_{C_\alpha - C_\alpha}$, $U_{C_\alpha - SC}$, U_{SC-SC} and U_{ELE} describe the interaction between beads that are separated by at least two other units along the chain (e.g. 1-4, 1-5 ... interactions). The Debye-Hückel term, U_{ELE}, accounts for electrostatic interactions between the charged residues. The backbone-backbone (BB), backbone-side-chain (BS), and side-chain-side-chain (SS) interactions are modeled using Lennard-Jones (LJ) potentials,

$$U_{C_a - C_a} = \sum_{i=1}^{N-3} \sum_{j \geq i+3}^{N} \epsilon_{BB} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right], \tag{2}$$

$$U_{C_a - SC} = \sum_{i=1}^{N} \sum_{j' \geq i+2 \text{ or } j' \leq i-2} \epsilon_{BS} \left[\left(\frac{\sigma_{ij'}}{r_{ij'}} \right)^{12} - 2 \left(\frac{\sigma_{ij'}}{r_{ij'}} \right)^{6} \right], \tag{3}$$

$$U_{SC-SC} = \sum_{i'=1}^{N-3} \sum_{j' \geq i'+1}^{N} \epsilon_{SS} \left| \epsilon_{i'j'} - 0.7 \right| \left[\left(\frac{\sigma_{i'j'}}{r_{i'j'}} \right)^{12} - 2 \left(\frac{\sigma_{i'j'}}{r_{i'j'}} \right)^{6} \right], \tag{4}$$

where N is the number of amino acids, indices i and j refer to residue numbers, where unprimed symbols (i and j) indicate backbone beads or alpha carbon atoms. The primed counterpart (i' and j') indicate SC beads of the corresponding residue, r_{ij} is the distance between beads i and j, σ_{ij} is the sum of the radii of two interacting beads and corresponds to the minimum of

the LJ potential. The parameters $\epsilon_{i'j'}$ are taken from the statistical potential [48]. There are only three adjustable parameters in the SOP-IDP energy function. They are ϵ_{BB}, ϵ_{BS}, and ϵ_{SS}, which specify the interaction strengths involving the backbone and SCs. We calculate them by minimizing the differences in the calculated and experimental radii of gyration (R_g) for a test set of proteins. We then use the optimized values to predict the SAXS profiles for several IDPs, which are not in the traing set.

Optimization procedure: The values of ϵ_{BB}, ϵ_{BS}, and ϵ_{SS} are calculated using a minor modification of a parameter optimization (POP) procedure designed to use experimental results to refine force field parameters [49, 50]. Briefly, consider an observable $\hat{O}_n(\Gamma)$, which is a function of the conformational space, Γ, and assume for simplicity that the observable does not depend on the energy of the system. The average of this quantity is,

$$\langle \hat{O}_n \rangle = \frac{\int d\Gamma \hat{O}_n(\Gamma) e^{-\beta E(\Gamma;\vec{p})}}{\int d\Gamma e^{-\beta E(\Gamma;\vec{p})}}, \tag{5}$$

where $\vec{p}$ is a set of parameters in the SOP-IDP energy function, $E(\Gamma;\vec{p})$ is the energy for the conformation Γ, and $\beta^{-1} = k_B T$.

Let $O_{\text{exp,n}}$ be the experimental value of the observable $\hat{O}_n(\Gamma)$. We compute R_g (which in this case is $O_{\text{exp,n}}$) for the training set containing five proteins: Histatin-5 (N=24), ACTR (N=71), Sic-1(N=90), K19 (N=94), and Osteopontin (N=273). The error in estimating a set of n_{obs} quantities is calculated using χ^2:

$$\chi^2 = \sum_{n=1}^{n_{\text{obs}}} \frac{[\langle \hat{O}_n \rangle - O_{\text{exp,n}}]^2}{u_n O_{\text{exp,n}}}, \tag{6}$$

where u_n's are constants that ensure that each element in the sum is dimensionless. Because we use R_g as a target, $u_n = 1$ nm. The idea is to find a set of parameters $\vec{p}^*$ which minimizes χ^2. We used Newton's method [51] to obtain the smallest value of χ^2. This calculation requires the first and second derivatives of the observable with respect to the parameters of the energy function for every step of the optimization procedure. Following the approach outlined previously [49, 50], the derivatives are computed analytically. Their values are obtained by averaging them over a sample of conformations obtained for a given parameter set. After a step, a new set of conformations is sampled with the new parameters, and the procedure is repeated until a satisfying agreement with experiments is achieved. We implemented this procedure with a slight modification. Instead of resampling the conformations after every step (which formally is the correct procedure), we re-weighted the sampled conformations using the new parameters, and took a new step without performing any new simulation. This procedure improves the performance of the algorithm. However, re-weighting becomes inaccurate if the new parameters change significantly from the original values. In order to avoid this scenario, we implemented a confining function that limits the maximum change for each parameter during the re-weighting-based optimization. After a few steps, we re-run the simulations with the new parameter set, making sure that there is only a small difference between the observables obtained via re-weighting the configurations sampled with the previous parameters, and those obtained after sampling with the refined parameters. We repeated this procedure a few times, until convergence was attained. We then used the optimal parameters for simulating the other IDP sequences. The list of IDPs is given in Tables S1 and S2 in the Supporting Information.

Simulations: We carried out Langevin dynamics simulations with the SOP-IDP force field using a customized version of the OpenMM code.[52] The equations of motion were integrated using the leap-frog algorithm, using a time-step, $h = 0.05\tau$. Here, $\tau = a_0\sqrt{\frac{m_0}{e_0}}$ is the natural

unit of time, with $m_0 = 1$ Da, $a_0 = 1$ Å being the length-scale, and $e_0 = 1$ kcal/mol being the energy scale. To enhance the sampling efficiency, the simulations were carried out in the underdamped limit, where the viscosity of water was reduced by a factor of 100.[53] For each IDP, simulations were carried out for 2×10^9 steps. The first 3×10^8 steps were considered as equilibration time, and statistics were collected every 10,000 steps from the rest of the trajectory.

SAXS profiles: We calculated the SAXS profiles using all the conformations generated in the simulations using the Debye formula,

$$I(q) = \sum_{i=1}^{N_{tot}} \sum_{j=1}^{N_{tot}} f_i(q) f_j(q) \frac{\sin(q r_{ij})}{q r_{ij}}. \tag{7}$$

In Eq. 7, N_{tot} is the total number of interaction sites, and includes both the C_α and SC beads. The q-dependent form factors, $f_i(q)$, for the backbone and side-chains are taken from Table S3 of ref [54]. Because Glycine is not explicitly represented as a SC in our model, we calculated the experimental form factor using $f_{Gly}(q) = f_{BB}(q) + f_{Gly-SC}(q)$, which operationally means that the distance from another residue to the Glycine BB is equal to the distance from that residue to the Glycine SC. This is not an unreasonable assumption, since the distance between BB and SC within a Glycine residue is ≈ 1.1 Å (length of a C-H bond for methane, see for instance ref. [55]). It should be noted that not all SAXS experiments report the values of I(0). Moreover, there are considerable uncertainties in extracting $I(0)$ from experiments. Thus, we adjusted $I(0)$ to obtain the best agreement between simulation and experiment for $I(q)$.

Radius of Gyration: The radius of gyration, R_g, was calculated using:

$$R_g = \sqrt{\frac{1}{N_{tot}} \langle \sum_{i=1}^{N_{tot}} (r_i - r_{CM})^2 \rangle}, \tag{8}$$

where r_i is the position of bead i, and r_{CM} is the center-of-mass coordinate, and $\langle ... \rangle$ is the ensemble average.

Hydrodynamic Radius: We calculated the hydrodynamic radius, R_h, using the standard polymer-physics formula,

$$\frac{1}{R_h} = \frac{1}{N_{tot}^2} \left\langle \sum_{i \neq j} \frac{1}{r_{ij}} \right\rangle. \tag{9}$$

Contact maps: Residues i and j are assumed to be in contact if the distance between their corresponding side-chains is less than or equal to 0.8 nm. Probability contact maps were generated from the time-dependent inter-residue matrices using the following logistic function,

$$p_{ij}(t) = \frac{1}{1 + exp\left[\beta \left(r_{ij}(t) - r_0\right)\right]}. \tag{10}$$

In the above equation, $\beta = 500\,\mathrm{nm}^{-1}$, and $r_0 = 0.8\,\mathrm{nm}$.The ensemble-averaged probability maps, $\langle p_{ij} \rangle$ are computed by taking an average over all the snapshots N_s in trajectories N_j.

Hierarchical clustering of conformational ensembles: In previous studies [43, 46], we showed that IDP conformational ensembles are highly heterogeneous. Therefore, relying on ensemble-averages alone to ascertain sequence-specific behavior could be misleading, and it does not give a complete picture of the conformational space accessible to IDPs. The representative conformations populating the IDP ensembles were determined using a hierarchical clustering

procedure, based on a pairwise distance metric, $D_{ij} = 1 - \chi_{ij}$, where χ_{ij}, the structural overlap between conformations i and j, is given by,

$$\chi_{ij} = \frac{1}{N_{pairs}} \sum_{a,b}^{N_{pairs}} H(d - |r^i_{a,b} - r^j_{a,b}|). \tag{11}$$

In the above equation, $r^i_{a,b}$ and $r^j_{a,b}$ are the distances between the coarse-grained sites a and b in snapshots i and j, respectively, and H is the Heaviside step function. The tolerance d, which accounts for thermal fluctuations, is taken to be 0.2 nm. In Eq. 11, the sum is performed over all the sites that are separated by more than two covalent bonds, and N_{pairs} is the number of such pairs. The distinct conformational clusters were identified using the Ward variance minimization [56] criterion, as available within the *scipy* module [57].

The contact map for a given cluster, k is visualized in terms of the difference map, $\Delta p^k_{ij} = p^k_{ij} - \langle p_{ij} \rangle$, where p^k_{ij} is the probability map for cluster k, and $\langle p_{ij} \rangle$ is the ensemble-averaged probability map. By definition, Δp^k_{ij} is bound between +1 and –1. Here, we restrict the probability contacts to positive values of Δp^k_{ij} in order to assess how the contact between residues i and j are enhanced as compared to the ensemble-average.

Paramagnetic relaxation enhancement: NMR paramagnetic relaxation enhancement (PRE) measurements provide local structural information of the ensemble, which is a consequence of the $\langle r^{-6} \rangle$ dependence of the PRE between the paramagnetic center and the nucleus of interest, which are usually the backbone amide protons [58]. In the experiments, a paramagnetic probe (such as a nitroxide spin-label) is attached to a mutated cysteine side chain of the protein. The PRE is usually reported as the intensity ratio between two ^{15}N-^{1}H heteronuclear single quantum coherence (HSQC) spectra in the presence and absence of the probe. Using the conformations generated in the simulations, we calculated the PRE signal from the probe (p) to the amino acid (i) as follows [32]:

$$\Gamma_2^{p,i} = K \left\langle r_{p,i}^{-6} \right\rangle \left[4\tau_c + \frac{3\tau_c}{1 + (\omega\tau_c)^2} \right], \tag{12}$$

where $K = 1.23 \times 10^{-32}$ cm^{6}s^{-2} is a constant for the nitroxide radical probe, τ_c is the correlation time of the protein reorientation (taken to be 2 ns, which is within the range estimated using atomistic simulations [27]), and $\omega/2\pi$ is the Larmor frequency (700 MHz). Because $(\omega\tau_c)^2 \approx 0.02$ is small, it follows from Eq. 12 that $\Gamma_2^{p,i}$ is roughly $\propto \tau_c$, thus making it sensitive to the precise value of τ_c.

Due to the coarse-grained nature of the SOP-IDP model, we approximated $r_{p,i}$ as the distance between the side-chain bead (where the probe attached) to the virtual site that lies between two i^{th} and $(i-1)^{\text{th}}$ C$_\alpha$ (to mimic the location of the ith amide group) (see Fig. 6 of main text). The calculated PRE is converted to the intensity ratio using,

$$\frac{I_{para}}{I_{dia}} = \frac{R_2 \exp(-\Gamma_2^{p,i} t_d)}{R_2 + \Gamma_2^{p,i}}, \tag{13}$$

in order to compare directly with experiments. In the above equation, t_d is the total INEPT time of the HSQC experiment (10 ms), and R_2 is the intrinsic transverse relaxation rate estimated from the reduced spectra (4 s^{-1}).

3 Results and Discussion

Optimal values of interaction energy scales for side-chains and backbone: The values of ϵ_{BB}, ϵ_{BS}, and ϵ_{SS} (in unit of kcal·mol^{-1}) in Eqs 2-4 after optimizations are:

$$\epsilon_{BB} = 0.104, \quad \epsilon_{SS} = 0.180, \quad \epsilon_{BS} = 0.137. \tag{14}$$

The corresponding values derived previously [43] are (in unit of kcal·mol^{-1}) $\epsilon_{BB} = 0.12$, $\epsilon_{SS} = 0.18$, and $\epsilon_{BS} = 0.24$. The largest change is in the value of ϵ_{BS}. Using the new set of parameters greatly improves the agreement of the Histatin-5 SAXS profiles (Fig. S1A, SI Appendix) while maintaining the overall correlation between the calculated and experimental values for the radius of gyration (see below and Fig. S1B).

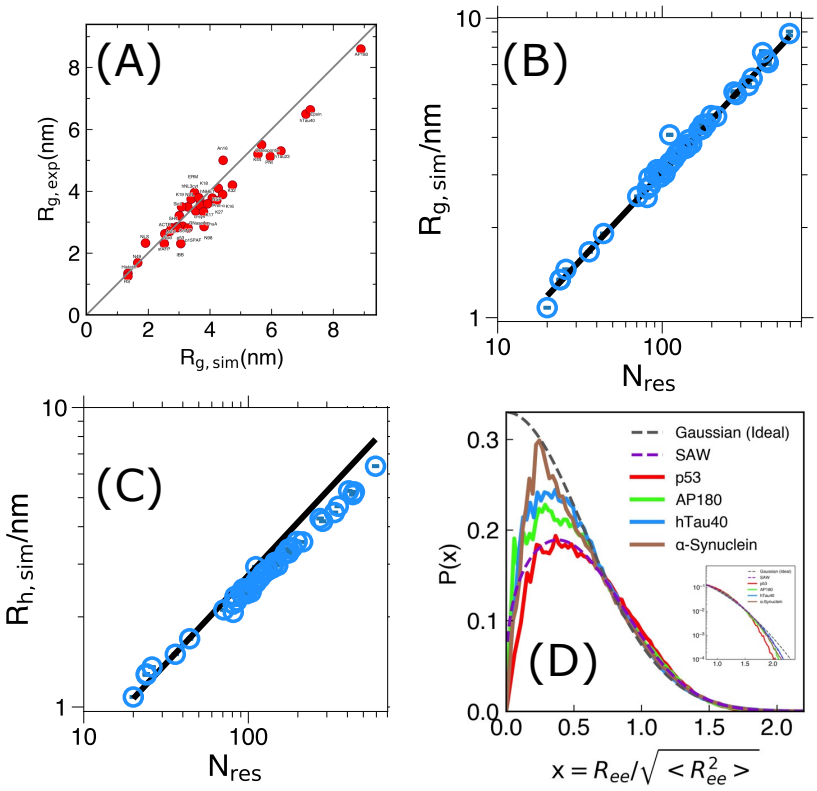

Figure 1: **Polymeric features of IDPs:** (A) Correlation plot between the radius of gyration calculated from SOP-IDP simulations ($R_{g,sim}$) and SAXS experiments ($R_{g,exp}$) for 36 IDP sequences with different lengths and sequence composition. The data are tabulated in Tables S1 and S2 (Supporting Information). The Pearson correlation coefficient is 0.96. (B) Scaling of $R_{g,sim}$ with the number of residues, N as $R_{g,sim} = a_g N^{\nu}$, with ν fixed at 0.588, and $a_g = (0.20 \pm 0.01)$ nm. (C) The calculated hydrodynamic radius, R_h compared to Flory's scaling law with $\nu = 0.588$ with $a_h = (0.18 \pm 0.01)$ nm. (D) The distributions of the scaled end-end distances for the four IDPs closely resemble the universal theoretical predictions for a self-avoiding random walk (SAW) (Eq.15).

IDPs behave as polymers in a good solvent: We first assess the efficacy of the SOP-IDP energy function by comparing R_gs with experiments for 36 IDP sequences (see Table S1 in the Supporting Information for a list). The calculated and experimental R_g values are in excellent agreement (Fig. 1A), with the Pearson correlation coefficient being 0.96 (Fig. S1B

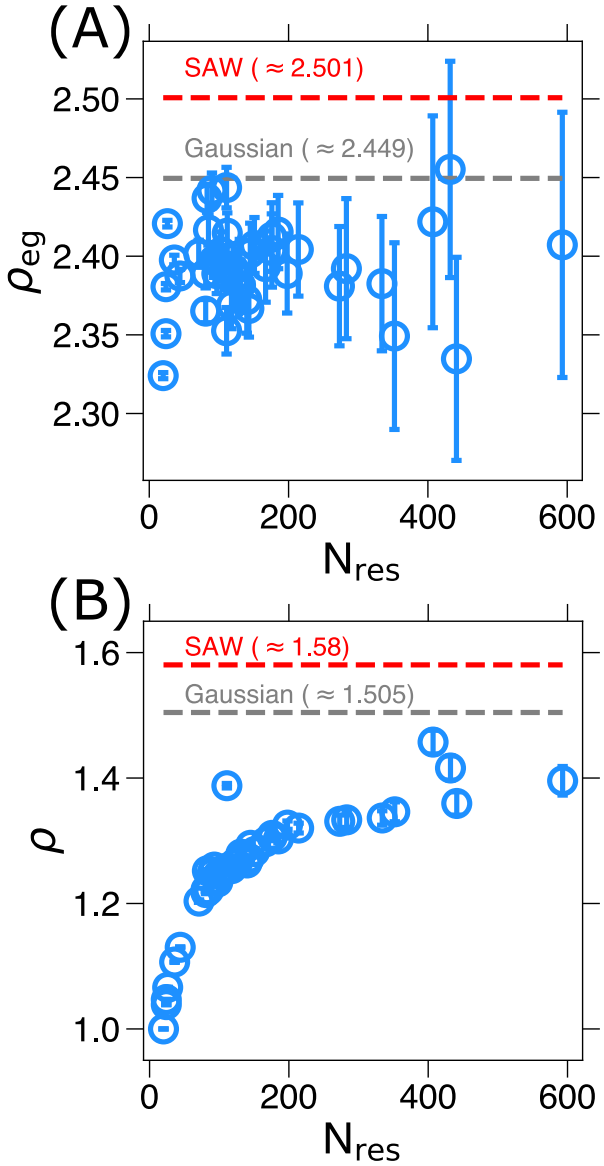

Figure 2: **Assessing the universality of ρ_{eg} and ρ:** (A) Dependence of $\rho_{eg} = R_{ee}/R_g$ on N calculated using simulations. The values for a Gaussian chain can be computed analytically [$R_g = R_{ee}/\sqrt{6}$ to the leading order, see ref. [59]], so $\rho_{eg} = \sqrt{6}$). For a SAW the result comes from lattice simulations [60]. (B) Same as (A) except $\rho = \sqrt{R_g}/R_h$ is plotted as a function of N. For a Gaussian chain, analytical calculations [$R_h = 3/8(\pi/6)^{1/2}R_{ee}$ to the leading order, see ref. [61]) give $\rho = 8/(3\sqrt{\pi})$, whereas we compare with simulations on a lattice for the SAW [62].

in the Supporting Information). Based on the results described in Fig. 1B, and Fig. S1e, we surmise that the IDPs behave as homopolymers in a good solvent, that are usually modeled as Self Avoiding Walks (SAWs). (1) The slope of the plot of $\ln(R_g)$ versus $\ln(N)$ (N is the number of residues in a given IDP) is linear with a slope that is consistent with $\nu = 0.588$ (Fig. 1B), which is the result obtained for SAWs. Thus, IDPs mimic polymers in a good solvent. To corroborate this observation, we fit to a line R_g/N^ν for ν in the range $0.45 - 0.62$. The correct scaling exponent would correspond to the slope that is closest to zero. As shown in Fig S1C, this analysis suggests that the best value of ν is in the range $0.60 - 0.62$. Note that this result is *independent* of the pre-factor, and thus probes only the scaling behavior without the risk of compensation between pre-factor and exponent. Next, by fixing ν, we fit the pre-factor a_g for $R_g \sim a_g N^\nu$ and obtained $a_g \approx (0.20\pm0.01)\,\mathrm{nm}$. Note that a_g is non-universal, although the extracted value for IDP is close to that found for globular proteins in the unfolded state [63], generated at high denaturant concentrations. (2) In contrast, the same analysis conducted on the hydrodynamic radius, which is measurable using FCS experiments, suggests a scaling $\nu < 0.5$ (Fig. S1D). Indeed Fig. 1C shows that $a_h N^\nu$ systematically underestimates (overestimates) IDPs with large (small) N. This suggests that finite-size scaling could be significant for the hydrodynamic radius. Accordingly, $\rho = \frac{R_g}{R_h}$ does not match exactly the values expected for a polymer in good solvent (Fig. 2B). (3) The dependence of $R_{ee} \equiv \sqrt{\langle R_{ee}^2 \rangle}$ on N is also well described by the Flory scaling, with $R_{ee} \approx a_e N^\nu$, where the extracted parameter is $a_e \approx (0.48\pm0.03)\,\mathrm{nm}$ (see Fig S1e). (4) The calculated R_h values for AP180 ($N = 593$) and Epsin (N=432) agree well with the values obtained using FCS experiments (Table S2, Supporting Information).

End-to-end distributions exhibit a universal behavior: If IDPs belong to the SAW universality class, nominally used to characterize homopolymers in a good solvents, then the distribution $P(x)$ ($x = \frac{R_{ee}}{\langle R_{ee}^2 \rangle^{0.5}}$) should have a universal behavior [12, 13, 64]. The expression for $P(x)$ is,

$$P(x) = Cx^g exp(-\alpha x^\delta), \tag{15}$$

where C and α are constants, which are determined using the conditions $\int_0^\infty dx 4\pi x^2 P(x) = \int_0^\infty dx x^2 4\pi x^2 P(x) = 1$. [13] The value of g in Eq. 15, which is zero for a Gaussian chain, is the correlation hole exponent, and $\delta = \frac{1}{(1-\nu)}$. For the four IDPs in Fig. 1D the shapes of the distribution functions $P(x)$s resembles the SAW prediction, which is striking given the finite size of the proteins and sequence variations.

Ratios involving R_g, R_h, and R_{ee}: For a finite sized SAW, it is well established that there are corrections to scaling for $\langle R_g^2 \rangle$, $\langle R_{ee}^2 \rangle$. Keeping only the leading order correction in N, the scaling correction for R_g may be written as (see for example [62]),

$$R_g^2 \approx D_g^2 N^{2\nu}(1 + b_g N^{-\Delta_1}). \tag{16}$$

Similarly, $R_{ee}^2 \approx D_e^2 N^{2\nu}(1 + b_e N^{-\Delta_1})$. The simplest scaling correction to R_h should have a similar form albeit with an extra term [62],

$$R_h^{-1} = D_H N^{-\nu}(1 + b_h N^{-\Delta_1} + c_h N^{-(1-\nu)}). \tag{17}$$

In Eq. 16 and Eq. 17, the coefficients D_g, D_e, and D_h, which are the analogues of a_g, a_e, and a_h, are non-universal, as are the pre-factors b_g, b_e, b_h and c_h (Eq.16 and Eq. 17). However, for sufficiently large N, the ratios $\frac{D_e}{D_g} = \rho_{eg}$ and $\frac{D_g}{D_h} = \rho$ are universal for SAWs [60, 62, 65].

We first calculated ρ_{eg} and ρ for IDPs as a function of N. By keeping the corrections to scaling (Eq. 16 and Eq. 17) in mind, we draw the following conclusions from the plots shown in Fig.

2A for ρ_{eg} and Fig. 2B for ρ. (1) Off-lattice simulations for homopolymers for N=80 yields ρ_{eg} = 2.512 [66]. The most accurate estimate for ρ_{eg}, calculated using Monte Carlo simulations for a SAW on a cubic lattice with $N \gg 1$ ($\approx 10^6$), is ≈ 2.501 [62]. The exact result for a Gaussian chain is $\sqrt{6} \approx 2.449$. These bounds are reached for the IDPs (Fig. 2A). Instead, the ratio ρ_{eg} is nearly a constant, in accord with the small correction to scaling obtained for the radius of gyration and the end-to-end distance (see next section). This observation is consistent with our previous work. [43] (2) The dependence of $\rho = \frac{R_g}{R_h}$ on N is relatively smooth. Even for AP180 (N=593) the value of ρ is less than the exact result for the Gaussian chain (Fig. 2B). The strong dependence on protein size displayed in Fig. 2 suggest that finite size corrections may be relevant, which we examine next. Note that in both cases the theoretical values for the ratios ρ and ρ_{eg} are not reached, which correlates with the observations that, although Flory's scaling law provides an excellent estimate of the size of the IDPs, ultimately they are heteropolymers. Consequently, the local structural results in subtle deviations from Gaussian or self-avoiding homopolymer description.

Corrections to scaling: In order to utilize Eq. 16 and Eq. 17, we set $D_x = a_x$ (x =g ,e, or h). The values of a_x are obtained using the scaling of R_x as a function of N (Fig. 1). Using the values of a_g, a_e, and a_h, and setting $\Delta_1 = 0.528$ [60, 62], the only free parameter in Eqs. 16,17 are b_g, b_e, b_h and c_h. We determined the first two parameters by fitting the simulation results for R_g and R_e to Eq. 16. The results in Fig. 3A,B and the small values found for b_g and b_e (in fact essentially zero if one accounts for the estimated error bar) show that including finite-size scaling correction has nearly no effect on the quality of the fit.

A similar exercise using Eq. 17 using a two parameter fit (b_h and c_h) also gives excellent agreement with simulations (Fig. 3C). In this case, the contribution of the correction terms is much larger. Although the error bar is also large, the results suggest that there is a stronger finite-size effect for the hydrodynamic radius. It is worth emphasizing that Eq. 16 and Eq. 17 were obtained for SAWs. That these equations agree almost exactly with the simulation results show that globally the IDP sequences indeed mimic homopolymers in a good solvent.

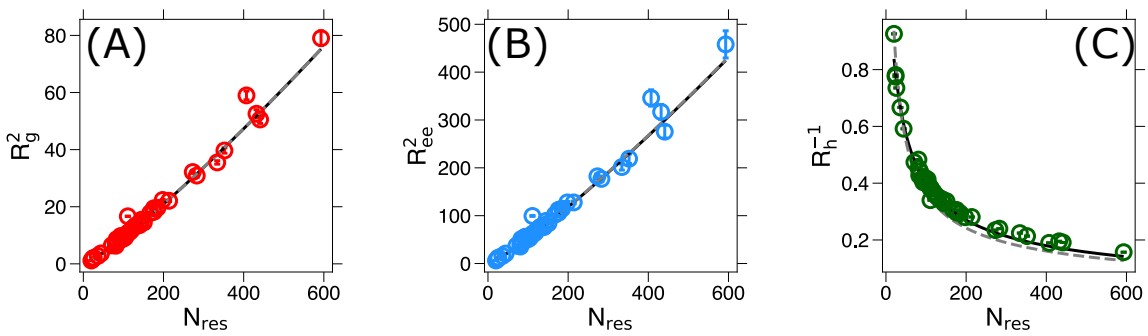

Figure 3: **Corrections to scaling for R_g, R_e, and R_h:** In all panels, the black lines show the refined fits obtained by including the finite size scaling corrections, whereas the dashed gray lines refer to the same fit shown in Fig. 1. The left and middle panels show the dependence of R_g^2 and R_{ee}^2 on N using Eq. 16. The units for R_g^2 and R_{ee}^2 are nm^2. The best fit values for b_g and b_e are -0.04 ± 0.45 and -0.06 ± 0.51, respectively. The one parameter fit gives excellent agreement with experiments. The right panel shows R_h^{-1} (in unit of nm^{-1}) on N. The two parameters in Eq. 17 ($b_h = -8.4 \pm 6.2$ and $c_h = 5.6 \pm 4.9$) were adjusted to reproduce the R_h calculated using simulations. The excellent agreement between simulations and small corrections to scaling further affirms that globally IDPs are similar to polymers in good solvent.

We should note that corrections to scaling for IDPs are very difficult to measure using experi-

ments. Besides the coefficients b_h and b_e are extremely small. Given the small values of these coefficients, let the relative correction to $R_g[= D_g N^{\nu}(1 + b_g/N^{\Delta_1})^{1/2} \approx D_g N^{\nu}(1 + 0.5 b_g/N^{\Delta_1}]$ scaling be $\delta_g = 0.5 b_g/N^{\Delta_1}$. The largest correction is expected for Histatin-5 (N=24). Using we find that $\delta_g \approx -3.6 \cdot 10^{-3}$ %. Thus, for IDPs the Flory scaling R_g is, for all practical purposes very accurate, especially considering the errors in estimating R_g using the Guinier approximation that depends on the range of q values used.

Calculated and experimental SAXS profiles are in good agreement: The calculated SAXS profiles (in the Kratky representation) for six IDPs, and their comparisons to experiments are shown in Fig. 4. The Kratky plots for the other IDP sequences are included in the Supporting Information (Figures.(S2-S6) in the Supporting Information). The SAXS profiles for the six IDPs in Fig. 4 are given in Fig. S7 in the Supporting Information. The calculated Kratky plots for AP180, FUS 1-108, and FUS 1-214 (shown as blue curves in Fig. 4) are predictions whose validation awaits future experiments. The calculated profiles match the experimental results well up to high q values for a large numbers of IDPs (Fig. 4 and the Figures (S2-S6) in the Supporting Information), despite substantial variations in sequence, length (N), and experimental conditions. We should emphasize that in these curves we only set an overall scale for the computed scattering profile to set the experimental values. The shape of the curve was not used to train the parameters, nor it was refined using procedures such as ensemble optimization, and thus constitutes a genuine prediction of the model.

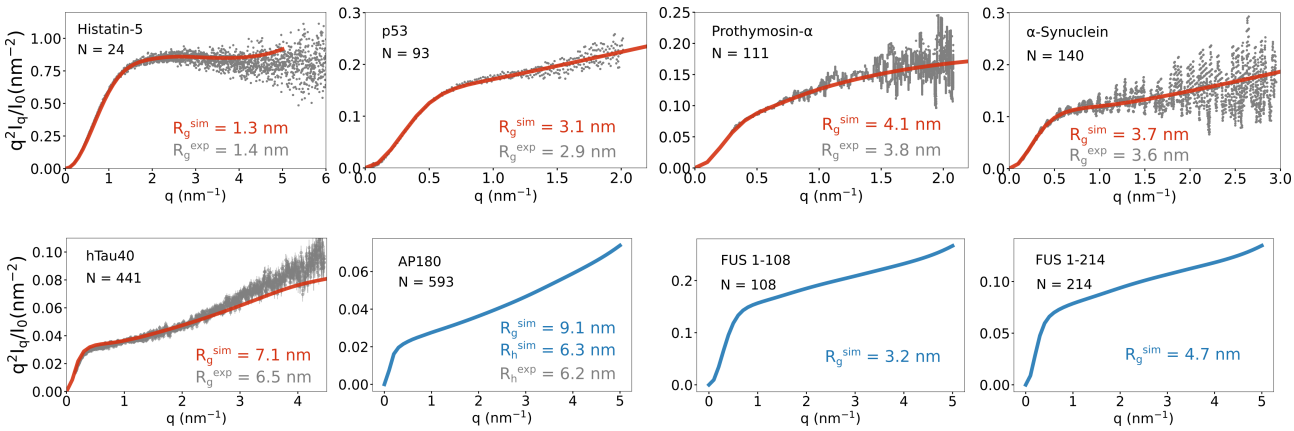

Figure 4: **Scattering Curves for IDPs:** Small angle X-ray scattering for 6 IDPs in Kratky plot (see Supporting Information for all other IDPs). Comparison between simulated (red) and experimental SAXS (grey) curves. The scattering profiles for AP180, FUS 1-108, and FUS 1-214 (shown in blue) serve as predictions.

Sequence-specific heterogeneity of IDP ensembles: Although IDPs mimic SAWs closely (Fig. 1, Fig. 2, and Fig. 3), the effect of sequence heterogeneity can be conjectured from the spread of the data. However, a much clearer and more detailed portrayal of sequence effects emerges by examining in detail the ensemble of conformations populated by the IDPs. We visualized the structurally heterogeneous ensemble in the form of a hierarchical partitioning into distinct clusters. In our previous studies, [43, 46] we showed that hierarchical clustering provides a convenient way to glean the finer features of the conformational landscapes that are masked when using ensemble-averages. In Fig. 5, we show the results from the clustering analyses for α-synuclein, htau40 and AP180. For each sequence, the minimal partitioning of the conformation space was determined from an elbow analysis of the largest distance jumps in the dendrograms as a function of the number of clusters. Fig. 5 shows the different cluster families are tidily segregated on the two-dimensional landscapes projected onto the R_g and R_{ee}

coordinates, implying that our clustering scheme is robust. In what follows, we provide some additional details related to the partitioning of the conformational space for three IDPs.

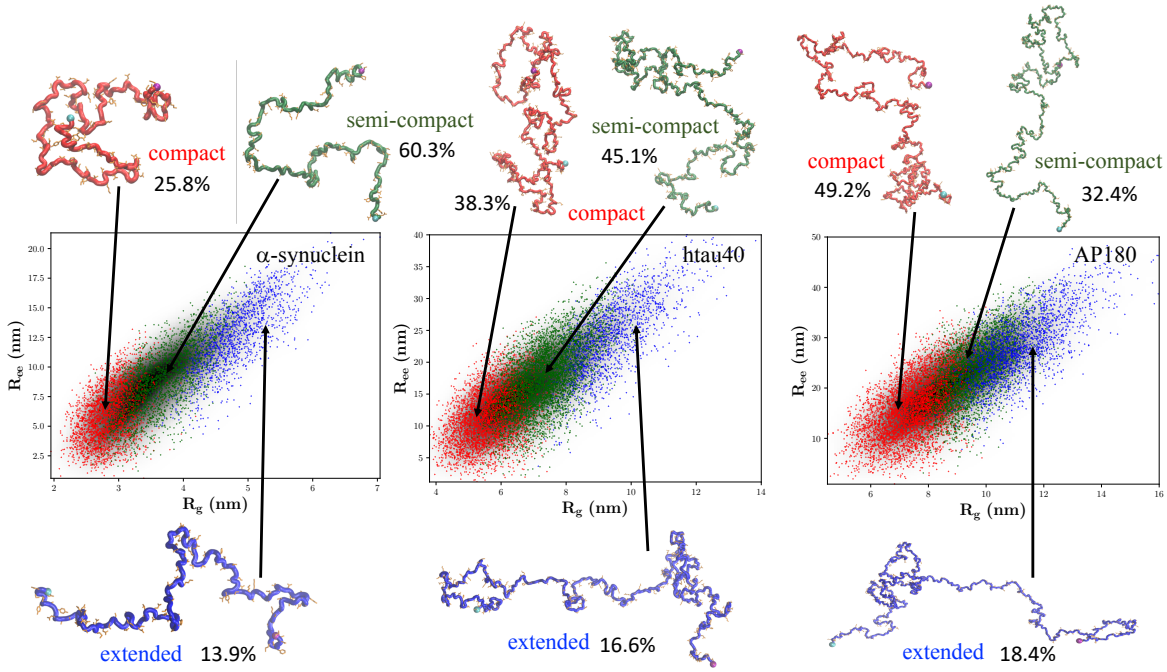

Figure 5: **Clustering of conformational ensembles:** Hierarchical clustering of the conformational ensembles for α-synuclein, htau40 and AP180. For each sequence, the conformational landscape is projected into the R_g and R_{ee} coordinates. The relative populations of the different clusters, and corresponding snapshots are shown superimposed on the two-dimensional landscapes. The backbones of these representative structures are rendered using different colors. To aid in visualization, the N-terminus residue is marked with a purple sphere, and the C-terminus is denoted as a cyan sphere. The same-color coding is used to distinguish the different clusters on the two-dimensional landscapes.

α-synuclein: α-synuclein is a 140-residue IDP, which has been extensively studied using experiments[67–72] and computer simulations, [43, 72–74] because of its key role in synaptic function, as well as its tendency to form aggregates that are implicated in Parkinson's disease and other synucleinopathies [75]. The α-synuclein sequence can be partitioned into three segments: the N-terminus (residues 1-40); the non-amyloid β component (NAC) (residues 41-95); and the C-terminus (residues 96-140). The NAC domain consists of an extended hydrophobic patch, which is a key driver of α-synuclein fibrillization. [76, 77]

The ensemble of α-synuclein conformations partition into three clusters, with semi-compact structures having a population of 60.3% (Fig. 5) . Compact structures are populated to a lesser extent, and have a population of 25.8%. Fully extended structures are only sparsely populated, and have an occupation probability of 13.9%. Conformations within the compact cluster exhibit different types of residue-residue contacts (Fig. S8A, Supporting Information). In particular, contacts within the NAC domain, which forms the core in α-synuclein fibrils, [78] appear most frequent. Some of the structures within this cluster also show long-range contacts between the N-terminus and the C-terminus, as well as the N-terminus and the NAC domain. Interestingly, contacts between the C-terminus and the NAC domain, which provide a protection mechanism against aggregation, [69, 79] are also evident in some structures within the compact cluster. Many conformations within the semi-compact cluster have weak long-range interactions be-

tween the NAC domain and residues 42-55 (often termed as the preNAC domain) within the N-terminus (Fig. S8B, Supporting Information). A recent study by Brockwell and coworkers revealed the presence of such N-terminus-NAC contacts using paramagnetic relaxation enhancement (PRE) measurements, and highlighted their key role in driving α-synuclein aggregation. [80] The contact map corresponding to the most-extended conformational cluster appear largely featureless, although some contacts seem to form transiently within the N-terminus, and the C-terminus and the NAC domains (Fig. S8C, Supporting Information). Our simulations capture the inherent plasticity of α-synuclein, and the partitioning of the conformational space is in accord with previous experimental reports. [67, 72, 81, 82]

htau40: The full-length human tau protein (htau40) is a 441 residue IDP. It is localized in the axons of neuronal cells, and regulates the assembly of microtubules.[83] The htau40 sequence can be partitioned into four domains: N-terminus (residues 1-150), proline rich domain (PRD) (residues 151-244), microtubule binding domain (MBD) (245-372), and the C-terminus (residues 372-441). The misfolding of tau and its different isoforms is linked to the formation of neurofibrillary tangles, a hallmark of various taupathies, including Alzheimer's and other diseases.[84] Because of their ability to undergo phase separation under different conditions, tau and its different truncated variants have been exploited as model systems to understand the regulatory mechanisms underlying the formation of membraneless organelles.[85]. Importantly, recent works suggest that the conformations relevant for phase separation and aggregation may be encoded within the monomer ensemble of tau,[86, 87] much like other amyloidogenic sequences.[46]

Like α-synuclein, there are three major clusters in the conformational ensemble of htau40, corresponding to compact, semi-compact and extended structures (Figure 3). The semi-compact structures dominate the equilibrium with a population of 45.1%. Compact structures are populated to a lesser extent (38.3%), while extended conformations are only sparsely populated (16.6%). Many conformations within the compact cluster exhibit enhanced contacts within the N-terminus, the PRD, and the MBD (Supporting Information, Figure S9A). The sequence of htau40 switches from predominantly acidic to mostly basic around residue 118. [88] This suggests that attractive electrostatic interactions might lead to local compaction of the chain. Accordingly, we observe a square-like pattern in the contact map around residues 100-150, which is clear in the compact ensemble (Supporting Information, Fig. S9A), visible in the semi-compact cluster (Supporting Information, Fig. S9B), and nearly absent for the more expanded ensemble (Supporting Information, Fig. S9C). Long-range contacts between the N-terminus and the PRD, which are key mediators of tau phase separation,[88, 89] exist in only a few conformations within the compact cluster (Supporting Information, Fig. S9A). Conformations within the most populated semi-compact cluster are devoid of any long-range contacts, and only local ordering is visible within the N-terminus, the PRD, and the MBD in some of the structures (Supporting Information, Fig. S9B). Fully extended conformations, which are only sparsely populated within the tau monomer ensemble, do not exhibit any residue-residue contacts (Supporting Information, Fig. S9C).

AP180: The endocytic protein AP180 plays an important role in clathrin assembly.[90, 91] AP180 is an adaptor protein of the clathrin-mediated endocytic pathway. Expressed primarily in neuronal cells, AP180 is responsible for linking phosphoinositide lipids to the clathrin coat and for recruiting SNAREs, the transmembrane proteins responsible for synaptic vesicle fusion, into clathrin-coated vesicles. [92] AP180 consists of a structured N-terminal ANTH (AP180 N-terminal Homology) domain, followed by an intrinsically disordered C-terminal domain of approximately 571 amino acids (Rattus norvegicus) [93]. The C-terminal domain of AP180 has been used as a model system to understand how IDPs induce membrane curvature.[94, 95]

The ANTH domain is responsible for recruiting phosphoinositide's and SNAREs, while the C-terminal domain participates in a network of weak, multi-valent contacts with clathrin and other components of the endocytic machinery [96]. The C-terminal domain has a moderately negative net charge per residue of -0.06, where the majority of charged residues are concentrated in the first third of the domain, which has a more significant net negative charge per residue of -0.11 [93].

We simulated a 593-residue fragment, consisting of the AP180 C-terminal domain (residues 23-593) with a HIS-tag linker (residues 1-22). The calculated value of the hydrodynamic radius (6.37 nm) is in near quantitative agreement with the estimate (6.0 nm) from fluorescence correlation spectroscopy experiments (Table S2 in the Supporting Information).[93] The R_g values from simulations and the Flory scaling relation are 8.89 nm and 8.15 nm, respectively. These specific comparisons suggest that AP180, like other IDPs simulated here, behave as homopolymers in good solvent.

Hierarchical clustering reveals that, unlike α-synuclein and htau40, the conformational ensemble of AP180 is dominated by compact structures, having an overall population of $\approx 49\%$. Semi-compact structures, which dominate the equilibrium population of both α-synuclein and htau40, are sampled less frequently for AP180 (population $\approx 32.4\%$). Extended conformations are only sparsely population ($\approx 18.4\%$). The long-range contacts between the different residues are not readily discernible from the contact maps of these three clusters (Supporting Information, Fig. S10), which result from a minimal partitioning of the conformational space. To gauge key structural details, we also performed clustering at a finer resolution. This divides the three major clusters (compact, semi-compact and extended) into additional subfamilies. We find that most structures within the compact cluster exhibit various types of long-range contacts, with a majority of them involving residues near the C-terminus (Supporting Information, Fig. S11A). The semi-compact cluster also consists of many conformations displaying long-range residue-residue contacts. However, these are mostly localized in the central segment ($\approx$ residues 210-270) of the peptide (Supporting Information, Fig. S11B). The conformations within the extended cluster are mostly RC-like, resulting in a featureless contact map (SI Appendix, Fig. S11C), but some of the subfamilies consist of structures exhibiting contacts between residues near the C-terminus (Supporting Information, Fig. S11D).

Overall, the results described in Fig. 5 illustrate that there are important details in the conformational ensemble of the ground states of IDPs that are masked in the global properties. To ensure that the hierarchical clustering scheme unambiguously reflects the sequence-specific features of IDPs, we simulated a SAW with $N = 140$ (see Supporting Information for details). The SAW ensemble appears more homogeneous, and is segregated into two classes (compact and extended) with nearly equal equilibrium populations (Supporting Information, Fig. S12). Therefore, the more nuanced partitioning of the conformational space displayed in Fig. 5 is a reflection of the sequence.

Paramagnetic relaxation enhancement (PRE) shows great recovery of local structures: In order to assess the accuracy of the model in predicting the local structures, we compared the calculated PRE with experimental measurements. PRE is an excellent tool to provide local structural information because of the long range ($\langle r^{-6} \rangle$) nature of the PRE effect, allowing for the detection of transient interactions up to 35 Å[58]. We choose α-synuclein to calculate PRE because of the wealth of available experimental and simulated data, which can be directly compared with our simulations.[27, 32, 67, 71, 97, 98]. Fig. 6 shows the intensity ratios of PRE effects for α-synuclein where the nitroxide probe is located at 4 different positions. The agreement between the calculated and experimental PRE is good. This suggests that the local structural features in our simulations reflect the experimental structural ensemble

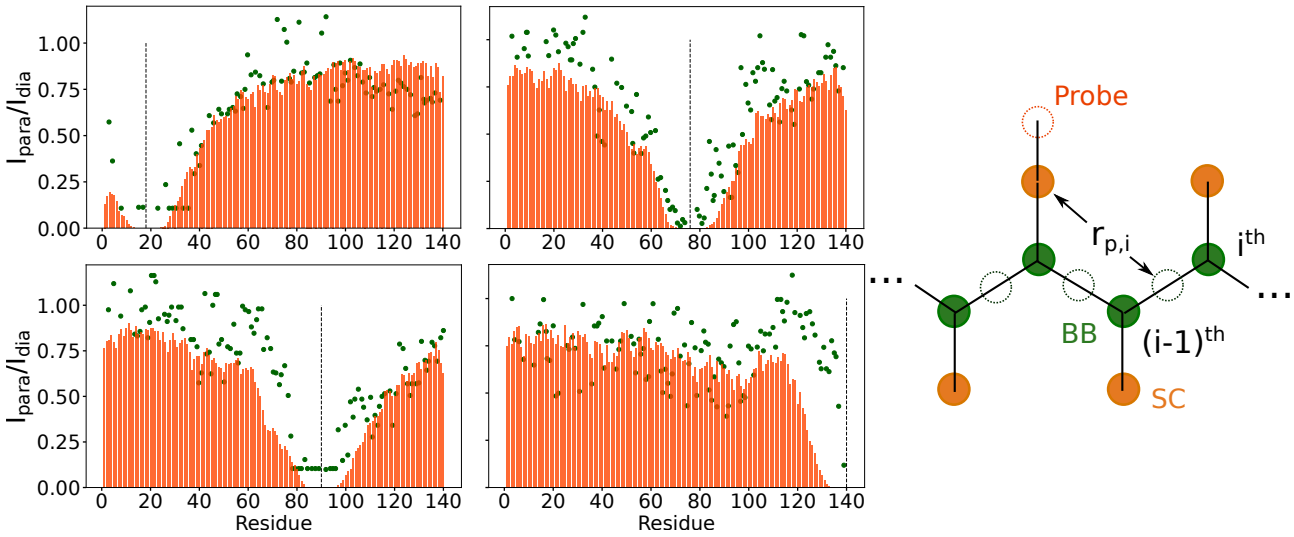

Figure 6: **Paramagnetic relaxation enhancement (PRE) of α-synuclein** Calculated PREs (orange) are in good agreement with experimental data (green dots) at 4 different locations of the probe (vertical dashed lines)[67, 71, 97]. Here, the distance is calculated from the side chain bead where the probe attached to the imaginary amide groups sits between the backbone C_α beads.

of α-synuclein.

Conclusions

Simulations using a transferable coarse-grained SOP-IDP model show that the SAXS profiles for thirty six IDPs are accurately reproduced. The calculated hydrodynamic radii for AP180, the largest IDP simulated here, and Epsin (N=432) agree quantitatively with FCS experiments (relative errors of about 6% and 9%, respectively). In addition, the agreement between simulations and paramagnetic relaxation enhancement measurements is excellent. With these achievements as a background, we summarize the lessons that are likely to be widely applicable for other IDPs. (i) The global properties, such as the radius of gyration and the mean end-to-end distance as a function of the number of residues obey the Flory scaling law accurately. As a result, finite-size scaling corrections are small, even for the 25-residue Histatin-5, in particular for the radius of gyration and the end-to-end distance. However, for the hydrodynamic radius finite-size corrections are required to improve the agreement with Flory's scaling law. Thus, the Flory scaling law suffices to estimate the sizes of IDPs without the need for experiments. (ii) The spread of the data around the theoretical curve signals that sequence-specific features provide small but measurable deviations around the expected IDP size. In addition, universal ratios between the radii of gyration and end-to-end distances, or between the radii of gyration and the hydrodynamic radii do not converge to the theoreticallay values expected for homopolymers. (iii) A fine-grained analysis further highlights and explains the sequence-dependent properties of IDPs. For instance, by probing the conformational ensemble for α-synuclein, htau40, and AP180 using hierarchical clustering, we found that the structure space partitions into clusters that have distinct properties. The extent of compaction between the clusters, and their weights are sequence dependent. It is likely that the populations of a specific structure are more likely to be functionally relevant, as has been shown in the aggregation of Aβ peptides [46] and Fused in Sarcoma [45]. (iv) Without tweaking any of the parameters, our simulations reproduce, semi

quantitatively, experimental paramagnetic relaxation measurement at four probe locations in α-synuclein. All these results show that the SOP-IDP model is transferable.

Comparison with other models: One can construct a large number of energy functions that can reproduce chosen experimental data. However, the space of energy function would likely narrow if the efficacy is assessed by comparison with multiple measurements. With the caveat that all empirical energy functions are arbitrary, we compare the SOP-IDP model with other models, which have been used to determine single-chain and multi-chain properties. The recent models [19, 24, 40, 41] use the single bead (referred to as C_α carbon model in protein folding) to represent amino acids. Because the R_g obtained in the initial version [19] produced compact structures, which disagreed with experiments, the stickiness parameter was tuned to obtain better agreement with R_g and PRE measurements [40]. In the Mpipi [41] model, which is also a single bead model, the interactions between the beads were constructed by matching to atomically detailed simulations. The calculated R_gs for 17 IDPs agreed well (correlation coefficient of 0.90) with experiments. In addition, the computed coexistence curves, as a function of temperature and concentration, for IDPs were in agreement with the available experiments. The SAXS profiles were not reported in previous studies. Given our finding that R_g is accurately predicted using the Flory scaling relation, even without finite size corrections, more stringent tests have to be performed to assess the validity of these models for single chain properties. At the very least, ii would be important to compare SAXS profiles for several IDPs.

Previous studies have shown that [29–33] the R_g estimates from simulations are sensitive to the details of the force field, with the error relative to experimental values being approximately 50% for some sequences (for e.g. Sic1 and α-synuclein modelled with the CHARMM36m force field, Table S3, Supporting Information). Reparametrization of the force field parameters [33] does rectify the over-compaction problem in certain cases. The SOP-IDP simulations fare better across the board (Table S3, Supporting Information). In Table S4 (Supporting Information), we show a systematic comparison of the R_g values obtained for α-synuclein, with different force fields and water models. Most all-atom force fields underestimate the R_g of α-synuclein, although reweighting the ensemble, which is erroneous , improves the agreement with experiment [27]. On the other hand, SOP-IDP accurately describes the size of α-synuclein without requiring any reweighting of the conformations. Its performance is comparable to AMBER19SB/TIP4P-D,[99] and the bead-necklace model,[100] which provide accurate estimates of α-synyuclein's R_g.

The SOP-IDP model differs in a few respects from other coarse-grained models. (1) Each amino acid uses two beads (C_α-SC in the protein folding context [101]). In contrast to the coarse-grained models [24, 40, 41], which have many parameters, the SOP-IDP has only three parameters whose values were determined by fitting the calculated values to the measured R_g for five proteins in the training set. (2) We simultaneously obtained agreement with experiments for R_g, $I(q)$, and PRE. None of the models have compared $I(q)$ with experiments. (3) The use of two beads, one for C_α and the other for side chains, is likely to better represent the properties of both single chains and condensates than one bead representation.

Supporting Information

A list of all IDP sequences considered in this work, along with the R_g values obtained from simulations, and those estimated from SAXS experiments; comparison of hydrodynamic radii, R_h from simulations, and those estimated from FCS measurements; comparison of the R_g estimates and SAXS profiles from the current and earlier force-field parameters; comparison of the R_g values obtained from SOP-IDP simulations, and those from all-atom simulations, for different

IDP sequences; R_g values of α-synuclein predicted using different force-fields; SAXS profiles of IDPs in the Kratky representation; scattering profiles of Histatin-5, p53, Prothymosin-α, α-synuclein, htau40 and FUS 1-214; dendrograms displaying the hierarchical organization of clusters in the α-synuclein and its corresponding SAW ensemble; residue-residue contact maps for the different conformational clusters of α-synuclein, htau40 and AP180.

Scripts and relevant input files that were used to generate the data, as well as detailed instructions regarding the implementation of the SOP-IDP force-field within OpenMM, are available from `https://github.com/tienhungf91/SOP_IDP2`.

Acknowledgement

We are grateful to Dr. Farkhad Maksudov for useful comments. We acknowledge the Texas Advanced Computing Center (TACC) for providing the necessary computing resources. This work was supported by a grant from the National Institutes of Health (GM-107703), the National Science Foundation (CHE 2320256), and a grant from the Welch Foundation (F-0019) administered through the Collie-Welch Regents Chair.
The authors declare no competing interest.

References

[1] Tompa, P. (2012) Intrinsically disordered proteins: a 10-year recap. Trends Biochem Sci 37:509–516.

[2] Deiana, A., Forcelloni, S., Porrello, A., Giansanti, A. (2019) Intrinsically Disordered Proteins and Structured Proteins with Intrinsically Disordered Regions Have Different Functional Roles in the Cell. PLoS ONE 14:e0217889.

[3] van der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D. T., Kim, P. M., Kriwacki, R. W., Oldfield, C. J., Pappu, R. V., Tompa, P., Uversky, V. N., Wright, P. E., Babu, M. M. (2014) Classification of Intrinsically Disordered Regions and Proteins. Chemical Reviews 114:6589–6631.

[4] Banani, S. F., Lee, H. O., Hyman, A. A., Rosen, M. K. (2017) Biomolecular condensates: organizers of cellular biochemistry. Nat Rev Mol Cell Biol 18:285–298.

[5] Alberti, S., Hyman, A. A. (2021) Biomolecular condensates at the nexus of cellular stress, protein aggregation disease and ageing. Nat Rev Mol Cell Biol 22:196–213.

[6] Holehouse, A. S., Pappu, R. V. (2018) Functional implications of intracellular phase transitions. Biochemistry 57:2415–2423.

[7] Hyman, A. A., Weber, C. A., Jülicher, F. (2014) Liquid-liquid phase separation in biology. Annu Rev Cell Dev Biol 30:39–58.

[8] Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K., Sharp, P. A. (2017) A Phase Separation Model for Transcriptional Control. Cell 169:13–23.

[9] G. O' Flynn, B., Mittag, T. (2021) The role of liquid-liquid phase separation in regulating enzyme activity. Curr Opin Struc Biol 69:70–79.

[10] Peeples, W., Rosen, M. K. (2021) Mechanistic dissection of increased enzymatic rate in a phase-separated compartment. Nat Chem Biol 17:693–702.

[11] Flory, P. J. (1953) Principles of Polymer Chemistry. Cornell University Press, Ithaca.

[12] de Gennes, P. G. (1979) Scaling Concepts in Polymer Physics. Cornell University Press, Ithaca and London.

[13] Rubinstein, M., Colby, R. H. (2015) Polymer physics. Oxford university press.

[14] Wang, J., Choi, J. M., Holehouse, A. S., Lee, H. O., Zhang, X., Jahnel, M., Maharana, S., Lemaitre, R., Pozniakovsky, A., Dreschel, D., Poser, I., Pappu, R. V., Alberti, S., Hyman, A. A. (2018) A Molecular Grammar Governing Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. Cell 174:688–699.

[15] Martin, E. W., Holehouse, A. S., Peran, I., Farag, M., Incicco, J. J., Bremer, A., Grace, C. R., Soranno, A., Pappu, R. V., Mittag, T. (2020) Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. Science 367:694–699.

[16] Pietrek, L. M., Stelzl, L. S., Hummer, G. (2020) Hierarchical Ensembles of Intrinsically Disordered Proteins at Atomic Resolution in Molecular Dynamics Simulations. J Chem Theory Comput 16:725–737.

[17] Shea, J. E., Best, R. B., Mittal, J. (2021) Physics-based computational and theoretical approaches to intrinsically disordered proteins. Curr Opin Struc Biol 67:219–225.

[18] Fawzi, N. L., Parekh, S. H., Mittal, J. (2021) Biophysical studies of phase separation integrating experimental and computational methods. Curr Opin Struc Biol 70:78–86.

[19] Dignon, G. L., Zheng, W., Kim, Y. C., Best, R. B., Mittal, J. (2018) Sequence determinants of protein phase behavior from a coarse-grained model. PLoS Comput Biol 14:e1005941.

[20] Mao, A. H., Crick, S. L., Vitalis, A., Chicoine, C. L., Pappu, R. V. (2010) Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. Proc Natl Acad Sci USA 107:8183–8188.

[21] Das, R. K., Ruff, K. M., Pappu, R. V. (2015) Relating sequence encoded information to form and function of intrinsically disordered proteins. Curr Opin Struc Biol 32:102–112.

[22] Wu, H., Wolynes, P. G., Papoian, G. A. (2018) AWSEM-IDP: a coarse-grained force field for intrinsically disordered proteins. J Phys Chem B 122:11115–11125.

[23] Latham, A. P., Zhang, B. (2019) Maximum entropy optimized force field for intrinsically disordered proteins. J Chem Theory Comput 16:773–781.

[24] Lotthammer, J., Ginell, G., Griffith, D., Emenecker, R., Holehouse, A. (2024) Direct prediction of intrinsically disordered protein conformational properties from sequence. Nat Methods 21:465–476.

[25] Huang, J., MacKerell Jr., A. D. (2018) Force Field Development and Simulations of Intrinsically Disordered Proteins. Curr Opin Struc Biol 48:40–48.

[26] Zerze, G. H., Zheng, W., Best, R. B., Mittal, J. (2019) Evolution of all-atom protein force fields to improve local and global properties. J Phys Chem Lett 10:2227–2234.

[27] Ahmed, M. C., Skaanning, L. K., Jussupow, A., Newcombe, E. A., Kragelund, B. B., Camilloni, C., Langkilde, A. E., Lindorff-Larsen, K. (2021) Refinement of α-Synuclein Ensembles Against SAXS Data: Comparison of Force Fields and Methods. Front Mol Biosci 8:654333.

[28] Shrestha, U. R., Smith, J. C., Petridis, L. (2021) Full structural ensembles of intrinsically disordered proteins from unbiased molecular dynamics simulations. Commun Biol 4:243.

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

[30] Rahman, M. U., Rahman, A. U., Liu, H., Chen, H. F. (2020) Comparison and evaluation of force-fields for intrinsically disordered proteins. J Chem Inf Model 60:4912–4923.

[31] Best, R. B., Zheng, W., Mittal, J. (2014) Balanced Protein-Water interactions improve properties of disordered proteins and non-specific protein association. J Chem Theory Comput 10:5113–5124.

[32] Piana, S., Donchev, A. G., Robustelli, P., Shaw, D. E. (2015) Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. J Phys Chem B 119:5113–5123.

[33] Robustelli, P., Piana, S., Shaw, D. E. (2018) Developing a molecular dynamics force field for both folded and disordered protein states. Proc Natl Acad Sci USA 115:E4758–E4766.

[34] Lincoff, J., Sasmal, S., Head-Gordon, T. (2019) The combined force field-sampling problem in simulations of disordered amyloid-β peptides. J Chem Phys 150:104108.

[35] Wang, A., Peng, X., Li, Y., Zhang, D., Zhang, Z., Li, G. (2020) Quality of force fields and sampling methods in simulating pepX peptides: a case study for intrinsically disordered proteins. Phys Chem Chem Phys 23:2430–2437.

[36] Cragnell, C., Rieloff, E., Skepö, M. (2018) Utilizing coarse-grained modeling and monte carlo simulations to evaluate the conformational ensemble of intrinsically disordered proteins and regions. J Mol Biol 430:2478–2492.

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

[38] Baidya, L., Reddy, G. (2022) pH Induced Switch in the Conformational Ensemble of Intrinsically Disordered Protein Prothymosin-α and Its Implications for Amyloid Fibril Formation. J Phys Chem Lett 13:9589–9598.

[39] Maity, H., Baidya, L., Reddy, G. (2022) Salt Induced Transitions in the Conformational Ensembles of Intrinsically Disordered Proteins. J Phys Chem B 126:5959–5971.

[40] Tesei, G., Schulze, T. K., Crehuet, R., Lindorff-Larsen, K. (2021) Accurate model of liquid-liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. Proc Natl Acad Sci USA 118:e2111696118.

[41] Joseph, J. A., Reinhardt, A., Aguirre, A., Chew, P. Y., O Russell, K., Espinosa, J. R., Garaizar, A., Collepardo-Guevara, R. (2021) Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. Nat Comput Sci 11:732–743.

[42] Emil Thomasen, F., Pesce, F., Ahrensback Roesgaard, M., Tesei, G., Lindorff-Larsen, K. (2022) Improving MARTINI 3 for Disordered and Multidomain Proteins. J Chem Theory Comput 18:2033–2041.

[43] Baul, U., Chakraborty, D., Mugnai, M. L., Straub, J. E., Thirumalai, D. (2019) Sequence effects on size, shape, and structural heterogeneity in Intrinsically Disordered Proteins. J Phys Chem B 123:3462–3474.

[44] Hyeon, C., Dima, R. I., Thirumalai, D. (2006) Pathways and kinetic barriers in mechanical unfolding and refolding of RNA and proteins. Structure 14:1633–1645.

[45] Kumar, A., Chakraborty, D., Mugnai, M. L., Straub, J. E., Thirumalai, D. (2021) Sequence Determines the Switch in the Fibril Forming Regions in the Low-Complexity FUS Protein and Its Variants. J Phys Chem Lett 12:9026–9032.

[46] Chakraborty, D., Straub, J. E., Thirumalai, D. (2020) Differences in the free energies between the excited states of Aβ40 and Aβ42 monomers encode their aggregation propensities. Proc Natl Acad Sci USA 117:19926–19937.

[47] D. Chakraborty, J. E. S., Thirumalai, D. (2023) Energy landscapes of Aβ monomers are sculpted in accordance with Ostwald's rule of stages. Sci Adv 9:eadd6921.

[48] Betancourt, M. R., Thirumalai, D. (1999) Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Prot Sci 8:361–369.

[49] Di Pierro, M., Elber, R. (2013) Automated optimization of potential parameters. J Chem Theory Comput 9:3311–3320.

[50] Di Pierro, M., Mugnai, M. L., Elber, R. (2015) Optimizing potentials for a liquid mixture: A new force field for a tert-butanol and water solution. The Journal of Physical Chemistry B 119:836–849.

[51] Nocedal, J., Weight, S. J. (2006) Numerical Optimization. Springer New York, NY.

[52] Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L. P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., Wiewiora, R. P., Brooks, B. R., Pande, V. S. (2017) OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. PLoS Comput Biol 13:e1005659.

[53] Honeycutt, J., Thirumalai, D. (1992) The nature of folded states of globular proteins. Biopolymers 32:695–709.

[54] Tong, D., Yang, S., Lu, L. (2016) Accurate optimization of amino acid form factors for computing small-angle X-ray scattering intensity of atomistic protein structures. J Appl Crystallogr 49:1148–1161.

[55] Pauling, L. (1988) General Chemistry. Dover.

[56] Ward, J. H. (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244.

[57] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat Meth 17:261–272.

[58] Clore, G. M., Iwahara, J. (2009) Theory, Practice, and Applications of Paramagnetic Relaxation Enhancement for the Characterization of Transient Low-Population States of Biological Macromolecules and Their Complexes. Chem Rev 109:4108–4139.

[59] Doi, M., Edwards, S. F. (1986) The Theory of Polymer Dynamics. Oxford University Press.

[60] Clisby, N. (2010) Accurate Estimate of the Critical Exponent ν for Self-Avoiding Walks via a Fast Implementation of the Pivot Algorithm. Phys Rev Lett 104:055702.

[61] Guttman, C. M., McCrackin, F. L., Han, C. C. (1982) Monte Carlo Calculations of the Hydrodynamic Radius at the Θ Point. Deviations from Analytical Gaussian Behavior. Macromolecules 15:1207–1210.

[62] Clisby, N., Dünweg, B. (2016) High-precision estimate of the hydrodynamic radius for self-avoiding walks. Phys Rev E 94:052102.

[63] Kohn, J. E., Millett, I. S., Jacob, J., Zagrovic, B., Dillon, T. M., Cingel, N., Dothager, R. S., Seifert, S., Thiyagarajan, P., Sosnick, T. R., Hasan, M. Z., Pande, V. S., Ruczinski, I., Doniach, S., Plaxco, K. W. (2004) Random-coil behaviour and the dimensions of chemically unfolded proteins. Proc Natl Acad Sci USA 101:12491–12496.

[64] Fisher, M. E. (1966) Shape of a self-avoiding walk or polymer chain. J Chem Phys 44:616–622.

[65] Dünweg, B., Reith, D., Steinhauser, M., Kremer, K. (2002) Corrections to scaling in the hydrodynamic properties of dilute polymer solutions. J Chem Phys 117:914–924.

[66] Honeycutt, J. D., Thirumalai, D. (1989) Static properties of polymer chains in porous media. J Chem Phys 90:4542–4559.

[67] Dedmon, M. M., Lindorff-Larsen, K., Christodoulou, J., Vendruscolo, M., Dobson, C. M. (2005) Mapping long-range interactions in α-synuclein using spin-label NMR and ensemble molecular dynamics simulations. J Am Chem Soc 127:476–477.

[68] Lee, J. C., Gray, H. B., Winkler, J. R. (2005) Tertiary contact formation in α-synuclein probed by electron transfer. J Am Chem Soc 127:16388–16389.

[69] Trexler, A. J., Rhoades, E. (2010) Single molecule characterization of α-synuclein in aggregation-prone states. Biophys J 99:3048–3055.

[70] Wu, K. P., Baum, J. (2010) Detection of transient interchain interactions in the intrinsically disordered protein α-synuclein by NMR paramagnetic relaxation enhancement. J Am Chem Soc 132.

[71] Bertoncini, C. W., Jung, Y. S., Fernandez, C. O., Hoyer, W., Griesinger, C., Jovin, T. M., Zweckstetter, M. (2005) Release of long-range tertiary interactions potentiates aggregation of natively unstructured α-synuclein. Proc Natl Acad Sci USA 102:1430–1435.

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

[73] Yu, H., Han, W., Ma, W., Schulten, K. (2015) Transient β-hairpin formation in α-synuclein monomer revealed by coarse-grained molecular dynamics simulation. J Chem Phys 143:243142.

[74] Graen, T., Klement, R., Grupi, A., Haas, E., Grubmuller, H. (2018) Transient Secondary and Tertiary Structure Formation Kinetics in the Intrinsically Disordered State of α-Synuclein from Atomistic Simulations. ChemPhysChem 19:2507–2511.

[75] Vekrellis, K., Rideout, H. J., Stefanis, L. (2004) The neurobiology of α-synuclein. Mol Neurobiol 30:1–21.

[76] Giasson, B. I., Murray, I. V., Trojanowski, J. Q., Lee, V. M. (2001) A hydrophobic stretch of 12 amino acid residues in the middle of alpha-synuclein is essential for filament assembly. J Biol Chem 276:2380–2386.

[77] Rodriguez, J. A., Ivanova, M. I., Sawaya, M. R., Cascio, D., Reyes, F. E., Shi, D., Sangwan, S., Guenther, E. L., Johnson, L. M., Zhang, M., Jiang, L., Arbing, M. A., Nannenga, B. L., Hattne, J., Whitelegge, J., Brewster, A. S., Messerschmidt, M., Boutet, S., Sauter, N. K., Gonen, T., Eisenberg, D. S. (2015) Structure of the toxic core of α-synuclein from invisible crystals. Nature 525:486–490.

[78] Tuttle, M. D., Comellas, G., Nieuwkoop, A. J., Covell, D. J., Berthold, D. A., Kloepper, K. D., Courtney, J. M., Kim, J. K., Barclay, A. M., Kendall, A., Wan, W., Stubbs, G., Schwieters, C. D., Lee, V. M., George, J. M., Rienstra, C. M. (2016) Solid-state NMR structure of a pathogenic fibril of full-length human α-synuclein. Nat Struc Mol Biol 23:409–415.

[79] Hoyer, W., Cherny, D., Subramaniam, V., Jovin, T. M. (2004) Impact of the Acidic C-Terminal Region Comprising Amino Acids 109-140 on α-Synuclein Aggregation in Vitro. Biochemistry 43:16233–16242.

[80] Doherty, C. P. A., Ulamec, S. M., Maya-Martinez, R., Good, S. C., Makepeace, J., Khan, G. N., van Oosten-Hawle, P., Radford, S. E., Brockwell, D. J. (2020) A short motif in the N-terminal region of α-synuclein is critical for both aggregation and function. Nat Struc Mol Biol 27:249–259.

[81] Frimpong, A. K., Abzalimov, R. R., Uversky, V. N., Kaltashov, I. A. (2010) Characterization of intrinsically disordered proteins with electrospray ionization mass spectrometry: conformational heterogeneity of α-synuclein. Proteins: Struc Func Bioinf 78:714–722.

[82] Ullman, O., Fisher, K. C., Stulz, C. M. (2011) Explaining the structural plasticity of α-synuclein. J Am Chem Soc 133:19536–19546.

[83] Weingarten, M. D., Lockwood, A. H., Hwo, S. Y., Kirschner, M. W. (1975) A protein factor essential for microtubule assembly. Proc Natl Acad Sci USA 72:1858–1862.

[84] Gao, Y. L., Wang, N., Sun, F. R., Cao, X. P., Yu, J. Y. (2018) Tau in neurodegenerative disease. Ann Transl Med 6:175.

[85] Boyko, S., Surewicz, W. K. (2022) Tau liquid-liquid phase separation in neurodegenerative diseases. Trends Cell Biol 32:611–623.

[86] Eschmann, N. A., Do, T. D., LaPointe, N. E., Shea, J. E., Feinstein, S. C., Bowers, M. T., Han, S. (2015) Tau Aggregation Propensity Engrained in Its Solution State. J Phys Chem B 119:14421–14432.

[87] Dong, X., Bera, S., Qiao, Q., Tang, Y., Lao, Z., Luo, Y., Gazit, E., Wei, G. (2021) Liquid-liquid phase separtion of tau protein is encoded at the monomeric level. J Phys Chem Lett 12:2576–2586.

[88] Boyko, S., Qi, X., Chen, T. H., Surewicz, K., Surewicz, W. K. (2019) Liquid-liquid phase separation of tau protein: The crucial role of electrostatic interactions. J Biol Chem 294:11054–11059.

[89] Wu, C., Zhao, J., Wu, Q., Tan, Q., Liu, Q., Xiao, S. (2021) Tau N-Terminal Inserts Regulate Tau Liquid-Liquid Phase Separation and Condensates Maturation in a Neuronal Cell Model. Int J Mol Sci 22:9728.

[90] Ahle, S., Ungewickell, E. (1986) Purification and properties of a new clathrin assembly protein. EMBO J 5:3143–3149.

[91] Morris, S. A., Schroder, S., Plessmann, U., Weber, K., Ungewickell, E. (1993) Clathrin assembly protein AP180: primary structure, domain organization and identification of a clathrin binding site. EMBO J 12:667–675.

[92] Miller, S. E., Sahlender, D. A., Graham, S. C., Honig, S., Robinson, M. S., Peden, A. A., Owen, D. J. (2011) The molecular basis for the endocytosis of small R-SNAREs by the clathrin adaptor CALM. Cell 147:1118–1131.

[93] Zeno, W. F., Thatte, A. S., Wang, L., Snead, W. T., Lafer, E. M., Stachowiak, J. C. (2019) Molecular mechanisms of membrane curvature sensing by a disordered protein. J Am Chem Soc 41:10361–10371.

[94] Busch, D. J., Houser, J. R., Hayden, C. C., Sherman, M. B., Lafer, E. M., Stachowiak, J. C. (2015) Intrinsically disordered proteins drive membrane curvature. Nat Commun 6:7875.

[95] Zeno, W. F., Thatte, A. S., Wang, L., Snead, W. T., Lafer, E. M., Stachowiak, J. C. (2019) Molecular mechanisms of membrane curvature sensing by a disordered protein. J Am Chem Soc 41:10361–10371.

[96] Owen, D. J., Collins, B. M., Evans, P. R. (2004) Adaptors for clathrin coats: structure and function. Annu Rev Cell Dev Biol 20:153–191.

[97] Salmon, L., Nodet, G., Ozenne, V., Yin, G., Jensen, M. R., Zweckstetter, M., Blackledge, M. (2010) NMR Characterization of Long-Range Order in Intrinsically Disordered Proteins. J Am Chem Soc 132:8407–8418.

[98] Esteban-M, S., Silvestre-Ryan, J., Bertoncini, C. W., Salvatella, X. (2013) Identification of fibril-like tertiary contacts in soluble monomeric α-synuclein. Biophys J 105:1192–1198.

[99] Pedersen, K. B., Flores-Canales, J. C., Schiott, B. Predicting molecular properties of α-synuclein using force fields for intrinsically disordered proteins. Proteins Struc Func 91:47–61.

[100] Fagerberg, E., Skepo, M. (2023) Comparative Performance of Computer Simulation Models of Intrinsically Disordered Proteins at Different Levels of Coarse-Graining. J Chem Inf Model 63:4079–4087.

[101] K.Klimov, D., Thirumalai, D. (2000) Mechanisms and kinetics of β-hairpin formation. Proc Natl Acad Sci USA 97:2544–2549.