THE JOURNAL OF
PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

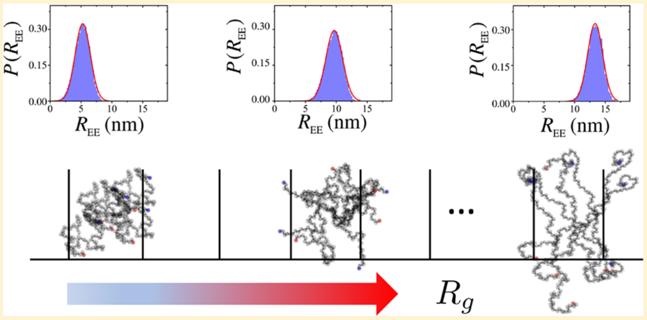
An Adequate Account of Excluded Volume Is Necessary To Infer Compactness and Asphericity of Disordered Proteins by Förster Resonance Energy Transfer

Jianhui Song,[†,¶] Gregory-Neal Gomes,[§,‡,¶] Claudiu C. Gradinaru,*[,§,‡] and Hue Sun Chan*[,†]

[†]Departments of Biochemistry and Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada
[§]Department of Physics, University of Toronto, Toronto, Ontario M5S 1A7, Canada
[‡]Department of Chemical and Physical Sciences, University of Toronto Mississauga, Mississauga, Ontario L5L 1C6, Canada

Supporting Information

ABSTRACT: Single-molecule Förster resonance energy transfer (smFRET) is an important tool for studying disordered proteins. It is commonly utilized to infer structural properties of conformational ensembles by matching experimental average energy transfer $\langle E\rangle_{\rm exp}$ with simulated $\langle E\rangle_{\rm sim}$ computed from the distribution of end-to-end distances in polymer models. Toward delineating the physical basis of such interpretative approaches, we conduct extensive sampling of coarse-grained protein chains with excluded volume to determine the distribution of end-to-end distances conditioned upon given values of radius of gyration R_g and asphericity A. Accordingly, we infer the most probable R_g and A of a protein disordered state by seeking the best fit between $\langle E\rangle_{\rm exp}$ and $\langle E\rangle_{\rm sim}$ among various (R_g,A) subensembles. Application of our method to residues 1−90 of the intrinsically disordered cyclin-dependent kinase (Cdk) inhibitor Sic1 results in inferred ensembles with more compact conformations than those inferred by conventional procedures that presume either a Gaussian chain model or the mean-field Sanchez polymer theory. The Sic1 compactness we infer is in good agreement with small-angle X-ray scattering data for R_g and NMR measurement of hydrodynamic radius R_h. In contrast, owing to neglect or underappreciation of excluded volume, conventional procedures can significantly overestimate the probabilities of short end-to-end distances, leading to unphysically large smFRET-inferred R_g at high [GdmCl]. It follows that smFRET Sic1 data are incompatible with the presumed homogeneously expanded or contracted conformational ensembles in conventional procedures but are consistent with heterogeneous ensembles allowed by our subensemble method of inference. General ramifications of these findings for smFRET data interpretation are discussed.



INTRODUCTION

Molecular biology has traditionally focused on highly ordered conformations, epitomized by the exquisitely folded structures in the Protein Data Bank. Recent advances, however, point to a more expansive—and arguably more balanced—relationship between biomolecular structure and function. Most notable of these developments is that crucial biological functions are served by intrinsically disordered proteins (IDPs).[1−5] Many IDPs do become significantly ordered upon binding, but some remain mostly disordered as parts of functional "fuzzy complexes".[6−11] More generally, even for globular proteins, the stability of their ordered folded states entails a balance with the proteins' disordered unfolded states,[12,13] and it is the energetics of the latter that drives a protein toward either the native functional conformation or dysfunctional aggregation. Thus, understanding and characterizing the dynamics and conformational diversity in protein disordered states are of paramount importance to molecular biology.

Single-molecule Förster resonance energy transfer (smFRET) has emerged as a versatile experimental tool for studying protein folding and IDPs.[14−16] The technique has shed light on fundamental issues in folding such as cooperativity,[17−19] transition paths,[20−22] and disease-causing aggregation.[23] It has been used extensively to determine the spatial dimensions of unfolded globular proteins[24−28] and IDPs,[16,23,29,30] the internal friction in these ensembles,[31] and the variation of conformational dimensions with chain length.[32] Förster resonance energy transfer (FRET) data also provided restraints for biomolecular modeling[33,34] and served as probes for IDP conformational distribution and switching.[35] For example, by combining chain simulations with a small number of smFRET intrachain distance constraints, conformational properties of the IDPs α-synuclein and tau implicated in

Received: September 18, 2015
Revised: November 10, 2015
Published: November 13, 2015

Alzheimer's and Parkinson's diseases can be determined accurately.[36] More recently, novel smFRET techniques were successfully developed to probe in vivo protein dynamics and folding kinetics in living cells.[37]

In view of flourishing applications, a reexamination of the physical basis for smFRET inference of conformational shape and dimensions of disordered protein ensembles is in order. smFRET typically provides data about the distribution of end-to-end distances R_{EE}. To infer properties arising from the entire chain molecule such as the radius of gyration R_{g} or hydrodynamic radius R_{h} from the limited spatial data on only two ends of the chain, a model of conformational distribution is assumed a priori. Hence the validity of the inference hinges on the presumed polymer model, which is often taken to be either the Gaussian chain[25,26,29] or models[27−29,32] based upon the homopolymer theory of Sanchez.[38] However, these presumed models are physically limited because the Gaussian model neglects excluded volume altogether and the mean-field treatment of excluded volume in the Sanchez model neglects the asphericity of polymer conformations.[39−42] One manifestation of possible shortcomings in current interpretation of smFRET data is the well-documented yet unresolved discrepany between the R_{g}'s of unfolded proteins inferred by smFRET and those measured by small-angle X-ray scattering (SAXS).[43−46]

Accordingly, we endeavored to assess conventional applications of polymer models to smFRET inference of chain dimensions and propose a different approach that does not presume a particular form of global conformational distribution. Our approach relies on extensive sampling of coarse-grained explicit-chain models with physical excluded volume,[47] which have proven useful in many IDP studies.[5,11,26,48−51] As an instructive example, we focus largely on the cyclin-dependent kinase (Cdk) inhibitor Sic1, which is crucial in the yeast cell cycle. Like many IDPs, Sic1 is a polyampholyte.[52] When sufficiently phosphorylated, Sic1 forms a fuzzy complex with the folded SCF ubiquitin ligase subunit Cdc4.[6,7,9,30] Previous NMR and smFRET measurements indicate that the conformational ensemble of Sic1 is heterogeneous and cannot be adequately described by a Gaussian chain model of homopolymers.[9,30] Consistent with this observation, we can now trace the unphysical overestimation of Sic1 R_{g} in high denaturant by conventional Gaussian- or Sanchez-based smFRET analyses to these methods' presumption of a homogeneous ensemble. In contrast, using our approach, a *most likely* average R_{g} and R_{h} obtained by matching smFRET data to the R_{EE} distributions in subensembles with restricted R_{g} ranges are much more in line with other experimental data on Sic1 compactness. Based on a physical model that accounts for excluded volume and the generally aspherical shapes of chain conformations, these findings establish a general framework for improved utilization of smFRET data to infer conformational properties of heterogeneous disordered protein ensembles, as will be discussed below.

■ MATERIALS AND METHODS

Protein Chain Model and Conformational Sampling. The C_{α} protein model in this study consists of a string of n beads sequentially connected by virtual bonds of length 0.38 nm with reference bond angle $\theta_0 = 106.3°$ coinciding with the angle most populated in the Protein Data Bank.[53] The potential energy $E = \sum_{i=2}^{n-1} \epsilon_{\theta}(\theta_i - \theta_0)^2 + (1/2)\sum_{i=1}^{n} \sum_{j=1}^{n} \epsilon_{\mathrm{ex}}(R_{\mathrm{hc}}/R_{ij})^{12}$, where $\epsilon_{\theta} = 10.0\ k_{\mathrm{B}}T$, θ_i is the virtual bond angle at bead i, k_{B} is the Boltzmann constant, T is absolute temperature, $\epsilon_{\mathrm{ex}} = 1.0\ k_{\mathrm{B}}T$ is the model protein's self-avoiding excluded-volume repulsion strength, and R_{ij} is the distance between beads i and j. The excluded-volume $(R_{\mathrm{hc}}/R_{ij})^{12}$ term is set to zero for $R_{ij} \geq 1.0$ nm. We consider several R_{hc} values for hard-core (hc) repulsion, including the $R_{\mathrm{hc}} = 0.4$ nm value commonly used in protein folding simulations[19,47] and the $R_{\mathrm{hc}} = 0.314$ nm value used in recent Sanchez theory approaches to FRET inference.[28] The latter R_{hc} is the radius of a sphere with volume equal to the average packing volume per amino acid in folded proteins.[54] Larger values of $R_{\mathrm{hc}} \geq 0.5$ nm are also tested in view of the possibility that nonspherical shapes of real amino acid residues may effectively increase R_{hc} in our single-bead model.

We conduct Monte Carlo sampling by the Metropolis criterion[55] at $T = 300$ K using a previously described efficient algorithm[11] involving equal probability for pivot and kink jumps.[56,57] The acceptance rate for these attempted chain moves (updates) is about 30%. For each simulation, the first 10^7 equilibrating updates are discarded, after which 10^9 updates are performed to sample 10^7 conformations for further analysis.

Measures of Conformational Shape and Size. Asphericity A as well as R_{EE}, R_{g}, and R_{h} are computed for the sampled conformations. As defined in ref 42

$$A \equiv 1 - \frac{3(\lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3)}{(\lambda_1 + \lambda_2 + \lambda_3)^2} \tag{1}$$

where λ_1, λ_2, and λ_3 (all ≥ 0) are the eigenvalues of the gyration tensor $S_{\alpha\beta} \equiv n^{-1}\sum_{i=1}^{n} (R_i - R_{\mathrm{cm}})_{\alpha}(R_i - R_{\mathrm{cm}})_{\beta}$, R_i and R_{cm} are position vectors of the ith bead and that of the centroid, respectively, α, β = 1, 2, 3 refer to the Cartesian components, and ${R_{\mathrm{g}}}^2 = n^{-1}\sum_{i=1}^{n} |R_i - R_{\mathrm{cm}}|^2 = \lambda_1 + \lambda_2 + \lambda_3$. The hydrodynamic radius R_{h} is computed using the formula[58] $1/R_{\mathrm{h}} = 1/n^2\langle\sum_{i\neq j} 1/R_{ij}\rangle$ based on Kirkwood theory[59] while neglecting the free-draining term. It is instructive to apply this formula to determine R_{h} for the $A = 1$ and $A = 0$ extreme cases. For an $A = 1$ rod of length N, $1/R_{\mathrm{h}} \approx (1/N)^2 \int_0^N \mathrm{d}x \int_0^N \mathrm{d}y\, |x - y|^{-1}$, where the integral is restricted to $|x - y| \geq 1$. It follows that $1/R_{\mathrm{h}} \approx (1/N)^2[N \ln N + (N - 1) \ln(N - 1)]$ and thus $R_{\mathrm{h}} \approx N/(2 \ln N)$ for sufficiently large N. For an $A = 0$ solid sphere of radius R, we assume for simplicity that the distribution of R_{ij} is that between two random points as given below in eq 8, viz., $P(r) = 3r^2 - 9r^3/4 + 3r^5/16$, where $0 \leq r = R_{ij}/R \leq 2$. In that case $1/R_{\mathrm{h}} \approx \int_c^2 \mathrm{d}r\, P(r)/r$ where $c \approx 2(\pi/\sqrt{18})^{1/3}N^{-1/3}$ is the minimum distance between two residues, resulting in $R_{\mathrm{h}} \approx 5R/6$ for $N \to \infty$ ($c = 0$) and $R_{\mathrm{h}} \approx 0.985R$ for $N = 99$. Interestingly, both of these derived results are very close to the intuitively expected $R_{\mathrm{h}} = R$ Stokes radius for an $A = 0$ solid sphere. The shape factor $\langle R_g^2\rangle^{1/2}/\langle R_{\mathrm{h}}\rangle \approx \sqrt{3/5} = 0.775$ thus estimated for an $A = 0$ solid sphere is also consistently similar to the exact shape factor $2^{1/2}(151 + 162/2^{1/2} + 68/3^{1/2} + 144/5^{1/2} + 96/6^{1/2})/729 = 0.792$ for the 27-mer simple cubic conformations configured in a 3 × 3 × 3 cube.[60] The ensemble average of the hydrodynamic radius is defined here as $\langle R_{\mathrm{h}}\rangle$. This average for our model is expected to be similar, though not identical, to the expression $(\langle {R_{\mathrm{h}}}^{-1}\rangle)^{-1}$ that appears in some theoretical analyses.[58] Since $\langle R_{\mathrm{g}}\rangle \approx (\langle {R_{\mathrm{g}}}^2\rangle)^{1/2}$ for our subensembles with narrow ranges of R_{g}, ensemble averages of $R_{\mathrm{g}}/R_{\mathrm{h}}$ with several slightly different definitions, including $R_{\mathrm{g}}^0/\langle R_{\mathrm{h}}\rangle$, $\langle R_{\mathrm{g}}\rangle/\langle R_{\mathrm{h}}\rangle$, and $(\langle {R_{\mathrm{g}}}^2\rangle)^{1/2}/\langle R_{\mathrm{h}}\rangle$ discussed below, are all referred to simply as shape factors in the present study.

Analysis of Experimental Data on Sic1. smFRET data on Sic1 was originally reported in ref 30. Experimental details

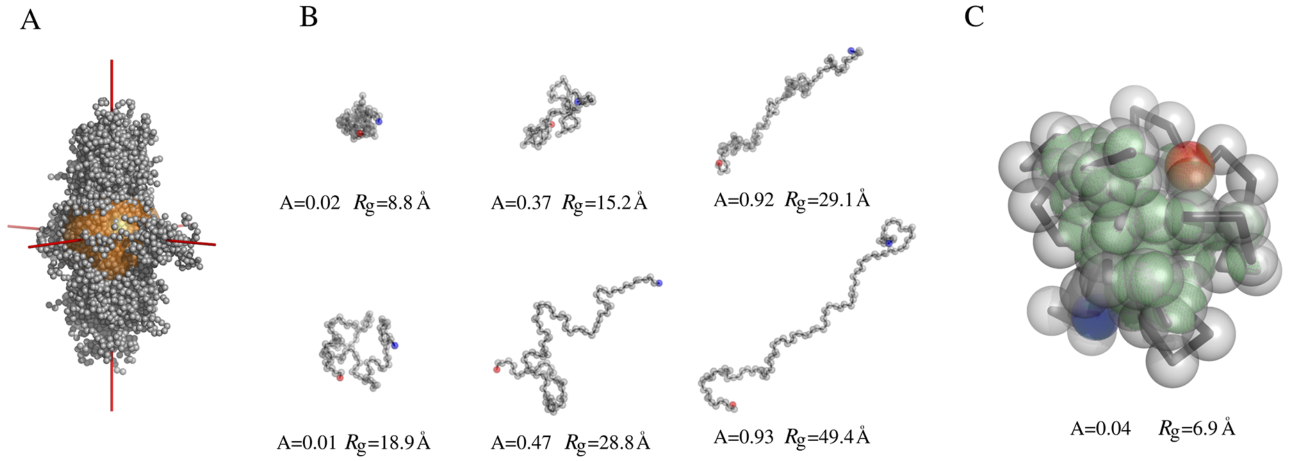


Figure 1. Protein chain models with and without excluded volume. Conformations shown are for n = 100. Each bead represents an amino acid residue. (A) One hundred randomly chosen SAWs. Conformational asphericity A is illustrated by aligning the longest principal axis of every SAW along the vertical axis. The root-mean-square $\sqrt{\langle R_g{}^2\rangle}$ of the full ensemble is given by the radius of the golden sphere. (B) Snapshots of sampled freely jointed Gaussian chains (top) and SAWs (bottom). From left to right are, respectively, highly compact, average, and extended conformations. For illustrative purposes, the R_g and A values of the middle conformations are chosen to coincide with the averages over the full ensemble. (C) Example of a highly compact Gaussian chain that violates excluded volume. The excluded volume of the corresponding SAW is indicated by the van der Waals spheres. Here the overlapping (forbidden) parts of excluded volume are depicted in green, whereas the N- and C-termini are shown, respectively, in blue and red. All SAWs in this figure are simulated using excluded volume radius R_{hc} = 0.4 nm.

of protein expression, purification, and mass spectrometry as well as FRET histograms fitting and analysis using a polyelectrolyte binding−screening model are provided in this reference. Bulk FRET data, observations from control FRET−fluorescence correlation spectroscopy (FCS) experiments, and control smFRET experiments at different pH, salt, and laser excitation powers can also be found in ref 30.

Relevant experimental Sic1 data are reanalyzed here using software we developed specifically for the present investigation using Labview (2009 32 bit, National Instruments Corp., Austin, TX, USA) and Matlab (Release 2013a, The MathWorks, Inc., Natick, MA, USA).

■ RESULTS

As explained above, we use a self-avoiding walk (SAW) to model protein molecules. Each amino acid residue is represented by a single spherical bead with excluded volume defined by a radius $R_{hc} \approx 0.4$ nm. To facilitate comparison with conventional Gaussian chain (CG) and Sanchez theory (ST) approaches, we consider homopolymers for which R_{hc} is the same for every bead. As described in Materials and Methods, three measures of conformational dimensions, R_{EE}, R_g, and R_h, are computed. Also calculated is asphericity A, which measures the deviation from a perfectly spherical conformation with isotropically positioned residues (A = 0), and by which the maximum deviation (A = 1) is identified with linear rod-like conformations. To probe the effects of excluded volume, the same quantities are computed in a freely jointed Gaussian chain model as controls. The SAW or Gaussian chain models are sampled without energetic bias to provide conformations with equal a priori probabilities (Figure 1). With such a representation of the unbiased full ensemble, inferred ensembles for protein disordered states can then be constructed by reweighting the conformations in accordance with a posteriori information from smFRET and/or other experiments. A broad range of chain lengths are studied to address general principles (Figure S1). Thermal motions of the dyes and their linkers were found to impact the inference of distance from smFRET data.[61] For Sic1, we use models with n = 100 residues in which 10 residues extra to that in Sic1 are added to provide an equivalent chain length to account for the two dye linkers.[24]

Conformational Asphericity in Explicit-Chain Protein Models. It is intuitive that residues in a randomly selected *individual* conformation are unlikely to be positioned isotropically (Figure 1A). Thus, polymer conformations are typically aspherical.[39−42,62−64] The average asphericity, $\langle A\rangle$, is around 0.4−0.5 for our SAW and Gaussian chain (GC) models, with wide variations in A in the full ensemble (Figure 1B and Figure S1A). For chains with up to 300 residues, $\langle A\rangle \approx 0.45$ for SAW, ≈0.4 for GC, and varies little with chain length (Figure S1A). Asphericity increases slightly with excluded volume because extremely compact GC conformations with small A values are incompatible with physical excluded volume (Figure 1C).

As expected, average R_h, R_g, and R_{EE} increase with chain length. Because of the repulsive effects of excluded volume, the increase is much steeper for SAW than for GC (Figure S1B−D; $N = n - 1$ is the number of bonds). Consistent with polymer theory,[59,65] both $\langle R_{EE}{}^2\rangle$ and $\langle R_g{}^2\rangle$ scale approximately as $N^{1.0}$ for GC and $N^{1.2}$ for SAW (Figure S1C,D). The SAW scaling with N exhibits a small deviation from linearity, signaling a slower approach to the large-N asymptotic regime for SAW than for GC. In contrast, the nonlinearity of average R_h with respect to N is prominent for both SAW and GC (Figure S1B). Part of the concave-downward trend should be attributable to contributions from aspherical conformations, as exemplified by the less-than-linear $R_h \approx N/(2 \ln N)$ scaling for A = 1 rod-like conformations (see Materials and Methods). Consequently, as chain length increases, R_h increases somewhat slower than R_{EE} or R_g. For the chain lengths we simulated, $R_h{}^2 \sim N^{0.9}$ for GC and $\sim N^{1.1}$ for SAW (Figure S1B).

The ratio $\langle R_g\rangle/\langle R_{EE}\rangle$ is of interest here because many smFRET applications entail inferring R_g from R_{EE}. This ratio ≈0.41−0.43 for the full SAW or GG ensemble, with a slightly lower value for SAW, and varies little with N (Figure S1E).

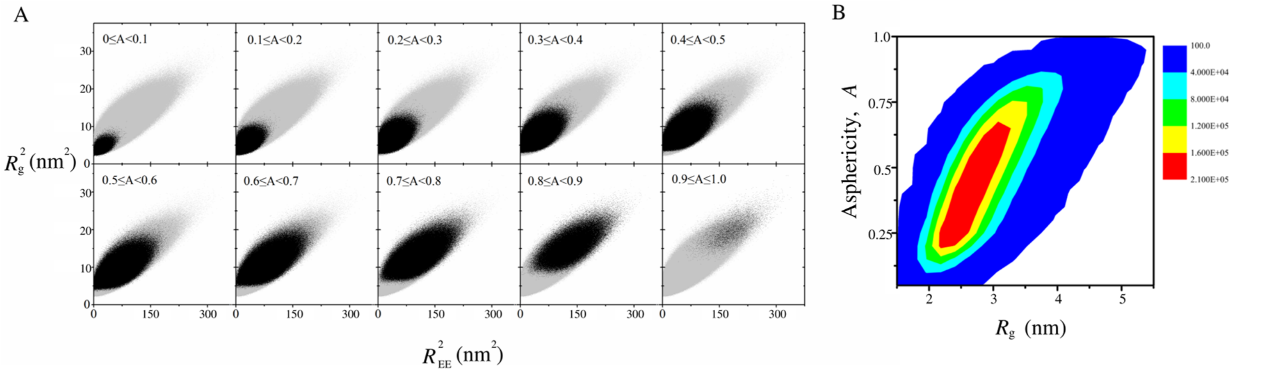


Figure 2. Variation and correlation among A, R_g, and R_{EE}. Simulation data are shown for n = 100 SAWs with R_{hc} = 0.4 nm. (A) R_g^2 versus R_{EE}^2 scatter plots for subensembles with different ranges of A values are shown in black. Included for comparison (gray) is the corresponding scatter for the full, all-A ensemble with 10^7 sampled conformations. (B) Contour plot of population as a function of A and R_g.

Although $\langle R_g\rangle \neq \langle R_g^2\rangle^{1/2}$ and $\langle R_{EE}\rangle \neq \langle R_{EE}^2\rangle^{1/2}$ in general, all simulated $\langle R_g\rangle/\langle R_{EE}\rangle$ values are quite close to the large-N, full-ensemble GC value of $(\langle R_g^2\rangle/\langle R_{EE}^2\rangle)^{1/2} = \sqrt{1/6} = 0.408$. Notably, the relationship between R_g and R_{EE} is very sensitive to asphericity (Figure S1E), with $\langle R_g\rangle/\langle R_{EE}\rangle$ decreasing with increasing A. This trend may be rationalized as follows. For an $A = 0$ solid sphere of radius R, $R_g = \sqrt{3/5}\,R$, $R_{EE} = \sqrt{6/5}\,R$ if we assume that the distribution of R_{EE} may be approximated by that between two randomly chosen points within the sphere (see below); thus $R_g/R_{EE} = \sqrt{1/2} = 0.707$. Now consider an $A = 1$ rod of length N. In this case, $R_{EE} = N$ and $R_g = N/\sqrt{12}$; thus $R_g/R_{EE} = \sqrt{1/12} = 0.289$, which is much smaller than the 0.707 value for the $A = 0$ solid sphere.

Since average R_g^2 and R_h are readily accessible by experiment, the ratio $\langle R_g^2\rangle^{1/2}/\langle R_h\rangle$, known as the shape factor, is commonly used to characterize conformational shape. Shape factor increases with asphericity (Figure S1F). This trend may be understood in terms of the above $A = 0, 1$ extreme cases. $R_h \approx R$ for an $A = 0$ solid sphere (Materials and Methods), whereas $R_h \approx N/(2 \ln N)$ for an $A = 1$ rod. Hence $\langle R_g^2\rangle^{1/2}/\langle R_h\rangle \approx \sqrt{3/5} = 0.775$ for $A = 0$, which is smaller than the value of $(\ln N)/\sqrt{3}$ for $A = 1$ when $N \geq 4$. This ~ln N scaling for $A \approx 1$ conformations also underpins the increase in the full-ensemble shape factor with increasing chain length N for both SAW and GC (Figure S1F).

Physical End-to-End Distances in Conformational Subensembles with Specific Compactness and Asphericity Are Key To Interpreting smFRET Data. The covariation of R_{EE}, R_g, and A in the n = 100 SAW model for Sic1 (Figure 2) shows that $A \approx 0$ is possible only with small R_g and R_{EE}, whereas $A \approx 1$ is possible only with large R_g and R_{EE} (Figure 2A, panels for $0 \leq A < 0.1$ and $0.9 \leq A \leq 1.0$). R_{EE} and R_g correlate with A, but the variation in A for a given R_{EE} or R_g is large except for extreme values of A (Figure 2). A similar trend is also observed for the GC model (Figure S2).

The distribution of end-to-end distance, $P(R_{EE})$, is of central importance to smFRET inference of conformational properties. The average FRET energy transfer

$$\langle E\rangle = \int \mathrm{d}R_{EE}\, \frac{R_0^{\,6}}{R_0^{\,6} + R_{EE}^{\,6}} P(R_{EE}) \qquad (2)$$

where R_0 is the Förster radius of the dye.[24,25] It follows that conformational properties may be inferred from a given experimental $\langle E\rangle_{exp}$ by first matching it with a simulated $\langle E\rangle_{sim}$ computed using eq 2 with the $P(R_{EE})$ of an appropriately chosen conformational ensemble. Inferred properties, then, are those of the matched ensemble.

Conventional ST- and CG-based smFRET inferences presume conformational ensembles with uniformly modulated compactness. The energetic effects of a protein's amino acid sequence are heterogeneous, however; their impact on different residues varies. From a functional standpoint, we expect this heterogeneity to be pronounced in native IDPs. Therefore, rather than presuming whatever conformational property that is of interest is always that of a homogeneous ensemble, in general one should endeavor to reweight the conformations in the full ensemble to achieve consistency with experimental smFRET data as has been performed for special cases.[66] A strength of smFRET experiments is their ability to resolve several subensembles.[16,30] Accordingly, a natural starting point of smFRET inference is to divide the full conformational ensemble into subensembles with narrow variations in the property of interest and determine the subensembles' $P(R_{EE})$s. Using these $P(R_{EE})$s as a basis set, the subensemble or combination of subensembles that is consistent with experiment may then be inferred by comparing $\langle E\rangle_{exp}$ with $\langle E\rangle_{sim}$'s computed using the subensemble $P(R_{EE})$s.

Here we develop such an approach for the conformational properties R_g and A. Conformations sampled in our simulations are sorted by these properties into (R_g,A) subensembles, and a set of distributions $P(R_{EE}|R_g,A)$ of R_{EE} conditioned upon narrow ranges of R_g and A are determined for the n = 100 SAW model for Sic1. Using this approach, R_h can also be inferred from the average R_h of each (R_g,A) subensemble (see below). Not surprisingly, as R_g and A increase, the center of the $P(R_{EE}|R_g,A)$ distribution shifts to higher R_{EE} (Figure S3). These distributions are approximated by Gaussians to faciliate computation of $\langle E\rangle_{sim}$ using eq 2. The Gaussian fitting parameters for $P(R_{EE}|R_g,A)$ are provided in Table S1 and Figure S4. Gaussian fitting parameters for distributions $P(R_{EE}|A)$ conditioned only upon A are given in Table S2 for n = 100; those for $P(R_{EE}|R_g)$ conditioned only upon R_g are given in Table S3 for n = 100 and in Tables S4−S7 for chain lengths n = 50, 75, 125, and 150. Gaussian fitting parameters for full-ensemble $P(R_{EE})$ of SAW with different excluded-volume radii (R_{hc}) are provided in Table S8.

DOI: 10.1021/acs.jpcb.5b09133

smFRET-Inferred Conformational Properties Vary Significantly Depending on the Presumed Polymer Model. Following eq 2, the average FRET efficiency of an (R_g,A) subensemble defined by a narrow range of A and of R_g is given by

$$E(R_g, A) \equiv \langle E\rangle(R_g, A) = \int dR_{EE}\, \frac{R_0^{\ 6}}{R_0^{\ 6} + R_{EE}^{\ 6}} P(R_{EE}|R_g, A) \tag{3}$$

A similar relation that replaces $P(R_{EE}|R_g,A)$ by $P(R_{EE}|R_g)$ in eq 3 applies to $E(R_g) \equiv \langle E\rangle(R_g)$ for a subensemble with a narrow range of R_g but is not restricted with respect to A. Although we mainly use $E(R_g)$ below for notational simplicity, the present subensemble formulation is readily generalized to include A or any set of conformational properties as conditions for the R_{EE} distribution. The construction of $E(R_g)$ may be viewed as a resolution of E into R_g components. This resolution may be applied to any polymer model, including SAW as well as explicit-chain GC models. With $E(R_g)$ in hand

$$\langle E\rangle_{sim} = \int dR_g\, P(R_g)\, E(R_g) \tag{4}$$

for any distribution $P(R_g)$ of radius of gyration. Thus, in principle one may arrive at an inferred distribution $P_{inf}(R_g)$ by minimizing $|\langle E\rangle_{exp} - \langle E\rangle_{sim}|$, in which case the inferred average R_g and average $R_g^{\ 2}$ would be given by

$$\langle R_g\rangle_{inf} = \int dR_g\, R_g P_{inf}(R_g),$$
$$\langle R_g^{\ 2}\rangle_{inf} = \int dR_g\, R_g^{\ 2} P_{inf}(R_g) \tag{5}$$

However, $\langle E\rangle_{exp}$ alone is clearly insufficient for a reliable inference of $P(R_g)$. Therefore, in contrast to CG or ST approaches that presume a particular form for $P(R_g)$, we first leave $P(R_g)$ open and focus instead on minimizing $|\Delta E(R_g)|$ where $\Delta E(R_g) \equiv \langle E\rangle_{exp} - E(R_g)$ to determine a "most probable" $R_g = R_g^0$ such that $E(R_g^0) = \langle E\rangle_{exp}$.

This method is illustrated in Figure 3 by $n = 100$ models for the Sic1-plus-dye system with $R_0 = 6.0$ nm corresponding to the dye pair in the Sic1 smFRET experiments.[30] For the SAW and explicit-chain GC models (Figure 3A,B), the bell-shape curves are $P(R_{EE}|R_g)$ distributions for $E(R_g) = 0.2$, 0.5, and 0.8, where R_{EE} is given by the bottom horizontal scale. The R_g values of the subensembles are marked by the vertical dashed lines referring to the top horizontal scale. For the Sic1 model with excluded volume (SAW, Figure 3A), R_g increases gradually with decreasing $E(R_g)$. In other words, as $\langle E\rangle_{exp}$ decreases, the inferred R_g^0 increases, but not dramatically. A similar trend is seen for the freely jointed GC subensemble model (SG, Figure 3B).

It is instructive to contrast our subensemble method to the conventional ST and CG approaches. CG does not consider $P(R_{EE}|R_g)$. Its starting point is the R_{EE} distribution conditioned upon a mean square R_g for a *full* Gaussian chain ensemble, viz.

$$P_{CG}(R_{EE}|\langle R_g^{\ 2}\rangle) = \frac{1}{2\sqrt{\pi}\langle R_g^{\ 2}\rangle^{3/2}} R_{EE}^{\ 2} \exp\left(-\frac{R_{EE}^{\ 2}}{4\langle R_g^{\ 2}\rangle}\right) \tag{6}$$

Accordingly, $\Delta E(\langle R_g^{\ 2}\rangle)$ is substituted for $\Delta E(R_g)$, and $\langle R_g^{\ 2}\rangle_{inf}$ is the $\langle R_g^{\ 2}\rangle$ that minimizes $\Delta E(\langle R_g^{\ 2}\rangle)$. The CG- and subensemble-inferred R_g's can be very different, with much larger CG-inferred R_g for small $\langle E\rangle_{exp}$ (Figure 3C). However, since the N-dependent $\langle R_g^{\ 2}\rangle$ becomes a free fitting parameter in CG while the physical chain length is constant, an assessment of the physical viability of the fitted $\langle R_g^{\ 2}\rangle$ is imperative. We will take up this issue below.

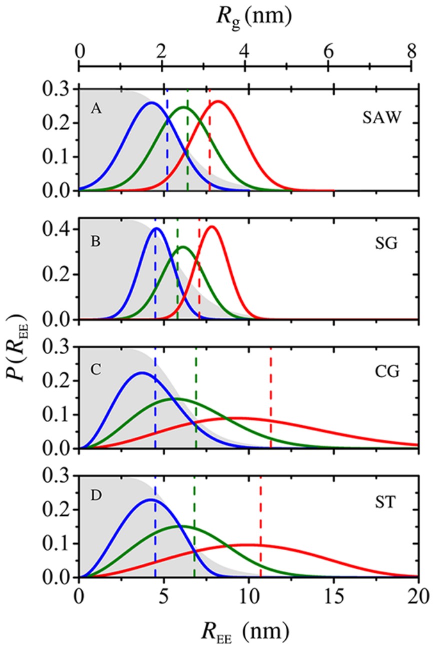


Figure 3. Relationship between R_g and R_{EE} distributions in methods of smFRET inference: (A) subensemble SAW, (B) subensemble Gaussian chain (SG), (C) conventional Gaussian chain (CG), and (D) Sanchez theory (ST). All data in this figure are for $n = 100$ (A–D), and $R_{hc} = 0.4$ nm in (A). The curves show R_{EE} distributions that yield average FRET efficiency $\langle E\rangle = 0.8$ (blue), 0.5 (green), and 0.2 (red) for $R_0 = 6.0$ nm in the four different inference methods. The shaded regions provide the profile for $(1 + R_{EE}^{\ 6}/R_0^{\ 6})^{-1}$ (arbitrary vertical scale) to underscore that, because of eq 2, smFRET inference is determined by the population of R_{EE} values inside the shaded region. The vertical dashed lines provide the R_g^0 or the average R_g's inferred by the $\langle E\rangle$ values (same color code). The inferred R_g's, given by the top horizontal scale, are for the subensembles with narrow ranges of R_g values in (A) and (B), and for the full ensemble in (C) and (D).

Whereas CG neglects excluded volume, ST is partially based on an approximate R_g distribution for chains with excluded volume:[38]

$$P_{ST}(R_g; \epsilon, R_g^{\Theta}) \propto (R_g^{\ 2})^3 \exp\left\{-\frac{7}{2}\left(\frac{R_g}{R_g^{\Theta}}\right)^2 + n\left[\frac{\epsilon\phi}{2} - \frac{1-\phi}{\phi}\ln(1-\phi)\right]\right\} \tag{7}$$

where ϵ is the mean-field intrachain contact energy that controls conformational compactness, $\phi \equiv (R_C/R_g)^3$ is the volume fraction, and R_C is the minimum R_g for maximally compact conformations. Following ref 28, we take $R_C = (0.3143)n^{1/3}$ nm and $R_g^{\Theta} = 0.225\sqrt{N}$ nm. This formula embodies both the Flory–Fisk approximate R_g distribution for GC conformations[67] (the exact mathematical form of which is more complicated[59,68]) and a mean-field treatment of excluded volume.[65] To obtain a distribution of R_{EE}, conventional ST approaches to FRET inference[27,28] apply the expression

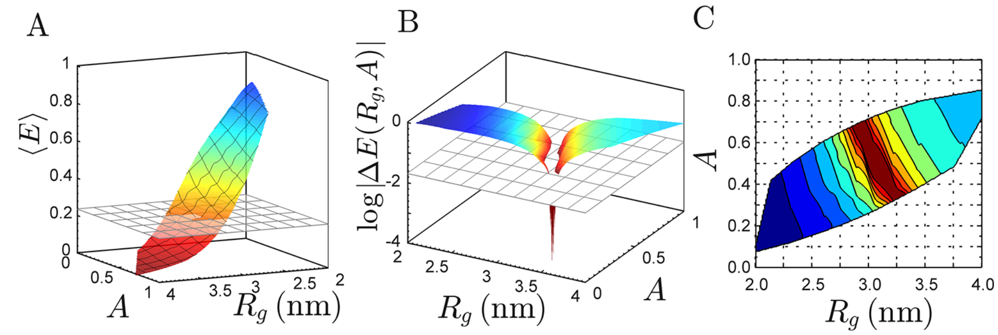


Figure 4. Inferring asphericity and radius of gyration by the subensemble SAW method. (A) $E(R_g,A)$ of simulated (R_g,A) subensembles is computed (eq 3) for $n = 100$ with $R_{hc} = 0.4$ nm. The curved surface for $E(R_g,A)$ is constructed by including interpolations between simulated values. All curved surfaces in this figure are color coded from red to blue to depict variation from low to high values of the quantity being represented. As an illustration, the intersecting horizontal plane in (A) is defined by the experimental FRET efficiency $\langle E\rangle_{exp}$ for Sic1 in [GdmCl] = 4 M, for which an inference on R_g and A will be sought in (B) and (C). (B) The curved surface here represents the logarithmic deviation $\log|\Delta E(R_g,A)|$, where $\Delta E(R_g,A) \equiv \langle E\rangle_{exp} - E(R_g,A)$, and $\langle E\rangle_{exp}$ is that in (A). The inferred R_g and A values are given by those for which the deviation falls below a certain threshold of experimental uncertainty $\sigma_{E,exp}$ of $\langle E\rangle_{exp}$. This threshold, estimated to be $\sigma_{E,exp} \approx 0.02$ in this case, is represented in (B) by the horizontal intersecting plane defined by $\log|\Delta E(R_g,A)| = \log \sigma_{E,exp} = \log(0.02) = -1.70$. (C) Contour plot of $|\Delta E(R_g,A)|$. The lowest level depicted by the contours represents $|\Delta E(R_g,A)| \le \sigma_{E,exp}$.

$$P_{ST}(R_{EE}|R_g) = (3r_{EE}^{\ 2} - 9r_{EE}^{\ 3}/4 + 3r_{EE}^{\ 5}/16)/\sqrt{5R_g^{\ 2}} \tag{8}$$

borrowed from the distribution of the distance between two random points within a solid sphere.[63] Here $r_{EE} \equiv R_{EE}/\sqrt{5R_g^{\ 2}}$. This expression is not part of the original Sanchez theory.[38] It neglects excluded volume and therefore overestimates probabilities for small R_{EE}'s because excluded volume is known to reduce the probability for two chain ends to be in close proximity[60,69,70] (see below). In any event, eq 8 is not used by itself in conventional FRET inference to derive $E(R_g)$, but is combined with eq 7 to form an R_{EE} distribution $P_{ST}(R_{EE}|\langle R_g^{\ 2}\rangle) = \int_{R_C}^{L/2} dR_g\ P_{ST}(R_{EE}|R_g)\ P_{ST}(R_g;\epsilon,R_g^{\Theta})$ conditioned upon the $\langle R_g^{\ 2}\rangle$ of the full ensemble, where $L/2 = 0.19\ N$ nm is the maximum R_g. Then $\langle R_g^{\ 2}\rangle = \int_{R_C}^{L/2} dR_g\ R_g^{\ 2}P_{ST}(R_g;\epsilon,R_g^{\Theta})$ is varied by changing ϵ to seek an $\langle R_g^{\ 2}\rangle$ such that $P_{ST}(R_{EE}|\langle R_g^{\ 2}\rangle)$ produces an $\langle E\rangle_{sim}$ that matches $\langle E\rangle_{exp}$ (refs 27 and 28). Notably, for $n = 100$, the R_g's thus inferred (Figure 3D) are similar to those by CG (Figure 3C).

Despite the fact that ST accounts for excluded volume while CG does not, their R_g inferences are similar because they both presume a homogeneous ensemble. In the absence of this presumption, it is possible for SAW and Gaussian chain subensembles (Figure 3A,B) with moderate R_g's to have a significantly decreased population of chains with short R_{EE}'s in the FRET-sensitive regime (shaded areas in Figure 3), demonstrating that small $\langle E\rangle_{exp}$ can be consistent with a moderate R_g in a heterogeneous ensemble. However, if a homogeneous GC or ST ensemble is enforced a priori, the only way to depopulate the FRET-sensitive regime significantly to conform to a small $\langle E\rangle_{exp}$ is to expand all conformations proportionately, including those that are already open, thus entailing a large increase in average R_g (Figure 3C,D).

The Exemplary Case of the IDP Sic1. We apply the subensemble SAW method to analyze experimental smFRET data on Sic1 (Figure 4). One advantage of this method is that it addresses the heterogeneity of asphericity, A, whereas such heterogeneity is neglected in CG and the mean-field ST. Because R_{EE} correlates positively with A (Figure 2A) and a large R_{EE} entails low E, $\langle E\rangle$ correlates negatively with A as expected (Figure 4A).

As outlined above, we use the simulated $P(R_{EE}|R_g,A)$ subensemble data (Figure S3 and Table S1) to compute $\langle E\rangle_{sim}$. Denoted as $E(R_g,A)$, this quantity is compared with $\langle E\rangle_{exp}$ to identify a narrow range of R_g and A consistent with the smFRET data to within the experimental uncertainty $\sigma_{\langle E\rangle_{exp}} \approx 0.02$ (Figure 4B). Here $\sigma_{\langle E\rangle_{exp}}$ is calculated using the bootstrap method[71] by fitting the FRET efficiency histograms of $B = 1000$ replicate data sets generated by random sampling with replacement from the original set of fluorescence bursts. The inferred most probable (R_g,A) range is represented by the dark-red area in Figure 4C. This region exhibits a negative correlation between R_g and A. In other words, to maintain the same $\langle E\rangle_{exp}$, or essentially the same $\langle R_{EE}\rangle$, more aspherical polymers must be more compact. Based on the given smFRET data alone, the range of possible A is quite large ($0.3 \lesssim A \lesssim 0.7$). However, additional experimental data can be used to narrow down the range of compatible A (see below).

Conventional Approaches Overestimate R_g of Sic1 at Low FRET Efficiency. We now examine in detail the SAW-inferred Sic1 properties and their dependence on the ionic denaturant guanidinium chloride (GdmCl; Figure 5). IDP conformational properties depend on the total charge and the arrangement of charges along the chain.[50,72] The N-terminal 90-residue region of Sic1 studied here contains 11 positively charged amino acids at neutral pH (i.e., fraction of charged residues = 0.12). Sic1 is a "well-mixed" polyampholyte with a relatively uniform distribution of charges along the sequence (a low κ parameter[50] of 0.16). Because Sic1 is more compact than a random coil and possesses significant secondary structure propensity,[9] it is unlikely that Sic1 can be adequately described by a homogeneous ensemble as presumed by the CG and ST approaches.

In a recent smFRET study,[30] Sic1 exposed to increasing [GdmCl] exhibited an initial increase in $\langle E\rangle_{exp}$ as the chain collapsed with increased electrostatic screening of intrachain charge−charge repulsion upon addition of the GdmCl salt. This increase was followed by a $\langle E\rangle_{exp}$ decrease upon further addition of [GdmCl] as the chain expands under GdmCl's denaturing effect through weakening hydrophobic interactions

DOI: 10.1021/acs.jpcb.5b09133

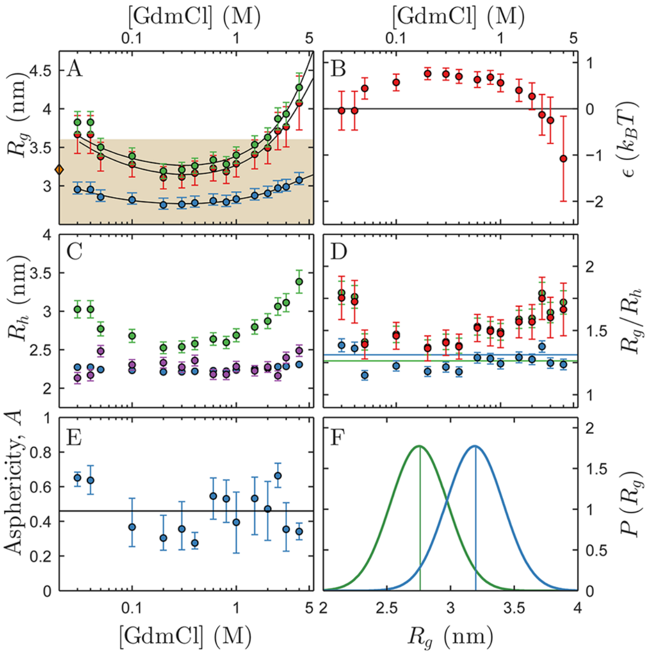


Figure 5. smFRET inference of Sic1 conformational properties. (A–E) Inferred and experimental quantites are given as functions of [GdmCl]. (A) Average R_g or R_g^0 (filled circles) inferred by CG (green), ST (red), and subensemble SAW (blue, R_{hc} = 0.4 nm). The shaded $1.69n^\nu < R_g < 2.20n^\nu$ region marks the expected range of an n = 100 fully unfolded protein ($R_g \approx 1.93\ n^\nu$, where $\nu = 0.598 \pm 0.028$ and R_g is in units of nanometers).[43] The error bars for CG- and ST-inferred average R_g are derived from $\sigma_{E,\text{exp}} = 0.02$, whereas the lower and upper error bar values for subensemble SAW are obtained by repeating the inference using, respectively, R_{hc} = 0.314 and 0.5 nm. The average R_g of Sic1 1−90 was measured using SAXS to be 3.21 ± 0.08 nm at zero denaturant [indicated by the orange diamond on the vertical axis in (A)].[9] Fits of the data to the analytical polyampholyte model are shown by the black curves. (B) Fitted intrachain interaction energy ϵ for the ST inference in (A). (C) Average R_h inferred by subensemble SAW (blue) and CG (green) are compared against R_h measured by FCS (purple). Here the simulated relation $\langle R_h\rangle = 0.791\langle R_g\rangle$ for the full ensemble of n = 100 Gaussian chains is used to obtain the CG R_h values from the inferred R_g values in (A). (D) The average R_g/R_h is the shape factor. Results shown for subensemble SAW (blue), CG (green) and ST (red) are obtained from the inferred R_g^0 or average R_g in (A) divided by the experimental R_h (purple circles in C). The horizontal lines mark the simulated $\langle R_g\rangle/\langle R_h\rangle$ ratios for the full ensemble (Figure S1F) of n = 100 SAWs (blue) or Gaussian chains (green). (E) Average asphericity A of Sic1 (blue circles with error bars) is estimated using the subensemble SAW shape factor in (D) and Figure S5. Included for comparison is the A = 0.46 horizontal line marking the average asphericity for the full ensemble of n = 100 SAWs. (F) Example R_g distributions consistent with the lowest (blue) and highest (green) $\langle E\rangle_{\text{exp}}$ for Sic1 ($\Delta E[P(R_g)] \le \sigma_{E,\text{exp}}$). Solid lines in the same color code are the corresponding inferred R_g^0's. The $\langle E\rangle_{\text{sim}}$'s (eq 4) computed using these distributions reproduce the corresponding $\langle E\rangle_{\text{exp}}$'s to within $\sigma_{E,\text{exp}}$.

and hydrogen bonding. Consistent with the expectation that Sic1 is not well described by a homogeneous ensemble, this study also showed the existence of at least two distinct conformational populations that are stable on the millisecond time scale.[30]

Conventional ST and CG overestimate Sic1 expansion at low [GdmCl] and high [GdmCl] because these approaches enforce a homogeneous ensemble in which a small $\langle E\rangle_{\text{exp}}$ can only be matched by stretching out all conformations, leading to unphysical inferred average R_g's that are larger than those measured by SAXS[9] or expected from the scaling relation for fully unfolded proteins[43] (shaded region in Figure 5A). In the CG approach, $\langle R_g^2\rangle$ is a fitting parameter. Because the total physical chain length N is a constant, $\langle R_g^2\rangle$ can only be increased in CG by increasing the Kuhn length, or the effective number, l_K, of residues moving in concert. At the same time, N_{eff}, the effective total number of Gaussian chain segments, must satisfy $N = N_{\text{eff}}l_K$, i.e., $N_{\text{eff}} = N/l_K$, to preserve the physical chain length. Figure 5A shows that an average $R_g \approx 4.3$ nm is needed for a homogeneously expanded Gaussian chain to conform to the small $\langle E\rangle_{\text{exp}}$ measured at [GdmCl] = 4 M. Since the C_α−C_α virtual bond length is ≈0.38 nm, this CG-inferred average R_g translates into an inferred $\sqrt{\langle R_{\text{EE}}{}^2\rangle} \approx \sqrt{6}(4.3) =$ 10.5 nm, or 10.5/0.38 = 27.7 C_α−C_α virtual bonds. Equating this to $\sqrt{N_{\text{eff}}}l_K = \sqrt{Nl_K}$ with N = 99 for our Sic1-plus-dye chain yields a Kuhn length $l_K \approx 7.8$ residues, which is clearly untenable. In a similar vein, the presumption of homogeneous expansion in the ST approach results in unphysical repulsive intrachain interaction energies for high [GdmCl] (Figure 5B). In contrast, the subensemble SAW method infers more compact conformations (smaller R_g) that are in line with SAXS data,[9] and shows significantly less modulation of R_g by [GdmCl] (Figure 5A).

The R_g, R_h, and Shape Factor of Sic1 Inferred by the Subensemble SAW Are Consistent with Other Experiments. We obtain inferred R_h from the computed $\langle R_h\rangle$ of the SAW R_g subensembles (Figure 5C). The resulting subensemble SAW-inferred R_h's (blue circles) are in excellent agreement with fluorescence correlation spectroscopy (FCS; purple circles). Notably, both the subensemble SAW-inferred and experimental FCS-determined R_h's show little dependence on [GdmCl]. In contrast, the CG-inferred R_h's (green circles) are significantly larger and exhibit a high sensitivity to [GdmCl] not observed in experiment. We further compare the subensemble SAW and conventionally inferred shape factors (Figure 5D). Shape factors (≈R_g/R_h; see Materials and Methods) are commonly used to characterize conformational shape. For Gaussian chains with length $N \to \infty$, $R_g/R_h \to 1.5$ (refs 58, 73); for $N \approx 100$, $R_g/R_h \approx 1.26$ (Figure S1F). Experimental shape factors are often less than the asymptotic limit of 1.5 for Gaussian chains even at high denaturant.[25,32,74] However, conventional ST- and CG-inferred R_g/R_h's for Sic1 at low [GdmCl] and high [GdmCl] exceed even the asymptotic limit (Figure 5D, red and green circles). In contrast, R_g/R_h's inferred by subensemble SAW (Figure 5D, blue circles) are in line with the expected value of 1.31 for SAWs with $N \approx 100$ (Figure S1F). In view of the aforementioned success of the subensemble SAW method, we also provide an estimate of the conformational asphericity of Sic1 by applying FRET and FCS data for R_g and R_h to the general relationship between shape factor and asphericity in Figure S5. Figure 5E shows that the inferred asphericity largely coincides with the full-ensemble $\langle A\rangle$ for SAW for [GdmCl] > 0.1 M but is noticeable higher at low salt.

From Most Probable Inferred R_g to a More Realistic R_g Distribution. As stated, the R_g inferred using SAW subensembles with narrow ranges of R_g may be viewed as the most probable R_g^0. Since it is unlikely that all conformations in a physical ensemble share the same R_g^0, it is reasonable to interpret R_g^0 as the most probable or average R_g of an inferred distribution $P_{\text{inf}}(R_g)$ (eq 5) that yields $\langle R_g\rangle_{\text{inf}} \approx R_g^0$ by taking a form approximately symmetric with respect to R_g^0. Such a $P(R_g)$

should lead to $\langle E\rangle_{\rm sim} \approx E(R_g^0)$ because $\Delta E(R_g) \approx -\Delta E(2R_g^0 - R_g)$ (Figure 4A) and thus the contributions from $P(R_g)$ and $P(2R_g^0 - R_g)$ to $\langle E\rangle_{\rm sim}$ (eq 4) approximately cancel. To illustrate this interpretation of R_g^0, Figure 5F shows two Gaussian distributions centered at the R_g^0's for our highest and lowest $\langle E\rangle_{\rm exp}$'s. These distributions are examples of possible $P_{\rm inf}(R_g)$s for Sic1 under the corresponding solvent conditions.

A Simple Polyampholyte Model for the Rollover of Protein Dimensions in Ionic Denaturant. All inferred R_g's in Figure 5A exhibit a rollover as noted above; i.e., R_g is higher for low and high [GdmCl]s than for intermediate [GdmCl]s. A similar rollover of conformational dimensions was observed for other IDPs such as IN and ProTα.[29] In that study, R_g([GdmCl]) was fit to a simple analytic polyampholyte model:

$$R_g = \sqrt{\frac{N}{6}}\,\alpha b\left(1 + \rho\frac{Ka}{1 + Ka}\right) \tag{9}$$

where α satisfies $\alpha^5 - \alpha^3 = (4/3)(3/2\pi)^{3/2}v'\sqrt{N}$, $v' = v + 4\pi f_+{}^2 l_B/(\kappa^2_D\, b^3)$, f_+ is fractional net charge on the IDP, vb^3 is the effective excluded volume of a neutral chain with the same N, l_B is the Bjerrum length, κ_D is the inverse Debye length, $b = 0.38$ nm is the virtual $C_\alpha - C_\alpha$ bond length, K and a are, respectively, the binding constant and activity of the denaturant, and ρ is the fractional change in R_g at high denaturant (large a) relative to that at low denaturant ($a \approx 0$). Details of this model can be found in refs 29, 30, and 75. Fitting of the CG- and ST-inferred Sic1 R_g's in Figure 5A leads to denaturant binding (K) and excluded volume (vb^3) similar to those reported for other IDPs,[29] but the fitted ρ is significantly higher (Table 1) because of the unphysically large R_g's inferred by these two conventional approaches. The rollover is much less prominent for the SAW-inferred R_g, but its [GdmCl] dependence cannot be fit by eq 9 unless v' is generalized to $v' = v + 4\pi\, C_{\rm el} f_+{}^2 l_B/(\kappa_D{}^2 b^3)$ with an additional fit parameter $C_{\rm el} \leq 1$. Physically, $C_{\rm el} < 1$ may indicate a combination of partial ionization of Sic1 and binding as well as electrostatic screening caused by the ions in GdmCl solutions.[76] $C_{\rm el} < 1$ may also reflect the subdued impact of electrostatics on the R_g of Sic1 because its dimensions are partly maintained by hydrophobicity.[9] As shown in Table 1, with $C_{\rm el} \approx 0.22$, values of the fitted K, ρ, and v for SAW are in line with those deduced previously for other IDPs.[29]

Table 1. Fitting Parameters for the Polyampholyte Model

inference method	K	ρ	v	$C_{\rm el}$	vb^3 [nm^3]
CG	0.1994	1.0191	6.2762	1.0000[a]	0.3444
ST	0.3123	0.7415	4.9310	1.0000[a]	0.2706
SAW	0.5770	0.2071	2.5907	0.2236	0.1422

[a] As in refs 29 and 30, $C_{\rm el}$ is not considered for CG and ST (eq 9); thus $C_{\rm el} = 1$ in $v' = v + 4\pi C_{\rm el} f_+{}^2 l_B/(\kappa_D{}^2 b^3)$ for these two cases.

DISCUSSION

Sic1 Data Help Delineate the Applicability of Conventional smFRET Inference Approaches. Our effort to assess methods for inferring conformational properties from smFRET data was prompted largely by data on Sic1,[30] for which conventional approaches did not appear to afford a physically reasonable interpretation. FRET efficiency informs about $P(R_{\rm EE})$. smFRET inference of conformational properties relies on our knowledge regarding the relationship between $P(R_{\rm EE})$ and the underlying conformational ensemble. It has been known from exact conformational enumeration and renormalization group analysis that excluded volume has a significant impact on $P(R_{\rm EE})$. For instance, the probability that two chain ends are close to each other, $P(R_{\rm EE}{\approx}0)$, can be significantly reduced by excluded volume. For the general scaling relation $P(R_{\rm EE}{\approx}0) \sim N^{-\nu}$, the Jacobson–Stockmayer exponent $\nu = 3/2$ applies for Gaussian chains but larger exponents $\nu \gtrsim 2$ and thus faster decreases of $P(R_{\rm EE}{\approx}0)$ with increasing N are entailed by excluded volume.[60,69,70] Consequently, when $\langle E\rangle_{\rm exp}$ is large indicating that the two chain ends are close to each other on average, the R_g's inferred by CG and SG (blue curves in Figure 3B,C) are smaller than that inferred using the more physical SAW model (blue curve in Figure 3A). For conventional ST, although excluded volume is accounted for approximately in the original Sanchez theory for R_g using a mean-field argument for spherically shaped conformations,[38] excluded volume and chain asphericity are neglected in the presumed conditional probability $P_{\rm ST}(R_{\rm EE}|R_g)$ (eq 8) introduced[27] for the ST inference approach. As a result, on one hand $P_{\rm ST}(R_{\rm EE}|R_g)$ is drastically inaccurate for large R_g (Figure S6). On the other hand, the full-ensemble distribution of $R_{\rm EE}$ in ST is shifted to lower $R_{\rm EE}$'s relative to that simulated using the more physical SAW model (Figure S7). Taken together, ST is inaccurate for both low and high $\langle E\rangle_{\rm exp}$'s (Figure 3), effectively restricting the FRET range in which reliable R_g inferences can be made to about 30–70%.

IDPs May Not Be Adequately Represented by Homogeneously Expanded or Contracted Conformational Ensembles. In addition to their neglect of excluded volume and conformational asphericity, a fundamental reason why conventional ST and CG cannot provide physically reasonable inferences for Sic1 is these approaches' presumption of a homopolymer-like homogeneous conformational ensemble. In fact, since full-ensemble $P(R_{\rm EE})$s for ST and SAW are not too dissimilar (Figure S7, bottom panels), even if homogeneous SAW ensembles that can account for excluded volume and asphericity were used for inference (as has been performed for other proteins[26]), unphysically large R_g's similar to those obtained by ST and CG would be inferred for Sic1 in high denaturant, i.e., at low FRET (see below). Experiments, however, show that Sic1 ensembles are heterogeneous, not homopolymer-like.[9,30] From a biological standpoint, while IDP conformational distributions can be modulated by solvent conditions,[77] it is unlikely in general that IDPs performing specific functions can be described adequately by homogeneous ensembles.[5] Even unfolded states of globular proteins conforming to the dimension scaling of SAW homopolymers[43,78,79] are not necessarily well described by homogeneous conformational ensembles.[80,81] If the conformational ensemble of an unfolded protein is not homogeneous, conventional smFRET inference method can lead to significantly inaccurate R_g, as has been demonstrated using a native-centric model that captures part of the sequence-dependent conformational heterogeneity.[48]

Subensemble SAW Inference as a First Step toward an Improved Account of Conformational Ensembles in smFRET Experiments. To relieve smFRET inference from the presumption of a homogeneous conformational ensemble, here we introduce a minimalist yet systematic subensemble SAW method that allows one to infer a most probable R_g. For readers interested in applying this method, relevant simulated data for several chain lengths are provided in Tables S3–S7. In essence, our method provides a resolution of FRET efficiency into R_g components. Such a resolution is useful for the construction of inferred R_g distributions and is readily generalizable to other conformational properties. A general comparison between the R_g's inferred using our subensemble SAW and conventional ST (Figure 6) indicates that the discrepancy between the two methods is largest for small $\langle E\rangle_{\text{exp}}$ (see also Figure 3), and the discrepancy increases with chain length (Figure S8). As we have explained, the unphysically large R_g inferred by ST for small $\langle E\rangle_{\text{exp}}$ originates from its presumption of a homogeneous conformational ensemble. In fact, as discussed above, if a homogeneous SAW ensemble is imposed (bottom panels of Figure S7), the inferred R_g's can also be unphysically large for small $\langle E\rangle_{\text{exp}}$ (circles in the middle panel of Figure 6), although they are not as high as the R_g's inferred by ST. In the case of a homogeneous SAW ensemble, the unphysicality is reflected by an unphysically large $R_{\text{hc}} > 7.0$ nm (Figure S7) that grossly overstates the excluded volume of an amino acid residue. In contrast, the subensemble SAW-inferred R_g's stay within reasonable physical limits, covering ranges of R_g values that are only slightly wider than that expected for $\sqrt{\langle R_g^2\rangle}$ of the full SAW ensemble (shaded regions in Figure 6). In comparison with ST, the steepness of the subensemble-SAW curves in Figure 6 reduces the variation in inferred R_g when large FRET changes are observed and therefore the method is less prone to yield unphysical results at the high and low ends of the FRET range, thus allowing a larger usable range of $\langle E\rangle_{\text{exp}}$.

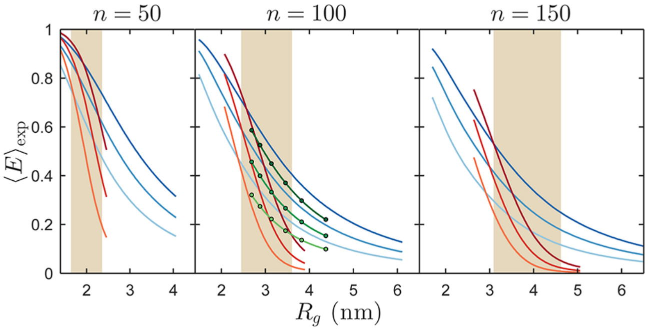


Figure 6. Comparing the subensemble SAW and ST approaches to smFRET inference of R_g. For a given $\langle E\rangle_{\text{exp}}$ (vertical axis), the R_g^0 inferred by subensemble SAW (red curves) and the average R_g inferred by ST (blue curves) for R_0 = 5.0, 6.0, and 7.0 nm (represented respectively by three curves with progressively darker hue) are provided for chain lengths n = 50, 100, and 150. The shaded regions demarcate the expected ranges for average R_g of fully unfolded proteins with the same n.[43] The implication of presuming a homogeneous SAW ensemble is explored further in the middle panel. Here, for n = 100, the inferred average R_g's of presumed full SAW ensembles with R_{hc} = 0.314, 0.4, 0.5, 0.6, 0.7, and 0.8 nm (see also Figure S7 and Table S8) are shown by green circles interpolated by curves with the same hue. As for subensemble SAW (red) and ST (blue), full-ensemble SAW results (green) for R_0 = 5.0, 6.0, and 7.0 nm are depicted in progressively darker hues.

CONCLUSIONS

The utility of the subensemble SAW method is hereby supported by its application to Sic1 that yielded R_g *and* R_h consistent with those determined by independent experimental techniques. This success brightens the prospect that the subdued sensitivity of subensemble SAW-inferred protein dimensions to change in denaturant, salt, pH, or solvent viscosity may offer clues for resolving the long-standing and puzzling discrepancy between FRET- and SAXS-measured R_g values of unfolded proteins.[44−46] Nonetheless, it is important to emphasize that it is a priori possible for the *real* underlying $P(R_g)$ to take certain functional forms such that it yields $\langle E\rangle_{\text{exp}}$ yet the average R_g does not coincide with the inferred most-probable R_g^0. Hence, to narrow down possibilities for ensemble properties in general and $P(R_g)$ in particular, complementary experiments[9,13] or theoretical simulations[36,66,82] are needed. In this regard, smFRET relaxation dynamics[31,83] data from different labeling positions,[48] different dye pairs, and three-color FRET should offer more direct information about $P(R_{\text{EE}})$ and thus additional constraints for inferring $P(R_g)$. To date, atomic simulations of IDPs are limited by the extreme sensitivity of predicted behaviors to the choice of force field[84] and the fact that IDP dimensions computed using common explicit-water molecular dynamics force fields tend to be significantly more compact than those determined experimentally.[85,86] Nonetheless, recent efforts in fine-tuning the strength of water−protein interactions[87,88] and the explicit water model itself[89] offer hope that the atomic force fields can be systematically improved to better capture IDP physics. Further developments of these experimental and computational techniques will be invaluable for our understanding of disordered protein states.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpcb.5b09133.

> Chain length dependence of shape and size measures, variation and correlation of A, R_g, and R_{EE} among Gaussian chains, distribution of end-to-end distance $P(R_{\text{EE}}|R_g,A)$ conditioned upon narrow ranges of A and R_g, parameters for fitted conditional end-to-end distance distributions, benchmarking the $P_{\text{ST}}(R_{\text{EE}}|R_g)$ expression (eq 8 in the main text) used in conventional ST approach to smFRET inference, etc. (PDF)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: claudiu.gradinaru@utoronto.ca (C.C.G.).
*E-mail: chan@arrhenius.med.toronto.edu (H.S.C.).

Author Contributions

¶J.S. and G.-N.G. contributed equally.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Financial support for this work was provided by a Natural Science and Engineering Research Council of Canada Discovery Grant to C.G. and a Canadian Institutes of Health Research Operating Grant to H.S.C. We are grateful for the allotments of computational resources from SciNet of Compute Canada.

REFERENCES

(1) Uversky, V. N.; Oldfield, C. J.; Dunker, A. K. Intrinsically disordered proteins in human diseases: Introducing the D^2 concept. *Annu. Rev. Biophys.* **2008**, 37, 215−246.

(2) Tompa, P. Intrinsically disordered proteins: A 10-year recap. *Trends Biochem. Sci.* **2012**, *37*, 509−516.

(3) Marsh, J. A.; Teichmann, S. A.; Forman-Kay, J. D. Probing the diverse landscape of protein flexibility and binding. *Curr. Opin. Struct. Biol.* **2012**, *22*, 643−650.

(4) Liu, Z.; Huang, Y. Advantages of proteins being disordered. *Protein Sci.* **2014**, *23*, 539−550.

(5) Chen, T.; Song, J.; Chan, H. S. Theoretical perspectives on nonnative interactions and intrinsic disorder in protein folding and binding. *Curr. Opin. Struct. Biol.* **2015**, *30*, 32−42.

(6) Nash, P.; Tang, X.; Orlicky, S.; Chen, Q.; Gertler, F. B.; Mendenhall, M. D.; Sicheri, F.; Pawson, T.; Tyers, M. Multisite phosphorylation of a CDK inhibitor sets a threshold for the onset of DNA replication. *Nature* **2001**, *414*, 514−521.

(7) Borg, M.; Mittag, T.; Pawson, T.; Tyers, M.; Forman-Kay, J. D.; Chan, H. S. Polyelectrostatic interactions of disordered ligands suggest a physical basis for ultrasensitivity. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 9650−9655.

(8) Tompa, P.; Fuxreiter, M. Fuzzy complexes: Polymorphism and structural disorder in protein-protein interactions. *Trends Biochem. Sci.* **2008**, *33*, 2−8.

(9) Mittag, T.; Marsh, J.; Grishaev, A.; Orlicky, S.; Lin, H.; Sicheri, F.; Tyers, M.; Forman-Kay, J. D. Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase. *Structure* **2010**, *18*, 494−506.

(10) Fuxreiter, M.; Simon, I.; Bondos, S. Dynamic protein-DNA recognition: Beyond what can be seen. *Trends Biochem. Sci.* **2011**, *36*, 415−423.

(11) Song, J.; Ng, S. C.; Tompa, P.; Lee, K. A. W.; Chan, H. S. Polycation-π interactions are a driving force for molecular recognition by an intrinsically disordered oncoprotein family. *PLoS Comput. Biol.* **2013**, *9*, e1003239.

(12) Dill, K. A.; Shortle, D. Denatured states of proteins. *Annu. Rev. Biochem.* **1991**, *60*, 795−825.

(13) Choy, W.-Y.; Forman-Kay, J. D. Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *J. Mol. Biol.* **2001**, *308*, 1011−1032.

(14) Schuler, B.; Hofmann, H. Single-molecule spectroscopy of protein folding dynamics−expanding scope and timescales. *Curr. Opin. Struct. Biol.* **2013**, *23*, 36−47.

(15) Gelman, H.; Gruebele, M. Fast protein folding kinetics. *Q. Rev. Biophys.* **2014**, *47*, 95−142.

(16) Banerjee, P. R.; Deniz, A. A. Shedding light on protein folding landscapes by single-molecule fluorescence. *Chem. Soc. Rev.* **2014**, *43*, 1172−1188.

(17) Huang, F.; Ying, L.; Fersht, A. R. Direct observation of barrier-limited folding of BBL by single-molecule fluorescence resonance energy transfer. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 16239−16244.

(18) Liu, J.; Campos, L. A.; Cerminara, M.; Wang, X.; Ramanathan, R.; English, D. S.; Muñoz, V. Exploring one-state downhill protein folding in single molecules. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 179−184.

(19) Knott, M.; Chan, H. S. Criteria for downhill protein folding: Calorimetry, chevron plot, kinetic relaxation, and single-molecule radius of gyration in chain models with subdued degrees of cooperativity. *Proteins: Struct., Funct., Genet.* **2006**, *65*, 373−391.

(20) Chung, H. S.; McHale, K.; Louis, J. M.; Eaton, W. A. Single-molecule fluorescence experiments determine protein folding transition path times. *Science* **2012**, *335*, 981−984.

(21) Chung, H. S.; Eaton, W. A. Single-molecule fluorescence probes dynamics of barrier crossing. *Nature* **2013**, *502*, 685−688.

(22) Zhang, Z.; Chan, H. S. Transition paths, diffusive processes, and preequilibria of protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 20919−20924.

(23) Elbaum-Garfinkle, S.; Cobb, G.; Compton, J. T.; Li, X.-H.; Rhoades, E. Tau mutants bind tubulin heterodimers with enhanced affinity. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 6311−6316.

(24) McCarney, E. R.; Werner, J. H.; Bernstein, S. L.; Ruczinski, I.; Makarov, D. E.; Goodwin, P. M.; Plaxco, K. W. Site-specific dimensions across a highly denatured protein; a single molecule study. *J. Mol. Biol.* **2005**, *352*, 672−682.

(25) Sherman, E.; Haran, G. Coil-globule transition in the denatured state of a small protein. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 11539−11543.

(26) Merchant, K. A.; Best, R. B.; Louis, J. M.; Gopich, I. V.; Eaton, W. A. Characterizing the unfolded states of proteins using single-molecule FRET spectroscopy and molecular simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 1528−1533.

(27) Ziv, G.; Haran, G. Protein folding, protein collapse, and Tanford's transfer model: Lessons from single-molecule FRET. *J. Am. Chem. Soc.* **2009**, *131*, 2942−2947.

(28) Hofmann, H.; Nettels, D.; Schuler, B. Single-molecule spectroscopy of the unexpected collapse of an unfolded protein at low pH. *J. Chem. Phys.* **2013**, *139*, 121930.

(29) Müller-Späth, S.; Soranno, A.; Hirschfeld, V.; Hofmann, H.; Rüegger, S.; Reymond, L.; Nettels, D.; Schuler, B. Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 14609−14614; Correction for Müller-Späth et al., Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 16693.

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

(31) Soranno, A.; Buchli, B.; Nettels, D.; Cheng, R. R.; Müller-Späth, S.; Pfeil, S. H.; Hoffmann, A.; Lipman, E. A.; Makarov, D. E.; Schuler, B. Quantifying internal friction in unfolded and intrinsically disordered proteins with single-molecule spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 17800−17806.

(32) 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*, 16155−16160.

(33) Hanson, J. A.; Brokaw, J.; Hayden, C. C.; Chu, J.-W.; Yang, H. Structural distributions from single-molecule measurements as a tool for molecular mechanics. *Chem. Phys.* **2012**, *396*, 61−71.

(34) Kalinin, S.; Peulen, T.; Sindbert, S.; Rothwell, P. J.; Berger, S.; Restle, T.; Goody, R. S.; Gohlke, H.; Seidel, C. A. M. A toolkit and benchmark study for FRET-restrained high-precision structural modeling. *Nat. Methods* **2012**, *9*, 1218−1225.

(35) Choi, U. B.; McCann, J. J.; Weninger, K. R.; Bowen, M. E. Beyond the random coil: Stochastic conformational switching in intrinsically disordered proteins. *Structure* **2011**, *19*, 566−576.

(36) Nath, A.; Sammalkorpi, M.; DeWitt, D. C.; Trexler, A. J.; Elbaum Garfinkle, S.; O'Hern, C. S.; Rhoades, E. The conformational ensembles of α-synuclein and tau: Combining single-molecule FRET and simulations. *Biophys. J.* **2012**, *103*, 1940−1949.

(37) König, K.; Zarrine-Afsar, A.; Aznauryan, M.; Soranno, A.; Wunderlich, B.; Dingfelder, F.; Stüber, J. C.; Plückthun, A.; Nettels, D.; Schuler, B. Single-molecule spectroscopy of protein conformational dynamics in live eukaryotic cells. *Nat. Methods* **2015**, *12*, 773−779.

(38) Sanchez, I. C. Phase transition behavior of the isolated polymer chain. *Macromolecules* **1979**, *12*, 980−988.

(39) Kuhn, W. Über die gestalt fadenförmiger moleküle in lösungen. (Concerning the shape of thread shapes molecules in solution). *Colloid Polym. Sci.* **1934**, *68*, 2−15.

(40) Šolc, K.; Stockmayer, W. H. Shape of a random-flight chain. *J. Chem. Phys.* **1971**, *54*, 2756−2757.

(41) Mazur, J.; Guttman, C. M.; McCrackin, F. L. Monte Carlo studies of self-interacting polymer chains with excluded volume. II. Shape of a chain. *Macromolecules* **1973**, *6*, 872−874.

(42) Rudnick, J.; Gaspari, G. The asphericity of random walks. *J. Phys. A: Math. Gen.* **1986**, *19*, L191−L193.

(43) 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 behavior and the dimensions of chemically unfolded proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 12491−12496.

(44) Yoo, T. Y.; Meisburger, S. P.; Hinshaw, J.; Pollack, L.; Haran, G.; Sosnick, T. R.; Plaxco, K. Small-angle x-ray scattering and single-molecule FRET spectroscopy produce highly divergent views of the low-denaturant unfolded state. *J. Mol. Biol.* **2012**, *418*, 226−236.

(45) Haran, G. How, when and why protein collapse: the relation to folding. *Curr. Opin. Struct. Biol.* **2012**, *22*, 14−20.

(46) Watkins, H. M.; Simon, A. J.; Sosnick, T. R.; Lipman, E. A.; Hjelm, R. P.; Plaxco, K. W. Random coil negative control reproduces the discrepancy between scattering and FRET measurements of denatured protein dimensions. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 6631−6636.

(47) Chan, H. S.; Zhang, Z.; Wallin, S.; Liu, Z. Cooperativity, local-nonlocal coupling, and nonnative interactions: Principles of protein folding from coarse-grained models. *Annu. Rev. Phys. Chem.* **2011**, *62*, 301−326.

(48) O'Brien, E. P.; 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.

(49) 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. U. S. A.* **2010**, *107*, 8183−8188.

(50) Das, R. K.; Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by linear sequence distribution of oppositely charged residues. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 13392−13397.

(51) Knott, M.; Best, R. B. Discriminating binding mechanisms of an intrinsically disordered protein via a multi-state coarse-grained model. *J. Chem. Phys.* **2014**, *140*, 175102.

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

(53) Levitt, M. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **1976**, *104*, 59−107.

(54) Miyazawa, S.; Jernigan, R. L. Estimation of effective interresidue contact energies from protein crystal structures: Qusai-chemical approximation. *Macromolecules* **1985**, *18*, 534−552.

(55) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of state calculation by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087−1092.

(56) Verdier, P. H.; Stockmayer, W. H. Monte Carlo calculations on dynamics of polymers in dilute solution. *J. Chem. Phys.* **1962**, *36*, 227−235.

(57) Lal, M. Monte Carlo computer simulation of chain molecules. I. *Mol. Phys.* **1969**, *17*, 57−64.

(58) Guttman, C. M.; McCrackin, F. L.; Han, C. C. Monte Carlo calculation of the hydrodynamic radius at the Θ point. Deviations from analytical Gaussian behavior. *Macromolecules* **1982**, *15*, 1205−1207.

(59) Yamakawa, H. *Modern Theory of Polymer Solutions*; Harper & Row: New York, 1971; pp 23−35, 269−285.

(60) Chan, H. S.; Dill, K. A. The effects of internal constraints on the configurations of chain molecules. *J. Chem. Phys.* **1990**, *92*, 3118−3135; Erratum: "The effects of internal constraints on the configurations of chain molecules" [J. Chem. Phys. 92, 3118 (1990)] . *J. Chem. Phys.* **1997**, *107*, 10353.

(61) Badali, D.; Gradinaru, C. C. The effect of Brownian motion of fluorescent probes on measuring nanoscale distances by Förster resonance energy transfer. *J. Chem. Phys.* **2011**, *134*, 225102.

(62) Theodorou, D. N.; Suter, U. W. Shape of unperturbed linear polymers: Polypropylene. *Macromolecules* **1985**, *18*, 1206−1214.

(63) Parry, M.; Fischbach, E. Probability distribution of distance in a uniform ellipsoid: Theory and applications to physics. *J. Math. Phys.* **2000**, *41*, 2417−2433.

(64) Rawdon, E. J.; Kern, J. C.; Piatek, M.; Plunkett, P.; Stasiak, A.; Millett, K. C. Effect of knotting on the shape of polymers. *Macromolecules* **2008**, *41*, 8281−8287.

(65) Chan, H. S.; Dill, K. A. Polymer principles in protein structure and stability. *Annu. Rev. Biophys. Biophys. Chem.* **1991**, *20*, 447−490.

(66) Kellner, R.; Hofmann, H.; Barducci, A.; Wunderlich, B.; Nettels, D.; Schuler, B. Single-molecule spectroscopy reveals chaperone-mediated expansion of substrate protein. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 13355−13360.

(67) Flory, P. J.; Fisk, S. Effect of volume exclusion on the dimensions of polymer chains. *J. Chem. Phys.* **1966**, *44*, 2243−2248.

(68) Fujita, H.; Norisuye, T. Some topics concerning the radius of gyration of linear polymer molecules in solution. *J. Chem. Phys.* **1970**, *52*, 1115−1120.

(69) des Cloizeaux, J. Short range correlation between elements of a long polymer in a good solvent. *J. Phys. (Paris)* **1980**, *41*, 223−238.

(70) Freed, K. F. *Renormalization Group Theory of Macromolecules*; John Wiley & Sons: New York, 1987.

(71) Efron, B. Bootstrap methods: Another look at the jack knife. *Ann. Stat.* **1979**, *7*, 1−26.

(72) Marsh, J. A.; Forman-Kay, J. D. Sequence determinants of compaction in intrinsically disordered proteins. *Biophys. J.* **2010**, *98*, 2383−2390.

(73) Grosberg, A. Y.; Kokhlov, A. R. *Statistical Physics of Macromolecules*; American Institute of Physics: Melville, NY, 1994.

(74) Wilkins, D. K.; Grimshaw, S. B.; Receveur, V.; Dobson, C. M.; Jones, J. A.; Smith, L. J. Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques. *Biochemistry* **1999**, *38*, 16424−16431.

(75) Ha, B.-Y.; Thirumalai, D. Conformations of a polyelectrolyte chain. *Phys. Rev. A: At., Mol., Opt. Phys.* **1992**, *46*, R3012−R3015.

(76) Muthukumar, M. *Polymer Translocation*; CRC Press: Boca, Raton, FL, 2011; pp 79−114.

(77) Levine, Z. A.; Larini, L.; LaPointe, N. E.; Feinstein, S. C.; Shea, J.-E. Regulation and aggregation of intrinsically disordered peptides. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 2758−2763.

(78) Ohkubo, Y. Z.; Brooks, C. L. Exploring Flory's isolated-pair hypothesis: Statistical mechanics of helix-coil transitions in polyalanine and the C-peptide from RNase A. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 13916−13921.

(79) Pappu, R. V.; Srinivasan, R.; Rose, G. D. The Flory isolated-pair hypothesis is not valid for polypeptide chains: Implications for protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 12565−12570.

(80) Plaxco, K. W.; Gross, M. Unfolded, yes, but random? Never! *Nat. Struct. Biol.* **2001**, *8*, 659−660.

(81) Fitzkee, N. C.; Rose, G. D. Reassessing random-coil statistics in unfolded proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 12497−12502.

(82) Best, R. B.; Hofmann, H.; Nettels, D.; Schuler, B. Quantitative interpretation of FRET experiments via molecular simulation: force field and validation. *Biophys. J.* **2015**, *108*, 2721−2731.

(83) Soranno, A.; Longhi, R.; Bellini, T.; Buscaglia, M. Kinetics of contact formation and end-to-end distance distributions of swollen disordered peptides. *Biophys. J.* **2009**, *96*, 1515−1528.

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

(85) Piana, S.; Klepeis, J. L.; Shaw, D. E. Assessing the accuracy of physical models used in protein-folding simulations: Quantitative evidence from long molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **2014**, *24*, 98−105.

(86) Skinner, J. J.; Yu, W.; Gichana, E. K.; Baxa, M. C.; Hinshaw, J.; Freed, K. F.; Sosnick, T. R. Benchmarking all-atom simulations using hydrogen exchange. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 15975−15980.

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

(88) 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*, 5543−5553.

(89) 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.