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

Gustavo Fuertes[1,3,#], Niccolò Banterle[1,#], Kiersten M. Ruff[4,#], Aritra Chowdhury[1], Davide Mercadante[7,8], Christine Koehler[1], Michael Kachala[3], Gemma Estrada Girona[1], Sigrid Milles[1], Ankur Mishra[9], Patrick R. Onck[9], Frauke Gräter[7,8], Santi Esteban-Martín[5,6], Rohit V. Pappu*[,4], Dmitri I. Svergun*[,3], Edward A. Lemke*[,1,2]

[1] European Molecular Biology Laboratory (EMBL), Structural and Computational Biology Unit, Meyerhofstrasse 1, 69117 Heidelberg, Germany.
[2] European Molecular Biology Laboratory (EMBL), Cell Biology and Biophysics Unit, Meyerhofstrasse 1, 69117 Heidelberg, Germany
[3] European Molecular Biology Laboratory (EMBL), Hamburg Outstation c/o DESY, Notkestrasse 85, 22607 Hamburg, Germany.
[4] Department of Biomedical Engineering and Center for Biological Systems Engineering, Washington University in St. Louis, One Brookings Drive, Campus Box 1097, St. Louis, Missouri 63130, United States
[5] Barcelona Supercomputing Center (BSC), Jordi Girona 29, 08034 Barcelona, Spain
[6] IDP Discovery Pharma SL, Barcelona Science Park, Baldiri i Reixac, 4, 08028 Barcelona, Spain.
[7] Heidelberg Institut für Theoretische Studien (HITS), Schloß-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany.
[8] Interdisciplinary Center for Scientific Computing (IWR), Im Neuenheimer Feld 368, 69120 Heidelberg, Germany.
[9] University of Groningen, Zernike Institute for Advanced Materials, Micromechanics section, Nijenborgh 4, 9747AG Groningen, The Netherlands

Equally shared first authors

*Correspondence to lemke@embl.de, svergun@embl-hamburg.de, pappu@wustl.edu

Supplementary table of contents:

Note S1. Sample preparation and characterization.

All constructs contained a TGT codon (coding for a cysteine) at the position of the second residue and a TAG amber codon (to encode a p-acetylphenylalanine or AcF, via amber supression) at the penultimate position (see protein sequences in **Table S2**). BBL was cloned as a N-terminal fusion protein of Intein-GFP-12His. N49, NLS, CSP, IBB, TRX, N98 and NSP were cloned as N-terminal fusion proteins of Intein-CBD-12His. NUS and NUL were sandwiched between an N-terminal hexahistidine-tag followed by a TEV cleavage site and a C-terminal intein-CBD tag.. All protein-encoding plasmids, under the control of the pBAD promotor, were co-transformed with a pEvol plasmid, containing an evolved aminoacyl-tRNA synthetase specific for p-acetylphenylalanine and its cognate tRNA (1), into *Escherichia coli* BL21 AI cells. Cells were grown in TB containing 50 µg/mL of ampicillin and 33 µg/mL of chloramphenicol at 37º C until an optical density at 600 nm between 0.2 was reached. At that moment, 1 mM AcF was added to the medium. When OD at 600 reached between 0.6-0.8, arabinose was added to a final concentration of 0.02% to induce protein expression. After 4 h at 37°C, cells were harvested by centrifugation. Cell pellets stored at -80 °C were resuspended in 4X PBS, pH=8 supplemented with urea 2M, TCEP 0.2M and PMSF 1mM (buffer A). Cells were disrupted by sonication.. The lysate was further clarified by centrifugation at 39000 g to remove any insoluble material. The supernatant was purified by immobilized metal affinity chromatography using Ni-NTA beads (2mL of slurry per L of culture). After an incubation of 2 h at 4°C, the beads were washed in buffer A containing 40 mM imidazole and proteins were eluted in buffer A containing 500 mM imidazole. Proteins were cleaved with 100 mM β-mercaptoethanol overnight to activate the intein moiety and split the target protein from the fusion partner (NUS and NUL were simultaneously cleaved with 100 mM of β-mercaptoethanol and TEV protease). Usually the proteins precipitated upon cleavage and were dissolved in urea 8 M and dialyzed against buffer A using membranes of 3 kDa cut-off. Proteins were incubated again with the Ni beads to remove uncleaved and fusion proteins. The flow-through was concentrated using centrifugal concentrating cells with a 3 kDa cut-off filter. Final purification was achieved by size-exclusion chromatography using a Superdex75 column using PBS 1X pH=7.4, Urea 2M and TCEP 0.2mM as eluent (buffer B). Fractions were analyzed by SDS-PAGE using 4-12% gradient gels with MES as running buffer and stained with Coomassie Blue. Pure fractions were pooled and exchanged to PBS pH=7.4, Urea 8 M, KCl 0.3 M, DTT 10 mM (buffer C), using 3 kDa cut-off filters (unlabeled samples). For labeling, samples were exchanged to an acetic acid/acetate buffer adjusted at pH=4.0 containing guanidinium chloride 4 M and NaCl 100 mM. Labeling with Alexa488 hydroxylamine was done using a 4x molar excess of dye over protein for 2 days at 65ºC (or 37ºC for 3 days in the case of IBB and TRX). After the reaction was completed, the proteins were exchanged into PBS buffer adjusted at pH=7.0 containing guanidinium chloride 4M and reduced with 10mM DTT. DTT was removed via 5 repeated buffer exchange steps using centrifugal concentrating cells. The freshly reduced protein was labelled with Alexa594 maleimide using a 2x molar excess of dye over protein for 2 h at room temperature. The reaction was quenched with 100 mM DTT and free unreacted dyes were finally removed by buffer exchange and subsequent gel filtration on a Superdex75 column using buffer B as a mobile phase. Labeled proteins were finally concentrated to >10 mg/mL and dialyzed against buffer C. A fraction of the proteins was labeled with either Alexa488 or Alexa594 for obtaining several spectroscopic parameters associated with the FRET measurements.. Protein concentration was determined by the BCA assay and confirmed by UV-Vis Absorbance and refractive

index analysis. Dye concentration was determined by UV-Vis absorbance using extinction coefficients of 71000 and 73000 $M^{-1}cm^{-1}$ for Alexa488 and Alexa594, respectively.

Note S2. Buffer composition.

The choice of the experimental conditions requires paying special attention to small details including: i) pH, ii) denaturant, iii) ionic strength and iv) reducing agent. This point is particularly important since in previous SAXS and smFRET studies a wide range of conditions have been employed (2) thus rendering a direct comparison of the results difficult given that protein size is highly sensitive to the physical and chemical properties of the medium in which they are dissolved (3, 4). To our knowledge, no systematic dye-labeled protein study has been performed by both, smFRET and SAXS, under the same experimental conditions. The pH of our solutions was carefully kept at 7.4 with a phosphate-saline buffer (5)(exceptionally, we used HEPES buffer for the NLS protein). Guanidinium chloride and urea are the two most popular chemical denaturants. Guanidinium chloride, but not urea, has an electrostatic effect and it has been shown that urea supplemented with high salt concentrations has actually a similar effect compared to guanidinium chloride (6). Unfortunately, guanidinium chloride possesses a high absorption in SAXS experiments and we noticed that the signal-to-noise ratio for our set of small proteins was low, making the acquisition of high quality SAXS profiles challenging. With all these considerations in mind, our “denaturing” buffer contains phosphate-saline buffer (PBS) adjusted at pH 7.4 containing 6 M urea, potassium chloride 0.3 M and DTT 10 mM and measurements were done at a temperature of 23°C. On the other hand, the “native” buffer contained only PBS pH=7.4 and 10 mM DTT. As mentioned, for NLS we substituted the phosphate-saline buffer by a combination of HEPES and sodium chloride: HEPES 25 mM pH=7.4, NaCl 150 mM, Urea 6 M, KCl 0.3 M, DTT 10 mM (unfolding buffer) and HEPES 25 mM pH=7.4, NaCl 150 mM, DTT 10 mM (native buffer). Additionally, for smFRET 0.002% v/v Tween-20 was included in the buffers to minimize protein adsorption on the glass walls.

Proteins were kept in PBS pH = 7.4 supplemented with urea 8 M and DTT 10 mM at -20 ºC. Immediately prior to the smFRET measurements, proteins were diluted into either denaturing buffer or native buffer. Immediately before SAXS measurements, samples were either adjusted to the composition of the denaturing buffer or dialyzed versus native buffer. Samples were centrifuged at 14000 g for 15 min at 4 ºC and the concentration in the supernatant was determined. Therefore, trace amounts of urea (in the order of nM for smFRET and aM for SAXS) may still remain in the buffers.

Note S3. smFRET methods.

Single-molecule fluorescence experiments were performed on a custom-built multiparameter spectrometer centered around a high-numerical-aperture water-immersion objective (60x, 1.27 NA) on a z-translator. Linearly polarized outputs ,at a frequency of 26.7 MHz, from a picosecond laser diode (LDH 485; Picoquant, Berlin, Germany), filtered through an excitation filter (482/18), and a white light laser (SuperkExtreme, NKT Photonics), filtered through a SuperK Varia tunable filter (NKT photonics) and subsequently an excitation filter of 572/15, were used to excite freely diffusing labeled proteins. The fluorescence emission from the donor and acceptor dyes was spatially filtered with a 0.1 mm pinhole, then spectrally separated into "green" (donor) and "orange" (acceptor) (emission filters 525/50, 620/60, multi-band fluorescence bandpass filter 488/568/660) fluorescence components which were again split into two components based on polarization using polarizing beam splitter cubes and finally detected with single photon counting detectors (τ-SPAD and PMA Hybrid detectors from Picoquant were used to detect the orange and green signal respectively). The laser pulses were alternated in order to probe the presence of the acceptor (7-9). Specifically, the "orange" laser was delayed with respect to the "green" laser by 25 ns. Laser synchronization was achieved using a computer-controlled multichannel picosecond diode laser driver (PDL 828 "Sepia II", Picoquant). Photon signals were acquired using a multichannel time-correlated single-photon counting module (Hydraharp400, Picoquant).

Acquired data were subjected to multiparameter fluorescence analysis (10-12). Single molecules were identified via a burst search algorithm on the lee-filtered photon stream (13, 14) and fluorescence intensities (I), lifetimes (τ) and anisotropies (r) were extracted from individual bursts. Data were analyzed with a custom-written program using Igor Pro (Wavemetrics, Lake Oswego, OR). The interphoton lag time threshold for burst selection was set to 90 microseconds and identified bursts were further subjected to a photon based selection criteria (70 photons for measurements in native conditions and 110 for that in denaturing conditions, to avoid spurious signal and to take into account the higher background in the latter case).

FRET efficiencies in a burst are related to the photon count by (15):

$$E_{FRET} = \frac{{I_A}^D}{\gamma {I_D}^D + {I_A}^D} \qquad \textit{Equation S1}$$

Similarly, stoichiometry of individual bursts were determined by:

$$S = \frac{\gamma {I_D}^D + {I_A}^D}{\gamma {I_D}^D + {I_A}^D + {I_A}^A} \qquad \textit{Equation S2}$$

where $I_x{}^y$ describes the corrected intensity detected from the donor (D) or acceptor (A) dye via donor or acceptor lasers, i.e., the green and orange lasers respectively (Note by green and orange lasers we refer to lasers exciting the green or orange dye and not the actual colour of the laser output, which are blue and yellow respectively). Raw intensities are corrected for background, leakage of donor signal into the acceptor channel and direct acceptor excitation from the green laser.. The factor γ corrects for the detection efficiency of acceptor and donor channels. Specifically, γ depends on the instrumental setup ($\gamma_{INSTRUMENT}$) and fluorophores (γ_{DYES}) as follows:

$$\gamma = \gamma_{INSTRUMENT} \cdot \gamma_{DYES} \quad \textit{Equation S3}$$

$$\gamma_{INSTRUMENT} = \eta_A / \eta_D \quad \textit{Equation S4}$$

$$\gamma_{DYES} = \phi_A / \phi_D \quad \textit{Equation S5}$$

Where η_A and η_D are the photon detection efficiencies of the acceptor channel and the donor channel, respectively and ϕ_A and ϕ_D are the quantum yields of the acceptor and donor dyes, respectively. The γ parameter can also be estimated from leakage and direct excitation corrected FRET efficiencies and stoichiometries, E_{app} and S_{app} (16) where:

$$E_{app} = \frac{{I_A}^D}{{I_D}^D + {I_A}^D} \quad \textit{Equation S6}$$

$$S_{app} = \frac{{I_D}^D + {I_A}^D}{{I_D}^D + {I_A}^D + {I_A}^A} \quad \textit{Equation S7}$$

One limitation of this approach is that it is only valid when the quantum yields of the donor and acceptor dyes do not vary much. To test if such a condition was met, we measured the ensemble lifetimes of all the singly labelled samples in a home built lifetime spectrometer consisting of a 485 nm picosecond pulsed laser QuixX 488 (Omicron-Laserage, Germany), an emission monochromator, PMA Hybrid detector (PicoQuant,Germany) and a Hydraharp module (PicoQuant,Germany) for photon counting. All ensemble lifetimes were measured with magic angle polarization condition, where the emission polarizer was set at 54.7° relative to the excitation polarization. Lifetimes were obtained by fitting the TCSPC decay traces to a convolution integral of a monoexponential function and the IRF (instrument response function), the latter being experimentally measured.

Fluorescence lifetime of a fluorophore is related to its quantum yield by:

$$\tau = \Gamma Q \quad \textit{Equation S8}$$

where τ is the fluorescence lifetime, Q the quantum yield and Γ the natural lifetime. If the natural lifetime is the same (which can be assumed for the same fluorophore even when in different environments), the ratio of their lifetimes represent the ratio of their quantum yields as shown below:

$$\frac{\tau_a}{\tau_b} = Q_a / Q_b \quad \textit{Equation S9}$$

where *a* and *b* represent two different fluorophore environments. Based on this fact, the quantum yield of the singly labelled samples were estimated via a ratiometric comparison of sample lifetime versus free Alexa 488 and 594 in PBS; the quantum yields of the latter two were taken as 0.92 and 0.66 respectively (this value is from the manufacturers and is consistent within error to the quantum yield values for the free dyes we had obtained). We found little variation among lifetimes and thus quantum yields between different samples in a given condition, with the exception of N49 (**Table S4A-B**, **FigS2 A-D**). However, there was a systematic decrease in lifetimes and thus quantum yields when conditions were changed from native to denaturing. Having confirmed minimal variation of quantum yields we felt

comfortable to extract gamma from E_{app} and S_{app}. A linear fit to a plot of $1/S_{app}$ vs E_{app} yields intercept a and slope b which relates to gamma in the following way (16):

$$\gamma = (a-1)/(a+b-1) \qquad \textit{Equation S10}$$

We estimated the gamma parameter separately for native and denatured datasets to be 0.77 and 0.87 respectively (**Figure S2 E-F**) but we excluded N49 from the analysis as it was an outlier in terms of quantum yields. As N49 behaves only marginally different than the other proteins in the set, the obtained average values were then applied to the whole data set, including N49.

The FRET data was further corrected for γ along with leakage and direct excitation and this dataset was used to extract FRET efficiencies (**Figure S3**). A fixed window of 0.4 stoichiometry units wide and covering the entire E_{FRET} range was used to first select the population of molecules showing FRET from E_{FRET} vs S histograms. The selected population was fitted with a single 2D Gaussian function and based on the parameters of this fit, the selected population was refitted with a 2D Gaussian function constraining the fit to $\pm$ 2σ (standard deviation) from the mean on the E_{FRET} axis. The means obtained from such fit were taken as the final $\langle E_{FRET} \rangle$ value and used in subsequent analyses (**Table S4**).

The transfer efficiency depends on the donor-to-acceptor distance r_{DA} with an inverse 6th power law dependence (17):

$$E(r_{DA}) = \frac{R_0^6}{R_0^6 + r_{DA}^6} \qquad \textit{Equation S11}$$

Where R_0 is the Förster radius, i.e. the distance at which the FRET efficiency is 50%. R_0 was calculated using the method described in (18) , according to:

$$R_0 = \sqrt[6]{\frac{9(ln10)\kappa^2\, J(\lambda)\, \Phi_D}{128\, \pi^5\, n^4\, N_A}} \qquad (R_0 \text{ in Å}) \qquad \textit{Equation S12}$$

Where the orientation factor κ^2 is assumed to be 2/3, $J(\lambda)$ is the spectral overlap integral between the donor emission and the acceptor excitation, Φ_D is the quantum yield of the donor (in the absence of the acceptor) and n is the refractive index of the medium measured to be 1.338 and 1.385 in native and denaturing buffers respectively. κ^2 can be assumed to be 2/3 when sufficient rotational averaging of the dyes exist; which is supposed to be the case as all our dyes have a C_5 flexible linker between the conjugating group and the chromophore. A direct evidence for sufficient rotational averaging comes from the very low anisotropy values for both the donor and acceptor dyes in all the conditions measured (**Table S5**), suggesting sufficient rotational averaging to allow approximation of $\kappa^2 = 2/3$.

Effect of multiple conformations on average FRET efficiencies.

Opposed to structured proteins where the distance between the two attachment points of the dyes is nearly fixed, disordered systems feature a distribution of donor-acceptor distances (r_{DA}). When the fluctuations of interdye distances occur on a timescale slower than the lifetimes, the observed FRET efficiency results from the averaging over multiple conformations weighted for their probability. The average measured FRET efficiency can then be calculated as:

$$\langle E_{FRET} \rangle = \int_0^\infty E(r_{DA}) P(r_{DA}) dr_{DA} \qquad \textit{Equation S13}$$

Unfortunately, a typical smFRET experiment does not contain enough information to retrieve a model-free distance distribution. The smFRET spectroscopist must therefore choose one (19) and the Gaussian chain model is arguably the most frequently cited model and will be discussed below. Other models can be found in **Note S8.**

Gaussian chain model.

In the Gaussian chain model the main underlying assumption is that the monomers occupy zero volume so that no part of the chain excludes another. Despite its simplicity it is commonly used for the analysis of IDPs and denatured proteins (20-23) and has even yielded satisfying results in systems were substantial amounts of residual structure were present (24). The distance distribution function between the donor and acceptor dyes, $P(r_{DA})$, takes the form:

$$P(r_{DA}) = 4\pi r_{DA}^2 \left(\frac{3}{2\pi \langle r_{DA}^2 \rangle}\right)^{3/2} exp\left(-\frac{3r_{DA}^2}{2\langle r_{DA}^2 \rangle}\right) \qquad \textit{Equation S14}$$

Where the root mean squared donor-to-acceptor distance $\sqrt{\langle r_{DA}^2 \rangle}$ we call simply $R_{E,L}$ (see exact definitions in **Table S1**). Considering $G=R_{E,L}^2/R_{G,L}^2$, substituting $\langle r_{DA}^2 \rangle = G \cdot \mathrm{R}_{\mathrm{G,L}}^2$ into **EQ. S14** yields:

$$P(r_{DA}) = 4\pi r_{DA}^2 \left(\frac{3}{2\pi\, G\, R_{G,L}^2}\right)^{3/2} exp\left(-\frac{3r_{DA}^2}{2\, G\, R_{G,L}^2}\right) \qquad \textit{Equation S15}$$

This equation together with **EQ. S11** and **EQ. S13** were used to fit E_{FRET} as a function of $R_{G,L}$ in order to get G (**Figure 2K**).

Note S4. SAXS methods.

Synchrotron radiation X-ray scattering data were collected on the EMBL P12 beamline at the PETRA III storage ring (DESY, Hamburg). Measurements were carried out at 23°C with 1-10 mg/mL solutions of labeled or unlabeled samples. The data were recorded using a 2M PILATUS detector (DECTRIS, Switzerland) at a sample-detector distance of 3.0 m and a wavelength of 0.1 nm, covering approximately the range of momentum transfer $0.05 < q < 4.50$ nm^{-1} ($q = 4\pi \sin\theta/\lambda$, where 2θ is the scattering angle). 20 frames of 50 ms each were collected and averaged. In general, no measurable radiation damage was detected. Data treatment was carried out using the ATSAS package (25). If curves obtained at different concentrations showed inter-particle interactions (see below) they were merged and extrapolated to infinite dilution using ALMERGE (26). SAXS profiles were featureless (**Figure S4A**), similar to those obtained for other IDPs and chemically denatured proteins, which has been rationalized as a consequence of the averaging over many conformations existing in the ensemble.

Size descriptors (R_G).

According to the Guinier law the R_G can be calculated as (27):

$$I(q) = I(0)exp\left(\frac{q^2R_G^2}{3}\right) \qquad \textit{Equation S16}$$

Where $I(0)$ is the forward scattering intensity (an equivalent relation is expressed by **EQ. 1** in the main text). Results from the Guinier analysis are shown in **Figure S4B** and **Table S3.** However, the Guinier plot is only linear over a restricted region of the scattering spectrum and for IDPs such regions may be even smaller. In order to facilitate the comparison among the different proteins we used a normalized structure factor:

$$P(q) = I(q) \,/\, I(O) \qquad \textit{Equation S17}$$

Alternatively, the protein dimensions can be inferred from the distance distribution function or P(r), which is a histogram of all interatomic distances within the protein weighted by the respective electron densities (28). P(r) was calculated as the Fourier transform of the scattering intensity using the program GNOM (29). R_G is then given by:

$$R_G = \sqrt{\frac{\int_0^{Dmax} r^2P(r)dr}{2\int_0^{Dmax} P(r)dr}} \qquad \textit{Equation S18}$$

Where D_{max} corresponds to the maximum diameter of the particle. P(r) distributions are shown in **Figure S4D**. R_G calculated from P(r) are also shown in **Table S3.** For further analysis we used the R_G extracted from the Guinier approximation.

Most proteins measured in native buffer (**Figure S5D** for unlabeled proteins and **S5F** for labeled proteins) show an increase in R_G as the concentration gets higher while for proteins unfolded in urea the R_G is approximately constant (**Figure S5C** for unlabeled proteins and **S5E** for labeled proteins) although negative inter-particle interactions can also be observed (e.g. for the protein NUS, blue points in **Figure S5C** and **S5E**). Therefore, for IDPs in native buffer we used the value of R_G extrapolated to

infinite dilution while for proteins denatured in the presence of urea, we took the R_G from the sample with the highest concentration in the dilution series (**Table S3**).

The "unfolded-ness" or random coil likeness can be qualitatively assessed by means of Kratky plots: globular macromolecules have well-shaped curves while denatured polypeptides lack such a peak and have a plateau or are slightly increasing at high angles (30). The dimensionless of the normalized Kratky plots in **Figure S4C** suggests that the proteins are unfolded under all conditions assayed.

The ensemble optimization method (EOM) was used to describe the conformational space of flexible proteins (25, 31). In this method, a large pool of explicit structures with atomic coordinates is generated and then a genetic algorithm is employed to select a subset a conformers that best fit the experimental SAXS profile. The initial pool was generated as described in **Note S5** (un-reweighted pool). The SAXS profiles of the sub-ensembles selected by EOM and the corresponding distribution of R_G are shown in **Figure S11A** and **S11B**, respectively**.** The corresponding average R_G values are shown in **Table S3.** FRET efficiencies (**Figure S11C**) were computed by means of **EQ. S11** by using the distances between the carbon alpha atom of the first and the last residues in the sequences and the R_0 shown in **Table S4**. Asphericity plots (**Figure S11D**) were generated as described in the next sub-section.

We note that we measured the experimental data on a high brilliance SAXS beamline dedicated to BioSAXS yielding very small minimum momentum transfer to compute reliable Guinier fits. For the angular range we started from $q \approx 0.05$ [1/nm]. This substantially reduced the noise in estimates of R_G values from Guinier analysis. We found that the R_G values assessed using indirect Fourier transformation analyses of the $P(r)$ function and the average values derived from the EOM ensembles were consistent with values obtained using the Guinier approximation (**Table S3).** Therefore, our results, both based on the R_G values and on the entire SAXS profiles clearly showed that the SAXS-determined overall sizes of urea-denatured IDPs systematically increased compared to the native IDPs. These results are in line with well documented difficulties in detecting changes of R_G in unfolding experiments (references 15, 18, 19 of the main text).

Shape descriptors (υ, δ^*).

One important parameter is the internal scaling exponent υ, which reflects the distribution on interatomic distances on different length scales (32). υ varies between 0 (sphere) and 1 (infinitely thin rod). In the polymer field, υ takes the values of 1/3, 1/2 and 3/5 for the limiting cases of the self-attracting chain, the theta chain and the self-avoiding chain, respectively (**Figure S7A**). The scattering profiles were fitted to a "generalized" Gaussian chain model (up to q=2.5 nm^{-1}) as implemented in the *SASfit* package (33) in order to get υ:

$$P_{GC}(q) = I_0 \frac{U^{\frac{1}{2\nu}}\Gamma\left(\frac{1}{2\nu}\right) - \Gamma\left(\frac{1}{\nu}\right) - U^{\frac{1}{2\nu}}\Gamma\left(\frac{1}{2\nu}U\right) + \Gamma\left(\frac{1}{\nu}U\right)}{\nu\, U^{\frac{1}{\nu}}} \qquad \textit{Equation S19}$$

Where the modified variable *U* equals:

$$U = \frac{(2\nu+1)(2\nu+2)}{6} q^2 R_G^2 \qquad \textit{Equation S20}$$

And $\Gamma(a,x)$ is the unnormalized incomplete gamma function and $\Gamma(a)$ is the gamma function. See https://www.ncnr.nist.gov/staff/hammouda/the_SANS_toolbox.pdf and https://kur.web.psi.ch/sans1/sasfit/sasfit.pdf for more information. Notice that the famous Debye approximation is recovered for υ=0.5 (34). Fits to the experimental SAXS profiles are shown in **Figure S4A** and the corresponding fitted values of υ can be found in **Table S6**.

Another important parameter is the asphericity δ^* that quantifies the deviation from spherical shape: δ^*=0 for a sphere and 1 for rod-like conformations. Theoretical scattering profiles corresponding to ellipsoidal bodies of semi-axes λ_1, λ_2 and λ_3 were simulated with the BODIES program (35). Asphericities were then calculated using **EQ. 6** from the main text and plotted in **Figure 7A**.

Note S5: Bridging the concentration gap between smFRET and SAXS.

In both smFRET and SAXS, proteins are freely diffusing in solution. The only difference between the two samples is the concentration range. smFRET necessitates the presence of at the most one labeled particle in the confocal observation volume at any given time, which translates into particle concentrations in the pM range (**Figure S5A** and **S5B**). In SAXS the scattering profile is build up from the contributions of many individual molecules requiring particle concentrations at least in the µM range (**Figure S5C** to **S5F**). Moreover, SAXS profiles are usually measured at different concentrations in order to exclude inter-particle interactions. While our data obtained with the unfolded proteins show essentially no concentration dependence, 5 out of 7 IDPs show a clear increase of R_G with concentration. While this result is in itself interesting as it points out to specific protein-protein interactions (e.g. aggregation) in IDPs which might be functionally relevant (36), at the same time it does not allow a direct comparison with smFRET data. In order to minimize the differences in the protein concentrations between the two techniques we used the following two approaches. On one hand, for the two IDPs that showed the highest effect with concentration (NLS and NUS) we measured smFRET of the double-labeled protein (pM) using an excess (µM) of their unlabeled counterparts. Such measurements are difficult since the photon background arising from fluorescent contaminations in the unlabeled protein sample may interfere with the actual photons coming from the labeled species. However, the calculated FRET efficiencies are similar (within 7 %) to the values obtained in the absence of unlabeled protein (**Figure S5B**). On the other hand, we add an excess of polyethylene glycol of 10 kDa (PEG, i.e. with a similar mass as the proteins used here), at the same concentrations used for the SAXS experiment, in order to test for molecular crowding effects (4). Again, no significant changes in the mean FRET efficiency values were observed. Despite the limitations of both approaches, our results suggest that no significant changes in the FRET efficiency occur in the concentration range between pM and µM. Therefore for the analysis of IDPs the SAXS profile obtained using either the extrapolated profile or the SAXS profile obtained at the lowest protein concentration can be safely compared to the smFRET dataset.

Note S6. Atomistic simulations

Atomistic simulations of N49, NLS, NUS, IBB, and NUL were conducted using the CAMPARI simulation package (http://campari.sourceforge.net) utilizing the ABSINTH implicit solvation and force-field paradigm (37). For each construct, three independent simulations were performed in spherical droplets with radii of 150 Å using parameters from the abs3.2_opls.prm parameter set. Neutralizing and excess Na^+ and Cl^- atoms were modeled explicitly with an excess NaCl concentration of 5 mM. Replica exchange was used to enhance sampling using the following temperature schedule: T = [280K, 300K, 320K, 340K, 360K, 380K, 400K]. Each simulation consisted of 6.15×10^7 steps of which the first 1×10^7 were taken as equilibration. Each step consisted of either a Metropolis Monte Carlo move or a temperature replica swap. The move set utilized included translational, pivot, concerted rotation, sidechain rotation, and proline puckering moves, the details of which have been published previously (38, 39). Temperature replica swaps were attempted every 5×10^4 steps. Trajectory frames were collected every 5×10^3 steps over the last 5.15×10^7 steps. This generated 1.03×10^4 frames that were later subjected to the addition of dyes (see below for details). Sequences were capped on the N-terminus by an acetyl unit and on the C-terminus by *N*-methylamide. Additionally, for each construct, p-acetylphenylalanine was replaced with a cysteine. This change was made since parameters were not available for p-acetylphenylalanine and so that both positions at which dyes were added post-facto have limited influence on the native ensembles.

Addition of implicit dyes to simulated ensembles

In order to implicitly add dyes to the simulated ensembles we utilized our in-house program COCOFRET. COCOFRET takes in a trajectory, dye rotamer libraries, residue positions at which to add the dyes, and R_0 and outputs the mean FRET efficiency for each frame that could incorporate the addition of dyes. Specifically, for each frame, 100 independent attempts were performed to attach Alexa594 and Alexa488 dyes to the cysteine residues at position 2 and the penultimate position, respectively, via a C5-linker and maleimide chemistry. Attachment of dyes was achieved by randomly selecting rotamers from the HandyFRET rotamer libraries and ensuring the carbon-sulfur-carbon angle was approximately ideal (40). Attaching the Alexa488 dye to a cysteine via a C5-linker and maleimide chemistry rather than to a p-acetylphenylalanine via a C5-linker and hydroxylamine chemistry as done in the experiments should have limited effects on the results extracted from the simulated ensembles (41). A given protein+dye conformation is accepted if no steric clashes are observed between the dye and the protein. Steric clashes are defined by any protein atom being within the solvation shell of any dye atom. The solvation shell for each dye atom was set to the default value of 5 Å except for the maleimide atoms which were set to 2 Å. This exception accounts for the connectivity of the protein and dyes. Then the set of accepted protein+Alexa488 and protein+Alexa594 conformations were combined. Protein+Alexa488+Alexa594 conformations were kept if no dye solvation shells were found to be overlapping and the FRET efficiencies of these conformations were calculated using the Förster formula. Then, the mean FRET efficiency was calculated, as well as the standard error. If the standard error was below 0.005, then the mean FRET efficiency for that frame was kept. If not, then the process was repeated until the standard error threshold was met or 10 iterations were performed. In the case where convergence was not reached, a mean FRET efficiency value was not recorded for that particular frame.

Reweighting of simulated ensembles to match experimental results

COPER, a maximum entropy method, was used to reweight simulated ensembles to match experimental results (42). COPER takes in mean experimental values (FRET efficiency and/or R_G^2) and the errors associated with those values and outputs the weights of each frame that yield a unique global solution which satisfies those constraints. The error for the FRET efficiencies was set to 0.02 (43). COPER was ran for each of the following temperatures: T = [300K, 320K, 340K, 360K, 380K]. In order to decide which temperature to analyze, the decrease in maximum entropy (ΔS) was calculated using:

$$\Delta S = S(\boldsymbol{p}^{post}) - S(\boldsymbol{p}^{prior}) = -\sum_{i=1}^{n_c} p_i^{post} \ln p_i^{post} - \ln(n_c) \qquad \textit{Equation S21}$$

Here, $\mathbf{p}^{post}$ is the vector of weights determined by COPER to match the mean experimental values, $\mathbf{p}^{prior}$ is the vector of equal weights determined by the number of conformations considered, n_c. The lowest temperature that yielded a mean ΔS > -1 was chosen for analyses of the conformational properties of each IDP (**Table S10**). Here, ΔS is averaged over three independent simulations. A value of ΔS > -1 corresponds to a mean free energy change of less than 1kT to the simulation potential function. Here, k is the Boltzmann constant and T is the temperature. Thus, a reweighted ensemble with ΔS > -1 implies that limited changes must be made to the potential function in order to match the experimental values.

Note S7: The effect of the dyes on R_G: parallel axes theorem.

Here we will show that the parallel axess theorem can be used to quantitatively describe the relationship between $R_{G,U}$, $R_{G,L}$ and $R_{E,L}$. This theorem gives the expression of the radius of gyration of a set of 2 particles (e.g. a protein with 1 label) with respective radii of gyration R_1 and R_2, when their centers of mass are separated by a distance d_{12} (as stated on page 276 of the book "Biomedical Applications of Synchrotron Radiation. E. Burattini and A. Balerma (Eds.), IOS Press (1996)")

$$R_T^2 = \frac{M_1}{M_1+M_2} R_1^2 + \frac{M_2}{M_1+M_2} R_2^2 + \frac{M_1 M_2}{(M_1+M_2)^2} d_{12}^2 \qquad \textit{Equation S22}$$

Multiplying both sides of the equation by $(\mathrm{M}_1 + \mathrm{M}_2)$ and simplifying gives:

$$(M_1 + M_2)\, R_T^2 = M_1\, R_1^2 + M_2\, R_2^2 + \frac{M_1 M_2}{(M_1+M_2)} d_{12}^2 \qquad \textit{Equation S23}$$

By similar reasoning a set of 3 particles (e.g. a protein with 2 labels) with R_1 being the radius of gyration of the protein, R_{23} the radius of gyration of the two dyes and d_{23} the distance between the centers of mass of the two labels, becomes equivalent to

$$(M_1 + M_2 + M_3)\, R_T^2 = M_1\, R_1^2 + (M_2 + M_3)\, R_{23}^2 \qquad \textit{Equation S24}$$

Where

$$(M_2 + M_3)\, R_{23}^2 = M_2\, R_2^2 + M_3\, R_2^2 + \frac{M_2 M_3}{(M_2+M_3)} d_{23}^2 \qquad \textit{Equation S25}$$

If the masses and the R_G of the two labels are identical ($\mathrm{M}_2 = \mathrm{M}_3$) then:

$$2M_2\, R_{23}^2 = M_2\, R_2^2 + M_2\, R_2^2 + \frac{M_2 M_2}{(M_2+M_2)} d_{23}^2 = 2M_2 R_2^2 + \frac{M_2}{2} d_{23}^2 = 2M_2 \left(R_2^2 + \frac{d_{23}^2}{4}\right) \qquad \textit{Equation S26}$$

Or equivalently:

$$R_{23}^2 = R_2^2 + \frac{d_{23}^2}{4} \qquad \textit{Equation S27}$$

Rewriting **EQ. S24** for our case gives:

$$m_L R_{G,L}^2 \;=\; m_U R_{G,U}^2 + m_{DYES} R_{G,DYES}^2 \qquad \textit{Equation S28}$$

Where m_x is the molecular weight of the species in the subscript x (unlabeled protein, labeled protein or dyes) and $R_{G,x}$ is the R_G of the species in the subscript x. The R_G of the two dye-linker moieties, $R_{G,DYES}$, is then given by (analogously to above **EQ. S27**):

$$R_{G,DYES}^2 = R_{G,DYE}^2 + \frac{R_{E,L}^2}{4} \qquad \textit{Equation S29}$$

Here, $R_{G,DYE}$ is the R_G of a single dye-linker moiety.

In order to test the validity of **EQ. S29**, Excluded Volume (EV) simulations, to which dyes were added explicitly post-facto, for 5 IDPs were performed using CAMPARI. Specifically, every 10th frame was extracted from a trajectory of 10,300 frames. Then, dyes were explicitly added to a given frame following the procedure outlined in **Note S6**. In this case, only one iteration was attempted per frame. Additionally, once a dye pair was successfully added that protein+dyes conformation was saved and the program moved on to the next frame. In this way, we created a trajectory of at most one protein+dyes conformation per frame. $R_{G,DYE}$ was determined by taking the root mean square of the R_G's extracted over the Alexa488 dye and averaging over the different IDP protein+dyes trajectories ($R_{G,DYE}$=0.66 nm). $R_{E,L}$ was taken to be the distance between the C19 atoms of Alexa488 and Alexa594, as defined by the HandyFRET AF488.pdb and AF594.pdb files (http://karri.anu.edu.au/handy/rl.html), respectively. A plot of **EQ. S29** together with the values of $R^2_{G,DYES}$ as a function of $R^2_{E,L}$ calculated from the CAMPARI EV protein+dyes trajectories for each IDP can be found in **Figure S1C**. Such results show good agreement between the theoretical and the computational models. Thus, **EQ. 3** together with **EQ. S28** and **EQ. S29** can be used to predict ΔR_G, defined as $R_{G,L} - R_{G,U}$, as a function of $R_{E,L}$ (**Figure S12A)**. Similarly, **Figure S12B** shows ΔR_G as a function of G. These plots illustrate the fact that for some values of $R_{E,L}$ dyes do not cause a measurable change in R_G, the minimum being observed near *G*~4.

Note S8. Other commonly used distance distributions.

Random points in a sphere model

One can relate the distribution of donor-to-acceptor distances to a distribution of R_G:

$$P(r_{DA}) = \int P(r_{DA}|R_G)P(R_G)dR_G \qquad \textit{Equation S30}$$

For which the conditional probability density function suggested by Ziv and Haran can be used (44):

$$P(r_{DA}|L) = \frac{1}{\delta\, R_{G,L}}\left[3\left(\frac{r_{DA}}{\delta\, R_{G,L}}\right)^2 - \frac{9}{4}\left(\frac{r_{DA}}{\delta\, R_{G,L}}\right)^3 + \frac{3}{16}\left(\frac{r_{DA}}{\delta\, R_{G,L}}\right)^5\right] \quad \textit{for } 0 \le r_{DA} < 2\delta R_{G,L} \qquad \textit{Equation S31}$$

Such function describes the distance distribution of two random points inside a sphere with radius $\delta\, R_G$ (where δ is a constant introduced to correct the statistics for the case of an ideal chain) and has been shown to be a reasonable approximation for unfolded polymers (45). Notice that δ and *G* are coupled to each other. For instance, *G*=6 when δ=$\sqrt{5}$, which is the value that has been assumed in previous uses of this model.

Self-avoiding random-walk (SARW) model.

The SARW model has been rarely employed in the protein field (19, 21). However, as the name suggests, such model has the particular advantage that excluded-volume effects are explicitly taken into account. In the SARW model the donor-acceptor distance distribution is described by:

$$P(r_{DA}) = \frac{a}{R_{E,L}}\left(\frac{r_{DA}}{R_{E,L}}\right)^{2+\theta} exp\left(-b\left(-\frac{r_{DA}}{R_{E,L}}\right)^{\omega}\right) \qquad \textit{Equation S32}$$

Where a and b are normalization constants whose values satisfy the relationship $\int P(r_{DA})dr_{DA} = \int r^2 P(r_{DA})dr_{DA} = 1$. θ and ω have values of 0.272 and 2.427 respectively (19, 46)

Model comparison.

In order to facilitate the comparison among the different models we plotted in **Figure S13** the distribution of distances as a function of the "normalized" distance ($r_{DA}/R_{E,L}$). It can be readily seen that all three theoretical models (Gaussian chain, Haran and SARW) display similar profiles. Such curves are similar to the distance distribution obtained with the CAMPARI simulations of NUS after reweighting to match the experimental $\langle E_{FRET}\rangle$ and $R_{G,U}^2$ obtained under denaturing conditions suggesting that simple polymer models can recapitulate the distances found in proteins unfolded in urea. However, the distribution obtained by reweighting against the values obtained for NUS under native conditions is clearly narrower and more skewed. Therefore, the limited applicability of the Gaussian chain model to describe the dimensions of native IDPs, due to the excessive variance of the distribution, also holds for the other two polymer models. This in turn highlights the crucial role played by simulations in defining a distance distribution which is compatible with both smFRET and SAXS observables.

Note S9. Bridging FRET and SAXS results with polymer theory

One of the main complications in comparing the results obtained from smFRET and SAXS is that the measured quantities are inherently different: smFRET measures the (average) distance between the two labeled sites (R_E) while SAXS provides a measure of the average of all inter-atomic distances within the molecule, i.e. the radius of gyration of the protein (R_G), and not just the ends. In order to compare the results from the two techniques, it is then necessary to establish a relation between these two quantities which can be represented by the ratio $G = \frac{R_E^2}{R_G^2}$ between the two distances. It is important to notice that such a relationship is far from being universal and it depends on, for example, both the conformation and the conformational distribution of the sample. In the following subsection we will summarize some key results in this direction developed in the context of polymer theory.

The simplest example which can be considered as an approximation of a disordered protein is the freely joined chain model. In such scenario the root mean squared distance between two monomers of a chain with a number of residues N_{RES}, bond length b and persistence length lp is (45):

$$R_{E,U} = \sqrt{2\ lp\ b}(N_{RES} - 1)^{1/2} \qquad \textit{Equation S33}$$

For proteins, *b*=0.38 nm i.e. the average distance between two consecutive C_α atoms (47) and *lp,* a measure of the stiffness of a polymers*,* is ~0.4 nm (45, 47). **EQ. S33** can be generalized to (45):

$$R_{E,U} = \sqrt{2\ lp\ b}(N_{RES} - 1)^{\upsilon} \qquad \textit{for } \upsilon \in \{0,1\} \qquad \textit{Equation S34}$$

Where υ is the Flory exponent (scaling exponent or correlation length exponent) equivalent to an excluded-volume parameter (48). The expression for the radius of gyration is then given by (45, 48, 49):

$$R_{G,U} = \sqrt{\frac{2\ lp\ b}{(2\upsilon+1)(2\upsilon+2)}}(N_{RES})^{\upsilon} \qquad \textit{Equation S35}$$

It follows from **EQ. S34** and **EQ. S35** that, in the long-chain limit ($N_{RES} \to \infty$), the square ratio between R_E and R_G is:

$$\frac{R_{E,U}^2}{R_{G,U}^2} = G = (2\upsilon + 1)(2\upsilon + 2) \qquad \textit{Equation S36}$$

G values for specific cases like the self-avoiding random walk, the ideal Gaussian chain and the self-attracting walk can be found in **Table S8**. Notably, for a random flight (a chain without excluded volume i.e. υ=0.5), **EQ. S36** simplifies to G=6, which is the well-known ratio that has been used once and again in the literature (20-22, 50) in order to convert between the two parameters for unfolded polymers in both the low and high-denaturant regime.

Since, according to experimental evidence, dyes do not alter the scaling behavior of the studied proteins (**Table S6**) a consequence that follows is that the ratio *G* is insensitive to the presence of the dyes:

$$\frac{R_{E,U}^2}{R_{G,U}^2} = \frac{R_{E,L}^2}{R_{G,L}^2} = \frac{R_E^2}{R_G^2} = G \qquad \textit{Equation S37}$$

Table S1. Parameter definition.

Definitions of key parameters used in this study. See sketch in Figure 1.

Abbreviation	Mathematical Definition ([1])	Explanation ([2])	Commonly used terms	Experimental technique
$R_{E,U}$	$\sqrt{\langle r_{fl}^2 \rangle}$	Root mean squared distance between the C_α atom of the first residue (*f*) and the C_α atom of the last residue (*l*) of the protein	End-to-end distance Residue-to-residue distance	None
$R_{E,L}$ ([4])	$\sqrt{\langle r_{DA}^2 \rangle}$	Root mean squared distance between the transition dipole moment of the donor fluorophore (*D*) and the transition dipole moment of the acceptor fluorophore (*A*). Fluorophores are attached to the first and last residues of the protein	Dye-to-dye distance Donor-to-acceptor distance, interdye distance	smFRET
$R_{G,U}$ ([4])	$\sqrt{\frac{1}{2n^2}\sum_{ij}^{n}\langle r_{ij}^2 \rangle}$	Root of half the mean squared distance between all pairs of atoms (*i,j*) of an unlabeled protein	Radius of gyration of the unlabeled protein	SAXS ([3])
$R_{G,L}$ ([4])	$\sqrt{\frac{1}{2N^2}\sum_{kl}^{N}\langle r_{kl}^2 \rangle}$	Root of half the mean squared distance between all pairs of atoms (*k,l*) of a labeled protein.	Radius of gyration of the labeled protein	SAXS ([3])

([1]) n is the number of atoms of an unlabeled protein; N is the number of atoms in a labeled protein. Alternative and equivalent definitions exist. In particular the radius of gyration can be defined as $R_G = \sqrt{\frac{\sum_i \|r_i\|^2 m_i}{\sum_i m_i}}$ where m_i is the mass of atom *i* and r_i is the position of atom *i* with respect to the center of mass of the molecule. This is for instance the definition implemented in the GROMACS tool gmx_gyrate. Another equivalent definition is given in **EQ. S18**.

([2]) For simplicity, in order to calculate $R_{E,L}$ from the CAMPARI ensembles of labeled proteins, the distance between the C19 atoms of Alexa488 and Alexa594, as defined by the HandyFRET AF488.pdb and AF594.pdb files (http://karri.anu.edu.au/handy/rl.html), respectively was used.

([3]) For a uniform polymer (a polymer sample composed of a single macromolecular species) the SAXS profiles reflect the root-mean squared radius of gyration: $R_G = \sqrt{\langle r_G^2 \rangle}$ where r_G is the radius of gyration of each individual molecule.

([4]) The relation between $R_{E,L}, R_{G,U}$ and $R_{G,L}$ is given by **EQ. S28** and **EQ. S29** (see parallel axes theorem in **Note S7**).

Table S2. Primary structure.

Sequence of the 10 proteins used in this study. *=AcF. Proteins were labeled with Alexa594 maleimide at position 2 and with Alexa488 hydroxylamine at the penultimate position. The number of residues probed by smFRET and SAXS are named N_{RES}^{FRET} and N_{RES}^{SAXS}. The plot shown in **Figure 1** was generated by calculating physico-chemical properties from the amino acid composition as follows. Mean charges were calculated by summing up all charged residues and dividing by the total number of residues. Mean hydrophobicities were calculated using the scale of Kyte and Doolittle as described in (51). Structures of the 3 folded proteins were taken from the PDB. Single structures of the 7 IDPs were taken from the ensembles simulated with CAMPARI.

Code	Protein sequence	N_{RES}^{FRET}	N_{RES}^{SAXS}
N49	GCQTSRGLFGNNNTNNINNSSSGMNNASAGLFGSKP*A	36	38
BBL	ACSPAIRRLLAEHNLDASAIKGTGVGGRLTREDVEKHLA*A	39	41
NLS	ACETNKRKREQISTDNEAKMQIQEEKSPKKKRKKRSSKANKPPE*A	44	46
CSP	ACGKVKFFDSKKGYGFITKDEGGDVFVHFSAIEMEGFKTLKEGQVVEFEIQEGKKGGQ*A	58	60
NUS	GCPSASPAFGANQTPTFGQSQGASQPNPPGFGSISSSTALFPTGSQPAPPTFGTVSSSSQPPVFGQQPSQSAFGSGTTPN*A	80	82
IBB	GCTNENANTPAARLHRFKNKGKDSTEMRRRRIEVNVELRKAKKDDQMLKRRNVSSFPDDATSPLQENRNNQGTVNWSVDDIV KGINSSNVENQLQAT*A	97	99
TRX	GCDKIIHLTDDSFDTDVLKADGAILVDFWAEWSGPSKMIAPILDEIADEYQGKLTVAKLNIDQNPGTAPKYGIRGIPTLLLF KNGEVAATKVGALSKGQLKEFLDAN*A	107	109
NUL	GCGFKGFDTSSSSSNSAASSSFKFGVSSSSSGPSQTLTSTGNFKFGDQGGFKIGVSSDSGSINPMSEGFKFSKPIGDFKFGV SSESKPEEVKKDSKNDNFKFGLSSGLSNPV*A	112	114
N98	GCFNKSFGTPFGGGTGGFGTTSTFGQNTGFGTTSGGAFGTSAFGSSNNTGGLFGNSQTKPGGLFGTSSFSQPATSTSTGFGF GTSTGTANTLFGTASTGTSLFSSQNNAFAQNKPTGFGNFGTSTSSGGLFGTTNTTSNPFGSTSGSLFGP*A	151	153
NSP	GCNFNTPQQNKTPFSFGTANNNSNTTNQNSSTGAGAFGTGQSTFGFNNSAPNNTNNANSSITPAFGSNNTGNTAFGNSNPTS NVFGSNNSTTNTFGSNSAGTSLFGSSSAQQTKSNGTAGGNTFGSSSLFNNSTNSNTTKPAFGGLNFGGGNNTTPSSTGNANT SNNLFGATANAN*A	176	178

Code	Protein	Fragment	Organism	UniprotKB	Coordinates
N49	Nucleoporin Nup49	121-154	S. cerevisiae	Q02199	CAMPARI
BBL	Dihydrolipoamide succinyltransferase	115-151	E. coli	P0AFG6	2WXC (52)
NLS	Heh2	99-140	S. cerevisiae	Q03281	CAMPARI
CSP	Cold-shock protein	3-58	T. maritima	O54310	3A0J
NUS	Nucleoporin 153 kDa	1313-1390	H. sapiens	P49790	PED2AAE (53)
IBB	Karyopherin alpha-2	3-97	H. sapiens	P52292	CAMPARI
TRX	Thioredoxin	2-106	E. coli	P0AA25	1XOB (54)
NUL	Nucleoporin 153 kDa	884-993	H. sapiens	P49790	CAMPARI
N98	Nup98	2-150	H. sapiens	P52948	CAMPARI
NSP	Nsp1	2-175	S. cerevisiae	P14907	CAMPARI

Table S3. R_G.

Table S3A. R_G of denatured proteins derived from SAXS. R_G were calculated from the SAXS profiles using three approaches: Guinier, P(r) and EOM (see **Note S4**). The q R_G range used to calculate the R_G from the Guinier fits is indicated. Guinier and P(r) plots are shown in **Figure S4B** and **S4D** respectively. Mean values ± standard deviation.

DENATURED	$R_{G,U}$ (nm)				$R_{G,L}$ (nm)		
Approach	GUINIER	q R_G range	P(r)	EOM	GUINIER	q R_G range	P(r)
N49	1.69±0.12	0.17-1.30	1.69±0.01	1.80	2.09±0.16	0.21-1.30	2.07±0.04
BBL	2.1±0.4	0.32-1.30	2.11±0.05		2.3±0.4	0.19-1.29	2.22±0.02
NLS	2.33±0.18	0.14-1.30	2.28±0.03	2.48	2.4±0.2	0.23-1.29	2.58±0.07
CSP	2.49±0.04	0.33-1.30	2.67±0.03		2.20±0.06	0.37-1.30	2.24±0.01
NUS	3.13±0.10	0.31-1.29	3.31±0.04	3.09	2.92±0.03	0.25-1.30	3.05±0.02
IBB	3.12±0.07	0.30-1.30	3.22±0.02	3.23	3.16±0.03	0.26-1.30	3.31±0.03
TRX	3.63±0.07	0.41-1.30	3.57±0.03		3.20±0.08	0.28-1.30	3.32±0.05
NUL	3.5±0.3	0.20-1.28	3.28±0.02	3.58	3.24±0.05	0.35-1.30	3.22±0.01

Table S3B. R_G of native IDPs derived from SAXS. R_G were calculated from the SAXS profiles using two approaches: Guinier and P(r) (see **Note S4**). Mean values ± standard deviation.

NATIVE	$R_{G,U}$ (nm)				$R_{G,L}$ (nm)		
Approach	GUINIER	q R_G range	P(r)	EOM	GUINIER	q R_G range	P(r)
N49	1.59±0.13	0.35-1.30	1.67±0.04	1.70	1.9±0.4	0.21-1.30	1.87±0.03
NLS	2.4±0.3	0.29-1.30	2.66±0.10	2.52	2.0±0.2	0.18-1.30	2.17±0.03
NUS	2.49±0.13	0.17-1.29	2.68±0.03	2.64	2.53±0.14	0.21-1.30	2.61±0.08
IBB	3.2±0.2	0.22-1.29	3.29±0.04	3.06	2.9±0.9	0.59-1.29	3.2±0.3
NUL	3.0±0.3	0.15-1.28	3.18±0.19	3.11	3.0±0.3	0.35-1.29	3.09±0.07
N98(*)	2.86±013	0.51-1.30	3.07±0.10		2.86±013	0.51-1.30	3.07±0.10
NSP(*)	4.1±0.3	0.24-1.30	3.98±0.06		4.1±0.3	0.24-1.30	3.98±0.06

(*) Only the unlabeled versions were actually measured. Because the dyes represent less than 10% of the mass of the protein. It was assumed that $R_{G,L}$~$R_{G,U}$

Table S4. smFRET parameters.

R_0 is the Förster distance, Φ_D is the quantum yield of the donor (in the absence of the acceptor), Φ_A is the quantum yield of the acceptor, $\langle E_{FRET} \rangle$ is the mean FRET efficiency, κ^2 is the orientation factor, n is the refractive index of the medium and γ is the detection efficiency correction factor. We note that in smFRET measurements, various error sources contribute to error, among the largest are introduced by R_0 and detection efficiency correction parameters. We work with an average values across the different biological specimens for denatured and native respectively and list all standard deviations in the table below. These standard deviations thus reflect what is likely due to biological or sample heterogeneity for the different proteins, which is the dominating imprecision contributing factor and certainly point to the fact that a precision of higher than 0.3 nm is certainly not reached in our data set. See **Note S3**.

Table S4A. Parameters used individually for IDPs under native conditions.

	NATIVE						
Protein	R_0 (nm)	Φ_D	τ_D(ns)	Φ_A	τ_A(ns)	$\langle E_{FRET} \rangle$	J ($M^{-1}cm^{-1}nm^4$)
ALEXA (free)	5.57	0.92	4.06	0.66	3.86		1.73E+15
N49	5.40	0.69	3.07	0.67	3.89	0.87	1.94E+15
NLS	5.56	0.89	3.91	0.70	4.09	0.79	1.80E+15
NUS	5.67	0.95	4.21	0.70	4.11	0.53	1.89E+15
IBB	5.56	0.88	3.89	0.71	4.18	0.5	1.82E+15
NUL	5.65	0.93	4.11	0.72	4.24	0.48	1.90E+15
N98	ND	0.95	4.22	0.72	4.26	0.64	
NSP	ND	0.96	4.21	0.73	4.22	0.45	
AVERAGE ± SD	5.61 ±0.05	0.93 ± 0.03	4.09 ± 0.14	0.72 ± 0.01	4.18 ± 0.06		(1.85± 0.045)E+15

Table S4B. Parameters used individually for denatured proteins.

	DENATURED						
Protein	R_0 (nm)	Φ_D	τ_D(ns)	Φ_A	τ_A(ns)	$\langle E_{FRET} \rangle$	J ($M^{-1}cm^{-1}nm^4$)
ALEXA (free)		0.84	3.71	0.66	3.84		
N49	5.25	0.74	3.28	0.65	3.78	0.66	1.76E+15
BBL	5.29	0.88	3.88	0.67	3.89	0.61	1.56E+15
NLS	5.35	0.85	3.76	0.66	3.87	0.48	1.72E+15
CSP	5.39	0.89	3.91	0.66	3.86	0.46	1.74E+15
NUS	5.40	0.88	3.89	0.66	3.88	0.29	1.76E+15
IBB	5.30	0.86	3.78	0.66	3.87	0.22	1.62E+15
TRX	5.35	0.87	3.85	0.65	3.82	0.27	1.68E+15
NUL	5.45	0.87	3.86	0.67	3.92	0.24	1.88E+15
N98		0.88	3.88	0.67	3.89	0.24	
NSP		0.88	3.88	0.67	3.89	0.20	
AVERAGE ± SD	5.37 ± 0.05	0.87 ± 0.01	3.85 ± 0.05	0.66 ± 0.005	3.88 ± 0.03		(1.72± 0.09)E+15

Table S4C. Parameters used globally for either native IDPs or denatured proteins:

Parameter	NATIVE	DENATURED
κ^2	2/3	2/3
n	1.338	1.385
γ	0.77	0.87
$Leakage$	0.125	0.09
$Direct\ Excitation$	0.145	0.124

Table S5. Donor and acceptor anisotropy.

Anisotropies were measured in an ensemble TCSPC spectrometer; by measuring emission count rates with an emission polarizer at 0° (I_{para}) and 90° (I_{perp}) relative to the excitation polarizer. Anisotropy (r) was calculated as:

$r = [I_{para} - GF\ I_{perp}]/ [I_{para} + 2GF\ I_{perp}]$

The G factor (GF) of the setup was calculated by the long-time tail matching method where the GF was adjusted till the anisotropy decay tail of a fast rotor became centered around 0 (In our case we used free Alexa 488 and Alexa 594 dyes in buffer for GF calibration).

Protein	Anisotropy (r)			
	Donor		Acceptor	
	DENATURED	NATIVE	DENATURED	NATIVE
N49	0.03	0.04	0.07	0.05
BBL	0.02	-	0.07	-
NLS	0.08	0.11	0.08	0.08
CSP	0.05	-	-	-
NUS	0.07	0.04	0.08	0.07
IBB	0.05	0.09	0.09	0.11
TRX	0.06	-	0.09	-
NUL	0.04	0.04	0.06	0.11
N98	0.05	0.06	0.09	0.11
NSP	0.04	0.04	0.07	0.08
AVERAGE ± SD	0.05±0.02	0.06±0.02	0.08±0.01	0.09±0.02

Table S6. Scaling exponent (υ).

$\boldsymbol{\upsilon}$ of the individual proteins calculated by fitting the SAXS profiles to EQ. S19 and S20 (see Note S4).

	DENATURED		NATIVE	
	UNLABELED	LABELED	UNLABELED	LABELED
N49	0.48	0.55	0.48	0.45
BBL	0.36	0.56	-	-
NLS	0.51	0.60	0.54	0.46
CSP	0.52	0.55	-	-
NUS	0.61	0.62	0.51	0.48
IBB	0.62	0.58	0.49	0.51
TRX	0.58	0.60	-	-
NUL	0.56	0.56	0.47	0.51
N98	-	-	0.50	-
NSP	-	-	0.61	-
AVERAGE ± SD	0.55±0.04	0.58±0.03	0.51±0.05	0.48±0.03
AVERAGE ± SD	0.57±0.03		0.50±0.04	

Table S7. $\boldsymbol{R_{E,L}}$.

$R_{E,L}$ were calculated from smFRET experiments (**Table S3**) for both denatured proteins (n=8) and native IDPs (n=7) using the Gaussian chain model (**Note S3**), or the simulations (n=5, **Note S5**). Note that the results from the Gaussian chain employ smFRET observables only while the simulations are consistent with both smFRET and SAXS data.

	$R_{E,L}$ (nm)			
Model	Gaussian chain model		Simulations	
	DENATURED	NATIVE	DENATURED	NATIVE
N49	4.9	3.6	4.8	3.9
BBL	5.3	-	-	-
NLS	6.3	4.3	6.0	4.6
CSP	6.5	-	-	-
NUS	8.4	6.2	8.2	6.1
IBB	9.6	6.4	8.4	7.0
TRX	8.7	-	-	-
NUL	9.2	6.6	8.6	6.9
N98	-	5.3	-	-
NSP	-	6.9	-	-

Table S8. G

Ratio between the squared values of R_G and R_E expected from theory (**Note S9**) and simulations (**Note S6**). Coefficient values ± standard deviation

	MODEL	G
THEORY	Rod	12
	Self-avoiding random walk (υ=0.6, MF)	7.04
	Ideal Gaussian chain (υ=0.5, MF)	6
	Self-attracting walk (υ=0.33, MF)	4.44
	Infinitely "monodisperse" polymer solution (γ=∞, RG)	2
	Sphere	1.31 [1]
EXPERIMENTS	$\langle E_{FRET} \rangle$- and $R_{G,L}$- (denaturing conditions)	7.1±0.5 [2]
	$\langle E_{FRET} \rangle$- and $R_{G,L}$- (native conditions)	4.3±0.4 [2]
SIMULATIONS	CAMPARI $R_{G,U}^2$- and $\langle E_{FRET} \rangle$- restrained (denaturing conditions)	6.6±0.2 [3]
	CAMPARI $R_{G,U}^2$- and $\langle E_{FRET} \rangle$- restrained (native conditions)	5.2±0.5 [3]
	KBFF version 2 (native conditions)	5.2±0.2 [4]
DATABASES	*NUS (*PED2AAE) (53) *(native conditions)*	6.49±0.06 [5]

*RG=renormalization group theory (55, 56), MF=mean-field theory (48)

[1]Calculated by combining the mean distance of amino acids within a protein (57) and the relation between the radius of a sphere and the radius of gyration.

[2] Calculated by fitting $\langle E_{FRET} \rangle$ globally as a function of $R_{G,L}$ under denaturing and native conditions to a distribution of distances according to a Gaussian chain model (**Note S3**).

[3]Averaged over the 5 studied proteins (N49, NLS, NUS, IBB and NUL).

[4]Averaged over the last 90 ns of the simulation trajectory (58). The Kirkwood-Buff force-field version 2 was used.

[5]Averaged over 5 ensembles of 200 conformers each. This ensemble was refined against NMR observables only.

Table S9. Swelling factors (α).

The swelling factor in R_G was calculated as $\alpha(R_{G,L}) = R^2_{G,L,D}/R^2_{G,L,N}$ where *N* represents native IDPs (**Table S3B**) and *D* represents denatured IDPs (**Table S3A**). Only $R_{G,L}$ values obtained via Guinier analysis were considered**.** The swelling factor in $R_{E,L}$ was calculated as $\alpha(R_{E,L}) = R^2_{E,L,D}/R^2_{E,L,N}$, where *N* represents native IDPs and *D* represents denatured IDPs (both taken from **Table S7**). $R_{E,L}$ values were obtained both via the Gaussian chain model (smFRET) and via molecular simulations restrained by R_G (SAXS) and E_{FRET} (smFRET).

	$\boldsymbol{\alpha(R_{G,L})}$	$\boldsymbol{\alpha(R_{E,L})}$	
SAMPLE	SAXS	smFRET	SAXS/smFRET/simulations
N49	1.21	1.87	1.51
NLS	1.44	2.18	1.70
NUS	1.33	1.86	1.81
IBB	1.19	2.24	1.44
NUL	1.17	1.95	1.55
AVERAGE ± SD	1.27±0.12	2.02±0.18	1.60±0.15

Table S10. Simulation details

Lowest simulation temperature that satisfied the criterion $\Delta S > -1$ for each IDP and fitting criteria. The ensembles extracted from these temperatures were used for the plots shown in the main text (**Figures 4-6**) and the supporting information (**Figures S9, S11, S13**).

IDP	Denatured Fit: R_G^2, $\langle E_{FRET} \rangle$	Native Fit: $\langle E_{FRET} \rangle$	Native Fit: R_G^2, $\langle E_{FRET} \rangle$
N49	340 K	300 K	340 K
NLS	300 K	300 K	320 K
NUS	380 K	320 K	360 K
IBB	320 K	300 K	320 K
NUL	340 K	300 K	320 K

Table S11. Scaling laws.

Scaling laws obtained by globally fitting our three experimental datasets: $R_{G,U}$, $R_{G,L}$ and $R_{E,L}$ to **EQ. 3, 7** and **8.**

DIMENSION	NATIVE IDPs	DENATURED PROTEINS
$R_{G,U}$	$0.30(N_{RES})^{0.50}$	$0.23(N_{RES})^{0.57}$
$R_{E,L}$	$0.63(N_{RES}+5)^{0.50}$	$0.61(N_{RES}+5)^{0.57}$
$R_{G,L}$	$0.27(N_{RES}+5)^{0.50}$	$0.22(N_{RES}+5)^{0.57}$

Figure S1. The fluorescent dye pair used in this study.

A) Structure of the donor dye Alexa488. B) Structure of the acceptor dye Alexa594. C) The radius of gyration of the two dye-linker moieties ($R_{G,DYES}$) as a function of the distance between the dyes ($R_{E,L}$). The results from the atomistic simulations are shown as colored dots. The solid line shows the theoretical expectation according to **EQ. S29**.

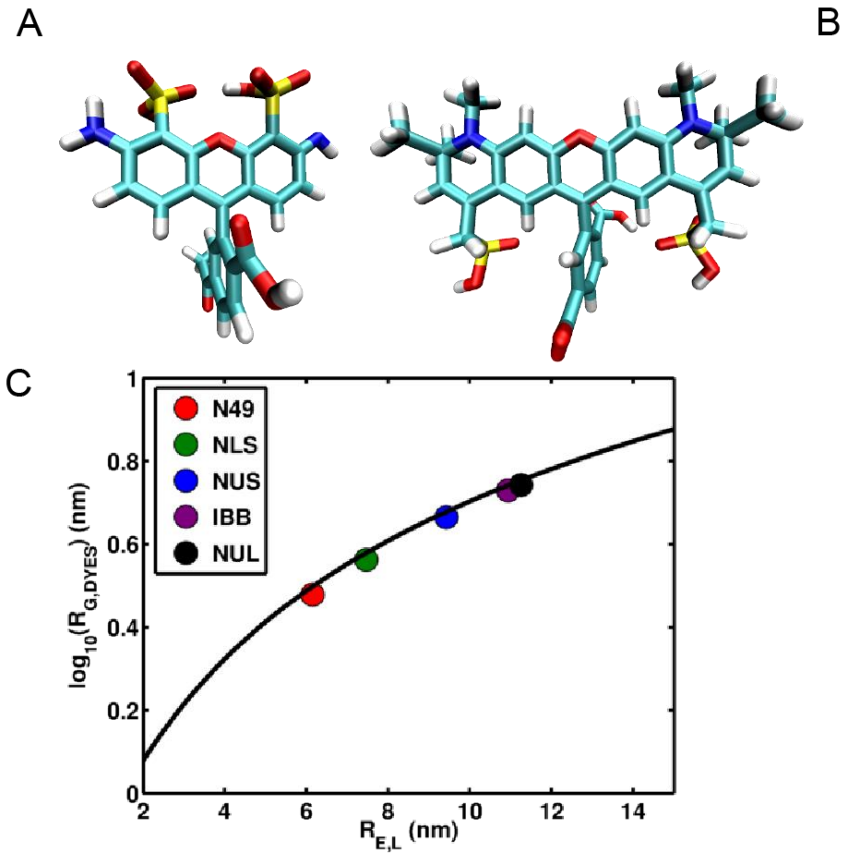


Figure S2. Gamma and quantum yield estimation.

A) Fluorescence lifetime decays of Alexa 488 free dye and conjugated to the studied proteins under native conditions. B) Fluorescence lifetime decays of Alexa 488 free dye and conjugated to the studied proteins in denaturing conditions. C) Fluorescence lifetime decays of Alexa 594 free dye and conjugated to the studied proteins under native conditions. D) Fluorescence lifetime decays of Alexa 594 free dye and conjugated to the studied proteins under denaturing conditions. E) Plot of $1/S_{(app)}$ vs $E_{(app)}$ for all the proteins (except N49) measured in native conditions for gamma estimation. F) Plot of $1/S_{(app)}$ vs $E_{(app)}$ for all the proteins (except N49) measured in denatured conditions for gamma estimation.

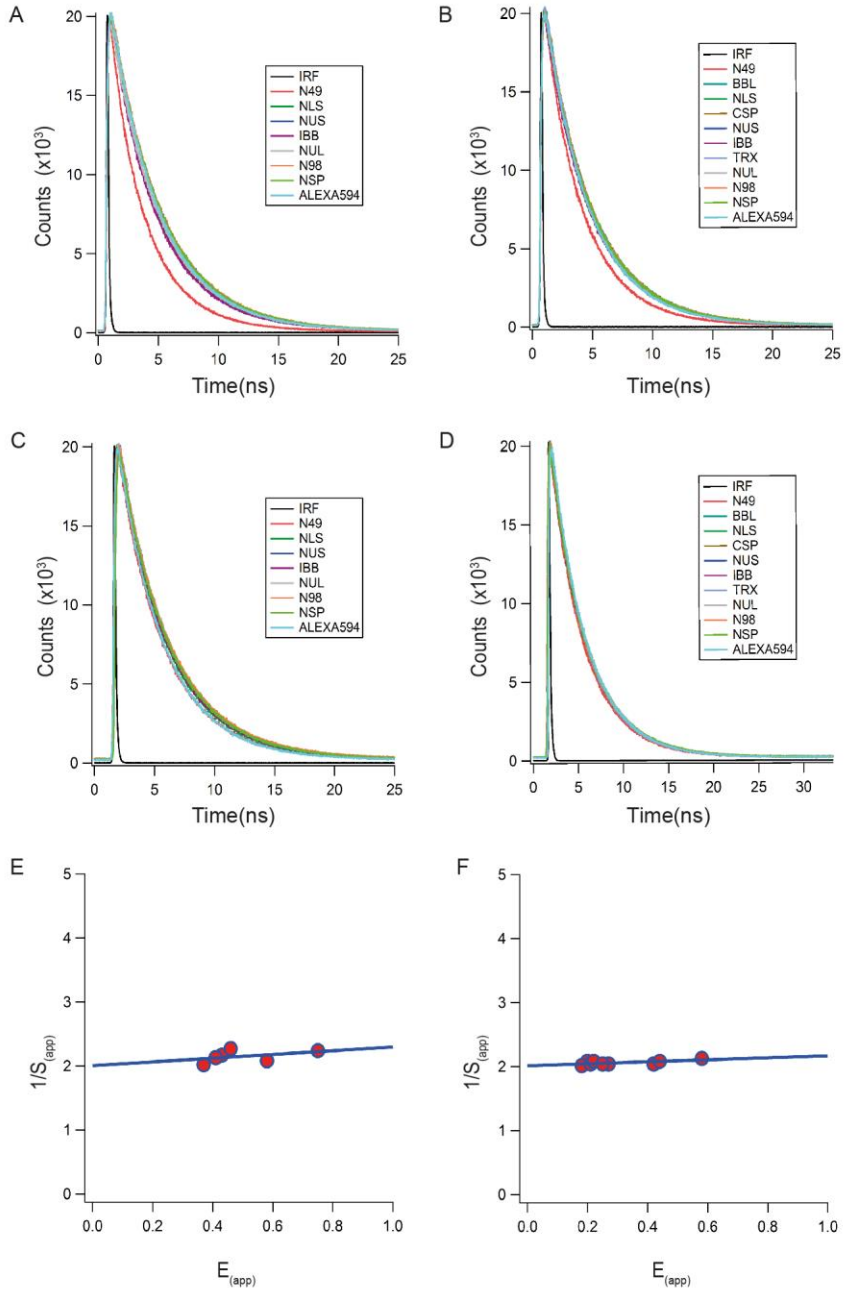


Figure S3. smFRET dataset

2D Stoichiometry (S) versus FRET efficiency (E_{FRET}) plots for all labeled proteins in the set measured by smFRET. Denatured proteins and IDPs measured in urea-containing buffer and IDPs measured in native buffer are shown in different columns. Mean FRET efficiencies $\langle E_{FRET} \rangle$ are shown in **Table S3**.

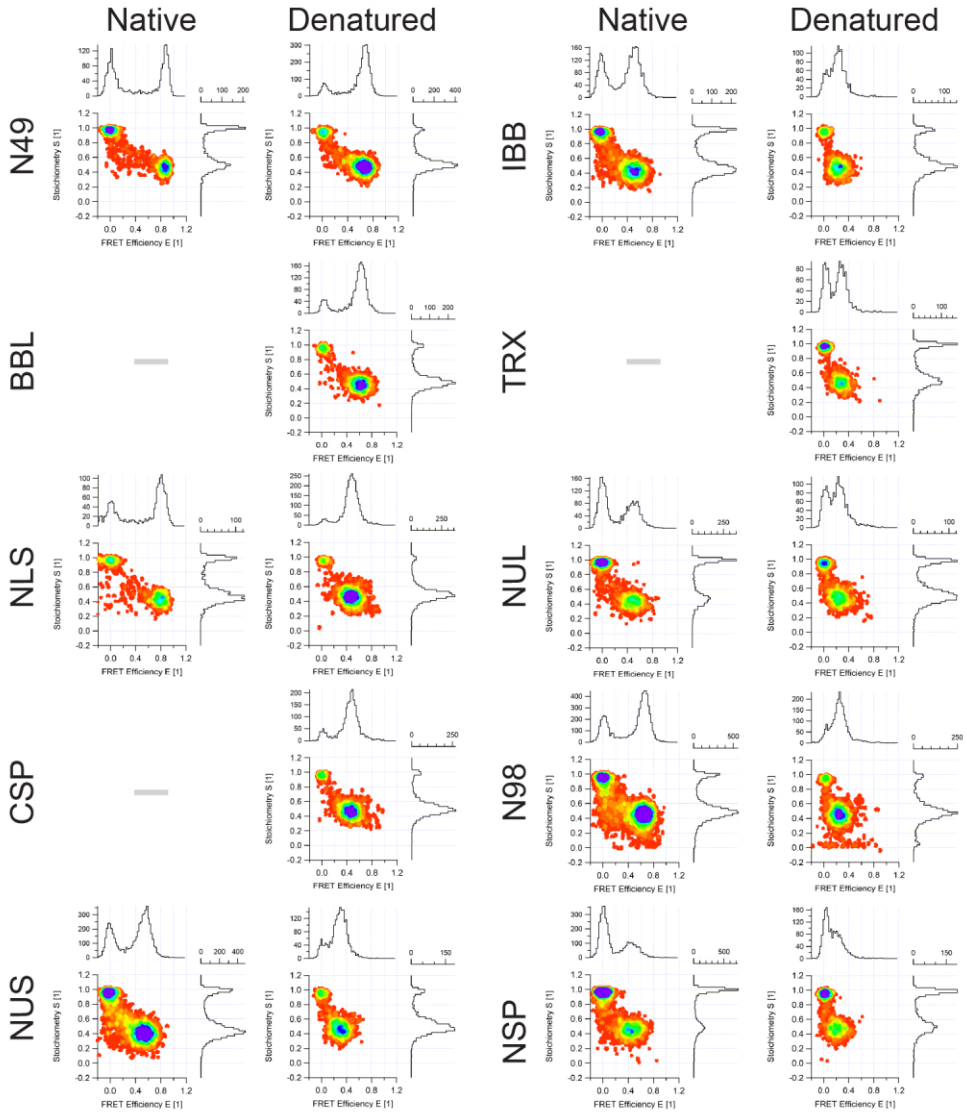


Figure S4. SAXS dataset.

A) Normalized SAXS scattering profiles of unlabeled (black lines) and labeled (red lines) samples. Dashed green blue lines show the fits to internal length scaling. B) Normalized Guinier plots of the same samples shown in A. Dashed cyan and yellow lines represent the Guinier fits of unlabeled and labeled proteins, respectively. C) Normalized Kratky plots of the same samples shown in A). D) Normalized distance distribution function, P(r), of unlabeled (black lines) and labeled (red lines) of the same samples shown in A) and B). See the values of $R_{G,U}$ and $R_{G,L}$ in **Table S3**. Denatured proteins measured in urea-containing buffer and IDPs measured in native buffer are shown in different columns. See **Note S4** for further details.

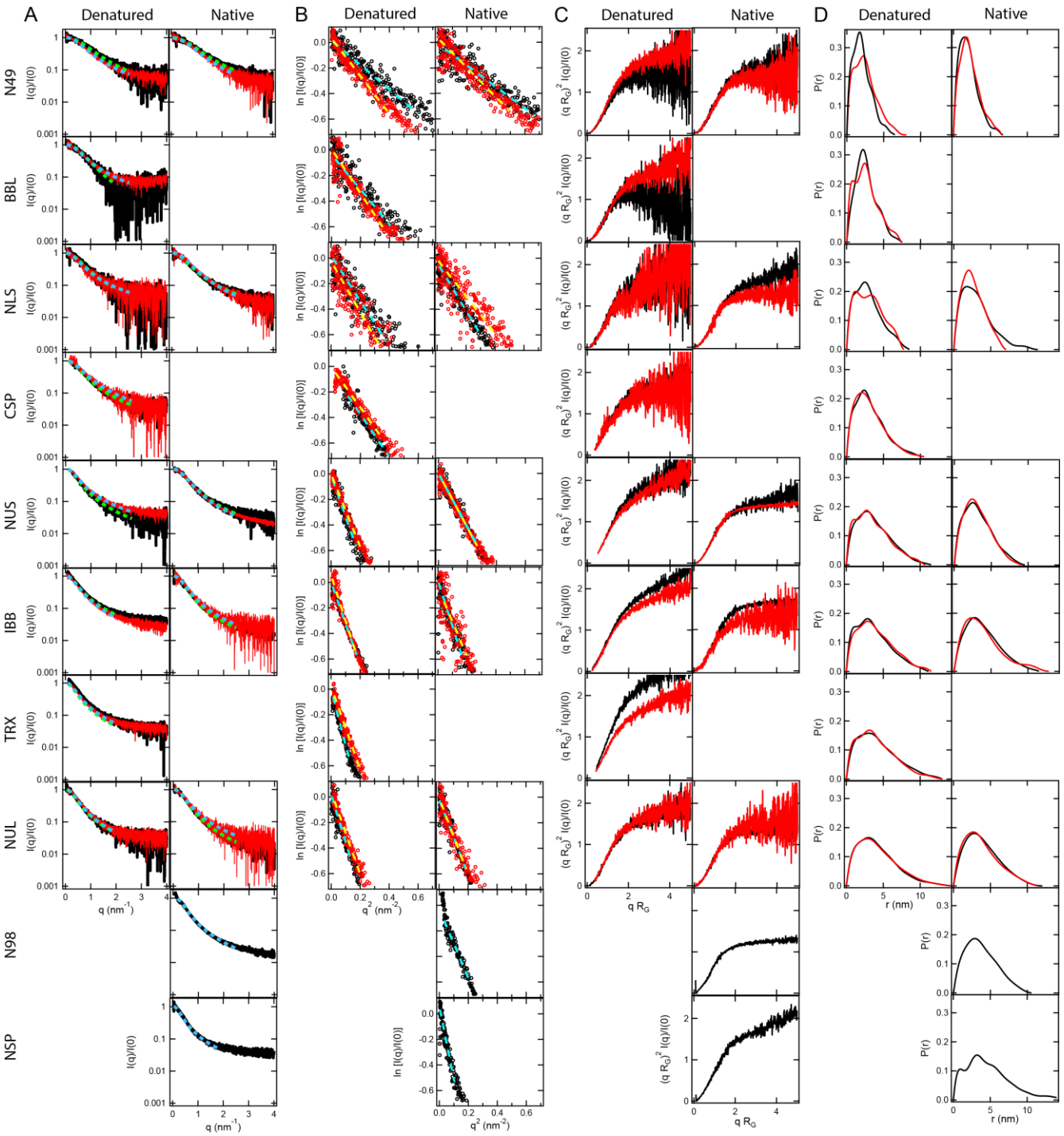

A
Denatured
Native
B
Denatured
Native
C
Denatured
Native
D
Denatured
Native
N49
BBL
NLS
CSP
NUS
IBB
TRX
NUL
N98
NSP
I(q)/I(0)
ln [I(q)/I(0)]
(q R_G)^2 I(q)/I(0)
P(r)
q (nm^-1)
q^2 (nm^-2)
q R_G
r (nm)

Figure S5. Effect of concentration.

A) $\langle E_{FRET} \rangle$ of unfolded proteins as a function of the concentration. Labeled proteins were measured at a concentration of ~10^{-7} mg/mL. B) $\langle E_{FRET} \rangle$ of IDPs in native buffer as a function of the concentration. Labeled proteins were measured at a concentration of ~10^{-7} mg/mL. For two IDPs (NLS and NUS) an excess of unlabeled protein was added at the indicated concentrations (squares). Alternatively, a solution of polyethylene glycol of 10 kDa was added at the indicated concentrations (open circles). Grey areas highlight the concentration regimes accessible to smFRET and SAXS. C) R_G of unlabeled proteins (n=8) measured by SAXS under denaturing conditions, D) R_G of unlabeled IDPs (n=7) measured by SAXS in native buffer, E) R_G of labeled proteins (n=8) measured by SAXS under denaturing conditions. F) R_G of labeled proteins (n=5) measured by SAXS in native buffer. The color code is as follows: N49 (black), BBL (red), NLS (orange), CSP (cyan), NUS (blue), IBB (magenta), TRX (violet), NUL (dark grey), N98 (green) and NSP (light grey). Lines show linear fits to the data. See **Note S6.**

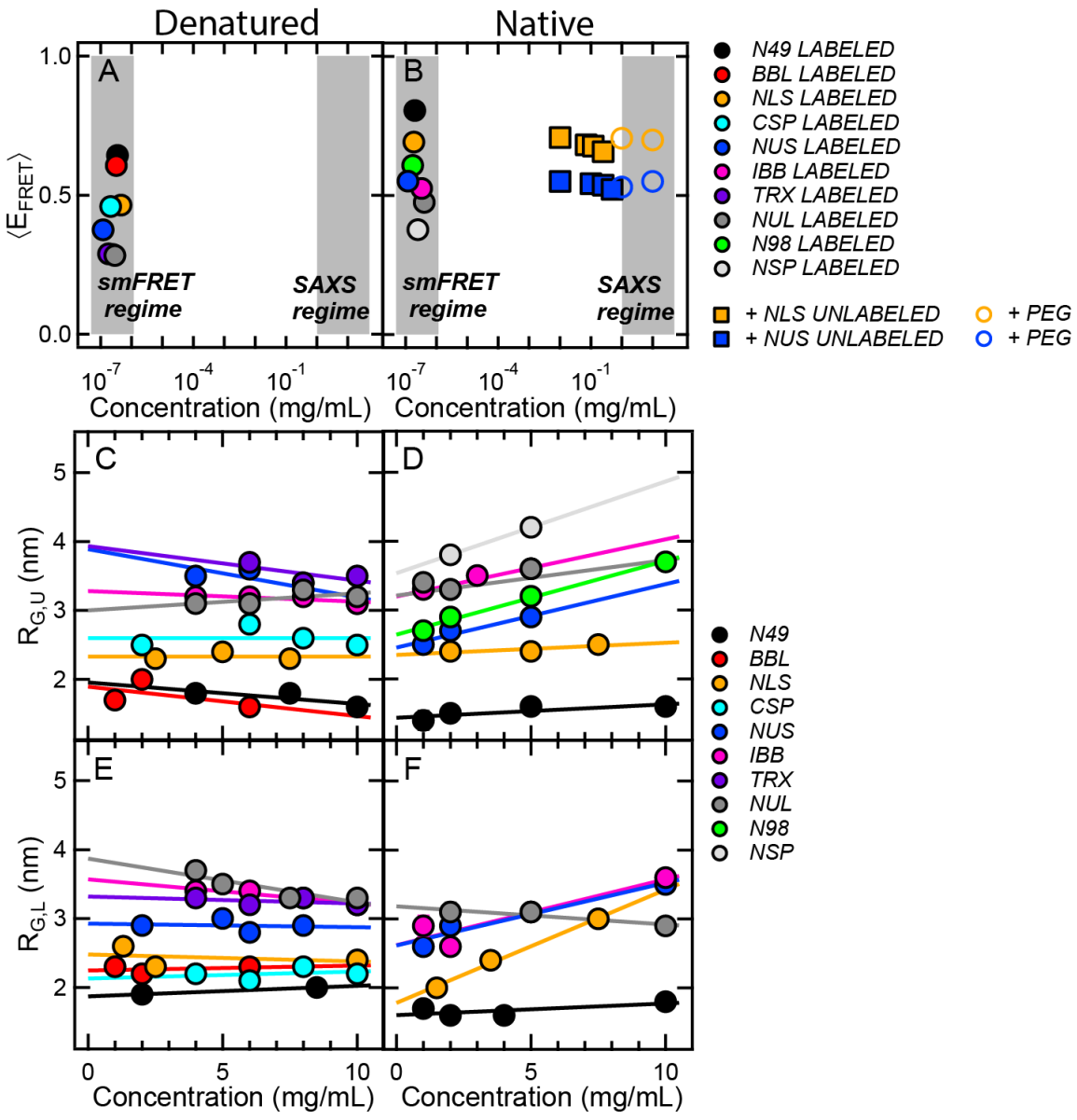

Denatured
Native
A
B
C
D
E
F
⟨E_FRET⟩
1.0
0.5
0.0
smFRET regime
SAXS regime
10^-7
10^-4
10^-1
Concentration (mg/mL)
R_G,U (nm)
R_G,L (nm)
5
4
3
2
0
5
10
N49 LABELED
BBL LABELED
NLS LABELED
CSP LABELED
NUS LABELED
IBB LABELED
TRX LABELED
NUL LABELED
N98 LABELED
NSP LABELED
+ NLS UNLABELED
+ NUS UNLABELED
+ PEG
+ PEG
N49
BBL
NLS
CSP
NUS
IBB
TRX
NUL
N98
NSP

Figure S6. Quantifying the contribution of the dyes (N_{DYES}).

A) The three measured datasets: $R_{G,U}$ (yellow squares), $R_{G,L}$ (red triangles) and $R_{E,L}$ (blue circles) of the 8 urea-unfolded proteins as a function of N_{RES}. B) The three measured datasets: $R_{G,U}$ (yellow squares), $R_{G,L}$ (red triangles) and $R_{E,L}$ (blue circles) of the 7 intrinsically disordered proteins in native buffer as a function of N_{RES}. Data were globally fitted to **EQ. 3**, **EQ. 7** and **EQ. 8** and the fits are shown as lines of the same color as the symbols. The fitted value of N_{DYES} is 5±3. The corresponding scaling laws are shown in **Table S11**. The shaded area shows the confidence interval (95%) of the study by Kohn et al (2) on chemically denatured proteins: ρ_G=0.19±0.03 nm and ν=0.60±0.03.

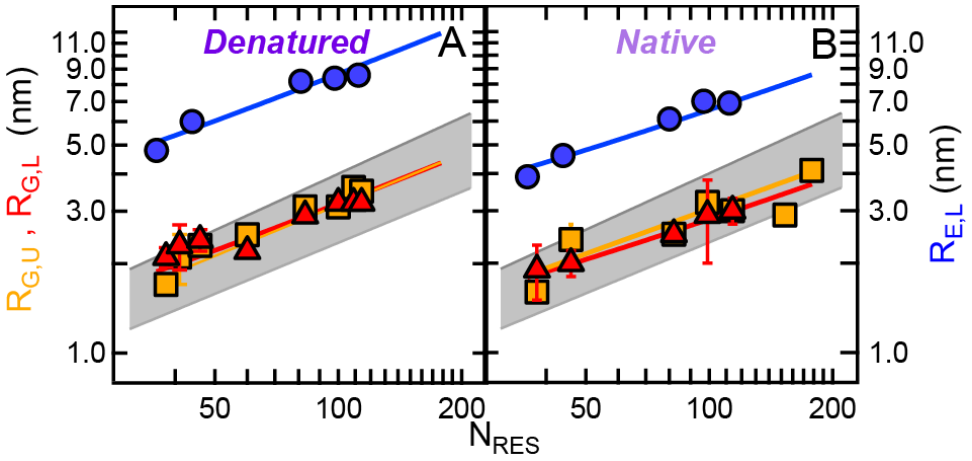


Figure S7. Shape information content in the SAXS profiles

The plots show the normalized SAXS profiles, logarithm of I(q) divided by I(0), as a function of the scattering vector *q* times R_G. Such way of plotting SAXS data effectively removes the size contribution to the scattering profile i.e. only shape information is present. The underlying color scale has been adapted from (59) and scaled between 0 (low ambiguity) and 100 (high ambiguity). A) SAXS and the scaling exponent υ: infinitely thin rod (black dashed line, υ=1), a self-avoiding chain (υ=3/5, solid gray line), a chain in theta solvent (υ=1/2, dotted gray line), a self-attracting chain (υ=1/3, dashed gray line) and a sphere (solid black line, υ=0). The curves were simulated on the basis of their theoretical form factors using **EQ. S19** and **S20** (**Note S4**). B) SAXS and the presence of fluorescent dyes: experimentally measured unlabeled NUS (black trace) and labeled NUS (red trace) profiles. C) SAXS and the ratio *G*: ensemble of NUS simulated with CAMPARI with $<E_{FRET}>$=0.8 and *G*=4 (magenta line), ensemble of NUS simulated with CAMPARI with $<E_{FRET}>$=0.2 and *G*=8 (black line with dash and dots). D) SAXS and refined ensembles: ensemble of unlabeled NUS arising from MD simulations run in water using KBFF (58) (black line with dash and dots). Previously reported MD simulations on protein NUS in aqueous solvent using the Kirkwood-Buff force-field were performed adopting the same protocol outlined in (58) but using version 2 of the force-field in which a CMAP correction of the ϕ and ψ dihedral angles is considered. $R_{G,U}$ and $R_{E,U}$ were calculated as described below and averaged over the last 80 ns of the trajectory. Ensemble of unlabeled NUS originating from Monte-Carlo simulations run with Flexible-Meccano and refined using NMR observables (chemical shifts) (53) (light blue line). The average R_G computed from the two ensembles ($R_{G,U}$=2.7 nm and 2.5 nm for the NMR and MD ensemble, respectively) are similar to the experimental R_G determined in this work (2.5 nm), while the mean $R_{E,U}$ values are 5.7 nm and 6.9 nm for the MD and NMR ensemble, respectively (58). As a consequence, the inferred values of *G* are different in each case: *G*= 6.5 for the NMR ensemble and *G*= 5.2 for the MD ensemble but the corresponding (computationally generated) SAXS curves are virtually indistinguishable from the measured SAXS curve. In C) and D) the profiles have been calculated from the atomic coordinates using CRYSOL and are not fits to the experimental profiles shown in B. The conclusion is that SAXS is very sensitive to υ but it is poorly sensitive to *G* and in particular to R_E. There is also no evidence that the presence of the dyes alters the shape of the ensemble. Therefore, R_E is an essential restraint to fully explore the conformational ensemble sampled by IDPs.

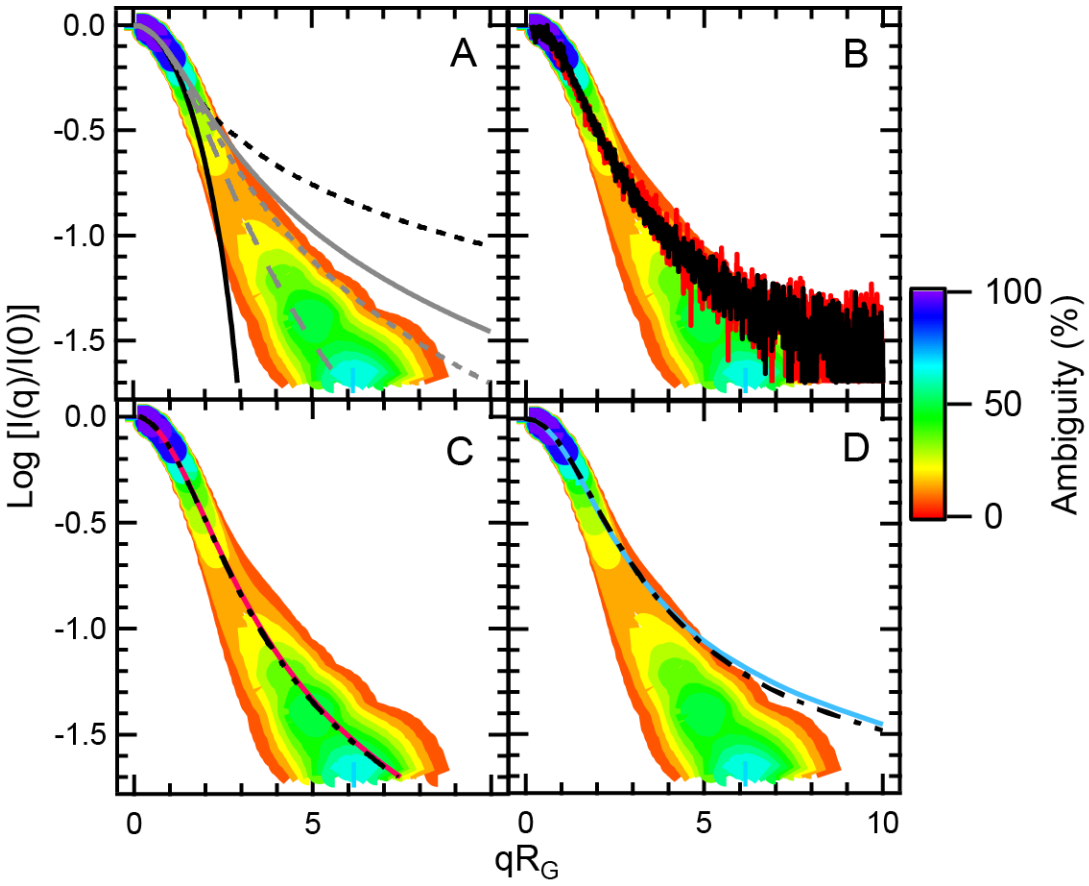

A
B
C
D
0.0
-0.5
-1.0
-1.5
0
5
10
Log [I(q)/I(0)]
qR_G
100
50
0
Ambiguity (%)

Figure S8. Test of decoupling between R_G and R_E by reweighting ensembles

A) Decrease from maximum entropy (ΔS) when simulated ensembles are reweighted to yield $R_{G,U}^2$ and each of the follow mean FRET efficiencies: $\langle E_{FRET} \rangle$=[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]. ΔS is calculated as described in **Note S6**. A ΔS value of -1 corresponds to a mean free energy change of 1kT in the simulated potential function. Here, k is the Boltzmann constant and T is the temperature. Thus, for ΔS>-1 reweighting factors are minimal. In other words, besides for the most extreme $\langle E_{FRET} \rangle$ values, simulated ensembles were generated with limited adjustments to the force-field in order to satisfy $R_{G,U}^2$ and the given efficiency. B) $\langle R_{G,SW}^2 \rangle$ calculated from each of the reweighted ensembles. All mean FRET efficiency values can yield $\langle R_{G,SW}^2 \rangle$ within error of the experimentally derived $R_{G,U}^2$ value. This suggests that a given R_G value can be consistent with the whole spectrum of mean FRET efficiencies. The red dashed line corresponds to $R_{G,U}^2$ and the grey box extends to the standard deviation associated with $R_{G,U}^2$. In A) and B) the error bars correspond to the standard error of mean over three independent simulations. C) $R_{G,SW}$ histograms extracted from the reweighted simulated ensembles with ΔS>-1. $R_{G,SW}$ histograms are similar regardless of the mean FRET efficiency constraint. D) δ^*_{SW} histograms extracted from the reweighted simulated ensembles with ΔS>-1. δ^*_{SW} quantifies the shape of a conformation, where $\delta^*_{SW} \rightarrow 0$ implies spherical conformations and $\delta^*_{SW} \rightarrow 1$ implies rod-like conformations. As the mean FRET efficiency constraints move to higher efficiencies, conformations become more spherical.

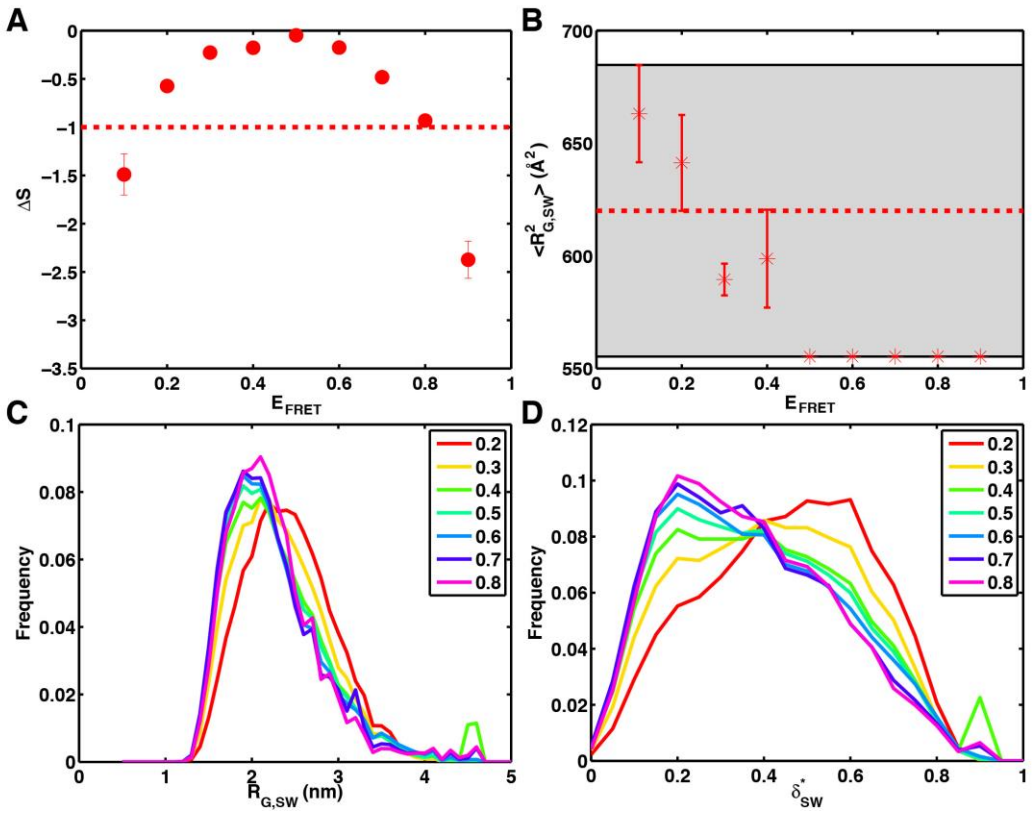


Figure S9. Sensitivity of R_G, R_E, δ^*, and G.

A) Two-dimensional histogram of R_E versus R_G for NUS ensembles reweighted to match the experimental $\langle E_{FRET} \rangle$ and $R_{G,U}^2$ values under native conditions. Lines denote G values that correspond to the following reference models: self-attracting walk (light grey), ideal Gaussian chain (grey), and the self-avoiding random walk (black). Within the region of high density (red colors) small changes in R_E (< 1 nm) can yield G values that span from those suggesting the protein adopts collapsed globular conformations (light grey) to those suggesting expanded random coil conformations (black). Thus, at small R_G values, G will be highly sensitive to the measured R_E. B) Relationship between R_G, R_E, δ^*, and G calculated from NUS ensembles reweighted to match the experimental $\langle E_{FRET} \rangle$ and $R_{G,U}^2$ values under native conditions. The binning was set to 0.4 nm and 0.1 nm in the R_E and R_G dimensions, respectively. The color of each square (bin) corresponds to the mean asphericity, $\langle \delta^* \rangle$, calculated over all (R_E, R_G) pairs that fall within this bin. The variance in $\langle \delta^* \rangle$ in the R_G dimension is smaller than the variance in $\langle \delta^* \rangle$ in the R_E dimension (colors are more similar for a given R_G slice, whereas there is a larger range of colors for a given R_E slice). Particularly, within the high density region shown in A), in contrast to G, shape quantified by δ^* shows limited changes as the R_E value changes. These results suggest δ^* is relatively insensitive to fluctuations that occur at the ends of a chain, whereas G is relatively sensitive to the fluctuations that occur at the ends of a chain. However, generally, an increase in $\langle \delta^* \rangle$ is correlated with an increase in R_G and R_E. C) Internal scaling profiles of different constructed ensembles used to test how the emergence of fluctuations at different sequence separations (|j-i|) effects R_G and δ^*. Here, the internal scaling plot for the NUS ensemble reweighted to match the experimental $\langle E_{FRET} \rangle$ and $R_{G,U}^2$ values under native conditions is shown in black. Additional NUS ensembles were reweighted to generate the remaining internal scaling profiles (red through pink colors), where each number refers to the |j-i| value at which fraying begins. The final R_E value of the frayed ensembles was set to either -1 nm or +1 nm from the reference ensemble (black). D) The fraction change in R_G and δ^* as a function of the position of end fraying. Here, fraction change is defined as $(R_{G,fray} - R_{G,ref})/R_{G,ref}$ where $R_{G,ref}$ is the value calculated from the reference NUS ensemble and $R_{G,fray}$ is the value calculated from the frayed ensemble. Both R_G and δ^* are relatively insensitive to spatial separation changes at large sequence separations (|j-i| > 60). However, δ^* becomes more sensitive to spatial separation changes at intermediate sequence separations (30 < |j-i| < 50) compared to R_G. Together, these results suggest R_G is most sensitive to smaller sequence separations, δ^* is sensitive to intermediate sequence separations, and R_E is sensitive to large sequence separations. Given that G is a measure of the ratio between R_E and R_G, G will be most sensitive to sequence separations of the most sensitive value, which for low to intermediate R_G values will be R_E.

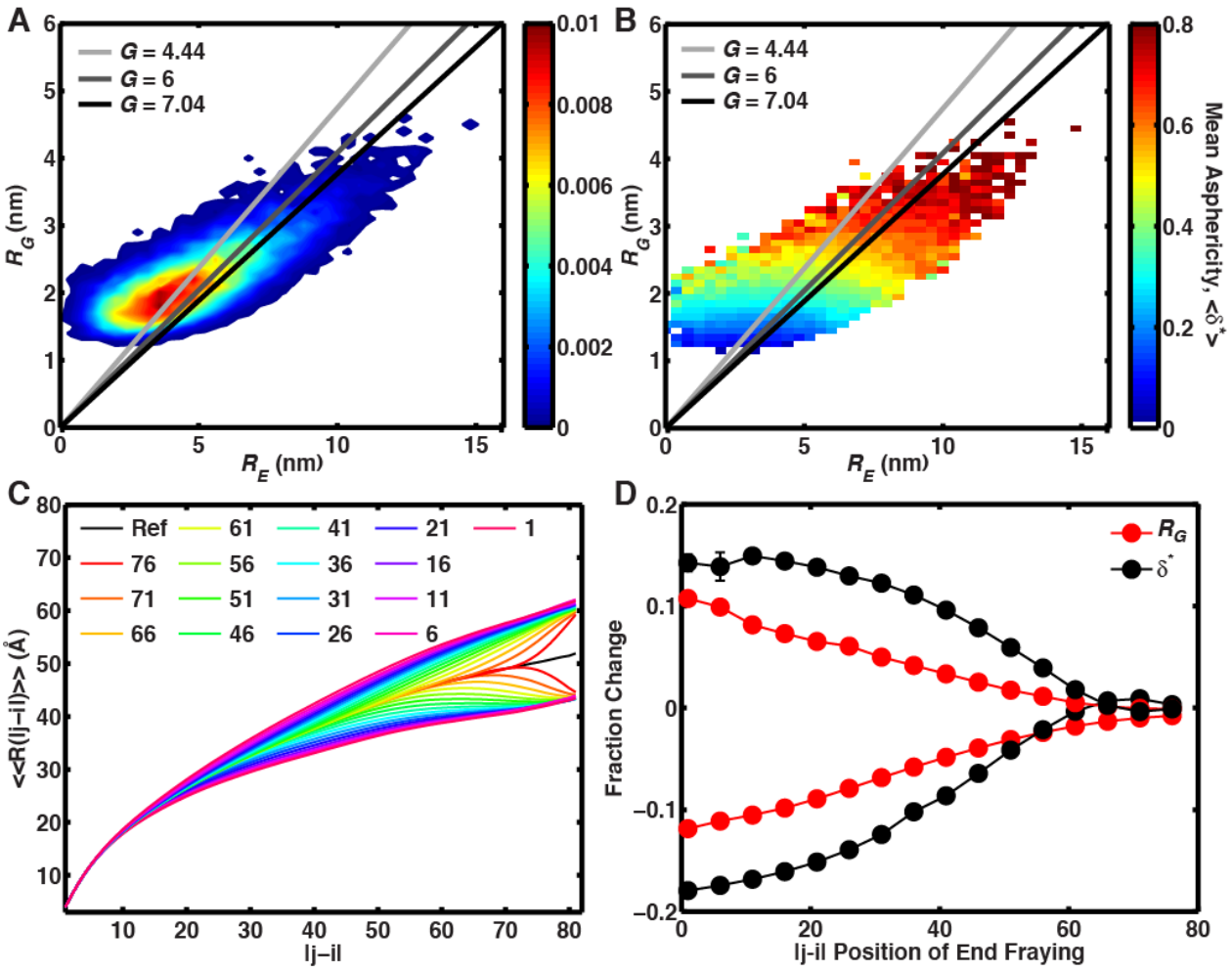

A
G = 4.44
G = 6
G = 7.04
R_G (nm)
R_E (nm)
B
Mean Asphericity, <δ*>
C
Ref
61
41
21
1
76
56
36
16
71
51
31
11
66
46
26
6
<<R(|j−i|)>> (Å)
|j−i|
D
R_G
δ*
Fraction Change
|j-i| Position of End Fraying

Figure S10: Relationship between shape as quantified by δ^* and G

A)-G) 2-dimensional histograms of G_{SW} versus δ^*_{SW}. H) Mean G_{SW} versus mean δ^*_{SW} for each mean FRET efficiency with $\Delta S>-1$. δ^*_{SW} quantifies shape using the conformation specific eigenvalues of the gyration tensor, whereas G_{SW} quantifies shape using the conformation specific R_G and R_E. Thus, G_{SW} depends largely on the distance between the ends of the chain, whereas δ^*_{SW} averages over the entire chain. This leads to a range of G_{SW} values that are consistent with a given δ^*_{SW} value and vice versa. However, the mean G_{SW} and δ^*_{SW} values are weakly coupled.

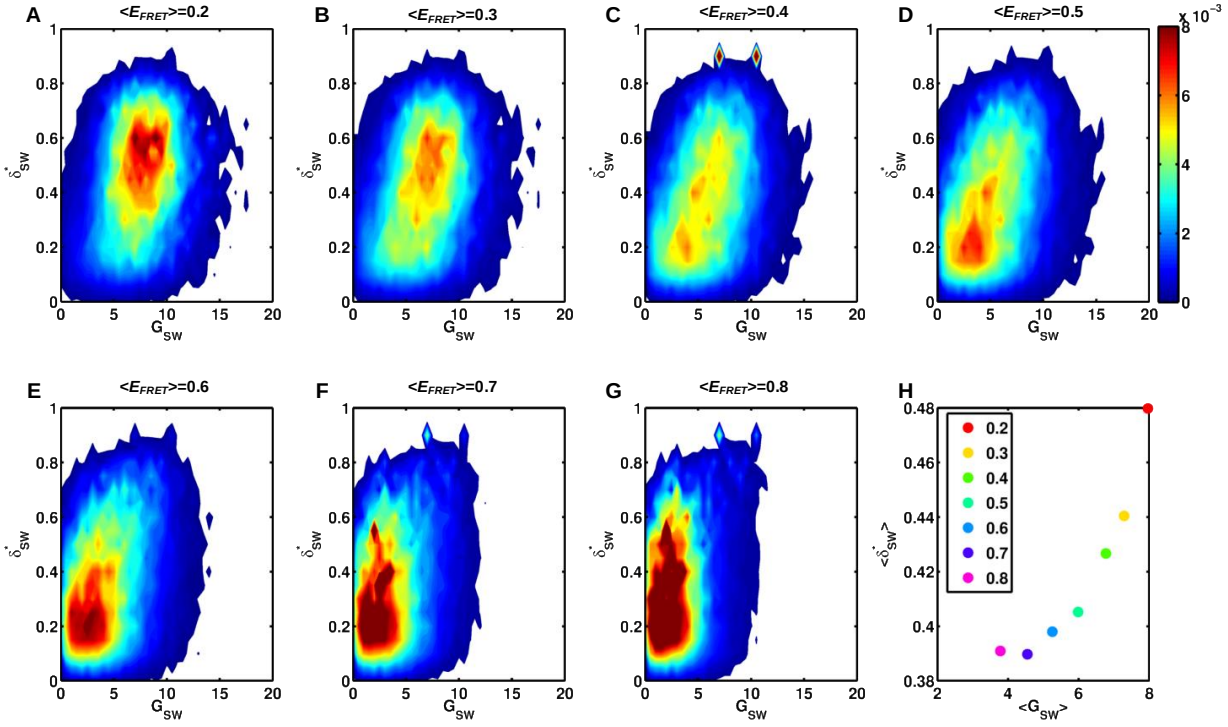


Figure S11. Comparison between EOM and reweighted ensembles. .

A large ensemble generated via CAMPARI simulations using the ABSINTH force-field was used as the pool from where individual conformers are selected. EOM ensembles (typically containing 20 structures) are generated so that the selected conformations collectively fit the full SAXS profile. Reweighted ensembles are generated by altering the probability of each and every conformer in the pool so that the average $R_{G,U}$ and $\langle E_{FRET} \rangle$ best matches the experimental $R_{G,U}$ and $\langle E_{FRET} \rangle$, respectively. Experimental data is shown in black. EOM ensembles and reweighted ensembles are shown in red and cyan colors, respectively. Denatured and native ensembles are shown as dashed and solid lines, respectively. EOM analysis suggest that the IDPs display significant heterogeneity of the conformational space, as manifested in the broad distributions of the EOM-selected ensembles compared to the random chain distributions. Note that reducing the EOM results to R_G distributions in a way to simplify the results and apparent "bimodality" in some R_G distributions in this figure does not mean "bimodality" in terms of specific configurations. Instead, such distributions can rather indicate that a system has high conformational heterogeneity and also that extended conformers are densely populated. Notably, significantly dissimilar heterogeneity is observed in the conformational spaces sampled by denatured vs native IDPs.

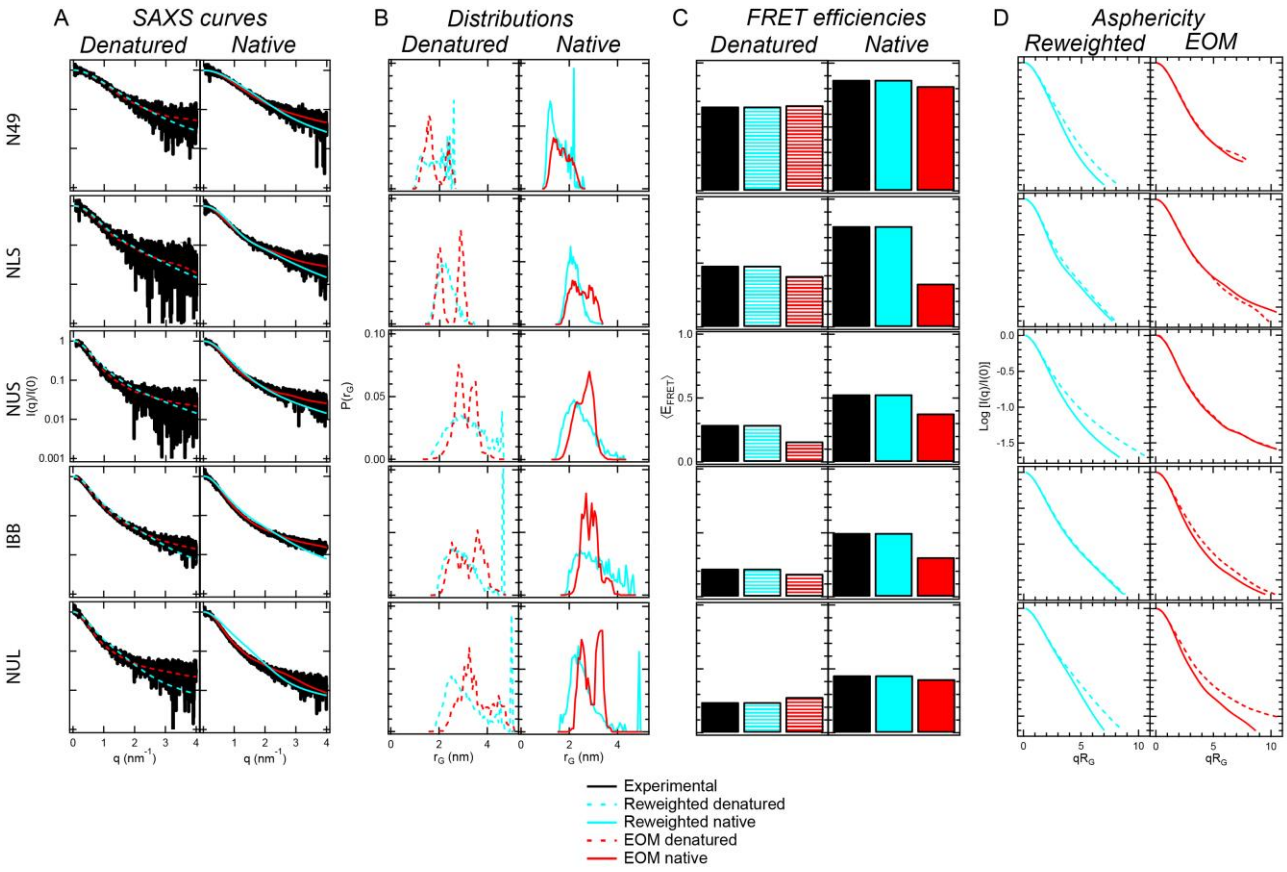


Figure S12. The effect of the dyes on protein size (R_G) depends on G.

A) Theoretical ΔR_G ($= R_{G,L} - R_{G,U}$) as a function of $R_{E,L}$. $R_{G,U}$ calculated from **EQ. 3** with ρ_G=0.2 nm, υ=0.6 and the corresponding N_{RES} (**Table S2**). $R_{G,L}$ was calculated from $R_{G,U}$ and $R_{E,L}$ according to the theoretical predictions from the parallel axes theorem (**EQ. S28** and **EQ. S29).** Notice that for many values of $R_{E,L}$ there is no increase in R_G (ΔR_G<0). B) Theoretical ΔR_G as a function of G. The graph has been generated from the plot in A) dividing $R_{E,L}$ by the corresponding $R_{G,L}$. Our calculated *G* values for denatured proteins (G=7.1) and IDPs (G=4.3) are shown as dark violet and light violet lines respectively. Note that for G=7.1, ΔR_G is well above zero while for G~4, ΔR_G ~0. The conclusion is that for denatured proteins in urea the dyes tend to locate preferentially towards the outer shell while for native IDPs, the ends of the polymer tend to be hidden near the center. The color code is as follows: N49 (black), NLS (orange) and NUS (blue), IBB (magenta) and NUL (dark grey). See **Note S7** for further details. C) Experimental ΔR_G as a function of N_{RES}. Native IDPs and denatured proteins are shown as light violet and dark violet points, respectively. Solid lines are linear fits to the experimental data points of the same color and are intended to guide the eye only. Notice that the increase in size due to the dyes is inversely related to the protein size i.e. the smaller the protein the larger the dye contribution, as intuitively expected.

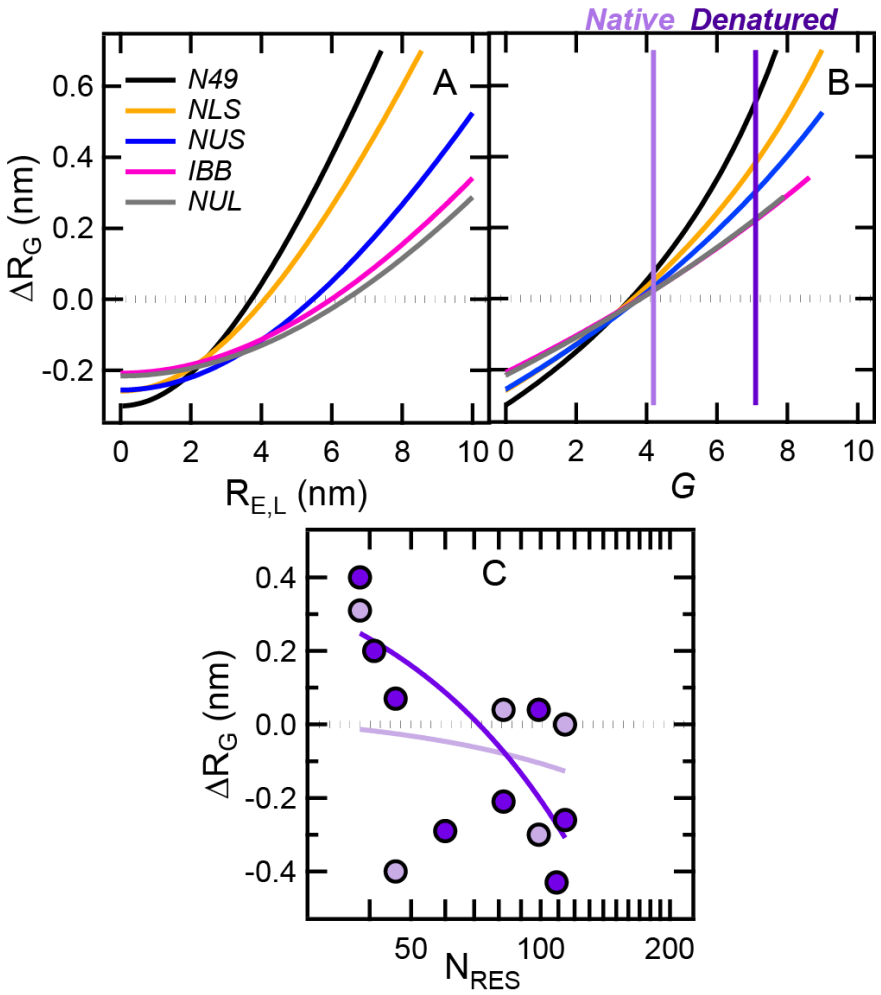
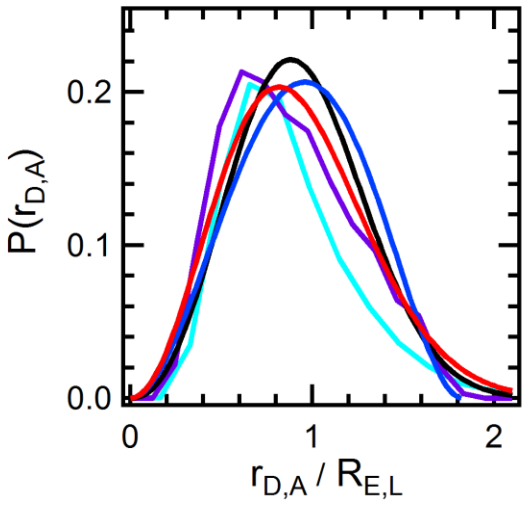


Figure S13. Comparison of different distributions.

Donor-acceptor distance distributions ($P(r_{D,A})$) plotted as a function of the normalized distance ($r_{DA}/R_{E,L}$). Gaussian chain model (black line), Haran model based on a distribution of points on sphere (blue line), SARW model (red line), NUS ensembles simulated with CAMPARI reweighted to match the experimental values ($\langle E_{FRET} \rangle$ and $R_{G,U}^2$) obtained for denatured conditions (violet line) and NUS ensembles simulated with CAMPARI reweighted to match the experimental values obtained for native conditions (cyan line). See **Note S8**.



Supplementary references.

1. Lemke EA (2011) Site-specific labeling of proteins for single-molecule FRET measurements using genetically encoded ketone functionalities. *Methods in molecular biology* 751:3-15.
2. Kohn JE*, et al.* (2004) Random-coil behavior and the dimensions of chemically unfolded proteins. *P Natl Acad Sci USA* 101(34):12491-12496.
3. Uversky VN (2009) Intrinsically Disordered Proteins and Their Environment: Effects of Strong Denaturants, Temperature, pH, Counter Ions, Membranes, Binding Partners, Osmolytes, and Macromolecular Crowding. *Protein J* 28(7-8):305-325.
4. Soranno A*, et al.* (2014) Single-molecule spectroscopy reveals polymer effects of disordered proteins in crowded environments. *P Natl Acad Sci USA*.
5. Liu J*, et al.* (2012) Exploring one-state downhill protein folding in single molecules. *P Natl Acad Sci USA* 109(1):179-184.
6. Muller-Spath S*, et al.* (2010) From the Cover: Charge interactions can dominate the dimensions of intrinsically disordered proteins. *P Natl Acad Sci USA* 107(33):14609-14614.
7. Kapanidis AN*, et al.* (2004) Fluorescence-aided molecule sorting: analysis of structure and interactions by alternating-laser excitation of single molecules. *P Natl Acad Sci USA* 101(24):8936-8941.
8. Muller BK, Zaychikov E, Brauchle C, & Lamb DC (2005) Pulsed interleaved excitation. *Biophysical journal* 89(5):3508-3522.
9. Kapanidis AN*, et al.* (2005) Alternating-laser excitation of single molecules. *Acc Chem Res* 38(7):523-533.
10. Eggeling C*, et al.* (2001) Data registration and selective single-molecule analysis using multi-parameter fluorescence detection. *J Biotechnol* 86(3):163-180.
11. Kudryavtsev V*, et al.* (2012) Combining MFD and PIE for accurate single-pair Forster resonance energy transfer measurements. *Chemphyschem : a European journal of chemical physics and physical chemistry* 13(4):1060-1078.
12. Sisamakis E, Valeri A, Kalinin S, Rothwell PJ, & Seidel CA (2010) Accurate single-molecule FRET studies using multiparameter fluorescence detection. *Methods in enzymology* 475:455-514.
13. Schaffer J*, et al.* (1999) Identification of single molecules in aqueous solution by time-resolved fluorescence anisotropy. *J Phys Chem A* 103(3):331-336.
14. Enderlein J*, et al.* (1997) A maximum likelihood estimator to distinguish single molecules by their fluorescence decays. *Chem Phys Lett* 270(5-6):464-470.
15. Gopich IV & Szabo A (2012) Theory of the energy transfer efficiency and fluorescence lifetime distribution in single-molecule FRET. *P Natl Acad Sci USA* 109(20):7747-7752.
16. Rubinstein M & Colby RH (2003) *Polymer Physics* (Oxford University Press, Oxford and New York).
17. Förster T (1948) Zwischenmolekulare Energiewanderung und Fluoreszenz. *Ann. Phys.* 437(1-2):55-75.
18. Ferreon AC, Gambin Y, Lemke EA, & Deniz AA (2009) Interplay of alpha-synuclein binding and conformational switching probed by single-molecule fluorescence. *P Natl Acad Sci USA* 106(14):5645-5650.
19. O'Brien EP, Morrison G, Brooks BR, & Thirumalai D (2009) How accurate are polymer models in the analysis of Forster resonance energy transfer experiments on proteins? *J Chem Phys* 130(12).
20. Schuler B, Lipman EA, & Eaton WA (2002) Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature* 419(6908):743-747.

21. Choi UB, McCann JJ, Weninger KR, & Bowen ME (2011) Beyond the Random Coil: Stochastic Conformational Switching in Intrinsically Disordered Proteins. *Structure* 19(4):566-576.
22. Kuzmenkina EV, Heyes CD, & Nienhaus GU (2005) Single-molecule Forster resonance energy transfer study of protein dynamics under denaturing conditions. *P Natl Acad Sci USA* 102(43):15471-15476.
23. Hoffmann A*, et al.* (2007) Mapping protein collapse with single-molecule fluorescence and kinetic synchrotron radiation circular dichroism spectroscopy. *P Natl Acad Sci USA* 104(1):105-110.
24. Fitzkee NC & Rose GD (2004) Reassessing random-coil statistics in unfolded proteins. *Proceedings of the National Academy of Sciences of the United States of America* 101(34):12497-12502.
25. Petoukhov MV*, et al.* (2012) New developments in the ATSAS program package for small-angle scattering data analysis. *J Appl Crystallogr* 45:342-350.
26. Franke D, Kikhney AG, & Svergun DI (2012) Automated acquisition and analysis of small angle X-ray scattering data. *Nucl Instrum Meth A* 689:52-59.
27. Guinier A (1939) La diffraction des rayons X aux tres petits angles: application a l'etude de phenomenes ultramicroscopiques.161-237.
28. Svergun DI, Koch MHJ, Timmins PA, & May RP (*Small angle x-ray and neutron scattering from solutions of biological macromolecules* First Edition. Ed pp ix, 358 pages.
29. Svergun DI (1992) Determination of the Regularization Parameter in Indirect-Transform Methods Using Perceptual Criteria. *J Appl Crystallogr* 25:495-503.
30. Receveur-Brechot V & Durand D (2012) How random are intrinsically disordered proteins? A small angle scattering perspective. *Current protein & peptide science* 13(1):55-75.
31. Bernado P, Mylonas E, Petoukhov MV, Blackledge M, & Svergun DI (2007) Structural characterization of flexible proteins using small-angle X-ray scattering. *Journal of the American Chemical Society* 129(17):5656-5664.
32. Johansen D, Trewhella J, & Goldenberg DP (2011) Fractal dimension of an intrinsically disordered protein: small-angle X-ray scattering and computational study of the bacteriophage lambda N protein. *Protein science : a publication of the Protein Society* 20(12):1955-1970.
33. Breßler I, Kohlbrecher J, & Thünemann AF (2015) SASfit : a tool for small-angle scattering data analysis using a library of analytical expressions. *J Appl Crystallogr* 48(5):1587-1598.
34. Debye P (1947) Molecular-weight determination by light scattering. *J Phys Colloid Chem* 51(1):18-32.
35. Theodorou DN & Suter UW (1985) Shape of Unperturbed Linear-Polymers - Polypropylene. *Macromolecules* 18(6):1206-1214.
36. Toretsky JA & Wright PE (2014) Assemblages: functional units formed by cellular phase separation. *J Cell Biol* 206(5):579-588.
37. Vitalis A & Pappu R (2009) ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. *Journal Of Computational Chemistry* 30(5):673-699.
38. Sanchez IC (1979) Phase-Transition Behavior of the Isolated Polymer-Chain. *Macromolecules* 12(5):980-988.
39. Vitalis A & Pappu RV (2009) ABSINTH: a new continuum solvation model for simulations of polypeptides in aqueous solutions. *J Comput Chem* 30(5):673-699.
40. Debye P (1946) The Intrinsic Viscosity of Polymer Solutions. *The Journal of Chemical Physics* 14(10):636-639.
41. Milles S*, et al.* (2012) Click strategies for single-molecule protein fluorescence. *Journal of the American Chemical Society* 134(11):5187-5195.

42. Leung HT*, et al.* (2016) A Rigorous and Efficient Method To Reweight Very Large Conformational Ensembles Using Average Experimental Data and To Determine Their Relative Information Content. *Journal of chemical theory and computation* 12(1):383-394.
43. Song J, Gomes GN, Gradinaru CC, & Chan HS (2015) An Adequate Account of Excluded Volume Is Necessary To Infer Compactness and Asphericity of Disordered Proteins by Forster Resonance Energy Transfer. *J Phys Chem B* 119(49):15191-15202.
44. Ziv G & Haran G (2009) Protein folding, protein collapse, and tanford's transfer model: lessons from single-molecule FRET. *Journal of the American Chemical Society* 131(8):2942-2947.
45. Hofmann H*, et al.* (2012) Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *P Natl Acad Sci USA* 109(40):16155-16160.
46. Fisher ME (1966) Shape of a Self-Avoiding Walk or Polymer Chain. *The Journal of Chemical Physics* 44(2):616-622.
47. Zhou HX (2004) Polymer models of protein stability, folding, and interactions. *Biochemistry* 43(8):2141-2154.
48. Flory PJ (1969) *Statistical mechanics of chain molecules* (Interscience Publishers) p 460.
49. Hammouda B (1993) Sans from Homogeneous Polymer Mixtures - a Unified Overview. *Adv Polym Sci* 106:87-133.
50. Merchant KA, Best RB, Louis JM, Gopich IV, & Eaton WA (2007) Characterizing the unfolded states of proteins using single-molecule FRET spectroscopy and molecular simulations. *Proceedings of the National Academy of Sciences of the United States of America* 104(5):1528-1533.
51. Uversky VN, Gillespie JR, & Fink AL (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* 41(3):415-427.
52. Neuweiler H, Johnson CM, & Fersht AR (2009) Direct observation of ultrafast folding and denatured state dynamics in single protein molecules. *P Natl Acad Sci USA* 106(44):18569-18574.
53. Milles S*, et al.* (2015) Plasticity of an Ultrafast Interaction between Nucleoporins and Nuclear Transport Receptors. *Cell* 163(3):734-745.
54. Jeng MF*, et al.* (1994) High-resolution solution structures of oxidized and reduced Escherichia coli thioredoxin. *Structure* 2(9):853-868.
55. Le Guillou JC & Zinn-Justin J (1977) Critical Exponents for the n-Vector Model in Three Dimensions from Field Theory. *Physical Review Letters* 39(2):95-98.
56. Witten TA & Schafer L (1978) 2 Critical Ratios in Polymer-Solutions. *J Phys a-Math Gen* 11(9):1843-1854.
57. Lund O*, et al.* (1997) Protein distance constraints predicted by neural networks and probability density functions. *Protein engineering* 10(11):1241-1248.
58. Mercadante D*, et al.* (2015) Kirkwood-Buff Approach Rescues Overcollapse of a Disordered Protein in Canonical Protein Force Fields. *J Phys Chem B* 119(25):7975-7984.
59. Petoukhov MV & Svergun DI (2015) Ambiguity assessment of small-angle scattering curves from monodisperse systems. *Acta Crystallogr D Biol Crystallogr* 71(Pt 5):1051-1058.