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

Mauro L. Mugnai,[†,‡] Debayan Chakraborty,[†] Abhinaw Kumar,[†] Hung T. Nguyen,[†] Wade Zeno,[¶] Jeanne C. Stachowiak,[§] John E. Straub,[‖] and D. Thirumalai[*,†,⊥]

†*Department of Chemistry, The University of Texas at Austin, 105 E 24th Street, Stop A5300, Austin, TX 78712, USA.*

‡*Present Address: Institute of Soft Matter Synthesis and Metrology, Georgetown University, Washington DC 20057, USA.*

¶*Mork Family Department of Chemical Engineering and Materials Science, University of Southern California, 3650 McClintock Ave, Los Angeles, CA 90089, USA.*

§*Department of Biomedical Engineering,The University of Texas at Austin, 105 E 24th Street, Stop A5300, Austin, TX 78712, USA.*

‖*Department of Chemistry, Boston University, MA 02215, USA.*

⊥*Department of Physics, The University of Texas at Austin, 2515 Speedway, Austin, TX 78712, USA.*

E-mail: dave.thirumalai@gmail.com

**Abstract**

The preponderance of Intrinsically Disordered Proteins (IDPs) in the eukaryotic proteome, and their ability to interact with each other, proteins, RNA, and DNA for functional purposes have made it important to quantitatively characterize their biophysical properties. Towards this end, we developed the transferable Self-Organized Polymer (SOP-IDP) model in order to calculate the properties of a number of IDPs. The calculated and measured radius of gyration values ($R_g$s) are in excellent agreement, with a correlation coefficient of 0.96. For AP180 and Epsin, the predicted and values obtained using Fluorescence Correlation Spectroscopy for the hydrodynamic radii ($R_h$s) are also in quantitative agreement. Strikingly, the calculated SAXS profiles for thirty six IDPs also nearly match the experiments. The dependence of $R_g$, the mean end-to-end distance ($R_e$), and $R_h$ obey the Flory scaling law, $R_\alpha \approx a_\alpha N^{0.59}$ ($\alpha = g$, $e$, and $h$), suggesting that globally IDPs behave as polymers in a good solvent. The values of $a_g$, $a_e$, and $a_h$ are 0.21 nm, 0.53 nm, and 0.16 nm, respectively. Surprisingly, finite size corrections to scaling, expected on theoretical grounds, for all the three quantities are negligible. Sequence dependencies, masked in ensemble properties, emerge through a fine structure analyses of the conformational ensembles using a hierarchical clustering method. Typically, the ensemble of conformations partition into three distinct clusters, with differing extent of population and structural properties. The subpopulations could dictate phase separation tendencies and association with ligands. Without any adjustments to the three parameters in the SOP-IDP model, we obtained excellent agreement with paramagnetic relaxation enhancement (PRE) measurements for $\alpha$-synuclein. The transferable SOP-IDP model sets the stage for a number of promising applications, including the study of phase separation in IDPs and interactions with nucleic acids.

# 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) It is suspected that a substantial fraction (possibly as high as 40%) of eukaryotic proteome are IDPs.[1–3] (2) IDPs play a central role in a variety of cellular functions because they are not only capable of self-association but also interact with multiple partners, including nucleic acids. (3) They organize themselves to form condensates without membranes, which are suspected to offer advantages in 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 solution, 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 reflect the exact sequence.[14,15]

Despite considerable progress in computer simulations of IDPs,[16–23] there is still a need to develop transferable models, with as few parameters as possible, whose accuracies can be assessed by direct unbiased comparison with different experiments. The availability of Small Angle X-ray Scattering (SAXS) profiles for a number of IDPs is often used as the basis for quantifying the accuracy of computational models. There are potentially a large number of energy functions that one could create in order to describe the shapes and conformational fluctuations in 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. Several coarse-grained computational models have been developed[19,22,24–29] with varying degree of success. Recently, we introduced the Self-Organized Polymer (SOP-IDP) model[30] for IDPs. The three parameters in the energy function (see below) were calibrated in order to reproduce the experimentally

determined radii of gyration ($R_g$s) for three IDPs. We then showed that the calculated scattering profiles for 24 other IDPs were in good agreement with SAXS profiles ($I(q)$s with $q$ being scattering wave vector) without 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. SA1 in the Supplementary Information (SI)). Although the calculated and experimental $R_g$s were in fair agreement (Fig. S1B in the SI) with each other (Pearson correlation coefficient is 0.83), in some instances the predicted $R_g$s were smaller than the experimental values by about 10%. The initial iteration[30] lacked the 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 found it necessary to recalibrate the parameters in order to reduce deviations from the experimentally measured $I(q)$ for Histatin-5.

The goal of this article is to use a different optimization procedure to recalibrate the three parameters in the SOP-IDP model. Simulations using the resulting energy function almost quantitatively reproduce the $I(q)$ profiles for 36 IDP sequences. We show that the IDPs behave like random coils, obeying the famous Flory scaling law for random coils. Strikingly, corrections to scaling for $R_g$, the mean end-to-end distance, $R_{ee}$, and the hydrodynamic radius, $R_h$ are small. Sequence effects are discernible only if the fine structural details are investigated using a hierarchical clustering scheme. In particular, the fractions of compact, semi-compact, and extended conformations vary depending on the sequence and external conditions. The proposed SOP-IDP model also reproduces the residue-specific paramagnetic relaxation enhancement (PRE) for $\alpha$-synuclein nearly quantitatively. As testable predictions, we report $I(q)$s AP180 and FUS. It is hoped that the model may be profitably used to investigate condensate formation in IDPs as well as protein-nucleic acid interactions.

# Methods

## Computational Details

**Energy function:** In the spirit of an earlier study,[31] and subsequent reports ,[30,32–34] each amino acid is represented by two beads, one representing the $C_\alpha$ atom. The other bead 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 along the backbone. The SOP-IDP interaction potential is,

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

The FENE potential ($U_{FENE}$) accounts for chain connectivity of the backbone $C_\alpha$ atoms and the chemical bonds between the $C_\alpha$ 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). On the other hand, $U_{C_a-C_a}$, $U_{C_a-SC}$, $U_{SC-SC}$ 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). Among these, 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. These are modeled using the Lennard-Jones (LJ) potentials,

$$U_{C_a-C_a} = \sum_{i=1}^{N_{BB}} \epsilon_{BB} \left[ \left(\frac{\sigma_i}{r_i}\right)^{12} - 2\left(\frac{\sigma_i}{r_i},\right)^{6} \right] \tag{2}$$

$$U_{C_a-SC} = \sum_{i=1}^{N_{BS}} \epsilon_{BS} \left[ \left(\frac{\sigma_i}{r_i}\right)^{12} - 2\left(\frac{\sigma_i}{r_i}\right)^{6} \right], \tag{3}$$

$$U_{SC-SC} = \sum_{i=1}^{N_{SS}} \epsilon_{SS} \left|\epsilon_i - 0.7\right| \left[ \left(\frac{\sigma_i}{r_i}\right)^{12} - 2\left(\frac{\sigma_i}{r_i}\right)^{6} \right], \tag{4}$$

where $\sigma_i$ is the sum of the radii of two interacting beads and corresponds to the minimum

of the potential. The parameters $\epsilon_i$ are taken from the Betancourt-Thirumalai statistical potential.[35] The three unknown parameters in the SOP-IDP energy function are $\epsilon_{BB}$, $\epsilon_{BS}$, and $\epsilon_{SS}$, which specify the interaction strengths involving the backbone and SCs. We evaluate their values 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 a number of other IDPs, which are not in the test set.

**Optimization procedure:** The values of $\epsilon_{BB}$, $\epsilon_{BS}$, and $\epsilon_{SS}$ are calculated using a minor modification of a parameter optimization (POP) procedure[36,37] in order to improve the agreement between simulations and experiments. Briefly, consider an observable $\hat{O}_n(\Gamma)$, which is a function of the conformational space, $\Gamma$, 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 $\Gamma$, 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 $n_{\text{obs}}$ quantities is calculated using $\chi^2$:

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

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 $\chi^2$. We used Newton's method[38] to obtain the smallest value of $\chi^2$. This requires the calculation of 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 in POP,[36,37] these derivatives are computed analytically. We introduced a slight modification. Instead of resampling the conformations after every step (which is formally the correct procedure), we re-weighed 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-weighing becomes inaccurate if the new parameters change significantly from the original values. In order to avoid this, we implemented a confining function that limits the maximum change for each parameter during the re-weighing-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-weighing 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 is reached. 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.[39] The equations of motion were integrated using the leap-frog algorithm, employing 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.[40] For each IDP, simulations were carried out for $2 \times 10^9$ steps. The first $3 \times 10^8$ steps were considered as equilibration time, and statistics were collected every 10,000 steps for the rest of the trajectory. All simulations are checked to give converged results for all calculated quantities.

**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 $C_\alpha$ and SC beads. The $q$-dependent form factors, $f_i(q)$, for the backbone and side-chains are taken from an earlier study[41] (we used the values from Table S3 in[41]). 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 1.5 Å(a C-H bond length). It should be noted 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 experimental results for I(q).

We also calculated the SAXS profiles using only the $C_\alpha$ representation (the two summations in Eq. (7) only run over the BB beads), using a different set of structure factors[41] (from Table S6). We found that the SAXS profiles calculated using the two different resolutions are almost identical, with a maximum deviation of around 4%.

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

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

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,

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

**Hierarchical clustering of conformational ensembles:** In previous studies,[30,33] 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, $D_{ij}$,

$$D_{ij} = 1 - \chi_{ij}, \tag{10}$$

where $\chi_{ij}$, the structural overlap between conformations $i$ and $j$, is given by,

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

In the above equation, $r_{a,b}^i$ and $r_{a,b}^j$ 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[42] criterion, as available within the *scipy* module.[43]

**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.[44] 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:[45]

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

where $K = 1.23 \times 10^{-32}$ cm$^6$s$^{-2}$ is a constant for the nitroxide radical probe, $\tau_c$ is the correlation time of the protein reorientation (taken to be 2 ns, which is within the range estimated using atomistic simulations[46]), 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 $\tau_c$.

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 $i^{\text{th}}$ and $(i-1)^{\text{th}}$ $C_\alpha$ (to mimic the location of the $i^{\text{th}}$ amide group) (see Fig. 6). 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$^{-1}$).

# Results:

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

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

The corresponding values derived previously[30] were (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 $\epsilon_{BS}$. Use of the new set of three parameters greatly improved the agreement between both SAXS profiles (Fig. S1 in the Supplementary Information (SI)) as well in the correlation between the calculated and experimental values for the radius of gyration (see below).

**IDPs behave as polymers in a good solvent:** We assess the efficacy of the SOP-IDP energy function by comparing $R_g$s 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, which is significantly higher than the previous value of 0.83 (Fig. S1B in the SI). Based on the results described in Fig. 1B and Fig. 1C, we surmise that the IDPs behave as homopolymers in a good solvent, that are usually modeled as Self Avoiding Walks (SAWs). (1) Plot of $\ln(R_g)$ versus $\ln(N)$ ($N$ is the number of residues in a given IDP) is linear with a slope consistent with $\nu \approx 0.59$ (Fig. 1B), which is the result obtained for SAWs, implying that IDPs mimic polymers in a good solvent. By fixing $\nu$, we fit the prefactor $a_g$ for $R_g \sim a_g N^{\nu}$ and obtained $a_g \approx 0.21\,\text{nm}$. 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,[47] generated at high denaturant concentrations. (2) The dependence of the hydrodynamic radius on $N$ ($R_h \approx a_h N^{\nu}$ with $a_h \approx 0.16\,\text{nm}$ obtained via data fitting) also conforms to the Flory scaling law (Fig. 1C). Based on the $R_g$ and $R_h$ scaling, we expect that $\rho = \frac{R_g}{R_h} = \frac{a_g}{a_h} \approx 1.35$. (3) The dependence of $R_e \equiv \sqrt{\langle R_{ee}^2 \rangle}$ on $N$ also follows the Flory scaling, with $R_e \approx a_e N^{\nu}$, where the fitted parameter here is $a_e \approx 0.53\,\text{nm}$. Thus, the ratio $\rho_{eg} = \frac{a_e}{a_g}$=2.501. However, we might expect that size ($N$) of IDPs should have important consequences in the values of $\rho$ and $\rho_{eg}$ (see below). (4) The calculated values of $R_h$ AP180 ($N = 593$) and Epsin ($N$=432) agree well with the values obtained using FCS experiments (Table S2 in the SI). The predicted $R_g \approx 0.21 N^{0.59}$= 9.08 nm for AP180 is very close to the value obtained using the SOP-IDP simulations (Table S2 in the SI).

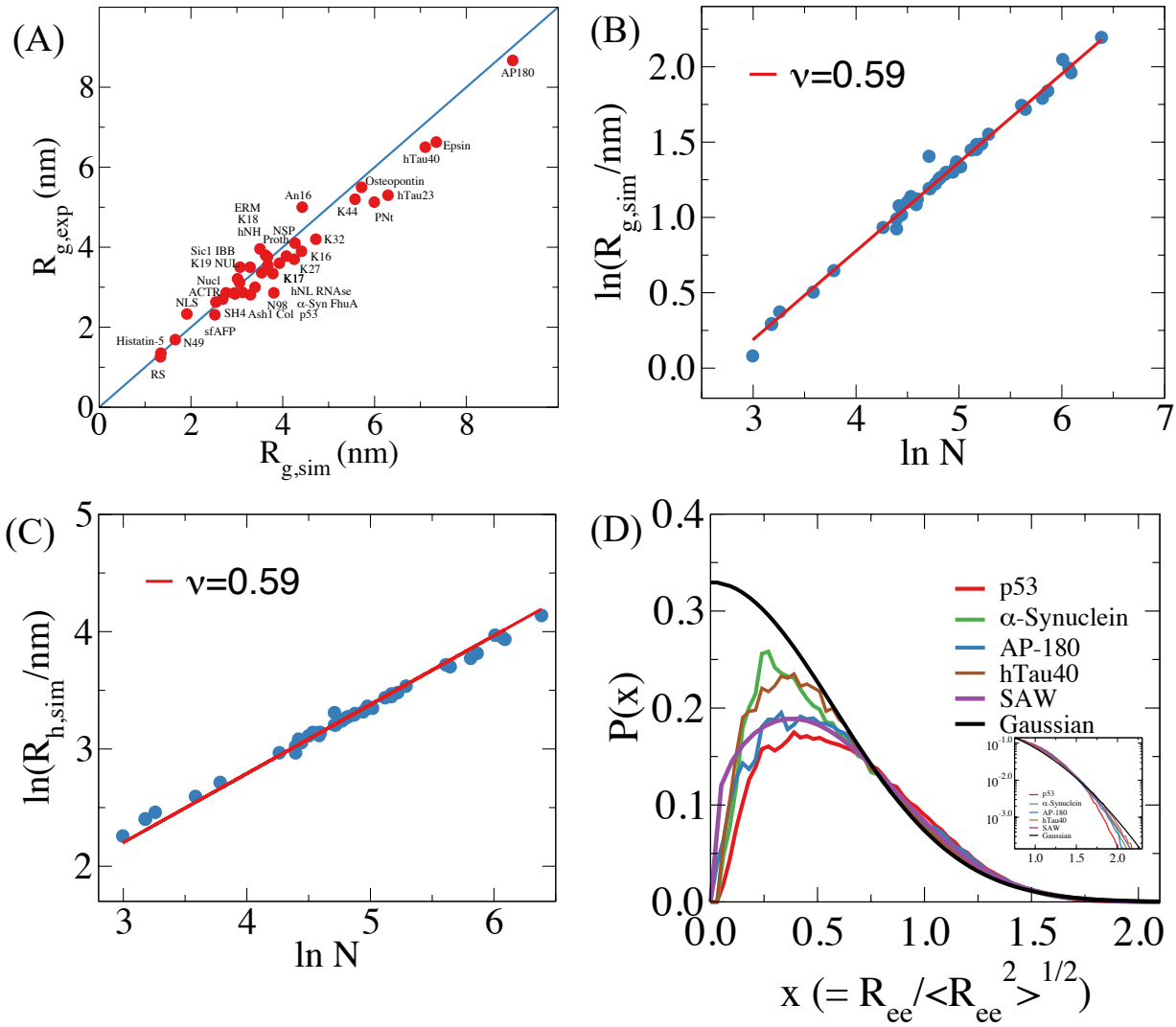


Figure 1: **Global IDP properties:** (A) Radius of gyration ($R_{g,sim}$) from SOP-IDP simulations versus $R_{g,exp}$ derived from SAXS experiments for the 36 IDP sequences with different lengths and sequence composition. The data are given in Table S1 in the SI. The Pearson correlation coefficient is 0.96. (B) Scaling of $R_{g,sim}$ with the number of residues, $N$ as $R_{g,exp} = a_g N^{\nu}$, with $\nu = 0.59$, and $a_g = 0.21\,\text{nm}$. (C) The calculated hydrodynamic radius, $R_h$ follows Flory's scaling law with $\nu = 0.588$ with $a_h \approx 0.156\,\text{nm}$. (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).

**End-to-end distribution exhibits 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.[12,13,48] The expression for $P(x)$ is,

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

where $C$ and $\alpha$ 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 $g$, 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 almost coincide with 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[49]),

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

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,[49]

$$\langle R_h^{-1} \rangle \approx D_h^{-1} N^{-\nu}(1 + b_h N^{-(\Delta_a - \nu)}), \tag{17}$$

where $\Delta_a$ and $b_h$ are fit parameters. Unlike the corrections to scaling for $R_g$ and $R_e$, the theoretical basis for Eq. 17 is less well-founded. 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 $b_g$, $b_e$, and $b_h$ (Eq.16 and Eq. 17). However, for sufficiently large $N$, the ratios $\frac{D_e}{D_g} = \rho_{eg}^2$ and $\frac{\sqrt{D_g}}{D_h} = \rho$ are universal for SAWs.[49–51]

We first calculated $\rho_{eg}$ and $\rho$ for IDPs as a function of $N$. By keeping the corrections to

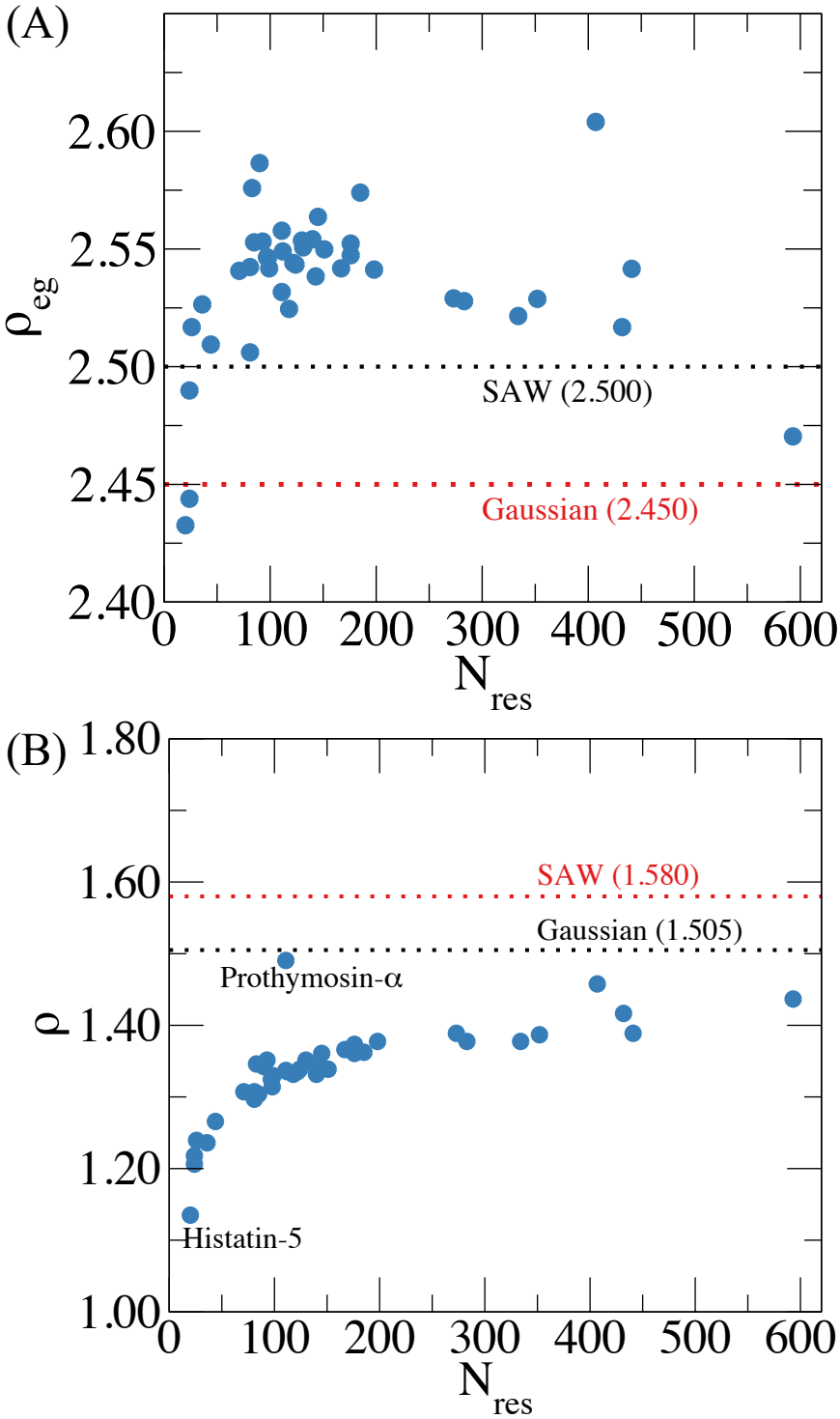


Figure 2: **Assessing the universality of $\rho_{eg}$ and $\rho$:** (A) Dependence of $\rho_{eg} = \sqrt{\frac{\langle R_{ee}\rangle^2}{\langle R_g\rangle^2}}$ on $N$ calculated using simulations. (B) Same as (A) except $\rho = \frac{a_g}{a_h}$ is plotted as a function of $N$.

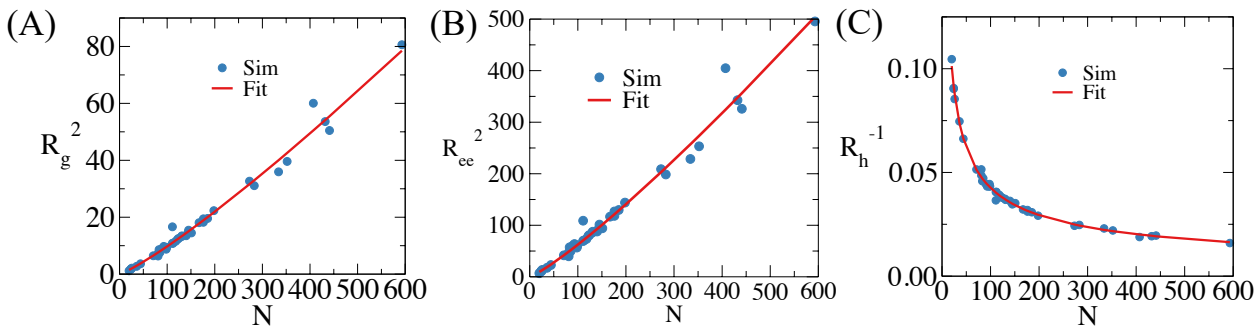


Figure 3: **Corrections to scaling for $R_g$, $R_e$, and $R_h$:** The left and middle panels show the dependence of $R_g^2$ and $R_e^2$ on $N$ using Eq. 16. The units for $R_g^2$ and $R_e^2$ are nm$^2$. The best fit values for $b_g$ and $b_e$ are 0.13 and 0.089, respectively. The one parameter fit gives excellent agreement with experiments. The right panel shows $R_n^{-1}$ (in unit of nm$^{-1}$) on $N$. The two parameters in Eq. 17 ($b_h = 2.29$ and $\Delta_a = 0.52$) 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.

scaling (Eq.16 and Eq. 17) in mind, we draw the following conclusions from the plots shown in Fig. 2A for $\rho_{eg}$ and Fig. 2B for $\rho$. (1) Off lattice simulations for homopolymers for $N$=80 yields $\rho_{eg} = 2.512$.[52] However, as $N$ increases $\rho_{eg}$ decreases only modestly. The most accurate estimate for $\rho_{eg}$, obtained using Monte Carlo simulations for a SAW on a cubic lattice with $N \gg 1$ ($\approx 10^6$), is 2.500.[49] The exact result for a Gaussian chain is $\sqrt{6} \approx 2.450$. Neither of these bounds are reached for the IDPs (Fig. 2A). Although the data in Fig. 2A are noisy (especially for $N < 300$), it does show that the SAW limit is reached from above, which is consistent with the correction to scaling ($b_g > 0$). (2) The dependence of $\rho = \frac{R_g}{R_h}$ on $N$ is relatively smooth. Even for AP180 ($N$=593) the value of $\rho$ is less than the exact result for the Gaussian chain (Fig. 2B). The results in Fig. 2 suggest that finite size corrections may be relevant, which we examine next.

**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 $a_g$ and $a_x$, the only free parameter in Eq. 16 ( Eq. 17) is $b_e$ ($b_h$). We determined these two parameters by fitting the simulation results for $R_g$ and $R_e$ to Eq. 16. The results in Fig. 3A for $R_g$ and Fig. 3B show that the equations, describing the scaling corrections,

are in quantitative agreement with simulations. A similar exercise using Eq. 17 using a two parameter fit ($b_h$ and $c_h$) also gives excellent agreement with simulations (Fig. 3C). 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.

We should note that corrections to scaling for IDPs are very difficult to measure using experiments. Besides the coefficients $b_h$ and $b_e$ are not that large. Although $b_h$ and $c_h$ are larger their magnitudes are similar and the result is that they make only small contributions to the Flory scaling. Let the relative correction to $R_g$ scaling be $\delta_g = \sqrt{\frac{b_g}{N^{\Delta_1}}}$. The largest correction is expected for Histatin-5 ($N$=24). Using we find that $\delta_g \approx 0.16$, which implies that the calculated value using Eq. 16 should yield $R_g \approx 1.46$ nm. The experimental value (1.35 nm given in Table S1 in the SI) is roughly 7% smaller. For AP180 the scaling correction to $R_g$ is about 4%. Thus, for IDPs the Flory scaling scaling is, for all practical purposes very accurate, especially considering the errors in obtaining $R_g$ using the Guinier approximation that depends on the range of $q$ values used.

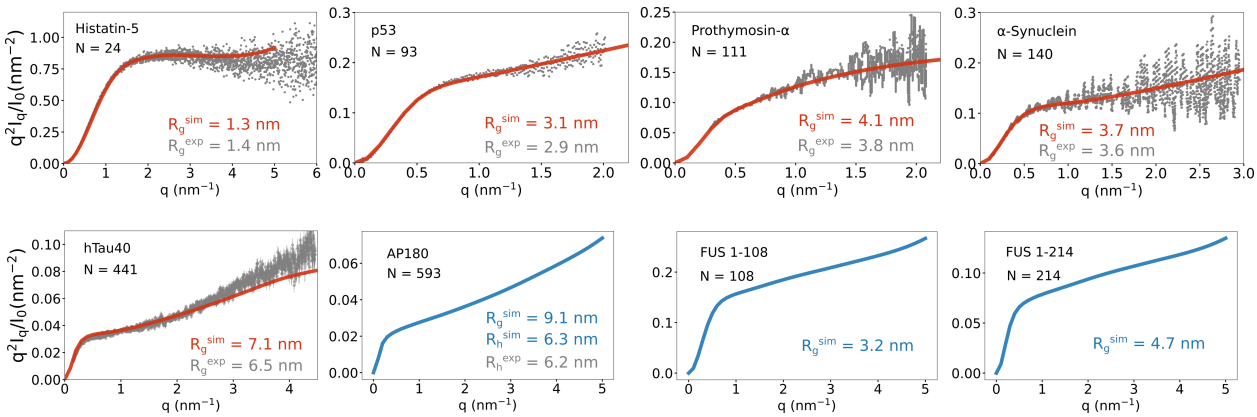


Figure 4: Small angle X-ray scattering for 6 IDPs in Kratky plot (see SI for all other IDPs). Comparison between simulated (red) vs. experimental SAXS (grey). The scattering profiles for AP180, FUS 1-108, and FUS 1-214 (shown in blue) serve as a predictions.

**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 SI). The SAXS profiles for the six IDPs in Fig. 4 are given in Fig. S7 in the SI. The calculated Kratky plots for AP180, FUS 1-108, and FUC 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 SI), despite substantial variations in sequence, length ($N$), and experimental conditions. We should emphasize that there was no attempt to fit the calculated SAXS data to experiments, or carry out ensemble optimization, indicating that our simulated structural ensembles are consistent with the ground truth.

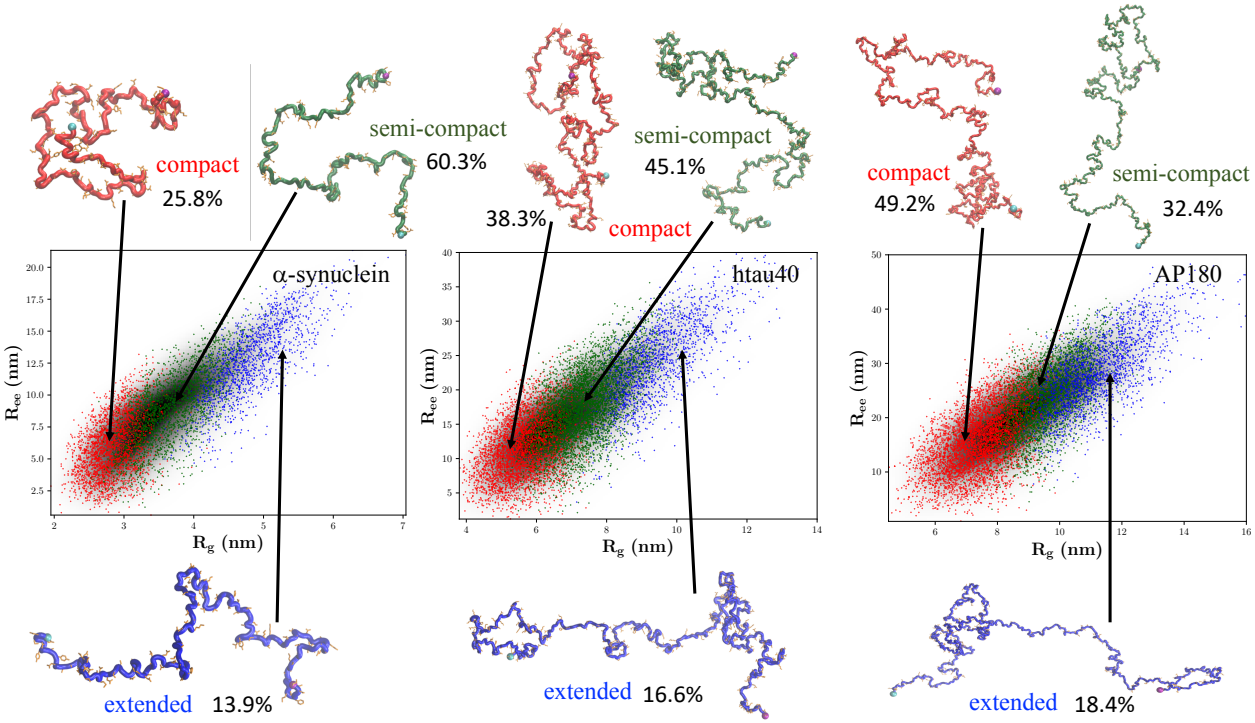


Figure 5: Hierarchical clustering of the conformational ensembles for $\alpha$-synuclein, htau40 and AP180. For each sequence, the conformational landscape is projected into the $R_g$ and $R_{ee}$ coordinates. The relative populations of the different clusters, and corresponding snapshots are shown superimposed 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.

**Sequence-specific heterogeneity of IDP ensembles:** If IDPs mimic SAWs so closely

(Fig. 1, Fig. 2, and Fig. 3) where do sequence effects manifest themselves at the monomer level? To answer this question, we examined 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,[30,33] 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 $\alpha$-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. As is evident from Fig. 5, 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.

**$\alpha$-synuclein:** $\alpha$-synuclein is a 140-residue IDP, which has been extensively studied using experiments[53–58] and computer simulations,[30,58–60] 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.[61] The $\alpha$-synuclein sequence can be partitioned into three segments: the N-terminus (residues 1-40); the non-amyloid $\beta$ 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 $\alpha$-synuclein fibrillization.[62,63]

As shown in Fig. 5, the ensemble of $\alpha$-synuclein conformations partition into three clusters, with semi-compact structures having a population of 60.3% . 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). In particular, contacts within the NAC domain, which forms the core in $\alpha$-synuclein fibrils,[64] appear most frequent. Some of the structures within this cluster also show long-range con-

tacts 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,[55,65] are also evident in some structures within the compact cluster. Many conformations within the semi-compact cluster have weak long-range interactions between the NAC domain and residues 42-55 (often termed as the preNAC domain) within the N-terminus (Fig. S8B). 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 $\alpha$-synuclein aggregation.[66] 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). Our simulations capture the inherent plasticity of $\alpha$-synuclein, and the partitioning of the conformational space is in accord with previous experimental reports.[53,58,67,68]

**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.[69] 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 Parkinson's disease.[70] 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.[71] Importantly, recent works suggest that the conformations relevant for phase separation and aggregation may be encoded within the monomer ensemble of tau,[72,73] much like other amyloidogenic sequences.[33]

Like $\alpha$-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). Long-range contacts between the N-terminus and the PRD, which have been shown to be key mediators of tau phase separation,[74,75] 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 (Fig. S9B). Fully extended conformations, which are only sparsely populated within the tau monomer ensemble, do not exhibit any residue-residue contacts (Fig. S9C).

**AP180:** The endocytic protein AP180 plays an important role in clathrin assembly.[76,77] 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.[78] 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).[79] The C-terminal domain of AP180 has been used as a model system to understand how IDPs induce membrane curvature.[80,81] 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.[82] 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.[79]

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 (6.3 nm) is in near quantitative agreement with the estimate (6.0 nm) from fluorescence correlation spectroscopy experiments (Table S2 in the SI).[79] From the scaling relationship between $R_h$ and $N$, we obtain $R_h \approx 6.6$ nm, which deviates from the experimental value by about 10%. The $R_g$ values from simulations and the Flory scaling relation are 9.0 nm and 9.1 nm, respectively. These specific comparisons confirm that AP180, like other IDPs simulated here, behave as homopolymers in good solvent.

Hierarchical clustering reveals that, unlike $\alpha$-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 $\alpha$-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 (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 (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 (Fig. S11B). The conformations within the extended cluster are mostly RC-like, resulting in a featureless contact map (Fig. S11C), but some of the subfamilies consist of structures exhibiting contacts between residues near the C-terminus (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 SI for details). The SAW ensemble appears more homogeneous, and is segregated into two classes (compact and extended) with

nearly equal equilibrium populations (Fig. S12). Therefore, the more nuanced partitioning of the conformational space displayed in Fig. 5 is a reflection of the sequence.

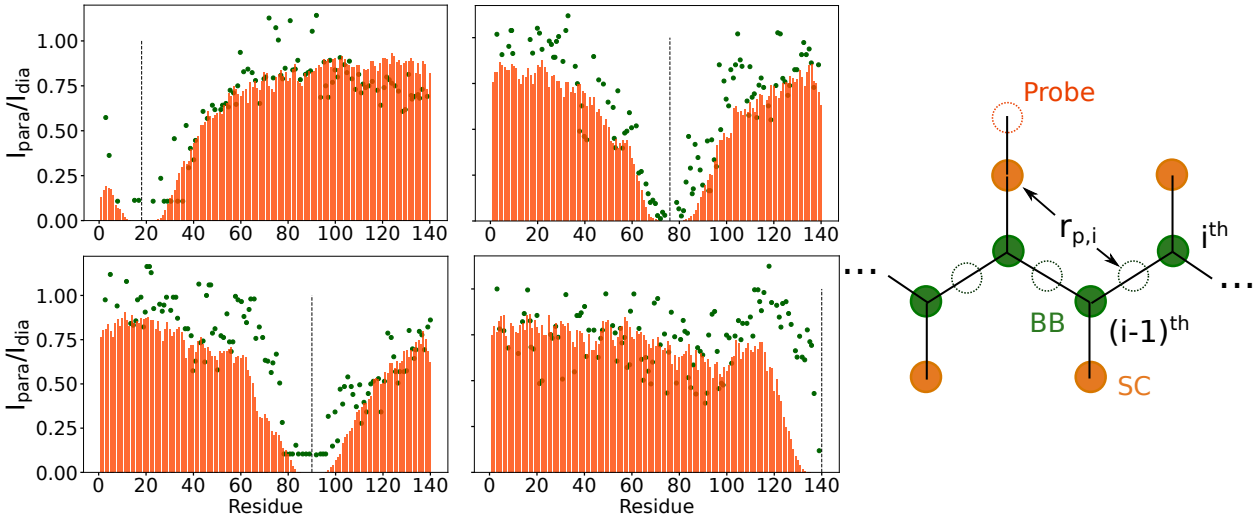


Figure 6: Paramagnetic relaxation enhancement (PRE) of $\alpha$-synuclein. The distance is calculated from the side chain bead where the probe attached to the imaginary amide groups that sit between the backbone C$\alpha$ beads. Calculated PREs (orange) are in good agreement with experimental data (green dots) at 4 different locations of the probe (vertical dashed lines).[53,57,83]

**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.[44] We choose $\alpha$-synuclein to calculate PRE because of the wealth of available experimental and simulated data, which can be directly compared with our simulations.[45,46,53,57,83,84] Fig. 6 shows the intensity ratios of PRE effects for $\alpha$-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 of $\alpha$-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. 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, the mean end-to-end distance, and the hydrodynamic radius as a function of the number of residues obey the Flory scaling law accurately. The shapes of the distributions of the end-to-end distance closely match the universal result for a polymer in a good solvent. (ii) For all practical purposes, the Flory scaling law for $R_g$, $R_e$, and $R_h$ is in excellent agreement with experiments and simulations (correlation between experiments and simulations for $R_g$ is 0.96). Thus, it is not surprising that correction to scaling for $R_g$, $R_e$, and $R_h$ is small even for the 24-residue Histatin-5. Thus, the Flory scaling law suffices to estimate the sizes of IDPs without the need for experiments. (iii) Only by analyzing the ensemble of conformations carefully some of the sequence-dependence emerge. For instance, by probing the conformational ensemble for $\alpha$-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 A$\beta$ peptides[33] and Fused in Sarcoma.[32] (iv) Without tweaking any of the parameters, our simulations reproduce, semi quantitatively, experimental paramagnetic relaxation measurement at four probe locations in $\alpha$-synuclein. All these results show that the SOP-IDP model is transferable.

We have trumpeted the triumph of the SOP model for IDPs. However, one could easily devise a large number of coarse-grained and all atom models, and train the parameters to obtain agreement with IDP properties. So in this sense, the model used here is by not unique. The virtue of the SOP-IDP model is that it has only three parameters, whose values

were calculated by matching the experimental and calculated values of $R_g$ for a few proteins in the training set. As the number of parameters in a model increases, it becomes difficult to calculate their optimum values, and then simulate a large number of sequences spanning a range of $N$, and environmental conditions. It is only by performing such simulations and comparing the results with experiments that the validity of any model can be quantitatively assessed.

## Supporting Information Available

A list of all IDP sequences considered in this work, along with the simulated values of $R_g$, 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; SAXS profiles of IDPs in the Kratky representation; scattering profiles of Histatin-5, p53, Prothymosin-$\alpha$, $\alpha$-synuclein, htau40 and FUS 1-214; dendrograms displaying the hierarchical organization of clusters in the $\alpha$-synuclein and its corresponding SAW ensemble; residue-residue contact maps for the different conformational clusters of $\alpha$-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`. This material is available free of charge via the Internet at `http://pubs.acs.org/`.

## 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 19-00033) and a grant from the Welch Foundation (F-

0019) administered through the Collie-Welch Regents Chair.

The authors declare no competing interest.

# References

(1) Wright, P. E.; Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell. Biol.* **2015**, *16*, 18–29.

(2) Wright, P. E.; Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **1999**, *293*, 321–331.

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

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

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

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

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

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

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

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

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

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

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

(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. A Molecular Grammar Governing Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell* **2018**, *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. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science* **2020**, *367*, 694–699.

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

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

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

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

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

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

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

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

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

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

(26) Baidya, L.; Reddy, G. pH Induced Switch in the Conformational Ensemble of Intrinsically Disordered Protein Prothymosin-$\alpha$ and Its Implications for Amyloid Fibril Formation. *J. Phys. Chem. Lett.* **2022**, *13*, 9589–9598.

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

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

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

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

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

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

(33) Chakraborty, D.; Straub, J. E.; Thirumalai, D. Differences in the free energies between the excited states of A$\beta$40 and A$\beta$42 monomers encode their aggregation propensities. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 19926–19937.

(34) D. Chakraborty, J. E. S.; Thirumalai, D. Energy landscapes of A$\beta$ monomers are sculpted in accordance with Ostwald's rule of stages. *Sci. Adv.* **2023**, *9*, eadd6921.

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

(36) Di Pierro, M.; Elber, R. Automated optimization of potential parameters. *J. Chem. Theory Comput.* **2013**, *9*, 3311–3320.

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

(38) Nocedal, J.; Weight, S. J. *Numerical Optimization*; Springer New York, NY, 2006.

(39) 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. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. 2017.

(40) Honeycutt, J.; Thirumalai, D. The nature of folded states of globular proteins. *Biopolymers* **1992**, *32*, 695–709.

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

(42) Ward, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244.

(43) 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, SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Meth.* **2020**, *17*, 261–272.

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

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

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

(47) 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. Random-coil behaviour and the dimensions of chemically unfolded proteins. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 12491–12496.

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

(49) Clisby, N.; Dünweg, B. High-precision estimate of the hydrodynamic radius for self-avoiding walks. *Phys. Rev. E* **2016**, *94*, 052102.

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

(51) Clisby, N. Accurate Estimate of the Critical Exponent $\nu$ for Self-Avoiding Walks via a Fast Implementation of the Pivot Algorithm. *Phys. Rev. Lett.* **2010**, *104*, 055702.

(52) Honeycutt, J. D.; Thirumalai, D. Static properties of polymer chains in porous media. *J. Chem. Phys.* **1989**, *90*, 4542–4559.

(53) Dedmon, M. M.; Lindorff-Larsen, K.; Christodoulou, J.; Vendruscolo, M.; Dobson, C. M. Mapping long-range interactions in $\alpha$-synuclein using spin-label NMR and ensemble molecular dynamics simulations. *J. Am. Chem. Soc.* **2005**, *127*, 476–477.

(54) Lee, J. C.; Gray, H. B.; Winkler, J. R. Tertiary contact formation in $\alpha$-synuclein probed by electron transfer. *J. Am. Chem. Soc.* **2005**, *127*, 16388–16389.

(55) Trexler, A. J.; Rhoades, E. Single molecule characterization of $\alpha$-synuclein in aggregation-prone states. *Biophys. J.* **2010**, *99*, 3048–3055.

(56) Wu, K. P.; Baum, J. Detection of transient interchain interactions in the intrinsically disordered protein $\alpha$-synuclein by NMR paramagnetic relaxation enhancement. *J. Am. Chem. Soc.* **2010**, *132*.

(57) Bertoncini, C. W.; Jung, Y. S.; Fernandez, C. O.; Hoyer, W.; Griesinger, C.; Jovin, T. M.; Zweckstetter, M. Release of long-range tertiary interactions potentiates aggregation of natively unstructured $\alpha$-synuclein. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 1430–1435.

(58) 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.

(59) Yu, H.; Han, W.; Ma, W.; Schulten, K. Transient $\beta$-hairpin formation in $\alpha$-synuclein monomer revealed by coarse-grained molecular dynamics simulation. *J. Chem. Phys.* **2015**, *143*, 243142.

(60) Graen, T.; Klement, R.; Grupi, A.; Haas, E.; Grubmuller, H. Transient Secondary and Tertiary Structure Formation Kinetics in the Intrinsically Disordered State of $\alpha$-Synuclein from Atomistic Simulations. *ChemPhysChem* **2018**, *19*, 2507–2511.

(61) Vekrellis, K.; Rideout, H. J.; Stefanis, L. The neurobiology of $\alpha$-synuclein. *Mol. Neurobiol.* **2004**, *30*, 1–21.

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

(63) 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. Structure of the toxic core of $\alpha$-synuclein from invisible crystals. *Nature* **2015**, *525*, 486–490.

(64) 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. Solid-state NMR structure of a pathogenic fibril of full-length human $\alpha$-synuclein. *Nat. Struc. Mol. Biol.* **2016**, *23*, 409–415.

(65) Hoyer, W.; Cherny, D.; Subramaniam, V.; Jovin, T. M. Impact of the Acidic C-Terminal Region Comprising Amino Acids 109-140 on $\alpha$-Synuclein Aggregation in Vitro. *Biochemistry* **2004**, *43*, 16233–16242.

(66) 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. A short motif in the N-terminal region of $\alpha$-synuclein is critical for both aggregation and function. *Nat. Struc. Mol. Biol.* **2020**, *27*, 249–259.

(67) Frimpong, A. K.; Abzalimov, R. R.; Uversky, V. N.; Kaltashov, I. A. Characterization of intrinsically disordered proteins with electrospray ionization mass spectrometry:

conformational heterogeneity of $\alpha$-synuclein. *Proteins: Struc. Func. Bioinf.* **2010**, *78*, 714–722.

(68) Ullman, O.; Fisher, K. C.; Stulz, C. M. Explaining the structural plasticity of $\alpha$-synuclein. *J. Am. Chem. Soc.* **2011**, *133*, 19536–19546.

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

(70) Gao, Y. L.; Wang, N.; Sun, F. R.; Cao, X. P.; Yu, J. Y. Tau in neurodegenerative disease. *Ann. Transl. Med.* **2018**, *6*, 175.

(71) Boyko, S.; Surewicz, W. K. Tau liquid-liquid phase separation in neurodegenerative diseases. *Trends. Cell. Biol.* **2022**, *32*, 611–623.

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

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

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

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

(76) Ahle, S.; Ungewickell, E. Purification and properties of a new clathrin assembly protein. *EMBO J.* **1986**, *5*, 3143–3149.

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

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

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

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

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

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

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

(84) Esteban-M, S.; Silvestre-Ryan, J.; Bertoncini, C. W.; Salvatella, X. Identification of fibril-like tertiary contacts in soluble monomeric $\alpha$-synuclein. *Biophys. J.* **2013**, *105*, 1192–1198.