Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements

Fuertes, Gustavo; Banterle, Niccolò; Ruff, Kiersten M; Chowdhury, Aritra; Mercadante, Davide; Koehler, Christine; Kachala, Michael; Estrada Girona, Gemma; Milles, Sigrid; Mishra, Ankur

Published in:
Proceedings of the National Academy of Sciences of the United States of America

DOI:
10.1073/pnas.1704692114

IMPORTANT NOTE: You are advised to consult the publisher’s version (publisher’s PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2017

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements


Published online July 17, 2017

Unfolded states of proteins and native states of intrinsically disordered proteins (IDPs) populate heterogeneous conformational ensembles in solution. The average sizes of these heterogeneous systems, quantified by the radius of gyration \(R_g\), can be measured by small-angle X-ray scattering (SAXS). Another parameter, the mean dye-to-dye distance \(R_{D\text{d}}\) for proteins with fluorescently labeled termini, can be estimated using single-molecule Förster resonance energy transfer (smFRET). A number of studies have reported inconsistencies in inference drawn from the two sets of measurements for the dimensions of unfolded proteins and IDPs in the absence of chemical denaturants. These differences are typically attributed to the influence of fluorescent labels used in smFRET and to the impact of high concentrations and averaging features of SAXS. By measuring the dimensions of a collection of labeled and unlabeled polypeptides using smFRET and SAXS, we directly assessed the contributions of dyes to the experimental values \(R_g\) and \(R_{D\text{d}}\). For chemically denatured proteins we obtain mutual consistency in our inferences based on SAXS vs. smFRET. IDPs. The average dimensions of these systems are quantified under native conditions are important for understanding the details of protein folding and the functions of IDPs. The average dimensions of these systems are quantified using the mean radius of gyration and mean end-to-end distance, measured by small-angle X-ray scattering (SAXS) and single-molecule Förster resonance energy transfer (smFRET), respectively, although systematic discrepancies emerge from these measurements. Through holistic sets of studies, we find that the disagreements arise from chemical heterogeneity that is inherent to heteropolymeric systems. This engenders a decoupling between different measures of overall sizes and shapes, thus leading to discrepant inferences based on SAXS vs. smFRET. Our findings point the way forward to obtaining comprehensive descriptions of ensembles of heterogeneous systems.
hydrochloride (14). Upon dilution of denaturants, proteins collapse and fold to form compact structures. An unresolved issue is the nature of the collapse transition (2, 4, 13, 15, 16). Inferences from smFRET measurements suggest that proteins, including IDPs, undergo continuous contraction as the denaturant concentration is decreased (4, 16, 17). The implication for protein folding is that the interactions between denaturants and chain units appear to have a homogenizing effect on the fluctuations that are smaller than the dimensions of unfolded proteins and IDPs. Also, with typical dye pairs, smFRET affords accurate estimates of distances that are limited to the range of ~2 nm to ~10 nm. In contrast, SAXS measurements do not require the attachment of labels and the measured scattering intensities are weighted averages over all of the protein molecules in solution, thus enabling direct investigations of chain dimensions. However, SAXS experiments require higher protein concentrations and the averaging over the conformations of all molecules in solutions makes it difficult to obtain assessments of conformational populations and insights regarding fluctuations that are smaller than the global dimensions of the protein. Here, we ask whether the discrepancies between inferences drawn from SAXS vs. smFRET measurements are due to the perceived weaknesses of the methods themselves or because the two methods provide complementary insights that have to be analyzed jointly to obtain a robust quantitative assessment of conformational features of heterogeneous systems.

We performed SAXS measurements on labeled and unlabeled IDPs as well as chemically denatured proteins. Inferences from these measurements were compared with those from smFRET measurements of labeled molecules. Atomic Monte Carlo simulations based on the ABSINTH (self-assembly of biomolecules studied by an implicit, novel, and tunable Hamiltonian) implicit solvation model (24) were used to generate quantitative insights into the joint analysis of SAXS and smFRET data. We made rigorous comparisons between $R_{GL}$, calculated from smFRET measurements and atomic simulations of dye-labeled proteins, and the values of $R_{GL}$ obtained from SAXS measurements. We find that the ads do not significantly influence the SAXS measurements, under either native conditions or denatured conditions. Instead, estimates of $R_G$ and $R_E$ yield different inferences because these quantities interrogate distinct length scales and are influenced by very different types of averaging. For finite-sized heteropolymers, we show that large changes in the equilibrium conformation can be detected with negligible changes in $R_G$ (22, 25). We discuss that such differences are minimized in long homopolymers and long block copolymers that are characterized by the chemical similarity of the interacting units (25). Accordingly, the estimates of $R_G$ and $R_E$ lead to mutually consistent inferences regarding conformational preferences and the physics of coil-to-globule transitions for long homopolymers (26). A similar robustness prevails for proteins in highly denaturing environments where preferential interactions between denaturants and chain units appear to have a homogenizing effect on the fluctuations that are smaller than the dimensions of unfolded proteins and IDPs.

Why do SAXS and smFRET lead to apparently conflicting inferences regarding the collapse transition and the nature of heterogeneous ensembles, especially under physiologically relevant conditions and away from high concentrations of denaturants? Both techniques have distinct strengths and weaknesses (15, 20–23). Initially, we discuss the collapse transition for protein folding and for IDPs have led to numerous debates (4, 9, 15, 20–23).

Results

The Protein Set, Labeling Scheme, and Experimental Design. We selected a set of 10 protein sequences with lengths between 38 residues and 178 residues, covering different amino acid compositions and physicochemical properties (Fig. 1E and SI Appendix, Table S2 and Note S1). Three of the 10 proteins fold to form stable structures under native conditions whereas the other 7 are IDPs that remain disordered in the absence of denaturant. To avoid potential uncertainties that can (30), but must not (31) arise from...
random labeling of proteins, we exploited the advantages of site-specific, unambiguous dual labeling. Specifically, the donor dye Alexa488 (SI Appendix, Fig. S1A) was attached via oxime ligation to the unnatural amino acid p-acetylphenylalanine, engineered at the penultimate position of the polypeptide chain using amber suppression technology (32). The acceptor fluorophore Alexa594 (SI Appendix, Fig. S1B) was reacted with a cysteine residue located at the second position via maleimide chemistry. Single-molecule measurements were made using the doubly labeled proteins under strongly denaturing conditions (6 M urea) and in (near)-native conditions with urea virtually absent (see buffer details in SI Appendix, Note S2 and experimental smFRET details in SI Appendix, Note S3).

SAXS measurements were performed using unlabeled and labeled samples (see experimental SAXS details in SI Appendix, Note S4). As an example of experimental results, we show the SAXS profiles (Fig. 2A and B) and Guinier fits (Fig. 2C and D) for the IDP NUS, under denaturing (Fig. 2A and C) and native conditions (Fig. 2B and D). The Rg is typically calculated from a plot of the SAXS intensity I(q) vs. the momentum transfer q, using the Guinier approximation (SI Appendix, Note S4):

$$I(q) = I(0) - q^2 R_g^2/3.$$  

Alternatively, Rg can be estimated from the pair-distance distribution function (SI Appendix, Note S4). RgLU and RgLL, calculated from either the Guinier approximation or the pair-distance distribution function were found to be similar to one another (values in SI Appendix, Table S3). Fig. 2 also shows the smFRET histograms (Fig. 2E and F) and the most common distance distribution functions used to infer RLL from \(E_{\text{FRET}}\) (Fig. 2G and H) corresponding to the same protein (NUS) under denaturing (Fig. 2E and G) and native (Fig. 2F and H) conditions. The peak at \(E_{\text{FRET}}\) near zero in the smFRET histograms arises from donor-only species (33), whereas the second population, originating from molecules containing an active donor–acceptor pair, appears at \(E_{\text{FRET}} \sim 0.55\) for native NUS. The parameter RLL quantifies the ensemble-averaged root mean-squared distance between the donor and acceptor dyes and it is related to \(E_{\text{FRET}}\) via

$$\langle E_{\text{FRET}} \rangle = \int_0^{\infty} \frac{1}{1 + (R_{LL}/R_0)^2} P(r_{DA}, R_{LL}) \, dr_{DA}. \quad [2]$$

Here, R0 or the Förster distance (the distance at which FRET efficiency is 50%) depends on the specific dye pair and it is usually around 5 nm (our measured values are in SI Appendix, Table S4); \(P(r_{DA}, R_{LL})\) is a probability distribution function that quantifies the likelihood of realizing values of interdye distances, within an interval \(r_{DA}\) and \(r_{DA} + dr_{DA}\) given a mean donor-to-acceptor distance of \(R_{LL}\). The form for \(P(r_{DA}, R_{LL})\) is unknown a priori and is usually chosen from a list of polymer models that includes the Gaussian chain model, the self-avoiding random walk (SARW) model, or a distribution of points inside a sphere of fixed diameter (34) (SI Appendix, Notes S3 and S8). These models are parameterized in terms of \(R_{LL}\), which reflects the contribution of the first (mean) and second (variance) moments of the distribution \(P(r_{DA}, R_{LL})\) (35).

Fig. 2 shows illustrative datasets from smFRET and SAXS measurements. The complete sets of data from smFRET measurements for all proteins and conditions are shown in SI Appendix, Table S4 (FRET parameters); SI Appendix, Table S5 (anisotropies); SI Appendix, Fig. S2 (gamma and quantum yields); and SI Appendix, Fig. S3 (FRET efficiencies). Similarly, the complete SAXS data are shown in SI Appendix, Fig. S4 A–D (SAXS profiles, Guinier plots, Kratky plots, and pair distance distribution function, respectively). Importantly, to deal with the fact that smFRET and SAXS measurements were performed at very different concentrations, we
carried out additional experiments to ensure that the large differences in concentration are not the source of discrepancies in inferences drawn from these measurements (SI Appendix, Note S5 and Fig. S5). Analyses of the datasets, which include information regarding $E_{\text{FRET}}$ (originating from smFRET), $R_{\text{G,L}}$(measured by SAXS), and $R_{\text{G,U}}$(also from SAXS), are presented in the following sections, first for denatured proteins and then for IDPs under native conditions.

**Measurements of $R_G$ and Estimates of $R_L$ from Measurements of ($E_{\text{FRET}}$) Yield Mutually Consistent Inferences for Denatured Proteins.** We performed SAXS experiments using labeled and unlabeled molecules to quantify the impact of fluorescent dyes on the global dimensions of flexible polymers. SI Appendix, Fig. S6A shows $R_{\text{G,L,D}}$(yellow points) and $R_{\text{G,U,D}}$(red points) calculated from the Guinier approximation as a function of the number of residues $(N_{\text{RES}})$ for eight proteins denatured in 6 M urea. Here, the letters $L$ and $U$ in the subscripts refer to labeled vs. unlabeled molecules and $D$ refers to denaturing conditions (and $N$ refers to native). Our dataset includes five IDPs and three proteins that fold autonomously. The differences between $R_{\text{G,L,D}}$ and $R_{\text{G,U,D}}$ were generally small, with a root mean-squared deviation (rmsd) of $\sim 0.3$ nm between both datasets. For flexible polymers, a scaling law governs the value of $R_G$ whereby

$$R_G \propto N_{\text{RES}}^{-\nu}.$$  

[3]

Here, $N_{\text{RES}}$ is the number of residues in the chain. The exponent $\nu$ quantifies the correlation length and is governed by the solvent quality. In good, theta (indifferent), and poor solvents the values of $\nu$ for long homopolymers are 0.59, 0.5, and 0.33, respectively (26). Scattering data for a given protein can be analyzed within an intermediate $q$ range to quantify $\nu$ (SI Appendix, Fig. S7A) because

$$I(q) \propto q^{-\nu/2}.$$  

[4]

For reference, the full form factor is shown in SI Appendix, Eqs. S19 and S20. An example of the fitting of SI Appendix, Eq. 4 to the experimental SAXS profile is shown in Fig. S4A for denatured NUS (all proteins can be found in SI Appendix, Fig. S4A). In 6 M urea, we find that $\nu = 0.55 \pm 0.04$ for unlabeled proteins. Within error, this value is similar to the value for labeled samples, $\nu = 0.58 \pm 0.03$ (SI Appendix, Table S6). These findings suggest that the dyes do not fundamentally alter the balance of chain–chain and chain–solvent interactions (SI Appendix, Fig. S7B), thus leaving the solvent quality unchanged. For the analysis that follows, we used an average value of $\nu$ $= 0.57 \pm 0.03$ for proteins in 6 M urea. This value for $\nu$ is in line with the expected value for the SARW model and the analysis of larger datasets from previous measurements (37, 38), which suggest that high concentrations of denaturants are good solvents for generic protein sequences (3, 23). To test whether smFRET measurements yield similar inferences regarding solvent quality, we calculated the values of $R_G = R_{\text{G,L,D}}/R_{\text{G,U,D}}^2$. For chains in a good solvent $G \sim 7$ (26), and obtaining such a value would require accurate estimates of $R_E$ from the smFRET data. In Fig. 2F we plot $(E_{\text{FRET}})$ against $R_{\text{G,L,D}}$, which is extracted from SAXS using exactly the same labeled proteins. The data were analyzed using a Gaussian chain model for the distribution of interdye distances (9–12), with $G$ as the fitting parameter (SI Appendix, Eq. S15 and Note S3). For denatured proteins we obtained $G_0 = 7.1 \pm 0.5$. This value is in line with theoretical expectations for a swollen chain in good solvent (39) and is larger than the value of 6 expected for random coils (40) in theta solvents ($R_{\text{G,L}}$ Values in SI Appendix, Table S7 and G values in SI Appendix, Table S8). Taken together, our analyses of SAXS and smFRET data yield mutually consistent inferences regarding solvent quality for denatured proteins in 6 M urea. Importantly, our data establish that the dyes do not materially impact the analysis of chain dimensions of denatured proteins.

**SAXS and smFRET Yield Discrepant Inferences Regarding IDP Dimensions in Native Conditions.** We applied the analyses described above to the set of seven IDPs under native conditions to calculate $\nu_N$ and $G_N$, Analysis of SAXS profiles for each of the labeled and unlabeled IDPs yielded similar values for $\nu_N$ (SI Appendix, Fig. S4A), suggesting that dyes do not have a major impact on the dimensions of IDPs under native conditions. The mean value of $\nu_N = 0.59 \pm 0.04$ (SI Appendix, Table S6) is in line with values reported for IDPs with similar compositional biases (3, 7, 41). This suggests that for a class of IDP sequences, the effects of chain–chain and chain–solvent interactions are, on average, mutually compensatory, thus unmasking statistics that are similar to those of chains in theta solvents (29, 41)—a result that has previously been described for unfolded protein ensembles under folding conditions (3). For $G$, we obtained a mean value of $G_N = 4.3 \pm 0.4$, and this is different from the value of 6 that is expected for chains in theta solvents (35, 39, 40). To test whether the anomalous value of $G$ reflects differences in the changes of $R_L$ vs. $R_E$, we quantified the swelling ratios that compare the dimensions in 6 M urea vs. native conditions. The swelling ratios are defined as

$$a(R_{\text{G,L}}) = \frac{R_{\text{G,L,D}}^2}{R_{\text{G,L,N}}^2} \text{ and } a(R_{\text{G,U}}) = \frac{R_{\text{G,U,D}}^2}{R_{\text{G,U,N}}^2}.  \text{ [5]}$$

The inferred values of $R_{\text{G,L}}$ of denatured IDPs ($R_{\text{E,L,D}}$) are considerably larger than those of native IDPs ($R_{\text{E,L,N}}$). However, the values of $R_{\text{G,L}}$ for denatured IDPs ($R_{\text{G,L,D}}$) are moderately yet systematically different from those of native IDPs ($R_{\text{G,L,N}}$). This is evidenced by larger values of $a(R_{\text{G,L}})$ vs. smaller values of $a(R_{\text{G,U}})$ (on average $2.02 \pm 0.18$ vs. $1.27 \pm 0.12$, respectively; individual values given in SI Appendix, Table S9). These findings are concordant with previous results, which point to disagreements between inferences from SAXS/small-angle neutron scattering (SAXS/SANS) and smFRET measurements at low denaturant concentrations (15, 21). SAXS measurements of labeled vs. unlabeled molecules rule out the dyes as the source of the discrepancy. One possible reason for the discrepancy is a result of the smFRET measurements, which are presented in Discussion, we are left with three other possible sources for the observed discrepancies: (i) the nature of the averaging that goes into the calculation of $R_E$ is likely to make this quantity relatively insensitive to small changes in solvent quality (20), especially for heteropolymers that transition between coil-like ensembles corresponding to $\nu \sim 0.59$ and $\nu \sim 0.5$ (42); (ii) because $R_E$ quantifies the average distance between a pair of residues, as opposed to an average over all interresidue distances, it is possible that this quantity is more sensitive to fluctuations due to the dangling ends of chains (43) and (iii) it is also possible that the inferred values of $R_E$ are subject to errors due to assumptions of a Gaussian chain model for the distribution of interdye distances. Each of these factors contributes to the discrepancies between inferences drawn from analysis of SAXS vs. smFRET data. We demonstrate this by analyzing conformational distributions extracted from atomistic simulations that accounted for the presence of fluorescent dyes.

**Source of the Discrepant Inferences Regarding the Extent of Collapse Observed Using SAXS vs. smFRET.** We performed all-atom Metropolis Monte Carlo thermal replica exchange simulations for five of the IDPs, using the ABSINTH implicit solvation model and force-field paradigm (24). This combination has proved to be useful for the analysis of conformationally heterogeneous IDPs (42, 44). Details of the simulations are described in SI Appendix, Note S6. For each sequence, we used the measured values of $(E_{\text{FRET}})$ in native conditions to generate reweighted ensembles that match the experimental data. Then, we selected the ensemble corresponding to the lowest simulation temperature (SI Appendix, Table S10) that best matched the experimental observable of interest (more details in SI Appendix, Note S6). To calculate $(E_{\text{FRET}})$, we incorporated atomistic descriptions of rotamers of fluorescent dyes into the reweighted ensembles. For each conformation of a specific sequence, we placed roughly 10$^3$ distinct dye rotamers in different mutual orientations and distances and calculated FRET efficiencies for each conformation. This process was repeated across the entire ensemble to calculate $(E_{\text{FRET}})$ across the ensemble. Conformations were
reweighted based on the agreement between the measured and calculated values of \( \langle F_{\text{FRET}} \rangle \). The reweighting of ensembles based on experimental data was performed using COPPER (45), which is a maximum-entropy reweighting method that attempts to give conformations similar weights while simultaneously attempting to match an experimental observable or a set of experimental observables.

Fig. 3A shows the values of \( R_G \) and \( R_E \) that were extracted from the unbiased ensembles (denoted as \( R_{E,G} \)) and the ensembles reweighted to match \( \langle F_{\text{FRET}} \rangle \) \( \langle R_{E,G,\text{SW}} \rangle \) reweighted ensembles corresponding to native conditions. The subscript \( S \) refers to values obtained from simulations and \( W \) refers to cases where the simulation values were weighted to match an experimental observable. Here, \( R_G \) was calculated as the distance between the \( C_n \) atoms of the first and last residues and \( R_E \) was calculated only over the protein atoms. The reweighting procedure revealed an interesting decoupling between the values of \( R_G \) and \( R_E \). Ensembles that were reweighted to match \( \langle F_{\text{FRET}} \rangle \) showed minimal changes between \( R_{E,G} \) and \( R_{E,G,\text{SW}} \) and large changes between \( R_{E,G} \) and \( R_{E,G,\text{SW}} \) (Fig. 3B). This is consistent with the idea that large changes to \( \langle F_{\text{FRET}} \rangle \) and hence \( R_E \) are compatible with minimal changes to \( R_G \). If true, then the discrepant inferences between SAXS and smFRET measurements must originate in the ability to decouple measures of specific pairwise distances such as \( R_E \) from the averaging over the square of all pairwise distances, which is the case with \( R_{E,G} \).

To put the proposed decoupling between \( R_G \) and \( R_E \) on a quantitative footing, we reweighted the NUS ensembles at 360 K to match the experimentally derived \( R_{E,G} \) and one of the following target values for mean FRET efficiencies: \( \langle F_{\text{FRET}} \rangle = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9] \). Here, \( R_{E,G} \) in the simulations is the weighted mean square of the \( R_G \) values calculated over the protein atoms alone. If \( R_G \) and \( R_E \) can be decoupled, then ensembles should be generated that satisfy a single value of \( R_{E,G} \) and a range of values of \( \langle F_{\text{FRET}} \rangle \). Indeed, we find that with the exception of the most extreme \( \langle F_{\text{FRET}} \rangle \) value (0.1), NUS ensembles can be generated that match \( R_{E,G} \) and a given \( \langle F_{\text{FRET}} \rangle \) value with minimal changes to the force field (SI Appendix, Fig. S8A and B and Note S6). This suggests that, under certain conditions, an entire spectrum of \( \langle F_{\text{FRET}} \rangle \) and therefore multiple \( R_E \) values are consistent with a given \( R_G \) value (22). This result is consistent with the finding that large differences in \( G \) are virtually indistinguishable by SAXS (SI Appendix, Fig. S7C). Such a result emerges from the combination of two effects: (i) at low to intermediate values of \( R_G \), small changes in \( R_E \) (1 mm) can lead to large changes in \( G \) (SI Appendix, Fig. S8A) and (ii) large, potentially informative fluctuations at the ends of chains have little effect on the global conformational properties measured by SAXS (SI Appendix, Fig. S9A and D).

The preceding findings do not imply that the ensembles generated to match different \( \langle F_{\text{FRET}} \rangle \) values have the same conformational properties. To make this point, we characterized the overall shapes of polymers and scaling of internal distances for ensembles of NUS that match the experimentally derived \( R_{E,G} \) and one of the following target values for mean FRET efficiencies: \( \langle F_{\text{FRET}} \rangle = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9] \), and one of the following conformation-specific and ensemble averaged values of susceptibility, \( \delta^* \), that is given in terms of the eigenvalues, \( \lambda_1, \lambda_2 \) and \( \lambda_3 \) of conformation-specific gyration tensors (46, 47).

\[
\delta^* = 1 - \frac{3(\lambda_1^2 + \lambda_2^2 + \lambda_3^2)}{(\lambda_1 + \lambda_2 + \lambda_3)^2}
\]

For rod-like conformations \( \delta^* \approx 1 \) and for a perfect sphere \( \delta^* \approx 0 \) (26, 47). Distributions of \( \delta_{\text{SW}} \) (SI Appendix, Fig. S8D) show that \( \delta_{\text{SW}} \) decreases as \( \langle F_{\text{FRET}} \rangle \) increases, whereas distributions of \( R_{E,G} \) are similar for all \( \langle F_{\text{FRET}} \rangle \) values (SI Appendix, Fig. S8C). The decrease in \( \delta_{\text{SW}} \) observed with decreasing \( R_G \) suggests that ensembles become more spherical to account for the same \( R_G \) albeit with smaller \( R_E \) values. SI Appendix, Fig. S10 shows a comparison of shape characterization in terms of \( G \) and \( \delta^* \). These parameters are weakly coupled although, on average, an increase in \( \langle G \rangle \) implies an increase in \( \delta^* \). The weak coupling results from the fact that \( G \) is highly sensitive to large fluctuations at the ends of chains, whereas \( \delta^* \) only mildly sensitive to such fluctuations and changes in \( \delta^* \) depend on the sequence separation at which the fluctuations emerge (SI Appendix, Fig. S9B–D). To extract further insights regarding the distributions of internal distances, we calculated internal scaling profiles that serve as formal order parameters in more nuanced theories of coil-to-globule transitions (48).

Internal scaling profiles quantify the mean spatial separation between all residues \( i \) and \( j \) that are \([i–i]\) apart along the linear sequence. Fig. 3C shows that all ensembles, irrespective of the target \( \langle F_{\text{FRET}} \rangle \) value used for reweighting, show similar scaling in spatial separation for \([i–i]\) < 40. However, the spatial separations start to diverge from one another at larger sequence separations. These internal scaling profiles highlight an important point: based on Lagrange’s theorem (59) we know that the mean-squared \( R_G \) can be written as the mean-squared sum over all internal distances (definition in SI Appendix, Table S1). Thus, if a majority of internal distances change negligibly, then the value of \( R_G \) will change minimally. In contrast, the overall shape shows intermediate changes and distances corresponding to larger sequence separations will show large fluctuations (SI Appendix, Fig. S9C and D).

Because we measured \( R_G \) and \( \langle F_{\text{FRET}} \rangle \) for each IDP under native and denatured conditions, we can analyze the ensembles that were reweighted to match both experimental observables. Fig. 4 shows the two-dimensional histograms of \( R_{E,G} \) vs. \( \delta_{\text{SW}} \) for ensembles reweighted to match both \( R_{E,G} \) and \( \langle F_{\text{FRET}} \rangle \) for each IDP under native (Fig. 4 F–J) and denatured (Fig. 4 A–E) conditions. For all IDPs, \( \delta_{\text{SW}} \) increases under denaturing conditions, indicating that the ensembles become less spherical. This is consistent with the larger \( G \) values extracted from denatured compared with native conditions. Internal scaling plots of the

---

Fig. 3. Simulated ensembles reweighted to match \( \langle F_{\text{FRET}} \rangle \) suggest decoupling between \( R_G \) and \( R_E \). (A) \( R_G \) and \( R_E \) values extracted from unbiased \( R_{E,G} \) and \( R_{E,G,\text{SW}} \) and reweighted \( R_{E,G,\text{SW}} \) ensembles for N49, NLS, NUS, IBB, and NUL. Here, reweighted ensembles refer to the ensembles generated by reweighting to \( \langle F_{\text{FRET}} \rangle \) values under native conditions. Error bars indicate the SEM over three independent simulations. The experimental \( R_{E,G} \) values determined under native conditions are plotted for reference. \( R_{E,G,\text{SW}} \) is used as a reference given that for the simulated ensembles \( R_G \) is calculated only over the protein. (B) The relative change in \( R_G \) and \( R_E \) between unbiased and reweighted ensembles calculated as: \( \Delta R_{G,E} = R_{E,\text{SW}} - R_{E,G,\text{SW}} \) and \( R_{E,G,\text{SW}} \) respectively. (C) Internal scaling plots for NUS simulated ensembles reweighted to match \( R_{E,G} \) and one of the following \( \langle F_{\text{FRET}} \rangle \) values: [0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]. For every pair of residues at given sequence separation (\( |j-i| \)) the average through-space distance for that given sequence separation (\( \langle \rho(j-i) \rangle \)) is plotted. Here, \( i \) and \( j \) are the residue positions. FRC denotes the internal scaling profile of the Flory random coil (Gaussian chain) reference and EV denotes the internal scaling profile of the excluded volume coil reference.
simulated ensembles (Fig. 4 K–O) show that the denatured ensembles diverge from native ensembles to prefer larger spatial separations for larger sequence separations. The sequence separation at which this divergence occurs is specific to each IDP sequence, thus highlighting the contribution of sequence-specific interactions to chain deformations under denaturing conditions. To visualize the change in shape between native and denatured ensembles, we extracted 100 representative conformations with the highest weights for NUS when reweighted to match the experimental observables under either native (Fig. 5A) or denatured (Fig. 5B) conditions. The results show that NUS adopts more elongated and less spherical conformations under denaturing conditions compared with native conditions.

We also note that simulations can be used to estimate the error associated with inferences of \( R_{\text{E,FRET}} \) from smFRET that are based on the use of the Gaussian chain or other generic polymer models for \( P(R_{\text{E,FRET}}, R_{\text{E,L}}) \) (49). Fig. 2 G and H shows the distance distributions corresponding to the Gaussian chain model together with the distance distributions obtained from the simulations by restraining the ensembles to match \( <E_{\text{FRET}}> \) and \( R_{\text{E,FRET}} \). The results suggest that the Gaussian chain model tends to overestimate \( R_{\text{E,L}} \) for denatured proteins and underestimate \( R_{\text{E,FRET}} \) for IDPs under native conditions (SI Appendix, Table S7). These results are consistent with the findings of O’Brien et al. (49) and Borgia et al. (23). Accordingly, the final \( a(R_{\text{E,L}}) \) values (SI Appendix, Table S9) are overestimated.

### Analysis of the Full SAXS Profiles Beyond \( R_q \) and \( q^* \)

If ensembles of chemically denatured proteins display larger asphericities compared with the native IDPs, then this should be discernible in the SAXS data as well. We tested this by performing a model-independent comparison of the experimental data. Indeed, if one computes a size-independent version of scattering profiles by plotting \( \log(q^*)/|j-i| \) vs. \( q R_q \) (Fig. 6A), then the curves corresponding to bodies with changing asphericity display a rather systematic trend, from the right (aspherical polymers) to the left (spherical polymers) of the plot. We plotted the experimental data for unlabeled native and chemically unfolded proteins (Fig. 6 B–F). For the two smallest proteins N49 and NLS, the differences are within the level of statistical noise, whereas the three larger proteins display a systematic shift of the size-independent scattering patterns from the right (higher asphericity for chemically denatured proteins) to the left (more spherical shapes for IDPs under native conditions). The results of this analysis are important because they were obtained solely from the experimental data.

We further tested the proposed change in asphericity, using size-independent maps of scattering profiles that were generated using the reweighed ABSINTH ensembles. CRYSOL (50) was used to convert each conformation to a SAXS profile and these were combined to generate the final weighted SAXS profile. The profiles generated from the reweighed ABSINTH ensembles consistently show an increase in asphericity for denatured IDPs compared with native IDPs (SI Appendix, Fig. S11D). This recapitulates the direct calculations of

---

**Fig. 4.** Quantification of the shape \( |\delta^*|_{\text{SW}} \), size \( R_{\text{G,SW}} \), and scaling of simulated ensembles reweighted to match both \( <E_{\text{FRET}}> \) and \( R_{\text{E,FRET}} \) for native and denatured conditions. (A–E) Two-dimensional histograms of \( R_{\text{G,SW}} \) vs. \( |\delta^*|_{\text{SW}} \) extracted from simulated ensembles reweighted to match both \( <E_{\text{FRET}}> \) and \( R_{\text{E,FRET}} \) for native conditions. (F–J) Two-dimensional histograms of \( R_{\text{G,SW}} \) vs. \( |\delta^*|_{\text{SW}} \) extracted from simulated ensembles reweighted to match both \( <E_{\text{FRET}}> \) and \( R_{\text{E,FRET}} \) for denatured conditions. (K–O) Internal scaling plots comparing the native (N) and denatured (D) profiles from simulated ensembles reweighted to match both \( <E_{\text{FRET}}> \) and \( R_{\text{E,FRET}} \). FRC and DEN denote the internal scaling profiles generated from the Flory random coil and excluded volume coil references, respectively. Error bars denote the SEM over three independent simulations.
Because the scaling behavior of $R_{G,L}$ depends on the actual number of amino acids in both the polypeptide chain ($N_{RES}$) and $N_{DYES}$, we rewrite Eq. 3 as follows for $R_{G,L}$ (a similar reasoning can be used for $R_{E,L}$):

$$R_{G,L} = \sqrt[3]{G} \rho_G (N_{RES} + N_{DYES})^{0.5}$$

and

$$R_{E,L} = \sqrt[3]{G} \rho_E (N_{RES} + N_{DYES})^{0.5}.$$  

Here, the preexponential factors $\rho_G$ and $\rho_E$ are related to the size of the repeating unit. Whereas dye labeling does not substantially affect $R_G$, as detected by SAXS, we can perform a global fit of the six experimental datasets to extract the contributions that dyes make to $R_{E,L}$ for both denatured proteins and native IDPs. This allowed us to obtain estimates of $N_{DYES} = 5 \pm 3$ (SI Appendix, Fig. S6 and Table S11).

### Discussion

SAXS and smFRET are two powerful experimental tools that provide useful insights regarding disordered systems such as IDPs and unfolded ensembles of autonomously foldable proteins (7, 54). However, the two measurements yield discrepant inferences when going from denatured to native conditions, with SAXS detecting minimal changes and smFRET suggesting discernible reduction in $R_{E,L}$ as measured by an increase in $<E_{FRET}>$. We obtained good agreement between inferred $R_{E,L}$ and $R_{G,L}$ values at high denaturant concentrations in terms of the scaling behavior and inferred solvent quality. However, we find a clear “mismatch” in inferences regarding chain sizes in the absence of denaturant: either the inferred values of $R_{E,L}$ appear to be too small or the measured values of $R_{G,L}$ are too large.

Our insights were derived by combining experimentally derived $R_G$ values and mean FRET efficiencies with simulations that also include the effects of dyes. A major conclusion from the simulations is that many disordered ensembles with substantially different $R_G$ can have similar values of $R_{E,L}$ (Figs. 3–6). This result was also demonstrated by Song et al. (22) for heteropolymers.

**Estimating $N_{RES}$ from SAXS and smFRET Data.** Although our work raises caution regarding the use of generic polymer models when analyzing smFRET data for heteropolymers, these models afford the practical convenience required to obtain quick estimates of $R_{E,L}$ from measured $<E_{FRET}>$ values for IDPs as well as denatured states. It is useful to quantify the contribution that dyes ($N_{DYES}$) make in terms of equivalent residues to the polypeptide chain ($N_{RES}$). Previous estimates of $N_{DYES}$ have varied from 0- to 20-residue equivalents (3, 52, 53). Given direct access to $R_{G,L}, R_{G,U}$, and estimates of $R_{E,L}$ we can quantify $N_{DYES}$ using these data.
systems and it implies that the discrepant expansion factors inferred from SAXS and smFRET measurements are not a consequence of any intrinsic weaknesses of these methods. Instead, they represent a fundamental decoupling between $R_G$, a globally averaged quantity, and $R_E$, as well as other distances between dangling ends that are not averaged across the entire sequence. This decoupling is amplified in finite-sized heteropolymeric sequences in the absence of denaturant.

Advanced theories that account for the effects of chain connectivity to describe excluded volume effects demonstrate that chains can undergo nonuniform expansion/compaction (55). The dangling ends of chains (43) experience fewer restrictions on fluctuations. Hence, inferences regarding chain dimensions can be different when quantified in terms of $R_G$ vs. $R_E$ or distances near the ends of chains. The use of $R_G$ and $R_E$ as equivalent measures of chain dimensions dates back to Flory-style mean-field theories that reduce polymers to collections of uncorrelated monomers or Kuhn segments (5). This is a powerfully simplifying approach that affords convenient analytical descriptions. In contrast, Lifshitz-style theories recognize the decoupling between $R_G$ and $R_E$ and rely on the radial density profile (equivalent to the internal scaling profile) as an order parameter for coil-to-globule transitions (36, 57).

Effects of Dyes in smFRET Measurements. For native and denatured conditions we showed that the behavior of labeled proteins is not different from that of their unlabeled counterparts, at least in terms of the scaling of internal distances manifested by similar values of $\kappa$ for unlabeled and labeled samples (SI Appendix, Table S6). The parallel axes theorem is a useful theoretical construct to describe the relationship between $R_G,U$, $R_G,L$, and $R_G,E$. A full theoretical treatment of this can be found in SI Appendix, Note S7. The main conclusion from this analysis is that for many values of $R_G,U$, dyes do not cause a measurable change of $R_G,U$ relative to $R_G,U$ (SI Appendix, Note S7 and Fig. S12A). However, as $G$ increases, the difference between $R_G,U$ and $R_G,E$ is predicted to increase, with larger changes observed for shorter chain lengths (SI Appendix, Fig. S12B). This prediction is consistent with the experimental trends we observe (SI Appendix, Fig. S12C). Combining the results from SAXS and smFRET with simulations, we estimated the contribution of dyes to $R_G,E$ expressed in terms of extra residues as $N_{DYES} = 5 \pm 3$ (SI Appendix, Fig. S6). Such a value is likely to be generally useful for smFRET analysis, irrespective of the particular fluorescent dye pair used, because the actual size of each fluorophore is limited by diffusion on the inferred distances (SI Appendix, Eq. S29 and Fig. S1C). To further rule out the possibility of artifacts due to the dyes themselves, we discuss potential sources of errors in our experimental design and broader implications.

Case A. Dyes might experience hindered rotations such that the orientation parameter $\kappa^2$, and hence the Förster distance $R_E$, deviates from the isotropic averaging condition (58). We tested this via anisotropy measurements. The low values we observe for anisotropies (0.01, SI Appendix, Table S8) support free dye rotation under all assayed conditions. Therefore, it appears to be reasonable to assume that rotational averaging is allowed, and thus the assumption of $\kappa^2 = 2/3$ in the FRET equation is valid (SI Appendix, Table S4).

Case B. The dyes might be drawn toward one another through cohesive forces. The analysis of scaling exponents should make such an effect easy to detect. We do not observe such a trend under either denaturing conditions or native conditions (SI Appendix, Fig. S6).

Case C. It is known that the dynamics of dyes can affect $E_{FRET}$ measurements (9, 31, 33, 59-63). For unfolded proteins of similar size and in similar solvents to the ones studied here (including NUS), chain reconfiguration times have been shown to be in the range of ~100 ns (3, 64), which is well above the donor lifetimes of ~4 ns and well below the transit times through the confocal volume, ~1 ms. As a result, a major role of dynamics in the measured intensity-based $E_{FRET}$ values seems unlikely. Taken together, we conclude that the dyes alone cannot explain the large changes to $R_G,U$ that we observe upon protein denaturation in contrast to the modest changes of $R_G,U$ (SI Appendix, Table S9).

Choice of Polymer Models for Analyzing smFRET Data. Our findings highlight the need for caution in cooping models for distributions of $R_E$ or $R_G$ that have been designed for infinitely long flexible homopolymers—a point that has been made in previous studies as well (22, 23, 49). Flory’s mean-field theory (5) yields a value of $G = 6$ for $\nu = 0.5$ in theta solvents and SARWs yield $G \sim 7$ for $\nu \sim 0.6$ (SI Appendix, Note S9 and Table S8). The values of $\nu$ (0.57) and the inferred values of $G$ (6.6) for denatured proteins are in accord with the values for SARWs. For the native dataset we obtained $G_N = 5.2$ and $\nu_N = 0.5$, respectively, when we used smFRET, SAXS, and simulations. This result suggests that according to the inferred value of $G$, IDPs under native conditions deviate from the Gaussian chain model, whereas the inferred scaling exponent suggests congruence with the statistics of the Gaussian chain model. In SI Appendix, Note S8 and Fig. S13 we show that the same issue persists when using other polymer models, thus highlighting the role of simulations in inferring self-consistent sets of distances and the need for caution in using generic polymer models for estimating $R_E$ from measured FRET efficiencies, especially in the absence of denaturants. To overcome difficulties associated with the choice of generic polymer models, O’Brien et al. (49) proposed a self-consistency test that requires the measurement of FRET efficiencies by attaching dyes along different internal positions within a sequence. They showed that the use of multiple, independent measurements provides a rigorous test of the polymer model that is used to extract distance estimates from measured FRET efficiencies.

Connections to Recent Studies. The discrepant inferences drawn from SAXS and smFRET measurements have stimulated numerous debates and independent investigations. Discrepancies were recently reported for nonbiological homopolymers like polyethylene glycol (PEG) (21). This study compared $R_G$ values from SANS experiments to $R_G$ values derived from smFRET. Unlike our study, the impact of dyes was not directly investigated as this would have required SANS measurements on the denatured state of the spectrin domain as a function of denaturant concentration. They focused their measurements on the denatured state of the spectrin domain in high concentrations of denaturants are consistent with those of Schuler and coworkers (23, 65, 66).
Working Hypothesis for the Decoupling Between $R_G$ and $R_e$. Flexible polymers can be described using the thermal blob model. $R_G$ and $R_e$ for a thermal blob will scale as $g^2$, where $g$ is the number of residues per blob (67). By definition, the blob is a length scale where the intrablob interactions and blob-solvent interactions are counterbalanced. The blob size is approximately five to seven residues for most IDPs (41). In mean-field theories for polymers in dilute solutions, there are two interrelated parameters to consider: the surface tension per blob ($\gamma_B$) and the effective pairwise interactions between blobs (67). Depending on solvent quality, $\gamma_B$ will be positive (poor solvent), zero (theta solvent), or negative (good solvent) and the pairwise interblob interactions will respectively be, negative, zero, or positive. All blobs are identical in homopolymers, and hence all interactions are uniform and a single parameter suffices to describe the overall chain statistics. Accordingly, in theta and good solvents, $R_G$ and $R_e$ will provide equivalent descriptions of chain behavior.

For heteropolymers, blobs can be quite different from one another, and this depends on the amino acid composition and sequence patterning (68, 69). The chain could have blobs that encode negative, zero, or positive values of $\gamma_B$ and these will in turn modulate the pattern of interblob interactions. Attractions can screen repulsions and this can give rise to relatively uniform density profiles that make $R_G$ inert to changes in solution conditions but they will be manifest as differences in distances across specific length scales (Fig. 4). The effects of heteropolymericity can be captured as an interaction matrix as opposed to a single interaction parameter, and the key question is whether the variance across the values within the interaction matrix is smaller than, equivalent to, or larger than thermal energy. This variance will encode the extent of convergence or divergence between measures of chain dimensions averaged across the entire sequence ($R_e$) and measures that probe specific length scales, such as $R_G$. The blob-based analysis explains why despite water being a poor solvent for polypeptide backbones (29, 70), we now know that the apparent solvent quality for real IDPs deviates from that for backbones and is actually governed by charge and proline contents as well as the patterning of charged and proline residues (3, 17, 41, 42, 68, 69, 71).

Conclusion and Perspective

Given the high cost required to perform complete SAXS experiments with dye-labeled samples and the small contribution of the commonly used dyes to the total protein size, it is both impractical and unnecessary to measure SAXS profiles for labeled molecules on a routine basis. We have shown that, for many IDPs, $R_{G,U}$ will be a reasonable approximation to $R_{G,L}$. Given the diversity of IDP sequences (68), it should be stressed that our measured values of $G_N$ and $\delta_v^*$ are unlikely to be universal. Therefore, $R_G$ and $R_e$ should be determined for each combination of solution condition and IDP through independent quantification of $R_{G,U}$ by smFRET and $R_{G,L}$ by SAXS or the measurement of multiple internal distances for different sequence separations by smFRET (3, 34) or through the joint use of intramolecular three-color FRET measurements (58). For SAXS measurements, this includes estimates of $R_G$ (7) combined with analysis of protein shape preferences from the entire SAXS profile. These measurements can be augmented using methods such as anomalous SAXS (59) that introduce gold labels along the chain for extracting intramolecular distances. Measurements when complemented with computer simulations as performed here and in other efforts (66) can help in converting experimental observables into self-consistent molecular models of the conformational ensembles. The relevance of our work goes beyond IDPs under native conditions. In the protein-folding field there is lingering controversy over the earliest folding events arising from dissimilar FRET and SAXS experiments (15, 34); suggestions have been put forward for chain collapse preceding the folding transition—a view largely supported by FRET measurements—whereas the alternative position is that collapse is intimately coupled with the folding transition—a view supported by SAXS measurements. Based on our data, we propose that the earliest events are likely to be changes in shape (26, 46, 72) within the unfolded ensembles upon dilution from denaturant before folding and the formation of stable local as well as nonlocal contacts; decreased asphericity may be what smFRET measurements pick up as a “collapse” transition. This would be difficult to detect by SAXS using only $R_G$, but the full SAXS profile might be more useful for detecting changes in asphericity and directly estimating the correlation length via the scaling exponent $\nu$.

Therefore, we propose that the joint use of smFRET and SAXS, together with other structural biology methods, and the support of computational tools and advanced theories will improve our understanding of heterogeneous conformational ensembles.

Materials and Methods

In total, 10 proteins (abbreviated as N49, BBL, NLS, CSP, NUS, IBB, TRX, NUL, N98, and NSP) bearing a cysteine residue at the second position and the noncanonical amino acid p-acetylphenylalanine at the penultimate position were expressed recombinantly in Escherichia coli BL21 AI cells, purified, and double labeled with Alexa488 hydroxylamine and Alexa594 maleimide. Proteins were measured in two PBS buffer conditions: “denaturing” (in presence of 6 M urea) and “native” (with urea absent). SmFRET was done on a custom-built multiparameter spectrometer, using picomolar concentrations of labeled proteins. FRET efficiencies were analyzed burst-wise. SAXS profiles of labeled and unlabeled proteins at different concentrations (micromolar and beyond) were measured at the BioSAXS P12 beamline of Petra III (DESY). The scattering profiles were analyzed in full to obtain size (mean radius of gyration and its distribution) and shape (asphericity, correlation length) information. Molecular simulations of labeled proteins were performed using the COMPASS package with the ABINIT solved solvation model and force-field paradigm. Experimental observables were used to extract the conformational space sampled by the simulated ensembles. Comprehensive descriptions of the protein expression, purification, labeling, smFRET and SAXS measurements, atomistic simulations, and theoretical considerations are described in detail in SI Appendix, Notes S1–S9, Tables S1–S11, and Fig. S1–S13.

ACKNOWLEDGMENTS. We thank Ben Schul, Robert Best, Andrea Soranno, and Hue-Sun Chan for insightful discussions. G.F. was supported by the EMBL Interdisciplinary Postdocs (EIPOD) Programme and a postdoctoral fellowship “Vall-i-d” from the Conselleria d’Educació, Formació i Occupació de la Generalitat Valenciana. M.K. acknowledges support from European Comission (the 7th Framework Programme) Marie Curie Grant IDPyNMR (Contract 264257). E.A.L. acknowledges funding by the Deutsche Forschungsgemeinschaft, especially the Emmy Noether program. R.V.P. acknowledges support from the US National Institutes of Health through Grants R01NS056114 and R01NS089932. Access to the P12 synchrotron beamline at Petra-3 has received funding from the European Community 7th Framework Programme (FP7-2007-2013) under BioStruct-X (Grant 283570).


