

Inferring properties of disordered chains from FRET transfer efficiencies

Wenwei Zheng,^{1,a),b)} Gül H. Zerze,² Alessandro Borgia,³ Jeetain Mittal,²
Benjamin Schuler,^{3,4,a)} and Robert B. Best^{1,a)}

¹Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-0520, USA

²Department of Chemical and Biomolecular Engineering, Bethlehem, Pennsylvania 18015, USA

³Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

⁴Department of Physics, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

(Received 28 September 2017; accepted 10 January 2018; published online 14 February 2018)

Förster resonance energy transfer (FRET) is a powerful tool for elucidating both structural and dynamic properties of unfolded or disordered biomolecules, especially in single-molecule experiments. However, the key observables, namely, the mean transfer efficiency and fluorescence lifetimes of the donor and acceptor chromophores, are averaged over a broad distribution of donor-acceptor distances. The inferred average properties of the ensemble therefore depend on the form of the model distribution chosen to describe the distance, as has been widely recognized. In addition, while the distribution for one type of polymer model may be appropriate for a chain under a given set of physico-chemical conditions, it may not be suitable for the same chain in a different environment so that even an apparently consistent application of the same model over all conditions may distort the apparent changes in chain dimensions with variation of temperature or solution composition. Here, we present an alternative and straightforward approach to determining ensemble properties from FRET data, in which the polymer scaling exponent is allowed to vary with solution conditions. In its simplest form, it requires either the mean FRET efficiency or fluorescence lifetime information. In order to test the accuracy of the method, we have utilized both synthetic FRET data from implicit and explicit solvent simulations for 30 different protein sequences, and experimental single-molecule FRET data for an intrinsically disordered and a denatured protein. In all cases, we find that the inferred radii of gyration are within 10% of the true values, thus providing higher accuracy than simpler polymer models. In addition, the scaling exponents obtained by our procedure are in good agreement with those determined directly from the molecular ensemble. Our approach can in principle be generalized to treating other ensemble-averaged functions of intramolecular distances from experimental data. <https://doi.org/10.1063/1.5006954>

I. INTRODUCTION

It is increasingly realized that disorder plays a key role in biology, exemplified by the rich variety of functions performed by intrinsically disordered proteins (IDPs), from transcription regulators to molecular chaperones.¹ Studying the structure, dynamics, and function of these disordered polypeptide chains, as well as other disordered biopolymers, requires new experimental methods, and new methods of interpreting the data, since they are inherently averaged over a broad ensemble of heteropolymer configurations.^{2–4} Examples of experiments which can help address this challenge include nuclear magnetic resonance (NMR), which can provide short-range structural information via scalar coupling, chemical shift, and nuclear Overhauser effect (NOE) data,⁵ as well as long-range information by paramagnetic relaxation enhancement (PRE) measurements,⁶ and global information from diffusion coefficient measurements; light-scattering⁷ and two-focus

fluorescence correlation spectroscopy (FCS),^{4,8} which also yield diffusion coefficients and hence the hydrodynamic radius; small-angle X-ray (or neutron) scattering (SAXS or SANS), which directly gives information on inter- and intramolecular pair distance distributions;⁹ and Förster resonance energy transfer (FRET), in particular single-molecule FRET, which probes the distribution of distances between pairs of chromophore labels attached to the molecule of interest, as well as the associated dynamics, and often enables structured and unstructured subpopulations to be separated.^{10,11} Obtaining as detailed as possible a picture of the disordered ensemble would ideally combine information from all of these experiments, if available. This can be done most comprehensively via an explicit ensemble description of the disordered state, either by reweighting of an existing molecular simulation^{2–4,12–17} or by performing an ensemble structural refinement with a very large number of replicas of the system.^{18–23} However, a simpler approach is frequently useful. For example, one may have limited experimental data available so that the result from ensemble simulation or reweighting methods would be strongly dependent on the quality of the simulation and force field used. In this case, a more straightforward analysis of the data, which does not involve running extensive

^{a)}Electronic addresses: wenweizheng@asu.edu; schuler@bioc.uzh.ch; and robertbe@helix.nih.gov.

^{b)}Current address: College of Integrative Sciences and Arts, Arizona State University, Mesa, Arizona 85212, USA.

simulations, may be preferred. In this paper, we consider how to infer ensemble properties from the most common quantities available from FRET data, namely, the mean ratiometric transfer efficiency²⁴ and fluorescence lifetimes.²⁵

As will be discussed in more detail below, we are considering the situation in which the FRET efficiency between donor and acceptor chromophores may be considered to depend only on the distance R between them. In this case, the average transfer efficiency can be obtained by integrating over a given distribution of distances $P(r)$, ideally characterized by a small number of parameters, since often only the mean transfer efficiency is available as an observable. Perhaps the most widely used distribution is a Gaussian (ideal) chain, which has the advantage of having only a single adjustable parameter, usually taken to be the mean-squared distance, $R^2 \equiv \langle r^2 \rangle$, between the FRET donor and acceptor.²⁶ The average radius of gyration, R_g , can in turn be estimated from R using the properties of the polymer model employed. However, it has been recognized that using a Gaussian chain may distort the apparent inter-chromophore distance R when the unfolded protein concerned is close to the excluded volume (EV) limit (scaling exponent $\sim 2/3$).^{27–29} In this case, the very broad $P(r)$ for a Gaussian chain contains an unphysical contribution at shorter distances which would tend to increase the apparent mean FRET efficiency (Fig. 1). In order to match the experimental data for a polypeptide with the properties of a pure EV chain (with no attractive intramolecular interactions), an artificially enlarged R^2 would need to be chosen. This artifact is widely acknowledged, and it has been proposed that using a self-avoiding walk (SAW) model is a better description for protein chains at high denaturant concentrations, where they are typically close to the EV limit.^{28–31} In a recent study of an unfolded and an intrinsically disordered protein in chemical denaturants, we indeed found that applying a SAW model to FRET data at high denaturant concentrations led to inferred average properties (R and R_g) close to those obtained by a global analysis integrating both FRET and SAXS data in an ensemble description based on atomistic simulations, while the result from using a Gaussian chain overestimated these properties.⁴ On the other hand, at low denaturant concentrations, R and R_g inferred from a Gaussian chain model were closer to those from a global analysis than the SAW model. This trend is physically expected, given that unfolded proteins in water often have a scaling exponent characteristic of the Θ -state^{31–33} ($\sim 1/2$), where an ideal chain description is expected to work best.^{31,34} This implies that, while using the same polymer model to infer average distances and R_g from FRET data over a range of solvent conditions may seem consistent, doing so may distort the results because the one chosen model is not valid over the complete range of conditions.

For a given protein, it is in principle possible to determine the polymer scaling exponent which best describes its properties under given conditions of solvent and temperature, for example, by attaching FRET labels at different positions, separated by a variety of chain lengths.³¹ This would allow the appropriate distribution $P(r)$ to be identified. However, the range of accessible sequence separations is limited by the Förster radii of suitable dye pairs, and producing multiple labeling variants presents a considerable additional

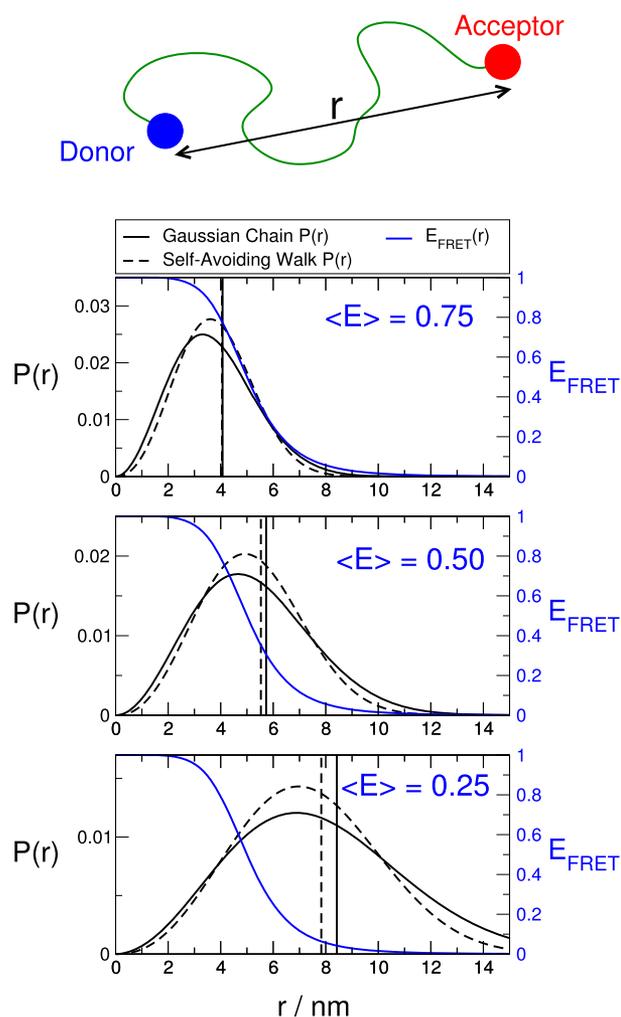


FIG. 1. Determining ensemble properties from mean FRET efficiencies. The top panel illustrates the distance r probed between donor and acceptor attached to a biopolymer of interest. In the lower panels, Gaussian chain and self-avoiding walk end-end distance distributions $P(r)$ are shown for scenarios in which the mean FRET efficiency is 0.75 (top), 0.5 (middle), and 0.25 (bottom). Vertical lines indicate the root mean square end-end distances $R \equiv \langle r^2 \rangle^{1/2}$ for the corresponding $P(r)$. $R_0 = 5$ nm.

experimental burden. Ideally, we desire a method that can be used to infer accurate ensemble properties of the chain from a minimum of available experimental data. In this paper, we propose such a method, based on an approximate theoretical distribution parameterized by the scaling exponent ν of the chain. We test the proposed method against synthetic FRET data calculated from known conformational ensembles. These ensembles are generated using an implicit-solvent model including both excluded volume repulsion as well as realistic attractive interactions within the chain³⁵ or using an all-atom explicit-solvent model with an optimized force field for IDPs.³⁶ We further test the method using experimental FRET data of proteins for which SAXS data are also available⁴ so that the reference radius of gyration can be determined more accurately by two different experimental methods. In all cases considered, we find that our method yields a radius of gyration in good agreement with the reference value, while simultaneously giving an accurate estimate of the polymer scaling exponent.

II. METHODS

To set the stage, we first provide a basic description of the principles behind a FRET experiment (Fig. 1, top panel). A donor and an acceptor chromophore are covalently linked to the molecule of interest at specific sites. The donor chromophore is optically excited, and the excitation energy is either emitted as a photon or transferred to an acceptor chromophore, which then emits a photon. Non-radiative processes can also contribute to the donor or acceptor decay, but they are corrected for in the experiments, so we neglect them here. The efficiency of energy transfer is related to both the distance between the chromophores and their relative orientation. We consider only the situation in which the chromophores are rapidly sampling different orientations on the time scale of the average duration of the donor excited state (or donor lifetime),³⁷ as is frequently the case, especially for unfolded and intrinsically disordered proteins.²⁷ In this situation, the orientational contribution is averaged out, and the transfer efficiency of a given protein configuration with chromophores separated by distance r is given by

$$E(r) = \frac{1}{1 + (r/R_0)^6}, \quad (1)$$

where R_0 is the spectroscopically determined Förster radius.³⁸ Because the transfer efficiency is an average over a highly heterogeneous conformational ensemble, a commonly used method of analysis averages $E(r)$ over a distribution of distances $P(r; \{\xi_i\})$ computed from a polymer model characterized by a set of parameters $\{\xi_i\}$,

$$\langle E \rangle = \int_0^\infty P(r; \{\xi_i\}) E(r) dr. \quad (2)$$

A complementary piece of information is the donor fluorescence lifetime,²⁵ i.e., the mean time between excitation of the donor and emission of a donor photon, which can be related to the mean transfer efficiency $\langle E \rangle$ and its variance σ_c^2 by^{39,40}

$$\langle \tau \rangle = \tau_D \left[1 - \langle E \rangle + \frac{\sigma_c^2}{1 - \langle E \rangle} \right], \quad (3)$$

where the variance of the transfer efficiency, σ_c^2 , is

$$\sigma_c^2 = \int_0^\infty E(r)^2 P(r) dr - \langle E \rangle^2. \quad (4)$$

Note that this analysis requires the inter-dye distance relaxation time to be much longer than the donor fluorescence lifetime, which is usually the case for the chromophores used and the chain lengths accessible in single-molecule FRET investigations of unfolded and intrinsically disordered proteins.²⁷

In the case of FRET measurements on less heterogeneous systems, such as between chromophores attached to a folded protein, the distance distribution is reasonably narrow and hence sufficiently well described by its mean and variance. However, the end-end distance distribution sampled by an unstructured biopolymer is asymmetric and extremely broad so that the shape and the tails of the distribution can have a significant effect on the mean transfer efficiency. To illustrate this effect, we show in Fig. 1 example distance distributions for two polymer models, a Gaussian chain and a SAW, corresponding

to the same mean FRET efficiency. As is evident, the root-mean-square end-end distance $\langle r^2 \rangle^{1/2} \equiv R$ of these distributions can differ considerably between the two polymer models when the mean efficiency is identical, particularly at low $\langle E \rangle$, i.e., very expanded chains. Consequently, using a Gaussian chain to interpret a given set of variations in $\langle E \rangle$ would lead to larger inferred variations in R than use of a SAW, a fact which is now well appreciated.^{4,28,29} While the question of which of these (or other) models is a more accurate reflection of the true distribution has received some attention, neither model is likely to provide a fully adequate description of $P(r)$ across all conditions.⁴

In this paper, we propose to use a more general form for the $P(r)$ of a self-avoiding walk (a polymer which cannot cross itself) that can accommodate a variation of the scaling exponent ν ,^{41,42}

$$P(r) = A \frac{4\pi}{R} \left(\frac{r}{R} \right)^{2+g} \exp \left[-\alpha \left(\frac{r}{R} \right)^\delta \right]. \quad (5)$$

In the above expression, $g = (\gamma - 1)/\nu$ ⁴² in three dimensions ($\gamma \approx 1.1615$),⁴³ $\delta = 1/(1 - \nu)$,⁴¹ and the constants A and α are determined, for given values of ν and R , from the conditions $\int_0^\infty P(r) dr = 1$ and $\int_0^\infty P(r) r^2 dr = R^2$, as required for normalization, and from the definition of R , respectively. For comparison, the distance distribution for a Gaussian chain is

$$P(r) = \left(\frac{3}{2\pi} \right)^{3/2} \frac{4\pi}{R} \left(\frac{r}{R} \right)^2 \exp \left[-\frac{3}{2} \left(\frac{r}{R} \right)^2 \right]. \quad (6)$$

Note that Eq. (5) does not reduce to Eq. (6), even when $\nu = \frac{1}{2}$, due to the factor $\left(\frac{r}{R} \right)^g$, which attenuates $P(r)$ at short distances, as required for a real chain. A value of $\delta > 2$ reduces the contribution of distances greater than R relative to the Gaussian chain. Strictly speaking, Eq. (5) was derived for a real chain in good solvent, but our analysis below indicates that it provides a useful approximation even outside the good solvent regime.

Eq. (5) has two adjustable parameters, ν and R . If both $\langle E \rangle$ and $\langle \tau \rangle$ are available, it may in principle be possible to determine both parameters directly by fitting the model $P(r)$ in Eq. (5) via Eqs. (2) and (3). In the more common situation that only $\langle E \rangle$ is available, we require some additional information. Here, we exploit the knowledge that the mean end-end distance in unfolded proteins approximately follows the scaling law,

$$R = bN^\nu. \quad (7)$$

The prefactor b has been estimated for proteins to be approximately 0.55 nm, a value also supported by the analysis of our simulation data (see below).³¹ With this closure relation, it is possible to solve for both R and ν . Note that Eq. (7) can also be applied to other biopolymers besides unfolded proteins by using an appropriately determined prefactor.

The above discussion outlines how the average end-end distance R can be determined, but often one would like an estimate of the radius of gyration R_g , which is the main quantity obtained from scattering experiments. Note that we use the symbol R_g to refer to the root mean square ensemble average over the radii of gyration r_g of individual conformations, i.e.,

$R_g \equiv \langle r_g^2 \rangle^{1/2}$. For this conversion, we employ the approximate relation⁴⁴

$$\lambda = \frac{R^2}{R_g^2} = \frac{2(\gamma + 2\nu)(\gamma + 2\nu + 1)}{\gamma(\gamma + 1)}. \quad (8)$$

Equations (5) and (8) for self-avoiding walks are approximations, and strictly speaking, only applicable to homopolymers with large chain length. Therefore, the accuracy of the extracted parameters needs to be tested for a model which is a closer approximation to a real heteropolymeric protein chain (or another biopolymer). For this purpose, we utilize both molecular simulations and experimental data.

III. RESULTS AND DISCUSSION

A. Testing the model with implicit solvent simulation data

First, we use implicit-solvent molecular simulation data for three proteins to test the model, including a mutant of the 116-residue R17 spectrin domain^{4,46} which stays unfolded even at low denaturant concentration, the 79-residue IDP activator for thyroid hormone and retinoid receptor (ACTR)^{4,47} (see details in the [supplementary material](#)), and a 100-residue poly-Val peptide. The last sequence is included as a reference homopolymer. While an implicit solvent model is clearly an approximation, we note that the force field used (ABSINTH³⁵) has been shown to capture sequence-dependent variations of

the radius of gyration (and hence, of the scaling exponent) reasonably well in previous studies.^{48,49} Dyes are not present in the simulations in the current work as the effect of dyes on the protein dimensions has been discussed in previous publications.^{50–52} The positions for end-end distances are thus assumed to be the C $^\alpha$ carbon atoms of the terminal residues. By running simulations at different temperatures, we obtained a set of ensembles with different polymer scaling exponents ν covering a wide range from ~ 0.40 to ~ 0.65 [Figs. 2(a)–2(c)]. We determined the exponents from fitting to the dependence of intramolecular distances on sequence separation as described in the [supplementary material](#). Note that due to the finite length of the chain, the scaling exponent can exceed the theoretical EV limit of 0.588.⁵³ The observable $\langle E \rangle$ can be back-calculated from the average of the individual frames of each ensemble (a Förster radius of 5.4 nm is assumed in all cases both for generating synthetic $\langle E \rangle$ and for inferring ensemble properties from it). We utilize $P(r)$ from different polymer models to obtain ν , R , and R_g from this value of $\langle E \rangle$ and then compare them with those determined directly from the simulation coordinates. The polymer models tested are the Gaussian chain, the self-avoiding walk in the EV limit (SAW-EV), the worm-like chain (WLC, see [supplementary material](#) for details), and the SAW with variable ν [SAW- ν , Eqs. (6) and (7)].

In Figs. 2(d)–2(f), we first show that the distance is similarly well recovered for all the models when ν is close to 0.5;

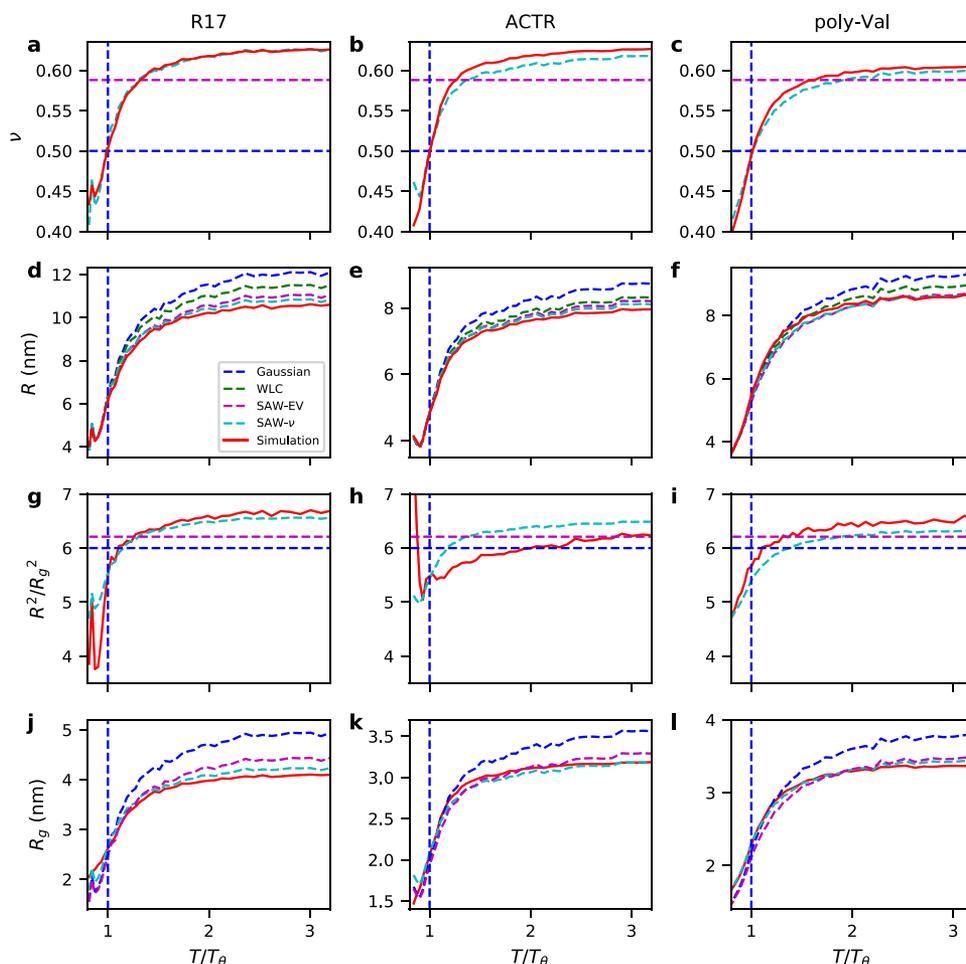


FIG. 2. Recovering ensemble properties from synthetic FRET data calculated from implicit solvent simulations. The recovered properties, obtained by fitting the FRET efficiency $\langle E \rangle$ using different polymer models, are compared with those of the original ensemble used to generate the data, for unfolded R17 (left), ACTR (middle), and poly-Val (right). In (a)–(c), we show the fitted scaling exponent ν , in (d)–(f) the end-end distance R , in (g)–(i) the ratio $\lambda = R^2/R_g^2$ derived from ν , and in (j)–(l) the resulting R_g obtained from R and ν . The legend shows the polymer model used. For the purposes of this plot, T_θ is defined as the temperature for which $\nu = 1/2$. The exponent ν in the simulations is determined from fitting internal distance scaling.⁴⁵

however, for more expanded chains, this is no longer the case and a deviation from the mean distance directly calculated from the simulation is observed. In particular, the Gaussian chain model overestimates the distance most, followed by the WLC and SAW-EV. SAW- ν still overestimates the distance by about 2% in both R17 and ACTR for $T/T_\theta = 3$ but provides the most accurate values. The increased accuracy of the variable- ν model is due to a more realistic $P(R)$, especially when the chain is more expanded than the excluded volume limit for long chains ($\nu > 0.588$), as shown in Fig. 3 and Fig. S1 of the [supplementary material](#). These results suggest that the SAW- ν model provides a good approximation not only for homopolymers but also for finite-length heteropolymers and that a well-chosen $P(R)$ is important for accurately determining chain dimensions from the FRET efficiency.

Importantly, SAW- ν gives the scaling exponent ν in addition to the distance in the procedure of fitting the FRET efficiency. This distinguishes it from the other polymer models used with a constant ν . In Figs. 2(a)–2(c), we show that the polymer scaling exponent recovered from SAW- ν is in remarkable agreement with that calculated directly from the simulations (see [supplementary material](#) for details of calculating ν). This observation suggests that the SAW- ν model is self-consistent in obtaining both R and ν . The variable ν , instead of a constant ν as in the Gaussian chain and SAW-EV, is necessary to describe the scaling behavior of real proteins in different solvent conditions.

To calculate R_g from R , one additional parameter is required, which is the ratio λ between R^2 and R_g^2 . This parameter is also dependent on the polymer model and has recently been suggested to vary significantly for unfolded proteins in different solvent conditions.^{4,54} In Figs. 2(g)–2(i), we show that the change in λ computed directly from the simulation is reproduced much better by the SAW- ν model [Eq. (8)] than by the constant ratios of 6 and 6.26,⁵⁵ for the Gaussian chain and excluded-volume chain models, respectively. As estimated from the SAW- ν model, using a constant λ will introduce an error in R_g of 5% for $\nu = 0.5$ and 2% for $\nu = 0.588$. The more accurate estimate of λ from the variable- ν model consequently leads to final estimates of R_g [Figs. 2(j)–2(l)] that are in much better agreement with those calculated from the simulations than the two most commonly used polymer models, the Gaussian chain model and SAW-EV. Thus, improvements in the

ν -dependent estimates of both $P(R)$ and λ help to infer a more accurate radius of gyration from the FRET efficiency.

While $\nu \in [0.5, 0.588]$ may seem to cover the likely situations for unfolded and disordered proteins, in the case of short chains, ν can lie outside this range, as has been found for real proteins.³¹ To our knowledge, a rigorous theoretical distribution covering the entire range of relevant solvent conditions for the finite-length chains of interest here is currently not available. One polymer model that can describe a $P(r)$ varying between theta solvent and excluded volume limits is that of Oono and Freed,⁵⁶ but the scaling exponent is only allowed to vary between 0.5 and 0.588, so it is difficult to apply it to finite-length chains. We therefore employ Eq. (5) as an approximation for analyzing single-molecule FRET data. Even though the equation is derived to describe a homopolymer, we do not see a significant difference in accuracy when applying the SAW- ν model to natural protein sequences (i.e., ACTR and R17) versus a homopolymer sequence (i.e., poly-Val). This observation suggests that unfolded/intrinsically disordered proteins with well-mixed sequences may still be described sufficiently well with models developed for homopolymers.

Since fluorescence lifetimes also report on the distance distribution $P(r)$ [Eq. (3)], they can be used as an additional or alternative observable to determine R and R_g . As for the FRET efficiency, we back-calculate the donor lifetime $\langle\tau\rangle$ from the simulations, obtain ν , R , and R_g using SAW- ν , and then compare them with their counterparts computed directly from the simulations (Fig. S2 of the [supplementary material](#)). The parameters obtained from $\langle\tau\rangle$ are very similar to those from the ratiometric FRET efficiency. This is not entirely surprising since the variance of the distance distribution is highly correlated with the average distance in the polymer model. The good agreement we observe for the parameters obtained independently from $\langle\tau\rangle$ and $\langle E \rangle$ reflects the good agreement between the shapes of $P(r)$ from the SAW- ν model and from the simulations (Fig. 3 and Fig. S1 of the [supplementary material](#)). Lifetime information can be very useful for identifying the presence of a broad distribution of rapidly interconverting distances in the sample in a model-free manner^{40,57} and provides a valuable experimental control,^{27,58} but the variances of the different polymer distance distributions are too similar to be discriminated reliably (Fig. 3). As a result, a combined analysis of $\langle\tau\rangle$ and $\langle E \rangle$ does not allow both the scaling

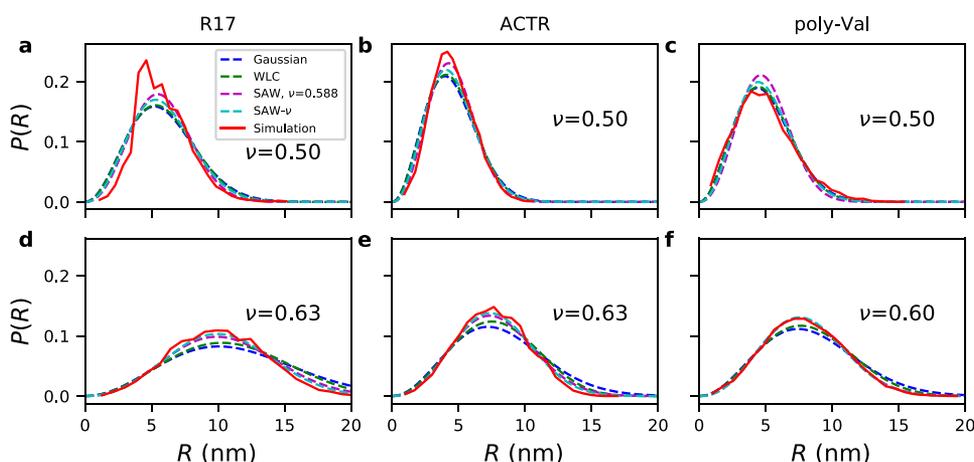


FIG. 3. Examples of distance distributions from different polymer models (see legend) and from the implicit solvent simulations (red) for scaling exponents of ~ 0.5 (top row) or ~ 0.6 (bottom row).

prefactor b and the scaling exponent ν to be determined independently. Here we thus limit our analysis to the mean transfer efficiencies.

Since one experimental observable, the mean transfer efficiency, only allows us to determine one free parameter of the model, we have used a fixed scaling prefactor b taken from earlier FRET experiments.³¹ While one might try to determine b by fitting the average distance R between residue pairs i and j as a function of their sequence separation $|i - j|$ (Fig. S3 of the [supplementary material](#)) to $R = b|i - j|^\nu$, correlations between b and ν make it difficult to accurately determine both parameters independently, especially for the limited range of sequence separations typically accessible for a single protein (a complication that is valid both for the simulations and, even more so, for the experimental data). A prefactor of $b = 0.55$ nm is suitable for recovering the properties of the original ensemble when used in our SAW- ν model. We show in Figs. S4 and S5 of the [supplementary material](#) the results for prefactors of 0.45 and 0.65 nm, respectively, to determine ensemble properties via the SAW- ν model. While the recovered end-end distances are quite robust and thus similar to those obtained using a prefactor of 0.55 nm (Fig. 2), a smaller prefactor tends to overestimate the ratio λ between R^2 and R_g^2 , whereas a larger prefactor underestimates λ , leading to corresponding errors in R_g . This observation suggests that a prefactor of 0.55 nm is close to optimal for use in the SAW- ν model.

B. Testing the model with all-atom explicit solvent simulation data

Our tests so far are based on simulations of unstructured proteins in implicit solvent. However, simulation models in which the solvent is represented explicitly are in principle the most realistic. We have taken advantage of a large recently obtained set of explicit solvent trajectories using the force field Amber03ws³⁶ for proteins, TIP4P/2005⁵⁹ for water, and (in some cases) KBFFs (the scaled Kirkwood-Buff force field)⁶⁰ for denaturants. Previous simulations with the Amber03ws model have resulted in good agreement with FRET,^{36,51} SAXS,^{36,61} and contact quenching⁶² experiments for diverse protein sequences. A particular strength is that it captures local structure formation well,^{36,63} in addition to global properties such as the radius of gyration. While there are some deficiencies with respect to the secondary structure propensity of individual amino acids,^{64,65} for the present purposes, these would only matter for sequences with a preponderance of those residues, i.e., not for the well-mixed sequences studied here. We used 39 trajectories in total from 30 proteins covering a chain length from 11 to 89 in different solvent conditions (i.e., water, urea, and GdmCl) (Table S1 of the [supplementary material](#)). These trajectories cover a wide range of scaling exponents from 0.39 to 0.63, which were determined by fitting the dependence of pairwise distances between C α atoms on sequence separation (Fig. S3 of the [supplementary material](#)). A full list of the simulations and proteins used is given in Table S1, with details in the [supplementary material](#). These force fields have been specifically developed with the aim of accurately representing ensembles of unfolded and disordered proteins and have been validated in several independent studies.^{36,60–63} We therefore believe that these

simulations generate more realistic conformational ensembles of unfolded/intrinsically disordered proteins than most other explicit solvent force fields, which typically yield structures that are too collapsed.⁶⁶ The simulation data have been collected either with unbiased MD simulations of length much greater than typical relaxation times for global parameters such as R_g , or via enhanced sampling methods, as detailed in the [supplementary material](#), in order to obtain representative samples of the equilibrium distributions. For each trajectory, the same framework as in testing the implicit solvent ensembles is applied.

In Fig. 4, we show the accuracy of recovering the scaling exponent, end-end distance, and radius of gyration using synthetic FRET efficiencies calculated from the end-end distances in the original trajectories, compared with the parameters directly obtained from the trajectories. All three polymer models (Gaussian chain, SAW-EV, and SAW- ν) show a good correlation between the recovered and directly calculated parameters [Figs. 4(b) and 4(d)]. However, only SAW- ν is able to capture the changing scaling exponent and ratio λ between R^2 and R_g^2 , and a comparison of the absolute errors of the three models shows that SAW- ν always performs best [Fig. 4(e)]. We have found no correlation between the chain length and the relative errors (Fig. S6 of the [supplementary material](#)), which suggests that SAW- ν , which was derived for long homopolymers, is not sensitive to the chain length once it

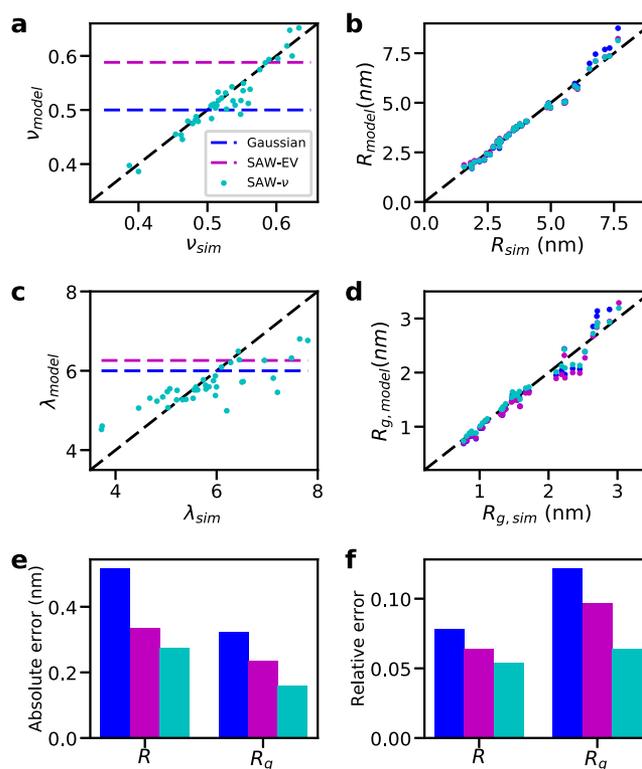


FIG. 4. Recovering ensemble properties from synthetic FRET data, with $\langle E \rangle$ calculated from explicit solvent simulations. We plot the recovered properties versus those of the original ensemble for (a) the scaling exponent ν , (b) the root-mean-squared end-end distance, R , (c) $\lambda = R^2/R_g^2$, and (d) the radius of gyration, R_g . Errors in (e) and (f) are computed as root mean square averages over the different proteins. The absolute error defined as $\langle (x_{\text{model}} - x_{\text{sim}})^2 \rangle^{1/2}$ is shown in (e) and the relative error $\langle ((x_{\text{model}} - x_{\text{sim}})/x_{\text{sim}})^2 \rangle^{1/2}$ is shown in (f). The legend shows the polymer model used, i.e., Gaussian chain model (blue), SAW-EV (magenta), and SAW- ν (cyan).

is greater than ~ 10 . However, the model starts to break down when the chain length is less than ~ 10 , as reflected by a poor fit of intrachain distances as a function of sequence separation to Flory's scaling law (Fig. S7 of the [supplementary material](#)).

C. Testing the model with experimental data

While we consider the simulation ensembles we have studied to be a realistic representation of unfolded states under various conditions, they are nonetheless based entirely on an empirical energy function. One would ideally like to use experimental ensembles, but determining such ensembles from experimental data alone is currently not possible. Instead, experimental data can be used in conjunction with a reasonably accurate force field to generate an ensemble, with the experimental data helping to correct deficiencies in the force

field. In a recent publication,⁴ we have performed both SAXS and FRET measurements on the unfolded R17 and the IDP ACTR. We used an ensemble reweighting method^{4,12} to obtain a unified molecular description of both experimental data sets, together with a simulation model (the ABSINTH force field³⁵). In these ensembles, the radius of gyration and end-end distance are strongly constrained by the SAXS and FRET data used to generate them and rely less on the properties of the force field. The reweighted ensembles from both SAXS and FRET experimental data therefore serve as a reference that can be used to test the accuracy of the SAW- ν model in interpreting real experimental data. This comparison can also tell us whether a good estimate of R_g could have been obtained by a simpler method than ensemble reweighting.

In Fig. 5, we present the comparison with the reweighted ensembles. Note that we have four data sets comprising

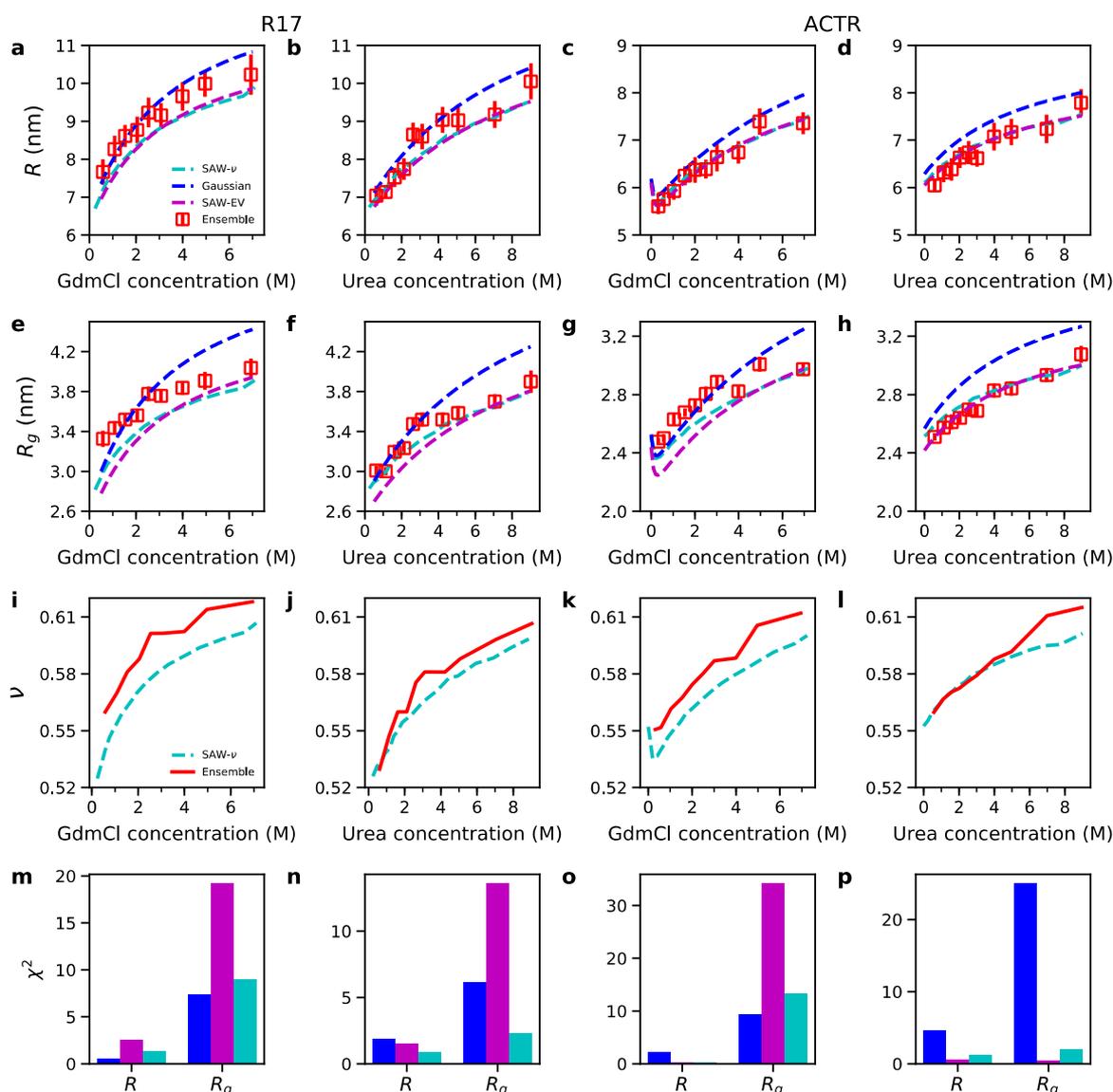


FIG. 5. Validation using ensembles determined from both SAXS and FRET data (proteins labeled with Alexa 488 and 594, $R_0 = 5.4$ nm) at different denaturant concentrations (GdmCl, urea). Ensemble properties estimated using polymer models from the FRET data alone are compared with those computed from molecular ensembles ("Ensemble" in legend) generated to match both SAXS and FRET data. We show R , R_g , and the scaling exponent recovered from FRET efficiency at each denaturant concentration, and χ^2 defined as $\langle ((x_{\text{ensemble}} - x_{\text{sim}})/\sigma_{\text{ensemble}})^2 \rangle$ of all denaturant concentrations for the proteins R17 and ACTR. The legend shows the polymer model used, i.e., Gaussian chain model is in blue; SAW-EV in magenta; and SAW- ν in cyan.

unfolded R17 in urea, unfolded R17 in guanidinium chloride (GdmCl), ACTR in urea, and ACTR in GdmCl. On average, the SAW- ν model provides the most accurate estimate of R_g and R over all experimental conditions (Table S2 of the [supplementary material](#)). The SAW-EV model performs much worse than the other two models in three cases, whereas the Gaussian chain model performs much worse in one case; this is likely to be a consequence of the respective systems being closer to the ideal- or EV-chain limits. By including the variability in the scaling exponent explicitly in the model, SAW- ν is better able to describe all of these cases, especially the ones exhibiting a crossover from near- Θ to good solvent conditions upon increasing denaturant concentration. The most significant discrepancy is a slight underestimation of the scaling exponent for the experiments in guanidinium chloride. However, the discrepancy is in the range of the deviations observed in the analysis of the large set of explicit-solvent simulations [Fig. 4(a)] and may thus reflect the limits of the method in terms of accuracy with a single prefactor. We have also compared the model with reference ensembles obtained only from FRET measurements, since the method is introduced here primarily for the analysis of FRET experiments. In Fig. S8 of the [supplementary material](#), we show that the SAW- ν yields a χ^2 less than 1 in most cases, suggesting that the method is appropriate to achieve an estimate of R and R_g within experimental error. In addition to R_g , the scaling exponent obtained from SAW- ν , which is a constant in the other models, is in reasonable agreement with that calculated directly from the reweighted ensemble. This suggests that SAW- ν captures the scaling behavior of the protein and that a ν -dependent $P(R)$ provides an improved description of unfolded or intrinsically disordered proteins compared to models assuming ν to be constant.

IV. CONCLUSION

We have demonstrated a straightforward scheme to estimate more accurately the end-end distance R and radius of gyration R_g directly from FRET measurements on unfolded proteins. The method makes use of an approximate expression for the $P(r)$ of an excluded volume chain due to des Cloizeaux.^{42,67} Not only is the scheme more accurate than polymer models commonly used to date but it also returns the scaling exponent of the chain. These advantages come partly via the assumption of a given scaling prefactor. The value we have chosen is based on experimental estimates, and its use has been validated here via the agreement between inferred properties such as R , R_g , R^2/R_g^2 and their true values in our simulation ensembles. Thus it appears that the variation of the optimal prefactor for typical IDP sequences is small; a possible direction for future improvement may be the inclusion of a sequence-dependent prefactor to accommodate unusual sequences and the effects of different co-solvents. For example, we find that a smaller prefactor of 0.45 nm is better suited for extracting properties of a poly-Ala sequence via SAW- ν (results not shown). Sequences rich in proline (polyproline being an extreme example) would on the other hand be expected to have a larger prefactor.^{37,68} Nonetheless, it appears that a common value of 0.55 nm is a good first

approximation for typical IDP sequence compositions, as the ones studied here. Similarly, in order to apply the method to other biomolecules, for example, RNA, it would be necessary to determine the prefactor for those cases. This could be done either experimentally³¹ or via simulation models such as those used here. In addition, our approach can also be applied to interpreting intramolecular distances from other experimental methods, if the system of interest can be described by the polymer model.

Like the Gaussian chain and SAW-EV models, the SAW- ν model is strictly applicable only to homopolymers. Proteins are heteropolymers, and it has been shown that the patterning of sequence features such as charge can have significant effects on their global properties,⁴⁹ which clearly would not be captured by any of the homopolymer theories discussed here. Heteropolymer effects have also been suggested to be important for explaining the discrepancy between SAXS and FRET experiments as applied to denatured proteins.^{54,69,70} Incorporating such effects would in general require a hybrid approach in which the experimental data are used to refine a reasonably accurate simulation ensemble. However, running such simulations adds computational cost and complexity to the analysis, making it less widely accessible. Provided that the protein under consideration has a reasonably well-mixed sequence, it should still be possible to approximate its global properties with a homopolymer theory. How well-mixed does the sequence need to be? We have studied here a wide range of intrinsically disordered and unfolded proteins, and we find empirically that indeed the approximate homopolymer theory applied here is sufficient to obtain remarkably accurate estimates in all cases considered. In fact, application of the SAW- ν model to a poly-Val homopolymer yields results of similar accuracy for R and R_g as those obtained for the natural protein sequences. Of course, a pronounced patterning of sequence properties as found in some IDPs⁷¹ likely would cause the method to fail. Signs of such cases could be a fitted scaling exponent outside the range [0.4-0.65] most commonly observed for unfolded and disordered proteins or a discrepancy between properties estimated from the ratiometric FRET efficiency and the donor fluorescence lifetime. The most stringent experimental test is to probe different segments of the chain by varying the positions of the FRET donor and acceptor in the protein and seeing whether they can be described globally with a single set of fit parameters.^{4,28,72,73} Discrepancies would suggest that a homopolymer $P(r)$ cannot provide a self-consistent interpretation of the data and that more detailed analysis is needed, including, e.g., atomistic simulations or ensemble reweighting, ideally combined with additional segment-specific experimental information, e.g., from NMR.^{3,73}

The SAW- ν method allows a scaling exponent to be determined using a FRET measurement from a single labeled variant of a protein, in addition to a more accurate estimate of the dye separation R and the radius of gyration R_g than established methods. This raises the question of which is the optimal pair of residues to label, from the perspective of this method. We suggest based on our results that choosing a pair such that their average separation R is comparable to the Förster radius R_0 yields the most accurate ensemble properties. Choosing an

average separation much larger or smaller than R_0 will make the results more sensitive to the accuracy of the tails of the distance distribution $P(r)$ and hence less robust than choosing R close to R_0 .

After submission of our work, a novel method of analyzing small-angle X-ray scattering (SAXS) experiments was published which also allows a scaling exponent to be determined from a single experiment.⁷⁴ An approach similar to the SAW- v model should also be applicable to analyzing SAXS experiments to obtain both the scaling exponent and R_g . We are currently investigating this possibility.

SUPPLEMENTARY MATERIAL

See [supplementary material](#) for supporting methods, figures and tables.

ACKNOWLEDGMENTS

We thank Andrea Soranno for many insightful discussions. R.B. and W.Z. were supported by the intramural research program of the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health. This work utilized the computational resources of the NIH HPC Biowulf cluster (<http://hpc.nih.gov>). B.S. was supported by the Swiss National Science Foundation. Research reported in this publication was supported in part by NIGMS R01GM118530. Use of the high-performance computing capabilities of the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation, Project No. TG-MCB120014, is also gratefully acknowledged.

- ¹P. E. Wright and H. J. Dyson, *Nat. Rev. Mol. Cell Biol.* **16**, 18–29 (2015).
- ²D. H. Brookes and T. Head-Gordon, *J. Am. Chem. Soc.* **138**(13), 4530–4538 (2016).
- ³M. Schwalbe, V. Ozenne, S. Bibow, M. Jaremko, M. Gajda, M. Ringkjøbing-Jensen, J. Biernat, S. Becker, E. Mandelkow, M. Zweckstetter, and M. Blackledge, *Structure* **22**, 238–249 (2014).
- ⁴A. Borgia, W. Zheng, K. Buholzer, M. B. Borgia, A. Schuler, H. Hofmann, A. Soranno, D. Nettels, K. Gast, A. Grishaev, R. B. Best, and B. Schuler, *J. Am. Chem. Soc.* **138**(36), 11714–11726 (2016).
- ⁵S. Meier, M. Blackledge, and S. Grzesiek, *J. Chem. Phys.* **128**, 052204 (2008).
- ⁶C. C. Jao, A. Der-Sarkissian, J. Chen, and R. Langen, *Proc. Natl. Acad. Sci. U. S. A.* **101**(22), 8331–8336 (2004).
- ⁷A. Nöppert, K. Gast, M. Müller-Frohne, D. Zirwer, and G. Damaschun, *FEBS Lett.* **380**, 179–182 (1996).
- ⁸T. Dertinger, V. Pacheco, I. Von der Hocht, R. Hartmann, I. Gregor, and J. Enderlein, *Chem. Phys. Chem.* **8**, 433–443 (2007).
- ⁹M. A. Graewert and D. I. Svergun, *Curr. Opin. Struct. Biol.* **23**, 748–754 (2013).
- ¹⁰X. Michalet, S. Weiss, and M. Jager, *Chem. Rev.* **106**(5), 1785–1813 (2006).
- ¹¹B. Schuler and H. Hofmann, *Curr. Opin. Struct. Biol.* **23**(1), 36–47 (2013).
- ¹²G. Hummer and J. Koefinger, *J. Chem. Phys.* **143**(24), 243150 (2015).
- ¹³O. F. Lange, N.-A. Lakomek, C. Fares, G. F. Schröder, K. F. A. Walter, S. Becker, J. Meiler, H. Grubmüller, C. Griesinger, and B. L. De Groot, *Science* **320**, 1471–1475 (2008).
- ¹⁴E. Boura, B. Rózycki, D. Z. Herrick, H. S. Chung, J. Vecer, W. A. Eaton, D. S. Cafiso, G. Hummer, and J. H. Hurley, *Proc. Natl. Acad. Sci. U. S. A.* **108**, 9437–9442 (2011).
- ¹⁵J. R. Allison, R. C. Rivers, J. C. Christodoulou, M. Vendruscolo, and C. M. Dobson, *Biochemistry* **53**, 7170–7183 (2014).
- ¹⁶C. K. Fisher, A. Huang, and C. M. Stultz, *J. Am. Chem. Soc.* **132**, 14919–14927 (2010).
- ¹⁷W. Boomsma, J. Ferkinghoff-Borg, and K. Lindorff-Larsen, *PLoS Comput. Biol.* **10**, e1003406 (2014).

- ¹⁸K. Lindorff-Larsen, R. B. Best, M. A. Depristo, C. M. Dobson, and M. Vendruscolo, *Nature* **433**, 128–132 (2005).
- ¹⁹R. B. Best and M. Vendruscolo, *J. Am. Chem. Soc.* **126**, 8090–8091 (2004).
- ²⁰J. Kuriyan, K. Osapay, S. K. Burley, A. T. Brunger, W. A. Hendrickson, and M. Karplus, *Proteins* **10**, 340–358 (1991).
- ²¹V. Venditti, T. Egner, and G. M. Clore, *Chem. Rev.* **116**(11), 6305–6322 (2016).
- ²²W. Rieping, M. Habeck, and M. Nilges, *Science* **309**, 303–306 (2005).
- ²³B. Roux and J. Weare, *J. Chem. Phys.* **138**, 084107 (2013).
- ²⁴A. A. Deniz, T. A. Laurence, M. Dahan, D. S. Chemla, P. G. Schultz, and S. Weiss, *Annu. Rev. Phys. Chem.* **52**, 233–253 (2001).
- ²⁵E. Sisamakias, A. Valeri, S. Kalinin, P. J. Rothwell, and C. A. M. Seidel, *Methods Enzymol.* **475**, 455–514 (2010).
- ²⁶P. J. Flory, *Principles of Polymer Chemistry* (Cornell University Press, Ithaca and London, 1953).
- ²⁷B. Schuler, A. Soranno, H. Hofmann, and D. Nettels, *Annu. Rev. Biophys.* **45**, 207–231 (2016).
- ²⁸E. O'Brien, G. Morrison, B. R. Brooks, and D. Thirumalai, *J. Chem. Phys.* **130**, 124903 (2009).
- ²⁹J. Song, G.-N. Gomes, C. C. Gradinaru, and H.-S. Chan, *J. Phys. Chem. B* **119**, 15191–15202 (2015).
- ³⁰J. E. Kohn, I. S. Millett, J. Jacob, B. Zagrovic, T. M. Dillon, N. Cingel, R. S. Dohager, S. Seifert, P. Thiagarajan, T. R. Sosnick, M. Z. Hasan, V. S. Pandey, I. Ruczinski, S. Doniach, and K. W. Plaxco, *Proc. Natl. Acad. Sci. U. S. A.* **101**, 12491–12496 (2004).
- ³¹H. Hofmann, A. Soranno, A. Borgia, K. Gast, D. Nettels, and B. Schuler, *Proc. Natl. Acad. Sci. U. S. A.* **109**(40), 16155–16160 (2012).
- ³²B.-Y. Ha and D. Thirumalai, *Phys. Rev. A* **46**(6), R3012–R3015 (1992).
- ³³A. H. Mao, S. L. Crick, A. Vitalis, C. L. Chicoine, and R. V. Pappu, *Proc. Natl. Acad. Sci. U. S. A.* **107**, 8183–8188 (2010).
- ³⁴K. A. Merchant, R. B. Best, J. M. Louis, I. V. Gopich, and W. A. Eaton, *Proc. Natl. Acad. Sci. U. S. A.* **104**, 1528–1533 (2007).
- ³⁵A. Vitalis and R. V. Pappu, *J. Comput. Chem.* **30**, 673–699 (2008).
- ³⁶R. B. Best, W. Zheng, and J. Mittal, *J. Chem. Theory Comput.* **10**, 5113–5124 (2014).
- ³⁷B. Schuler, E. A. Lipman, P. J. Steinbach, M. Kumke, and W. A. Eaton, *Proc. Natl. Acad. Sci. U. S. A.* **102**, 2754–2759 (2005).
- ³⁸T. Förster, *Ann. Phys.* **437**, 55–75 (1948).
- ³⁹I. V. Gopich and A. Szabo, *Proc. Natl. Acad. Sci. U. S. A.* **109**, 7747–7752 (2012).
- ⁴⁰H. S. Chung, J. M. Louis, and I. V. Gopich, *J. Phys. Chem. B* **120**, 680–699 (2016).
- ⁴¹M. E. Fisher, *J. Chem. Phys.* **44**, 616–622 (1966).
- ⁴²J. des Cloizeaux, *Phys. Rev. A* **10**(5), 1665–1669 (1974).
- ⁴³J. C. Le Guillou and J. Zinn-Justin, *Phys. Rev. Lett.* **39**(2), 95–98 (1977).
- ⁴⁴T. A. Witten and L. Schäfer, *J. Phys. A: Math. Gen.* **11**(9), 1843–1854 (1978).
- ⁴⁵H. T. Tran, A. Mao, and R. V. Pappu, *J. Am. Chem. Soc.* **130**, 7380–7392 (2008).
- ⁴⁶A. Borgia, B. G. Wensley, A. Soranno, D. Nettels, M. B. Borgia, A. Hoffmann, S. H. Pfeil, E. A. Lipman, J. Clarke, and B. Schuler, *Nat. Commun.* **3**, 1195 (2012).
- ⁴⁷S. J. Demarest, M. Martinez-Yamout, J. Chung, H. Chen, W. Xu, H. J. Dyson, R. M. Evans, and P. E. Wright, *Nature* **415**, 549–553 (2002).
- ⁴⁸R. Wuttke, H. Hofmann, D. Nettels, M. B. Borgia, J. Mittal, R. B. Best, and B. Schuler, *Proc. Natl. Acad. Sci. U. S. A.* **111**, 5213–5218 (2014).
- ⁴⁹R. K. Das and R. V. Pappu, *Proc. Natl. Acad. Sci. U. S. A.* **110**(33), 13392–13397 (2013).
- ⁵⁰R. B. Best, H. Hofmann, D. Nettels, and B. Schuler, *Biophys. J.* **108**(11), 2721–2731 (2015).
- ⁵¹W. Zheng, A. Borgia, K. Buholzer, A. Grishaev, B. Schuler, and R. B. Best, *J. Am. Chem. Soc.* **138**(36), 11702–11713 (2016).
- ⁵²G. H. Zerze, R. B. Best, and J. Mittal, *Biophys. J.* **107**, 1654–1660 (2014).
- ⁵³K. Kremer and K. Binder, *Comput. Phys. Rep.* **7**, 261–310 (1988).
- ⁵⁴G. Fuentes, N. Banterlea, K. M. Ruff, A. Chowdhury, D. Mercadante, C. Koehler, M. Kachala, G. E. Girona, S. Milles, A. Mishra, P. R. Onck, F. Gräter, S. Esteban-Martin, R. V. Pappu, D. I. Svergun, and E. A. Lemke, *Proc. Natl. Acad. Sci. U. S. A.* **114**(31), E6342–E6351 (2017).
- ⁵⁵L. Schäfer, *Excluded Volume Effects in Polymer Solutions: As Explained by the Renormalization Group* (Springer Science & Business Media, 2012).
- ⁵⁶Y. Oono and K. F. Freed, *J. Phys. A: Math. Gen.* **15**(6), 1931–1950 (1982).
- ⁵⁷S. Kalinin, A. Valeri, M. Antonik, S. Felekyan, and C. A. M. Seidel, *J. Phys. Chem. B* **114**(23), 7983–7995 (2010).

- ⁵⁸A. Soranno, B. Buchli, D. Nettels, R. R. Cheng, S. Muller-Spath, S. H. Pfeil, A. Hoffmann, E. A. Lipman, D. E. Makarov, and B. Schuler, *Proc. Natl. Acad. Sci. U. S. A.* **109**(44), 17800–17806 (2012).
- ⁵⁹J. L. F. Abascal and C. Vega, *J. Chem. Phys.* **123**, 234505 (2005).
- ⁶⁰W. Zheng, A. Borgia, M. B. Borgia, B. Schuler, and R. B. Best, *J. Chem. Theory Comput.* **11**, 5543–5553 (2015).
- ⁶¹J. Henriques, C. Cragnell, and M. Skepö, *J. Chem. Theory Comput.* **11**(7), 3420–3431 (2015).
- ⁶²G. H. Zerze, J. Mittal, and R. B. Best, *Phys. Rev. Lett.* **116**(6), 068102 (2016).
- ⁶³A. E. Conicella, G. H. Zerze, J. Mittal, and N. L. Fawzi, *Structure* **24**, 1537–1549 (2016).
- ⁶⁴R. B. Best, D. De Sancho, and J. Mittal, *Biophys. J.* **102**, 1462–1467 (2012).
- ⁶⁵F. Meng, M. M. J. Bellaiche, J.-Y. K. Kim, G. H. Zerze, R. B. Best, and H. S. Chung, “Highly disordered amyloid-beta monomer probed by single-molecule FRET and MD simulation,” *Biophys. J.* (in press).
- ⁶⁶S. Piana, J. L. Klepeis, and D. E. Shaw, *Curr. Opin. Struct. Biol.* **24**, 98–105 (2014).
- ⁶⁷H. T. Tran and R. V. Pappu, *Biophys. J.* **91**, 1868–1886 (2006).
- ⁶⁸R. B. Best, K. A. Merchant, I. V. Gopich, B. Schuler, A. Bax, and W. A. Eaton, *Proc. Natl. Acad. Sci. U. S. A.* **104**, 18964–18969 (2007).
- ⁶⁹K. M. Ruff and A. S. Holehouse, *Biophys. J.* **113**(5), 971–973 (2017).
- ⁷⁰J. Song, G. N. Gomes, T. Shi, C. C. Gradinaru, and H. S. Chan, *Biophys. J.* **113**(5), 1012–1024 (2017).
- ⁷¹R. K. Das, K. M. Ruff, and R. V. Pappu, *Curr. Opin. Struct. Biol.* **32**, 102–112 (2015).
- ⁷²A. Hoffmann, A. Kane, D. Nettels, D. E. Hertzog, P. Baumgartel, J. Lengefeld, G. Reichardt, D. A. Horsley, R. Seckler, O. Bakajin, and B. Schuler, *Proc. Natl. Acad. Sci. U. S. A.* **104**(1), 105–110 (2007).
- ⁷³M. Aznauryan, L. Delgado, A. Soranno, D. Nettels, J. R. Huang, A. M. Labhardt, S. Grzesiek, and B. Schuler, *Proc. Natl. Acad. Sci. U. S. A.* **113**(37), E5389–E5398 (2016).
- ⁷⁴J. A. Riback, M. A. Bowman, A. M. Zmyslowski, C. R. Knoverek, J. M. Jumper, J. R. Hinshaw, E. B. Kaye, K. F. Freed, P. L. Clark, and T. R. Sosnick, *Science* **358**(6360), 238–241 (2017).

Supporting Information for: “Inferring Properties of Disordered Chains from FRET Transfer Efficiencies”

Wenwei Zheng^{1,5}, Gul H. Zerze², Alessandro Borgia³, Jeetain Mittal²,
Benjamin Schuler^{3,4}, Robert B. Best¹

¹Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-0520, USA.

²Department of Chemical and Biomolecular Engineering, Bethlehem, PA 18015, USA

³Department of Biochemistry, ⁴Department of Physics, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland. ⁵Current address: College of Integrative Sciences and Arts, Arizona State University, Mesa, Arizona 85212, USA.

SI Text

Implicit solvent simulations. The method of generating molecular ensembles using the implicit solvent model ABSINTH¹ is the same as in our recent publication². Briefly, temperature replica exchange Monte Carlo was run for 1.3 million steps using the CAMPARI program, spanning a range of 280 to 2446 K with 64 replicas, starting from a fully extended configuration. The first 0.3 million steps were for equilibration, and protein configurations were recorded every 100 steps in the production run, for a total of 10,000 frames, at each temperature. The sequences of R17 SAXS/DLS and ACTR SAXS/DLS in Table S1 of Ref² are used.

Explicit solvent simulations.

39 all-atom explicit solvent simulation trajectories from multiple proteins in different solvent conditions are used to test the model. These simulations come from multiple previous publications, as listed in Table S1³⁻⁴, with the same Amber03ws force field⁵ for the protein, the TIP4P/2005 water model⁶ and KBFFs model⁷ for urea and GdmCl.

Simulations of C[AGQ]_nW peptide and ACTR were run using long simulations at constant temperature maintained by a Langevin thermostat and pressure of 1 bar maintained by a Parrinello-Rahman barostat⁸. The electrostatic forces were computed with particle-mesh Ewald summation⁹. Simulations of all other peptides were performed following the parallel-tempering in well-tempered ensembles (PTWTE) enhanced sampling technique using the Gromacs 4.6.7 simulation code¹⁰ and Plumed 2.1.1¹¹. The well-tempered ensemble (WTE)¹² amplifies the fluctuations of the potential energy, maintaining the same potential energy average, therefore, combined with the standard parallel tempering (PT)¹³, the combined method (PTWTE) reduces the number of replicas required to have a reasonable exchange acceptance¹². Initial coordinates are generated as extended conformers using the CHARMM program¹⁴ and briefly compacted in a gas phase simulation for 1000 steps. Solvated coordinates are equilibrated in NVT (T=300 K) ensembles for 200 ps followed by NPT (P=1 bar, T=300 K) equilibration for 200 ns. Prior to starting PTWTE simulations, unbiased simulation of each replica is performed for 200 ps to equilibrate the potential energy at each temperature. We applied PTWTE where potential energy of the system is biased using 500 kJ/mol sigma, 1.0 kJ/mol initial Gaussian height where Gaussian potentials are added at every 2000 steps with a bias factor of 20. Simulations are performed in NVT ensembles using a stochastic dynamics algorithm¹⁵ with a 1 ps⁻¹ friction coefficient. 12-16 replicas distributed between 300 K to 500 K are used depending on the system size to achieve an average 20-30% exchange acceptance. Only the 300K trajectories are used for analysis.

Details of equilibration length and production run length are shown in Table S1. The scaling exponents of all 35 trajectories are shown in Figure S4 and Table S1. Most of the foldable proteins remained unfolded on the time scale of simulation (i.e. NTL9 and Ubiquitin); for the remainder, only the unfolded parts of the trajectories were used for further analysis (i.e. GB1 and Trp-cage).

Scaling exponent. The scaling exponent of each molecular ensemble is obtained by fitting the average squared distances between all pairs of residues with a sequence separation of $|i-j|$ as a function of the sequence separation $|i-j|$ to the equation $\langle R_{ij}^2 \rangle = b^2|i-j|^{2\nu}$,¹⁶ in which the prefactor b has previously been determined to be approximately 0.55 nm for proteins¹⁷. In a previous work (Figure S16 of Ref. 4), we have shown that the scaling exponent can vary when changing the fitting range of $|i-j|$, however, there is a plateau when we start fitting from $|i-j|=0$ to $|i-j|=30$. So in the current work, we always fit the scaling exponent through full range of $|i-j|$.

Worm-like chain (WLC) model. The distance distribution function of a worm-like chain model¹⁸ can be approximated as

$$P(r) = C_1 \frac{4\pi r^2}{l_c^2 \left[1 - \left(\frac{r}{l_c}\right)^2\right]^{9/2}} \exp \left[-\frac{3l_c}{4l_p \left[1 - \left(\frac{r}{l_c}\right)^2\right]} \right]$$

in which $l_c = Nl_b$ is the contour length and C_l is the normalization factor¹⁹. Then by providing the FRET efficiency, the persistence length l_p and consequently the average squared distance can be solved for numerically.

SI Figures

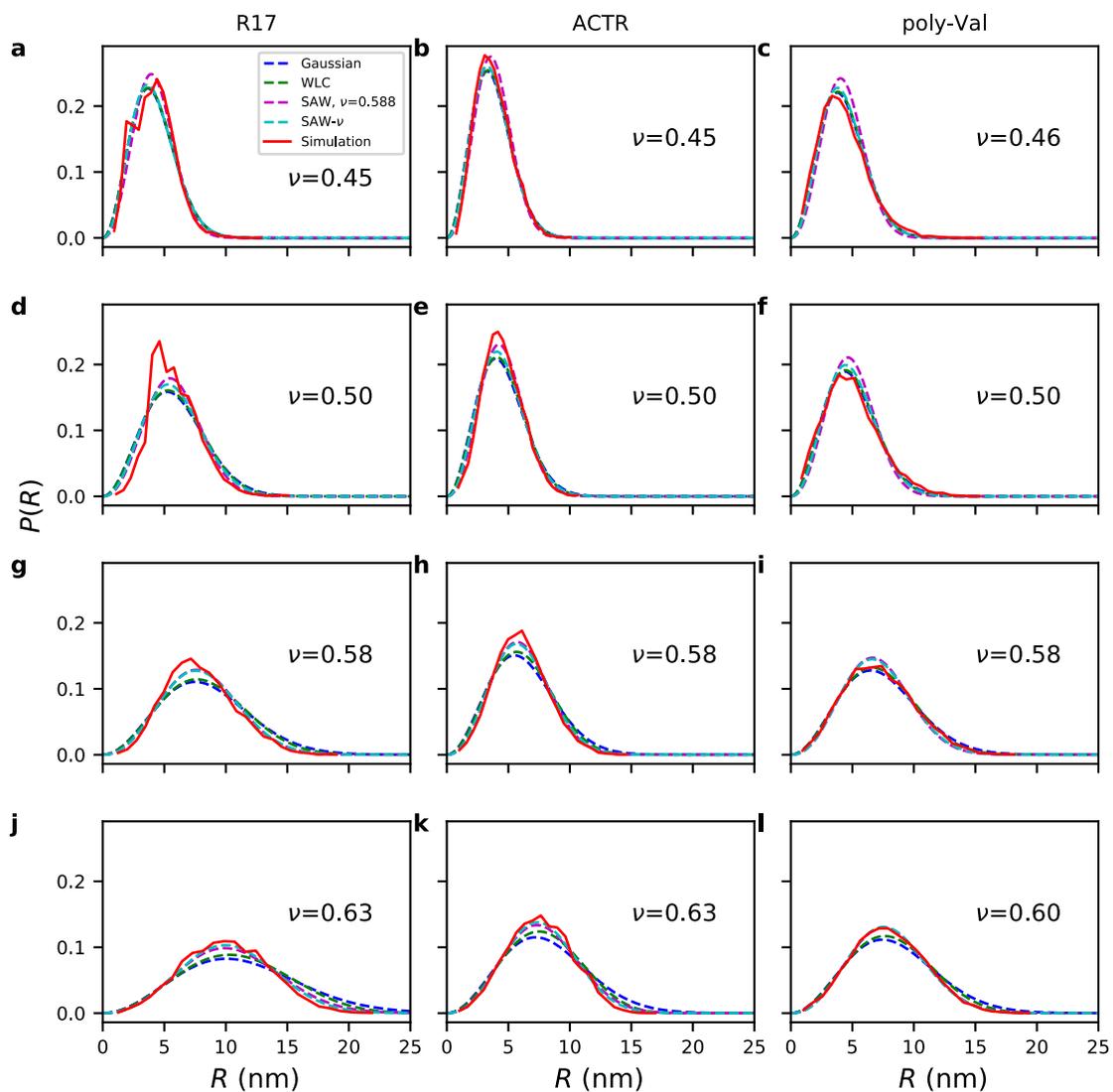


FIG. S1. Distance distributions recovered from FRET efficiency using different polymer models. The legend shows the polymer model used. Figures are arranged top to bottom in order of increasing polymer scaling exponent.

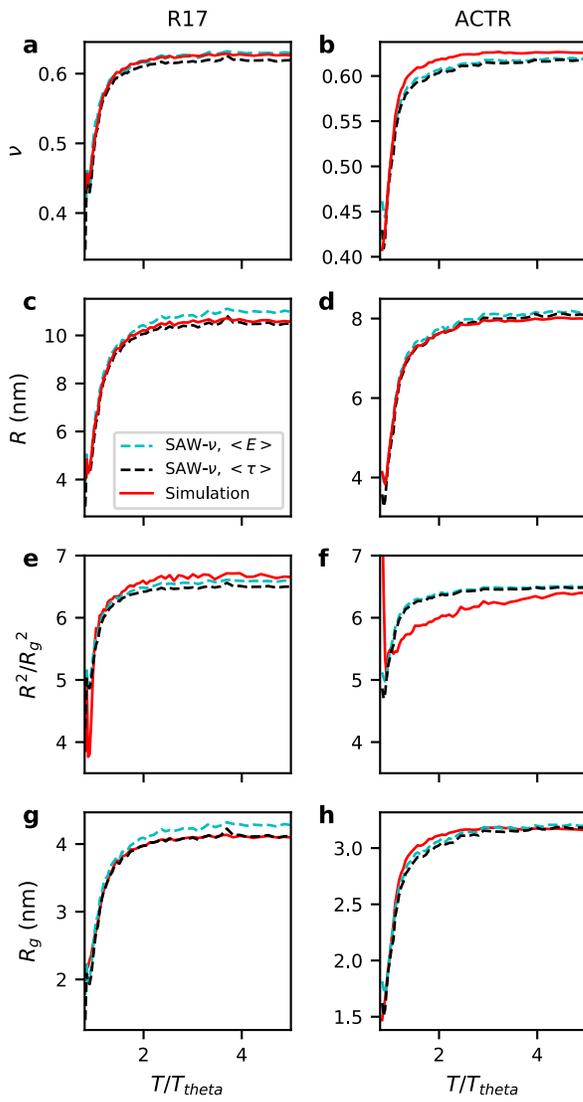


FIG. S2. Parameters recovered from FRET efficiency $\langle E \rangle$ (cyan) or from donor lifetime (black) based on the SAW- ν model in unfolded R17 (left) and ACTR (right), including R_g , R , scaling exponent ν , and the squared ratio of R and R_g .

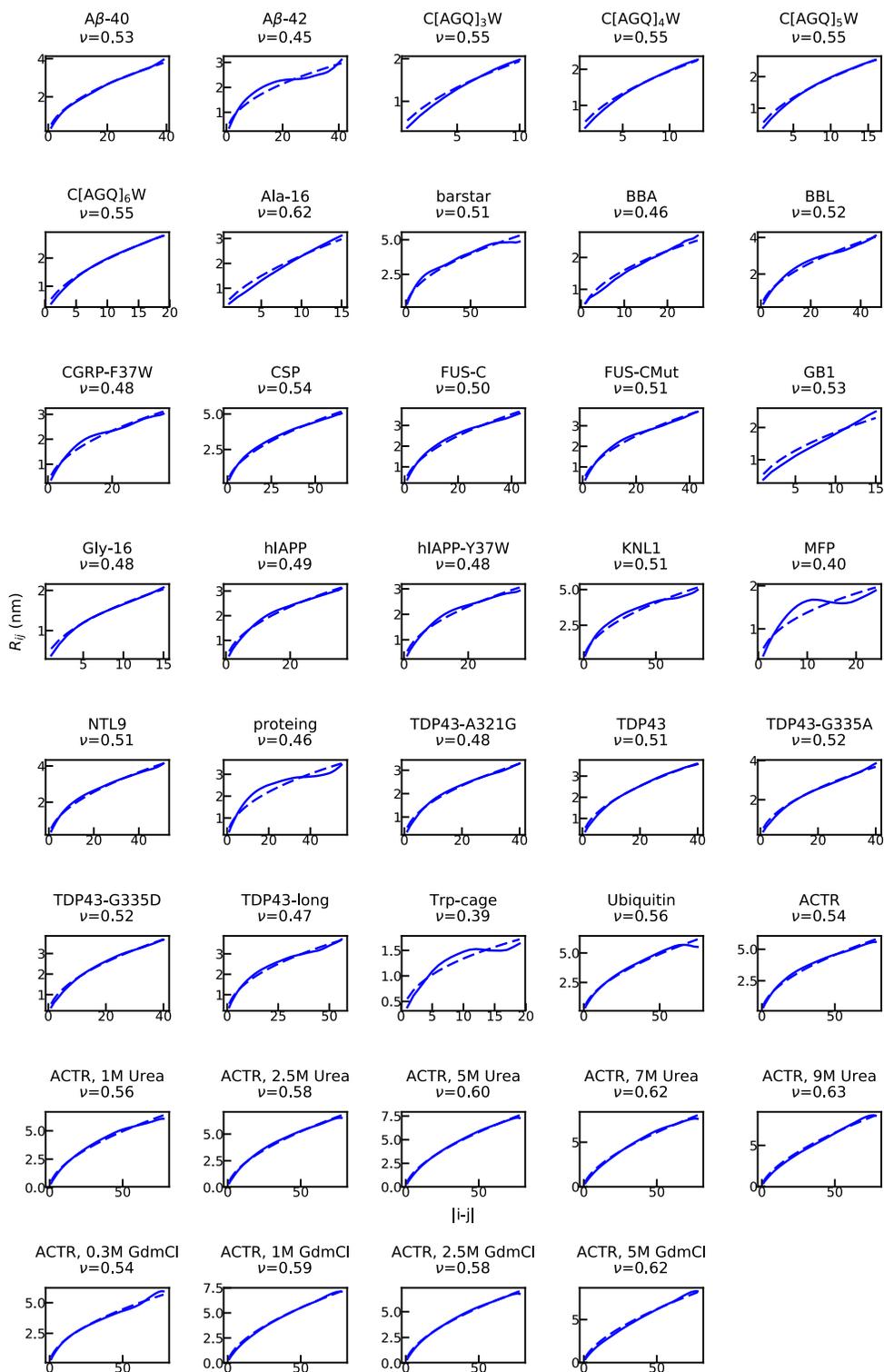


FIG. S3. The polymer scaling exponent ν obtained by fitting the distance as a function of sequence separation based on the all-atom explicit-solvent simulations with $R_{ij} = 0.55 \text{ nm } N^\nu$. The full range of $|i-j|$ is used for fitting. The data is shown as a solid line and the fit as a dashed line.

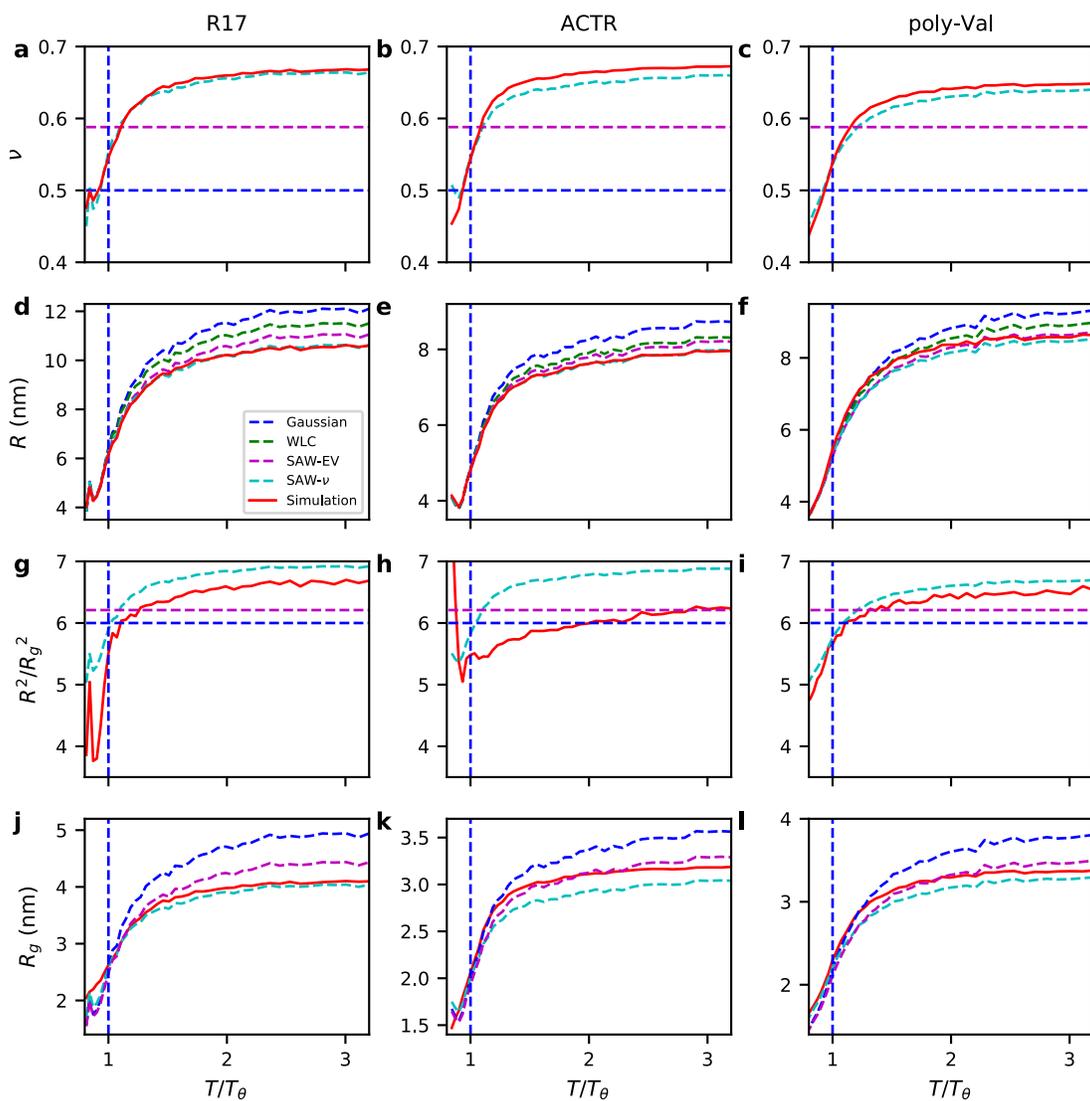


FIG. S4. Recovering ensemble properties from synthetic FRET data calculated from implicit solvent simulations using a different scaling prefactor of 0.45 nm as compared to 0.55 nm in FIG. 2 of the main text. The recovered properties, obtained by fitting the FRET efficiency $\langle E \rangle$ using different polymer models, are compared with those of the original ensemble used to generate the data, for unfolded R17 (left), ACTR (middle) and poly-Val (right). In (a), (b) and (c) we show the fitted scaling exponent ν , in (d), (e) and (f) the end-end distance R , in (g), (h) and (i) the ratio $\lambda = R^2/R_g^2$ derived from ν , and in (j), (k) and (l) the resulting R_g obtained from R and ν . The legend shows the polymer model used. For the purposes of this plot, T_θ is defined as the temperature for which $\nu = 1/2$.

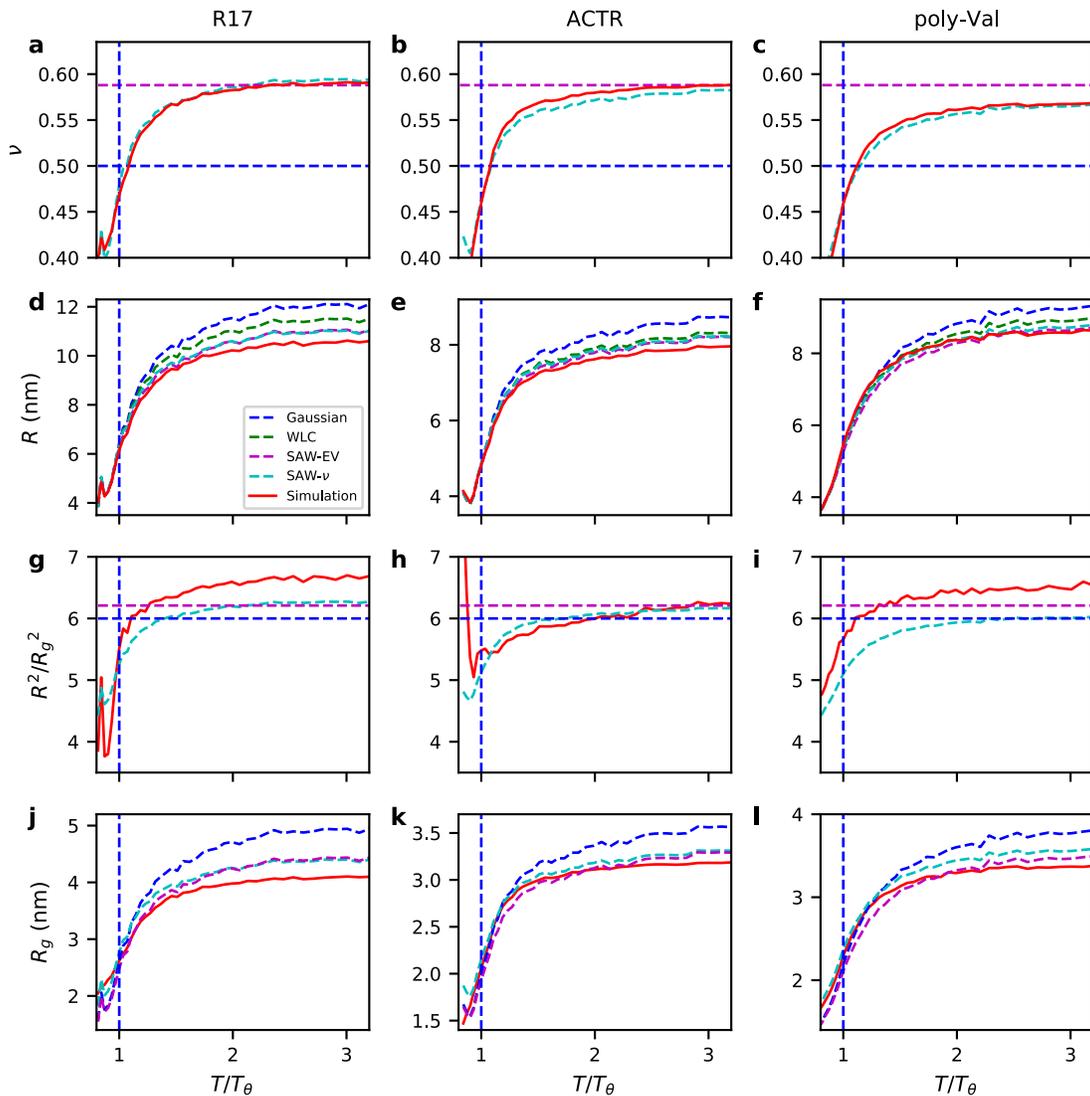


FIG. S5. Recovering ensemble properties from synthetic FRET data calculated from implicit solvent simulations using a different scaling prefactor of 0.65 nm as compared to 0.55 nm in FIG. 2 of the main text. The recovered properties, obtained by fitting the FRET efficiency $\langle E \rangle$ using different polymer models, are compared with those of the original ensemble used to generate the data, for unfolded R17 (left), ACTR (middle) and poly-Val (right). In (a), (b) and (c) we show the fitted scaling exponent ν , in (d), (e) and (f) the end-end distance R , in (g), (h) and (i) the ratio $\lambda = R^2/R_g^2$ derived from ν , and in (j), (k) and (l) the resulting R_g obtained from R and ν . The legend shows the polymer model used. For the purposes of this plot, T_θ is defined as the temperature for which $\nu = 1/2$.

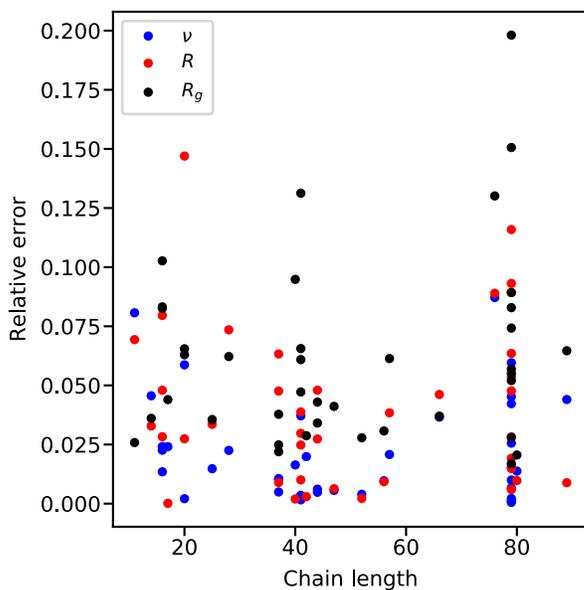


FIG. S6. The relative errors $|x_{\text{model}} - x_{\text{sim}}|/x_{\text{sim}}$ of SAW- ν to recover ν , R and R_g as a function of the chain length for all-atom explicit-solvent simulations.

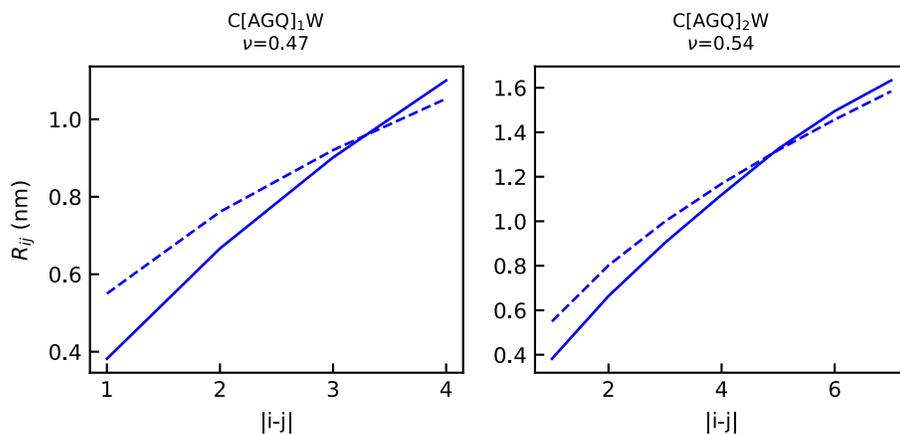


FIG. S7. The fit of intrachain distances as a function of sequence separation $|i-j|$ to Flory's scaling law when the chain length is smaller than 10. C[AGQ]₁W and C[AGQ]₂W have a chain length of 5 and 8, respectively.

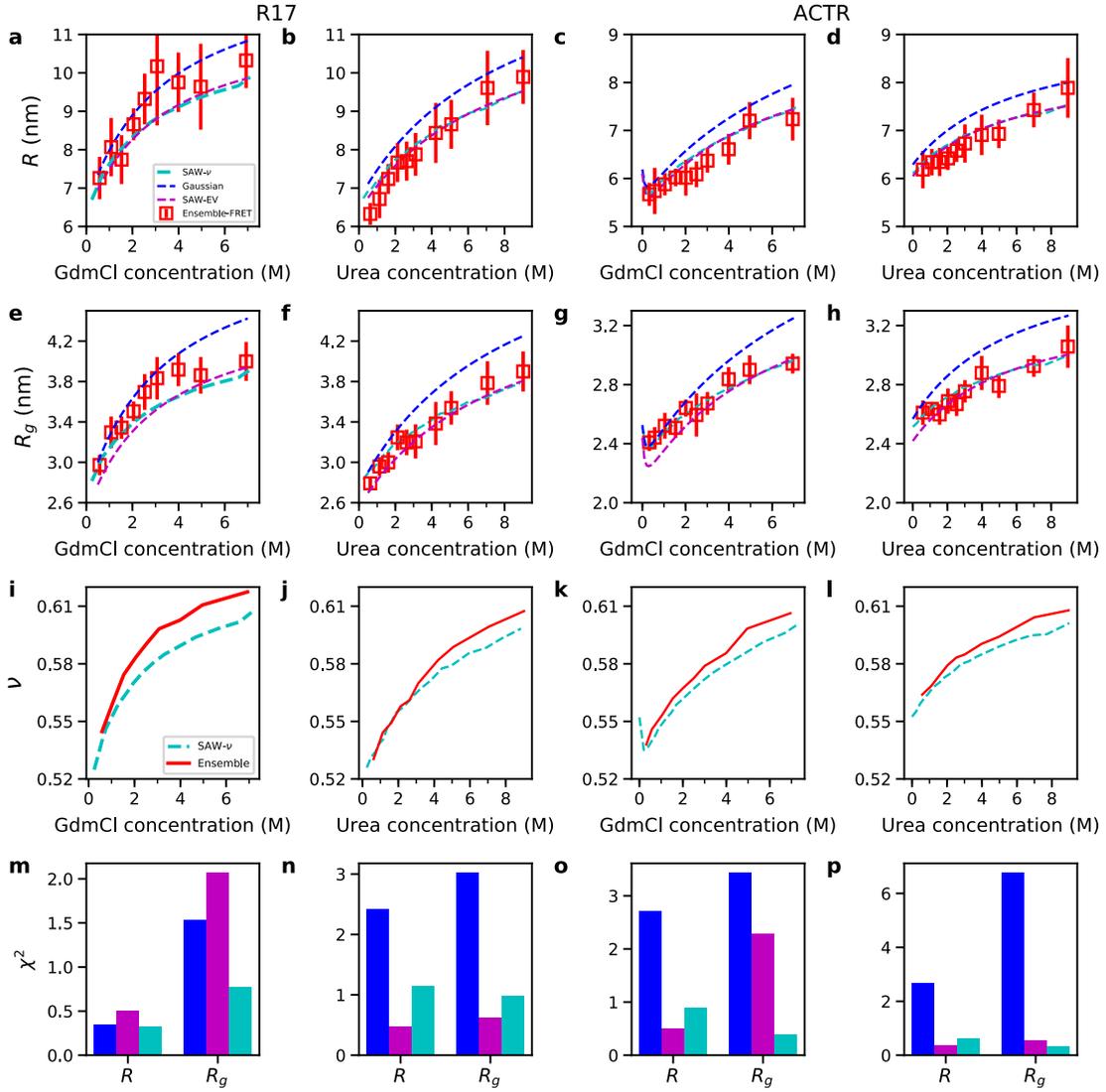


FIG. S8. Validation using molecular ensembles determined from *only* FRET efficiencies at different denaturant concentrations. Ensemble properties estimated using polymer models from the FRET data alone are compared with those computed from molecular ensembles (“Ensemble” in legend) generated to match both SAXS and FRET data. We show R , R_g and the scaling exponent ν recovered from FRET efficiency at each denaturant concentration for the proteins R17 and ACTR (ν is determined from the self-consistent SAW- ν method, or from internal distance scaling for the structural ensembles reweighted to match FRET data). $\chi^2 = \langle ((x_{\text{ensemble}} - x_{\text{sim}})/\sigma_{\text{ensemble}})^2 \rangle$. χ^2 for R_g is smaller than in Figure 4 because R_g is unconstrained by SAXS data in the original ensembles.

SI Tables

Table S1. List of details for all-atom simulation data sets used in the verification. All data without explicit citation are taken from Ref²⁰.

Protein	Number of residues	Equilibration length (ns)	Production run length (ns)	Scaling exponent
A β -40	40	50	150	0.53
A β -42	42	50	150	0.45
C[AGQ] ₃ W ³	11	20	2164	0.55
C[AGQ] ₄ W ³	14	20	2124	0.55
C[AGQ] ₅ W ³	17	20	2925	0.55
C[AGQ] ₆ W ³	20	20	2207	0.55
Ala-16	16	20	80	0.62
barstar	89	50	175	0.51
BBA	28	50	150	0.46
BBL	47	50	73	0.52
CGRP-F37W	37	50	250	0.48
CSP	66	50	182	0.54
FUS-C	44	50	150	0.50
FUS-SMut	44	50	150	0.51
GB1 (unfolded)	16		500	0.53
Gly-16	16	20	80	0.48
hIAPP	37	50	150	0.49
hIAPP-Y37W	37	50	150	0.48
KNL1	80	50	130	0.51
MFP	25	50	150	0.40
NTL9	52	50	50	0.51
protein G	56	50	150	0.46
TDP43-A321G	41	50	150	0.48
TDP43	41	50	150	0.51
TDP43-G335A	41	50	150	0.52
TDP43-G335D	41	50	150	0.52
TDP43-Long	57	50	150	0.47
Trp-cage (unfolded)	20		150	0.39
Ubiquitin	76	50	50	0.56
ACTR ⁴	79	200	1800	0.54
ACTR (1M Urea) ⁴	79	200	1800	0.56
ACTR (2.5M Urea) ⁴	79	200	1800	0.58
ACTR (5M Urea) ⁴	79	200	1800	0.60
ACTR (7M Urea) ⁴	79	200	1800	0.62
ACTR (9M Urea) ⁴	79	200	1800	0.63
ACTR (0.3M GdmCl) ⁴	79	200	1800	0.54
ACTR (1M GdmCl) ⁴	79	200	1800	0.59
ACTR (2.5M GdmCl) ⁴	79	200	1800	0.58
ACTR (5M GdmCl) ⁴	79	200	1800	0.62

Table S2. χ^2 between properties estimated using different polymer models and properties from the ensembles determined using both FRET and SAXS experimental data.

	R			R_g		
	Gaussian	SAW-EV	SAW-v	Gaussian	SAW-EV	SAW-v
R17 GdmCl	0.67	2.57	1.39	7.33	19.21	9.01
R17 Urea	1.92	1.52	0.87	6.18	13.60	2.33
ACTR GdmCl	2.19	0.24	0.36	9.33	34.21	13.39
ACTR Urea	4.62	0.65	1.19	25.02	0.46	2.06
Average	2.35	1.25	0.95	11.97	16.87	6.70

References

1. Vitalis, A.; Pappu, R. V., ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. *J. Comput. Chem.* **2008**, *30*, 673-699.
2. Borgia, A.; Zheng, W.; Buholzer, K.; Borgia, M. B.; Schuler, A.; Hofmann, H.; Soranno, A.; Nettels, D.; Gast, K.; Grishaev, A.; Best, R. B.; Schuler, B., Consistent View of Polypeptide Chain Expansion in Chemical Denaturants from Multiple Experimental Methods. *J Am Chem Soc* **2016**, *138* (36), 11714-26.
3. Zerze, G. H.; Mittal, J.; Best, R. B., Diffusive Dynamics of Contact Formation in Disordered Polypeptides. *Phys Rev Lett* **2016**, *116* (6), 068102.
4. Zheng, W.; Borgia, A.; Buholzer, K.; Grishaev, A.; Schuler, B.; Best, R. B., Probing the Action of Chemical Denaturant on an Intrinsically Disordered Protein by Simulation and Experiment. *J Am Chem Soc* **2016**, *138* (36), 11702-11713.
5. Best, R. B.; Zheng, W.; Mittal, J., Balanced Protein-Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *J Chem Theory Comput* **2014**, *10* (11), 5113-5124.
6. Abascal, J. L. F.; Vega, C., A general purpose model for the condensed phases of water: TIP4P/2005. *J. Chem. Phys* **2005**, *123*, 234505.
7. Zheng, W.; Borgia, A.; Borgia, M. B.; Schuler, B.; Best, R. B., Empirical Optimization of Interactions between Proteins and Chemical Denaturants in Molecular Simulations. *J Chem Theory Comput* **2015**, *11* (11), 5543-53.
8. Parrinello, M.; Rahman, A., Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.* **1981**, *52* (12), 7182-7190.
9. Darden, T.; York, D.; Pedersen, L., An N-log(N) method for Ewald sums in large systems. *J. Chem. Phys* **1993**, *98*, 10089-10092.
10. Hess, B.; Kutzner, C.; Van der Spoel, D.; Lindahl, E., GROMACS4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theor. Comput.* **2008**, *4* (3), 435-447.

11. Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G., PLUMED 2: New feathers for an old bird. *Comput Phys Commun* **2014**, *185* (2), 604-613.
12. Bonomi, M.; Parrinello, M., Enhanced Sampling in the Well-Tempered Ensemble. *Phys Rev Lett* **2010**, *104* (19).
13. Sugita, Y.; Okamoto, Y., Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett* **1999**, *314* (1-2), 141-151.
14. Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M., CHARMM: The Biomolecular Simulation Program. *J Comput Chem* **2009**, *30* (10), 1545-1614.
15. Goga, N.; Rzepiela, A. J.; de Vries, A. H.; Marrink, S. J.; Berendsen, H. J. C., Efficient Algorithms for Langevin and DPD Dynamics. *J Chem Theory Comput* **2012**, *8* (10), 3637-3649.
16. Flory, P. J., *Principles of Polymer Chemistry*. Cornell University Press: Ithaca and London, 1953.
17. Hofmann, H.; Soranno, A.; Borgia, A.; Gast, K.; Nettels, D.; Schuler, B., Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc Natl Acad Sci U S A* **2012**, *109* (40), 16155-60.
18. Thirumalai, D.; Ha, B. Y., In *Theoretical and mathematical methods in polymer research*, Grosberg, A. Y., Ed. Academia: New York, 1988.
19. O'Brien, E.; Morrison, G.; Brooks, B. R.; Thirumalai, D., How accurate are polymer models in the analysis of Förster resonance energy transfer experiments on proteins? *J. Chem. Phys.* **2009**, *130*, 124903.
20. Zerze, G. H.; Zheng, W.; Best, R. B.; Mittal, J., Evolution of all-atom protein force fields to improve local and global properties (manuscript in preparation). 2017.