Biophysical Journal
Article

Biophysical Society

Assessment of models for calculating the hydrodynamic radius of intrinsically disordered proteins

Francesco Pesce,[1] Estella A. Newcombe,[1] Pernille Seiffert,[1] Emil E. Tranchant,[1] Johan G. Olsen,[1] Christy R. Grace,[2] Birthe B. Kragelund,[1,*] and Kresten Lindorff-Larsen[1,*]
[1]Structural Biology and NMR Laboratory, The Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Copenhagen, Denmark and [2]Department of Structural Biology, St. Jude Children's Research Hospital, Memphis, Tennessee

ABSTRACT Diffusion measurements by pulsed-field gradient NMR and fluorescence correlation spectroscopy can be used to probe the hydrodynamic radius of proteins, which contains information about the overall dimension of a protein in solution. The comparison of this value with structural models of intrinsically disordered proteins is nonetheless impaired by the uncertainty of the accuracy of the methods for computing the hydrodynamic radius from atomic coordinates. To tackle this issue, we here build conformational ensembles of 11 intrinsically disordered proteins that we ensure are in agreement with measurements of compaction by small-angle x-ray scattering. We then use these ensembles to identify the forward model that more closely fits the radii derived from pulsed-field gradient NMR diffusion experiments. Of the models we examined, we find that the Kirkwood-Riseman equation provides the best description of the hydrodynamic radius probed by pulsed-field gradient NMR experiments. While some minor discrepancies remain, our results enable better use of measurements of the hydrodynamic radius in integrative modeling and for force field benchmarking and parameterization.

SIGNIFICANCE Accurate models of the conformational properties of intrinsically disordered proteins rely on our ability to interpret experimental data that report on the conformational ensembles of these proteins in solution. Methods to calculate experimental observables from conformational ensembles are central to link experiments and computation, for example, in integrative modeling or the assessment of molecular force fields. Benchmarking such methods is, however, difficult for disordered proteins because it is difficult to construct accurate ensembles without using the data. Here, we circumvent this problem by combining independent measures of protein compaction to test several methods to calculate the hydrodynamic radius of a disordered protein, as measured by pulsed-field gradient NMR diffusion experiments, and find the Kirkwood-Riseman model to be most accurate.

INTRODUCTION

Intrinsically disordered proteins and regions (here collectively termed IDPs) are highly flexible molecules in solution and they should therefore be described as ensembles of different conformations. The biological function of IDPs is often linked to their dynamics and therefore the knowledge of the conformational ensemble can be helpful in understanding their functions (1,2). Integrative modeling approaches are often used to study the conformational ensembles of IDPs (3–8). Here, experiments typically probe ensemble-averaged structural information, and are interpreted using computational methods to generate structures at atomic or coarse-grained resolutions.

A key property that describes the conformation of an IDP is its average dimension. For example, the expansion of an IDP determines its "capture radius" for binding (9) and is correlated with its propensity to phase separate (10,11). Until relatively recently, the force fields used in all-atom and certain coarse-grained molecular dynamics simulations led to conformational ensembles that were too compact (12–18). Experimentally, compaction may be probed by, for example, small-angle x-ray scattering (SAXS) (19), pulsed-field gradient (PFG) nuclear magnetic resonance (NMR) diffusion experiments (20), fluorescence correlation spectroscopy (21), and dynamic light scattering (22).

Submitted June 15, 2022, and accepted for publication December 9, 2022.
*Correspondence: bbk@bio.ku.dk or lindorff@bio.ku.dk
Editor: Samrat Mukhopadhyay.
https://doi.org/10.1016/j.bpj.2022.12.013

Comparison of experiments and simulations is often based on so-called forward models that enable the calculation of experimental observables (or close proxies) from atomic (or coarse-grained) coordinates. Forward models play a key role in integrative modeling and force field assessment. Developing accurate forward models for IDPs is, however, complicated by the lack of precise conformational ensembles that can be used to train and parametrize these models (23). Instead, forward models are generally developed and benchmarked for folded and relatively static proteins, whose structures may more easily and accurately be determined, but it is not always clear how well these models are transferable to highly dynamic, unfolded, and disordered proteins.

Here, we examine the accuracy of methods to calculate the hydrodynamic radius (R_h) of IDPs from conformational ensembles and the comparison to PFG NMR diffusion measurements. The R_h is a measure of the overall dimension of a protein as it represents the radius of a sphere that diffuses with the same translational diffusion coefficient (D_t) of the protein, and may conveniently be probed via PFG NMR experiments. In these, D_t is probed via monitoring the effects of a nonuniform magnetic field (defined by a gradient strength, G), in a spin echo NMR experiment. Depending on how far the protein has moved in the sample during a set diffusion time, Δ, different levels of signal decays are observed (20,24–26). In practice, this is often detected by integrating a specific region of the NMR spectrum to measure the signal intensity, I, and varying the gradient strength, G. This profile can then be fitted to the Stejskal-Tanner equation to obtain D_t (20):

$$I = I_0 e^{-G^2\gamma^2\delta^2\left(\Delta - \frac{\delta}{3}\right)D_t}. \tag{1}$$

Here, γ is the gyromagnetic ratio and δ is the length of the gradient. The value of R_h may then be obtained from D_t either via the Stokes-Einstein equation, when the solvent viscosity is known:

$$D_t = \frac{k_b T}{6\pi\eta R_\mathrm{h}}, \tag{2}$$

where k_b is the Boltzmann constant, T is the temperature, and η is the solvent viscosity, or, and most often used, indirectly using an internal reference with known R_h:

$$R_\mathrm{h}(\mathrm{protein}) = \frac{D_t(\mathrm{reference})}{D_t(\mathrm{protein})} R_\mathrm{h}(\mathrm{reference}). \tag{3}$$

Different forward models have been proposed to calculate R_h from atomic coordinates (27–29). In particular, HYDROPRO (27) and HYDROPRO-derived models (30) are widely used to compare R_h values obtained by PFG NMR diffusion experiments to conformational ensembles of IDPs from molecular simulations (15,31–35). Despite this, differences of ∼20% between the results provided by different forward models have been observed (30,36).

Given the widespread use of R_h measurements for constructing conformational ensembles and the potential for assessing and improving force fields, we decided to assess the accuracy of different forward models for R_h. Having an accurate forward model, for example, makes it possible to provide a more fine-grained assessment of force field accuracy. In the context of integrative modeling, a conformational ensemble of an IDP can be pushed to be either more expanded or more compact to fit the experimental data depending on the forward model used. The first step in our work was thus to generate conformational ensembles of IDPs that are accurate in terms of reproducing their overall dimensions, but without using the PFG NMR diffusion measurements. We did so by using state-of-the-art computational methods for sampling the overall dimensions of IDPs, and further used SAXS data to benchmark and improve the agreement with independent experimental data that also provide information on the average dimension of proteins in solution (Fig. 1). While SAXS and NMR diffusion experiments may probe different aspects of compaction (34), we assume that these differences are small and would vary between different proteins. Thus, we used the SAXS-refined conformational ensembles as input to different forward models for R_h and compared the results with experiments. As data for benchmarking the forward models, we chose 11 IDPs with varying lengths (24–441 residues) and amino acid compositions and recorded both SAXS and PFG NMR diffusion data unless data were already available in literature. We find that, for this diverse set of proteins, the Kirkwood-Riseman equation (28) gives a better agreement between our ensembles and the measured hydrodynamic radii.

MATERIALS AND METHODS

Protein purifications and experimental conditions

The growth hormone receptor intracellular domain

For SAXS experiments, the growth hormone receptor intracellular domain (GHR-ICD) (residues 270–620) was expressed and purified as described in (37). For NMR experiments, GHR-ICD was expressed as a $\mathrm{His_6}$-SUMO fusion protein ($\mathrm{His_6}$-SUMO-GHR-ICD) in *E. coli* BL21(DE3) cells, transformed by heat shock transformation. One liter of LB medium supplemented with 50 μg $\mu\mathrm{L}^{-1}$ kanamycin was inoculated with a preculture, grown at 37°C. At $\mathrm{OD_{600}}$ of 0.6–0.8, expression was induced by addition of 1 mM isopropyl β-D-1-thiogalactopyranoside and grown for 4 h. Cells were harvested by centrifugation at 5000 × g for 15 min at 4°C and the pellet resuspended in 25 mL lysis buffer (50 mM Tris-HCl, 150 mM NaCl, 10 mM imidazole, 10 mM β-mercaptoethanol [βME], 1 mM phenylmethylsulfonyl fluoride, 1 tablet ethylenediaminetetraacetic acid-free protease inhibitor [Roche Diagnostics, Copenhagen, Denmark] [pH 8]), and lysed using a French pressure cell disrupter (Constant Systems MC Cell Disrupter, Daventry, United Kingdom) at 20 kPsi. Lysate was cleared by centrifugation at 20,000 × g at 4°C for 20 min. The supernatant was incubated with 2 mL Ni-NTA resin (GE Healthcare, Brøndby, Denmark),

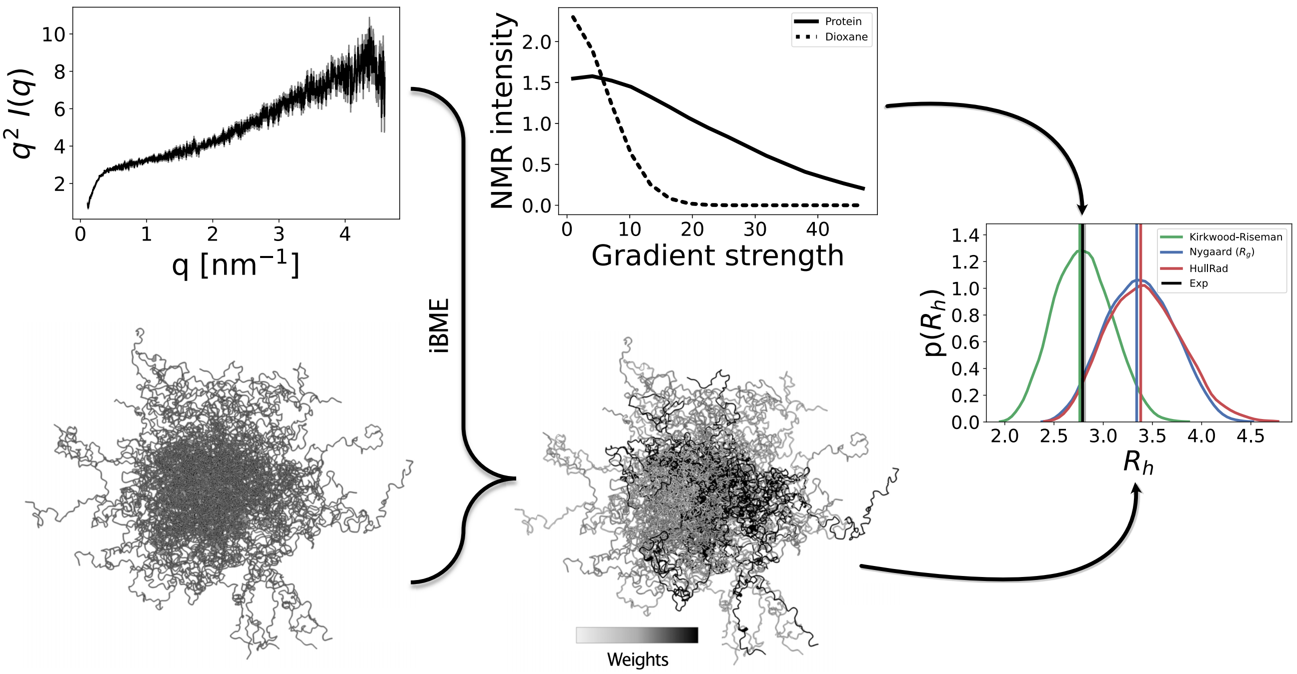

FIGURE 1 Overview of the approach. For each of the 11 proteins, we computationally generate a conformational ensemble and optimize the weights of each conformer to get a reweighted ensemble consistent with SAXS data. Then we compute the R_h from the reweighted ensemble with different forward models and compare the values to the R_h determined experimentally by PFG NMR experiments. To see this figure in color, go online.

equilibrated with buffer A (50 mM Tris-HCl, 150 mM NaCl, 10 mM imidazole [pH 8]) for 1 h at room temperature. The column was washed with 50 mL buffer B (50 mM Tris-HCl, 1 M NaCl, 10 mM imidazole, 10 mM βME [pH 8]), and His$_6$-SUMO-GHR-ICD was eluted with 15 mL buffer C (50 mM Tris-HCl, 250 mM imidazole, 10 mM βME [pH 8]). The elution was kept for further purification, while the flowthrough was reincubated with 2 mL Ni-NTA resin. His$_6$-SUMO-GHR-ICD was eluted. The His$_6$ -SUMO tag was off cleaved by adding 200 μg of the ULP1 protease and dialyzed overnight at 4°C against 3 L cleavage buffer (50 mM Tris-HCl, 150 mM NaCl, 10 mM βME [pH 8]). After cleavage, the His-SUMO tag was separated from GHR-ICD by incubating the sample with 2 mL Ni-NTA resin for 1 h. The flowthrough was collected and used for further purification by reversed-phase chromatography using a Resource RPC column (GE Healthcare), equilibrated in Milli-Q water with 0.08% trifluoroacetic acid (v/v),and eluted with a linear gradient from 0 to 100% of 70% acetonitrile (v/v), 0.1% trifluoroacetic acid (v/v). NMR experiments were recorded in 20 mM Na_2HPO_4/NaH_2PO_4, 150 mM NaCl, 10 mM βME (pH 7.3), 10% (v/v) D_2O, 0.25 mM DSS, 0.05% (v/v) dioxane, 0.02% NaN_3 and SAXS data in 20 mM Na_2HPO_4/NaH_2PO_4, 300 mM NaCl, 10× excess of DTT, 2% (v/v) glycerol as described, with protein concentration in the range from 1 to 6 mg mL^{-1} (38).

The human sodium-proton exchanger 6 intracellular distal domain

A modified pET-24b vector with an N-terminal His$_6$-SUMO tag was inserted with the sodium-proton exchanger 6 intracellular distal domain (NHE6cmdd) sequence (residues 554–669). BL21(DE3) *E. coli* cells were heat shock transformed with the finalized plasmid and incubated in LB medium for 45 min at 37°C, plated on agar containing 50 mg L^{-1} kanamycin, and incubated overnight at 37°C. Preheated LB medium (10 mL) with 50 mg L^{-1} kanamycin was inoculated with one colony and incubated overnight at 37°C and 200 rpm. The next day, the culture was added to 1 L of LB medium containing 50 mg L^{-1} kanamycin and incubated at 37°C and 200 rpm. For His$_6$ -SUMO-NHE6cmdd expression, β-D-1-thiogalactopyranoside was added to a final concentration of 0.5 mM at an OD$_{600}$ of 0.6–0.8. Cells were harvested after 4 h by centrifuging at 5000 × *g* for 20 min at 4°C. The cell pellet resuspended in 20 mL of Tris-HCl buffer (50 mM Tris-HCl [pH 8.0], 150 mM NaCl, 10 mM imidazole, 1 mM DTT) and cells lysed by 1 cycle of French Press at 25 kPsi (Constant Systems MC Cell Disrupter). The lysate was centrifuged at 4°C and 20,000 × *g* for 30 min and the supernatant applied to a gravity flow column with 4 mL preequilibrated Ni-NTA Sepharose resin (GE Healthcare). The column was washed with 50 mL of high-salt Tris-HCl buffer (50 mM Tris-HCl [pH 8.0], 1 M NaCl, 10 mM imidazole, 1 mM DTT) and bound protein eluted with 15 mL of high-salt Tris-HCl buffer (50 mM Tris-HCl [pH 8.0], 150 mM NaCl, 250 mM imidazole, 1 mM DTT). An aliquot of 100 μg His-ULP-1 was added, and the sample transferred to a presoaked dialysis bag with a 3.5 kDa cutoff, and dialyzed against 2 L of a low-salt Tris-HCl buffer (50 mM Tris-HCl [pH 8.0], 150 mM NaCl, 10 mM imidazole, 1 mM DTT) overnight at 4°C while stirred. The sample was applied to 4 mL Ni-NTA Sepharose resin and the flowthrough containing NHE6cmdd was collected. The NHE6cmdd was concentrated to <2 mL by centrifuging with 3 kDa cutoff spin filters (Amicon Ultra), before being loaded onto a 3 mL RPC column (Cytiva prepacked 3 mL SOURCE 15RPC) mounted on an Äkta Purifier system, preequilibrated with 50 mM NH_4HCO_3 (pH 7.8). A 0–100% linear gradient (20 column volumes) of 50 mM NH_4HCO_3 (pH 7.8), 70% (v/v) acetonitrile was used to elute the bound NHE6cmdd. Identity and purity of NHE6cmdd was confirmed by SDS-PAGE analysis and mass spectrometry. NMR data were recorded in 20 mM Tris-HCl (pH 7.4), 150 mM NaCl, 5 mM DTT, 0.1% (v/v) dioxane, 25 μM DSS, 10% (v/v) D_2O, 15°C, and SAXS data in 20 mM Tris-HCl (pH 7.4), 150 mM NaCl, 2% (v/v) glycerol, 5 mM DTT, 15°C. PFG NMR experiments were recorded with 150 μM (1.9 mg mL^{-1}) NHE6cmdd, while the SAXS experiments were recorded with 0.7, 1.2, and 1.6 mg mL^{-1} NHE6cmdd.

Prothymosin-α

Prothymosin-α (ProTα) was produced and purified as described in (39), where also PFG NMR diffusion experiments are reported. SAXS intensities were recorded in 1× TBSK (10 mM Tris, 0.1 mM ethylenediaminetetraacetic acid, 155 mM KCl [pH 7.4]) and 2% glycerol at 15°C. Protein concentrations for SAXS samples were 0.27, 0.74, and 1.6 mg mL^{-1}. Due to the absence of aromatic residues in the sequence of ProTα, the absorbance had to be measured at 214 nm. This was not possible in the TBSK buffer, where salts absorb most of the light at 214 nm. The concentration in the most diluted sample is calculated from the elution peak from a chromatogram in a reversed-phase run, where the protein is in water and acetonitrile and there is no background absorbance. The eluted fractions are then lyophilized and resuspended in 1× TBSK, and then concentrated. Due to the unfeasibility of measuring concentration from absorbance at 214 nm in TBSK buffer, we recovered the concentrations of the samples at 0.74 and 1.6 mg mL^{-1} from the intensity of the forward scattering of their SAXS profiles, using as reference the most diluted sample that had a known concentration. The intensity of the forward scattering was obtained by Guinier fit using the ATSAS package (40).

α-Synuclein

α-Synuclein (αSyn) was produced and purified as described in (41). NMR experiments were recorded in PBS buffer (20 mM Na_2HPO_4/NaH_2PO_4, 150 mM NaCl [pH 7.4]), 2% glycerol 10% D_2O, 0.25 mM DSS, 0.02% dioxane, 0.02% NaN_3, and recorded at 20°C. SAXS data are from (42).

ANAC046$_{172–338}$

The disordered region (residues 172–338) of the *Arabidopsis* NAC (no apical meristem, ATAF1/2, and cup-shaped cotyledon [CUC2]) transcription factor, ANAC046, was produced and purified as described in (41). NMR experiments were recorded in PBS buffer (20 mM Na_2HPO_4/NaH_2PO_4, 100 mM NaCl, 1 mM DTT, 0.02% dioxane, 0.02% NaN_3 [pH 7.0]) at 25°C, and SAXS data in 20 mM Na_2HPO_4/NaH_2PO_4, 100 mM NaCl, 5 mM DTT (pH 7.0), same temperature. Protein concentrations for SAXS samples were 1, 3, and 5 mg mL^{-1}.

Dss1

Deleted in split hand/split foot 1 protein (Dss1) from *S. pombe* (43) was produced and purified as in (41) in the presence of 5 mM βME. NMR experiments were recorded in Tris buffer (20 mM Tris, 150 mM NaCl, 5 mM DTT, 2% glycerol, 10% D_2O, 0.25 mM DSS, 0.02% dioxane, 0.02% NaN3) and recorded at 15°C and SAXS data recorded in 20 mM Tris, 150 mM NaCl 2% glycerol 5 mM DTT (pH 7.4), 15°C, protein concentrations were 1, 1.5, and 3 mg mL^{-1}.

hnRNPA1-LCD

The low complexity domain from hnRNPA1 (hereafter called A1) was produced and purified as described previously (10,44). NMR experiments were recorded in 20 mM HEPES, 150 mM NaCl, 0.02% dioxane, 10% D_2O at 25°C. Protein concentration was 70 μM.

Diffusion ordered NMR spectroscopy

Translational diffusion constants for each protein (50–150 μM) and the internal reference were determined by fitting peak intensity decays within 0.5 and 2.5 ppm (where protons belonging to methyl and methylene groups resonate) (26) from diffusion ordered spectroscopy experiments (45), using the Stejskal-Tanner equation (20). We used 1,4-dioxane (0.02–0.10% [v/v]) as internal reference, with an R_h value of 2.12 Å (24). Spectra (16 scans for αSyn, 64 scans for NHE6cmdd and A1, and 32 scans for the other proteins) were recorded on a Bruker 600 MHz equipped with a cryoprobe and Z-field gradient, and were obtained over gradient strengths from 2 to 98% (γ = 26,752 rad s^{-1} $Gauss^{-1}$) with a diffusion time (Δ) of 200 ms (299.9 ms for GHR-ICD and 50 ms for A1) and gradient length (δ) of 3 ms, except for NHE6cmdd where this was 2 ms, and A1 where it was 6 ms. Diffusion constants were fitted in Dynamics Center v2.5.6 (Bruker, Fällanden, Switzerland) and GraphPad Prism v9.2.0. Diffusion constants were used to estimate the R_h for each protein (46), with error propagation using the diffusion coefficients of both the protein and dioxane.

SAXS

Samples for SAXS were prepared in the same buffers as for the NMR experiment, leaving out D_2O and DSS, and in some cases adding 2% (v/v) glycerol, and using a range of protein concentrations. The samples were either dialyzed extensively into the buffer, or a final size-exclusion step into the buffer was done, and collecting either the dialysate or the SEC buffer for SAXS analyses. Buffer samples were run before and after the protein samples. SAXS data on Dss1, ProTα, and NHE6cmdd were collected at the DIAMOND beamline B21 (London, UK), using a monochromatic (λ = 0.9524 Å) beam operating with a flux of 2×10^4 photons/s. The detector was an EigerX 4M (Dectris, Baden, Switzerland). The detector to sample distance was set to 3.7 m. Samples were placed in a Ø = 1.5 mm capillary at 288 K during data acquisition. SAXS data on ANAC046 and GHR-ICD were collected at the EMBL bioSAXS-P12 beam line (λ = 0.124 nm, 10 keV) at the PETRA III storage ring (Hamburg, Germany) (47). Scattering profiles were recorded on a Pilatus 2M detector (Dectris) (47) following standard procedures and at 298 K. The resulting scattering curves were analyzed as an average of consecutive frames recorded for each sample (detected degenerate frames were removed). The averaged scattering curves of the buffer were subtracted from the averaged scattering curve of the samples. Finally, we scaled the buffer-subtracted curves to absolute scale with DATABSOLUTE, part of the ATSAS package (48), using water and empty capillary measurements, performed at the same temperature as the experiments.

Conformational ensembles

We generated conformational ensembles with two distinct methods, specifically Flexible-Meccano (hereafter FM) (49) and Langevin simulations with the CALVADOS (coarse-graining approach to liquid-liquid phase separation via an automated data-driven optimization scheme) M1 parameters for a C_α-based coarse-grained model (50).

FM generates conformations for the backbone atoms of IDP sampling from backbone dihedral potentials derived by disordered regions of entries in the PDB. We varied the number of conformers produced with the length of the proteins, to reflect the higher complexity of the ensembles for longer chains (Table S1).

Langevin simulations with CALVADOS were run for 1 μs with a 10 fs time step using OpenMM v7.5.1 (51). Trajectories were subsampled taking a frame every 50 ps according to the shortest observed lag-time resulting in a close-to-zero autocorrelation function of the R_g (Fig. S1), resulting in 20,000 frames per simulation. Temperatures of the experimental measurements were reproduced in the simulations, as well as the ionic strength, by means of the Debye-Hückel potential used in CALVADOS to describe electrostatic interactions.

For both FM and CALVADOS simulations we generated all-atom representations for the ensembles before SAXS calculations using PULCHRA (52) with default settings; these structures were also used to calculate the hydrodynamic radius when using centers of mass to represent the positions of the amino acids.

SAXS calculations

We calculated SAXS intensities using Pepsi-SAXS (53) as described recently (54). The scale factor and constant background were fitted as global parameters for all the conformers in an ensemble (see below), while the contrast of the hydration layer and the effective atomic radius were fixed (respectively, 3.34 e/nm^3 and 1.025 × r_m, where r_m is the average atomic radius of the protein).

Ensemble reweighting

We used the Bayesian/maximum entropy (BME) software (3) to improve the agreement of the conformational ensembles with the SAXS experiments by minimizing the functional $\mathcal{L}$ (4,6):

$$\mathcal{L}(\omega_1 \cdots \omega_n) = \frac{m}{2}\chi_r^2(\omega_1 \cdots \omega_n) - \theta S_{\mathrm{rel}}(\omega_1 \cdots \omega_n). \quad (4)$$

Here, m is the number of experimental data points, $(\omega_1 \cdots \omega_n)$ are the weights associated with each conformer of an ensemble, χ_r^2 measures the agreement between the calculated and experimental data, S_{rel} measures how much the optimized weights (i.e., the posterior distribution) diverge from the initial weights (i.e., the prior distribution), and θ is a parameter that sets the balance between minimizing χ_r^2 and maximizing S_{rel}. The value $\varphi_{\mathrm{eff}} = \exp(S_{\mathrm{rel}})$ indicates the fraction of the frames that effectively contributes to the averages calculated with the optimized weights. A low φ_{eff} means a considerable deviation from the initial ensemble and it can indicate overfitting and artifacts in the reweighted ensemble (3). By scanning different values for θ and plotting χ_r^2 versus φ_{eff}, it is possible to choose the optimal value for θ as the one located at the "elbow" of the curve, where the χ_r^2 reaches a plateau with the least amount of deviation from the initial weights.

For SAXS data, the iterative extension of BME (iBME) (54) enabled us to fit a scale factor (s) and constant background (cst) of the calculated SAXS profile by iterating least-squares fitting of experimental and calculated SAXS profiles and BME reweighting until convergence of the χ_r^2. In this approach, the χ_r^2 in the $\mathcal{L}$ functional is:

$$\chi_r^2(\omega_1 \cdots \omega_n, s, cst) = \frac{1}{m}\sum_q^m \frac{\left[\left(s\sum_j^n \omega_j I_q\left(x_j\right) + cst\right) - I_q^{\mathrm{EXP}}\right]^2}{\sigma_{\mathrm{BIFT},q}^2}, \quad (5)$$

where $I_q(x_j)$ is the calculated SAXS intensity at scattering angle q for the conformer x_j, I_q^{EXP} is the experimental SAXS intensity at scattering angle q, and $\sigma_{\mathrm{BIFT},q}$ is the error of the experimental intensity at scattering angle q normalized as described by Larsen and Pedersen (55). The Bayesian indirect Fourier transformation (BIFT) was used to compute the pair distance distribution function $p(r)$ from a model SAXS profile by minimizing the $\chi_{r,\mathrm{BIFT}}^2$ calculated against the experimental SAXS profile and maximizing a prior on the smoothness of the $p(r)$. Then the experimental errors were corrected according to $\sigma_{\mathrm{BIFT},q} = \sigma_q\sqrt{\chi_{r,\mathrm{BIFT}}^2}$. This procedure enabled a more direct comparison of χ_r^2 values from different systems.

Hydrodynamic radius calculation

We employed four distinct approaches to compute the R_{h} from a specific protein conformation:

1. The equation described by Nygaard et al. (30), who derived a sequence-length-(N)-dependent relationship between the radius of gyration of the C_α atoms $(R_{\mathrm{g},C_\alpha})$ and the R_{h}:

$$\frac{R_{\mathrm{g},C_\alpha}}{R_{\mathrm{h}}^{\mathrm{Nyg}}} = \frac{\alpha_1\left(R_{\mathrm{g},C_\alpha} - \alpha_2 N^{0.33}\right)}{N^{0.60} - N^{0.33}} + \alpha_3, \quad (6)$$

where the fitting parameters are $\alpha_1 = 0.216$ Å^{-1}, $\alpha_2 = 4.06$ Å, and $\alpha_3 = 0.821$. This expression for $R_{\mathrm{h}}^{\mathrm{Nyg}}$ was obtained by fitting the R_{h} calculated with HYDROPRO (27) as a means to have a more computationally efficient forward model, and which interpolates between the behavior for compact and expanded states.

2. The HullRad algorithm (29) that uses the convex hull method to predict hydrodynamic properties of proteins. The R_{h} computed with HullRad will hereafter be referred as $R_{\mathrm{h}}^{\mathrm{HR}}$.
3. The Kirkwood-Riseman equation (28,56,57): $R_{\mathrm{h}}^{\mathrm{KR}} = 1/\langle r_{ij}^{-1}\rangle_{i\neq j}$, where r_{ij} is the distance between the C_α atoms i and j. An alternative approach is to calculate $R_{\mathrm{h}}^{\mathrm{KR}}$ employing the center of mass of each residue instead of the C_α. The two strategies can lead to minor differences in the resulting R_{h} (see results and supporting material).
4. The linear fit proposed by Nygaard et al. (30) to approximate the R_{h} of HYDROPRO from $R_{\mathrm{h}}^{\mathrm{KR}}$: $R_{\mathrm{h}}^{\mathrm{Nyg-KR}} = 1.186R_{\mathrm{h}}^{\mathrm{KR}} + 1.03$ Å.

Once R_{h} was calculated for all conformers of an ensemble of size n, the average R_{h} ($\langle R_{\mathrm{h}}\rangle$) was calculated as $\langle R_{\mathrm{h}}\rangle = 1/n^{-1}\sum_i^n\left(1/R_{\mathrm{h},i}\right)$ (31,34). The transformation of the R_{h} of each conformer of the ensemble before averaging was done to reflect that the intensities measured by PFG NMR are proportional to $\exp(-R_{\mathrm{h}}^{-1})$ (Eqs. 1 and 3). The exponential transformation can be omitted because it does not change the calculated average (31). For simplicity, we hereafter refer to $\langle R_{\mathrm{h}}\rangle$ as R_{h}.

We calculated the χ^2 to compare R_{h} calculated with the models above with experiments across the 11 proteins. The reported errors of the experimentally determined values of R_{h} varied considerably across the different experiments. To avoid putting too much weight on a few experiments with the smallest estimated errors, we instead used the average relative error of R_{h} (~2%) in the calculation of χ^2.

Scripts and data used in this study are available at https://github.com/KULL-Centre/papers/tree/main/2022/rh-fwd-model-pesce-et-al.

RESULTS AND DISCUSSION

Proteins and experimental measurements

We collected a data set consisting of 11 IDPs of different lengths (spanning from 24 to 441 residues) and sequence features (net charge per residue, number of prolines, overall charge, etc.; see Table S1) with both SAXS and PFG NMR diffusion measurements. We measured PFG NMR diffusion and SAXS data for those proteins (see materials and methods) where data were not already available in literature (Table 1). We stress that, although different approaches exist to extract the R_{g} from a SAXS profile (63–65) (Table S2), we do not build our ensembles using the SAXS-derived R_{g} values; rather we use the SAXS intensities themselves. We note, however, that the average ratio of the R_{g} extracted from SAXS data and the R_{h} from PFG NMR for the 11 proteins is 1.2 (Table S2), in line with expectations from disordered Gaussian chains (66).

To minimize discrepancies related to the dimensions of IDPs being influenced by experimental conditions (for

TABLE 1 Data sets used, consisting of 11 IDPs with SAXS and PFG NMR measurements

Name	Length (residues)	R_{h} (nm)	SAXS
Hst5	24	1.28 ± 0.02 (58)	(58)
RS	24	1.19 ± 0.01 (59)	(15)
Dss1	71	1.70 ± 0.06	this study
Sic1	90	2.15 ± 0.1 (60)	(36)
ProTα	111	2.89 ± 0.08 (39)	this study
NHE6cmdd	116	2.67 ± 0.02	this study
A1	137	2.29 ± 0.06	(44)
αSyn	140	2.79 ± 0.03	(42)
ANAC046	167	3.04 ± 0.01	this study
GHR-ICD	351	5.08 ± 0.02	(38)
Tau	441	5.40 ± 0.2 (61)	(62)

example, temperature and ionic strength of the buffer), we aimed at having SAXS and PFG NMR diffusion measured in the same buffer and conditions. There are few exceptions, where we note some differences in the conditions at which PFG NMR and SAXS measurements were performed (Table S3). Buffers used for SAXS often contain glycerol to limit the radiation damage, which might in principle cause some discrepancies as glycerol was not present in some of the PFG NMR experiments. Previous work, however, suggests that small amounts (2%) of glycerol do not affect the compaction of the IDPs (67,68). Similarly, PFG NMR experiments of R_h use dioxane as an internal reference, and potential interactions between dioxane and the IDPs could also cause discrepancies between NMR and SAXS experiments. Previous work shows consistency across different internal reference compounds (69) and below we find good consistency between SAXS and PFG NMR data; together these results suggest that protein-dioxane interactions do not affect the experimental diffusion measurements. To examine this further, we recorded ^{1}H-^{15}N HSQC spectra of the ^{15}N-labeled ANAC046 alone and in presence of different concentration of dioxane. The resulting data show no changes in position or intensity of the peaks in the spectra (Fig. S2). Together, these observations support the notion that minor discrepancies between experimental conditions in SAXS and NMR experiments do not cause systematic differences.

We measured SAXS data for Dss1, ProTα, NHE6cmdd, and ANAC046 at different protein concentrations (Fig. S3). We then inspected the resulting intensities as a function of the scattering angle to check for signs of aggregation or interparticle repulsion in the small-angle region (19). In absence of these effects, we selected the SAXS profiles showing the lowest amount of experimental noise. Therefore, in further analyses we used SAXS data collected at 3 mg/mL for Dss1, 1.6 mg/mL for ProTα, 1.6 mg/mL for NHE6cmdd, and 5 mg/mL for ANAC046 (Fig. 2).

Agreement of the ensembles with SAXS data

As described above, our approach involved first generating conformational ensembles that were in agreement with SAXS data and then assessing four different forward models by calculating R_h from these ensembles. We based this procedure on the assumption that conformational ensembles that are generated by accurate physical models and that are in agreement with SAXS data will also be in good agreement with measurements of R_h. While there can be differences in the conformational properties probed by SAXS and PFG NMR (31,34), we expect that such differences are generally small and will be "averaged out" when examining a diverse set of proteins.

We generated ensembles for the 11 proteins using both FM (49) and Langevin simulations with the CALVADOS coarse-grained model (50), both of which are known to generate ensembles in good agreement with SAXS data (8,50,70–72). Forward models for SAXS data are relatively consistent with each other and many are based on the same physical principle and spherical harmonics approximation (53,73). Moreover, issues related to fitting free-parameters describing hydration layer and excluded volume

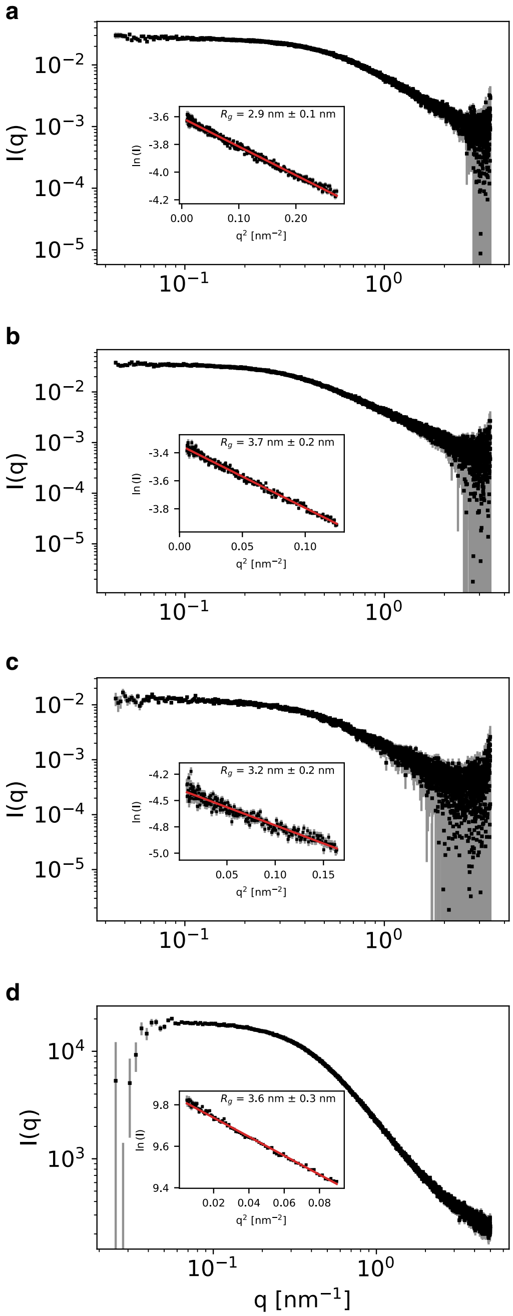

FIGURE 2 Experimental SAXS profiles for (*a*) Dss1, (*b*) ProTα, (*c*) NHE6cmdd, and (*d*) ANAC046. Experimental intensities are shown as black squares with gray error bars (representing the experimentally determined errors rescaled as described in the materials and methods). The inserts show the linear fit (*red line*) in the Guinier region (identified with the *autorg* tool of the ATSAS package (40)) for each profile, and the resulting R_g value. To see this figure in color, go online.

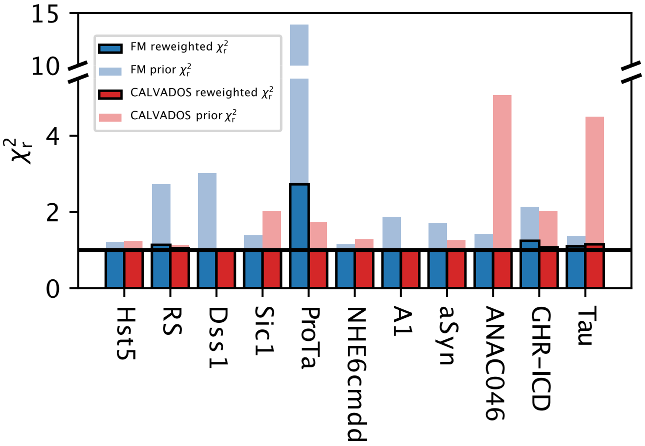

FIGURE 3 Results of reweighting the conformational ensembles generated with FM and CALVADOS against SAXS data. The χ^2_r of the prior ensembles (before reweighting) are shown as partially transparent bars, while the χ^2_r of the reweighted ensembles are shown as solid bars (*blue* for FM and *red* for CALVADOS). The horizontal black line delineates $\chi^2_r = 1$, that, given the use of scaled errors for experimental SAXS intensities (see materials and methods), denotes reweighted ensembles in good agreement with the experimental data and devoid of overfitting. To see this figure in color, go online.

in implicit-solvent-based SAXS forward models have recently been addressed also in the context of IDPs (54,74). We therefore calculated SAXS data from the conformational ensembles and compared these with experiments (Fig. 3, partially transparent bars). For both FM and CALVADOS we found good agreements ($\chi^2_r \approx 1$) in many cases. The largest outlier was the highly charged protein ProTα, where the FM ensemble did not provide as good a fit to the SAXS data (Figs. 3 and S4). Presumably the difference in agreement for ProTα arises because CALVADOS explicitly takes the effect of the charges into account.

We improved the agreement with the SAXS data further by using BME reweighting of the ensembles against the SAXS data (Fig. 3, *solid bars*). For all but the ProTα FM ensemble this led to excellent agreement with experiments with only minor levels of reweighting (Table S4). For the FM ensemble of ProTα we were also able to obtain a reasonably good fit, although at the cost of stronger reweighting and lower φ_{eff} (Table S4).

We analyzed the effect of the different priors (FM versus CALVADOS) and reweighting by examining the distribution of the R_g (Fig. 4). In most cases, we found very similar distributions both before and after reweighting and with the two different methods to generate the conformational ensembles. For the few ensembles with intermediate values of χ^2_r (in the range 2–5) before reweighting, we also observe minor adjustments in the R_g distributions due to reweighting. In general, the fit to the small-angle region of the SAXS profile was already good in most cases for the CALVADOS prior, indicating that this prior is highly efficient in reproducing the average chain dimensions (Fig. S5). We note that this is likely explained—at least in part—by the fact that CALVADOS was parameterized to reproduce R_g for IDPs. The deviations responsible for the higher χ^2_r of ANAC046 and Tau were decreased by reweighting the CALVADOS ensembles (Fig. S5).

To summarize, with the only exception of ProTα, the two priors provided similar levels of agreement with SAXS data (Fig. 3) and similar distributions of R_g (Fig. 4). Given that the CALVADOS prior provided a good estimate of the average chain dimensions even without reweighting and considering the specific case of ProTα, we below focus our further analyses on the ensembles generated by CALVADOS.

Comparison of forward models for the R_h

We tested four previously described forward models to compute R_h from atomic coordinates. These models are based on different principles, different ways of treating the hydration of proteins, and were developed for different types of molecules. We applied these models to the SAXS-reweighted CALVADOS ensembles, and compared the resulting ensemble-averaged values for R_h to the R_h from PFG

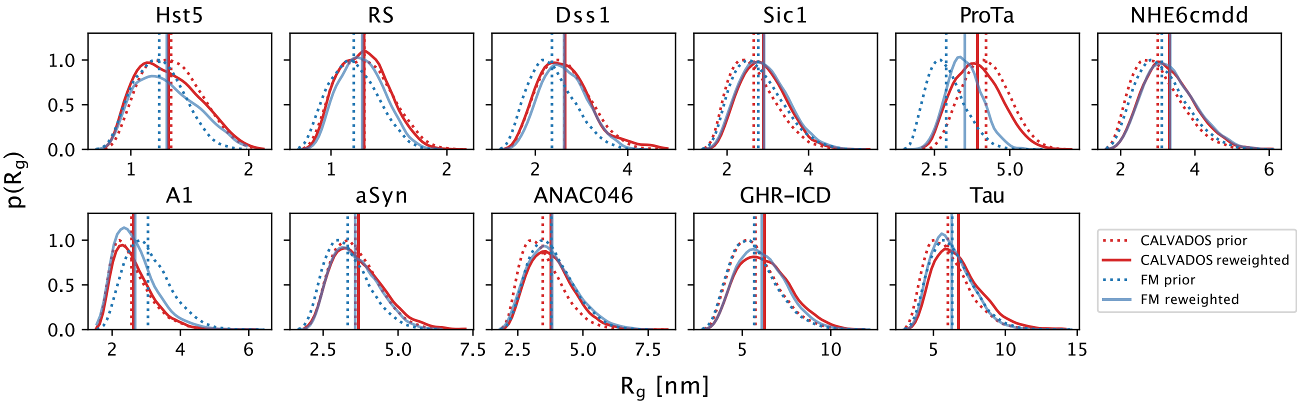

FIGURE 4 Probability distributions of the R_g calculated from the ensembles generated by CALVADOS (in *red*) and FM (in *blue*), both before (*dotted lines*) and after (*solid lines*) reweighting the ensembles against SAXS data. To see this figure in color, go online.

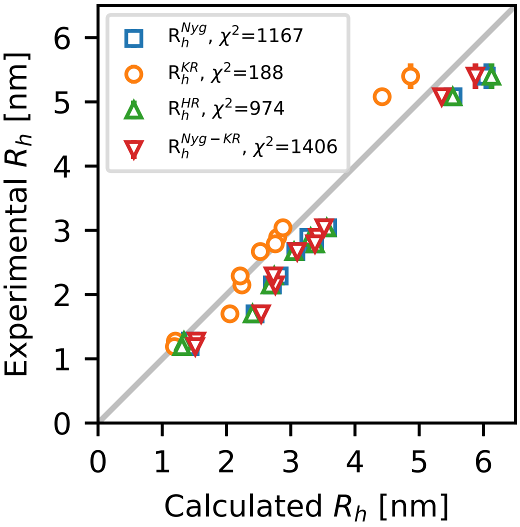

FIGURE 5 Ensemble-averaged R_h values calculated from the SAXS-reweighted CALVADOS ensembles, compared with the R_h determined by PFG NMR diffusion (error bars represent the standard error from fitting the NMR data). We tested four approaches to calculate the R_h from atomic coordinates: the R_g-dependent Nygaard equation (R_h^{Nyg}, in *blue*), the Kirkwood-Riseman equation (R_h^{KR}, in *orange*), HullRad (R_h^{HR}, in *green*), and the Nygaard correction to the Kirkwood-Riseman equation (R_h^{Nyg-KR}, in *red*). To see this figure in color, go online.

NMR diffusion experiments (Fig. 5). We find that all four models led to a high correlation between the calculated and experimental values of R_h. To quantify the accuracy of the models, we calculated the χ^2 across the 11 proteins, and find the Kirkwood-Riseman equation provided the best agreement with experiments (Fig. 5), with a χ^2 of 188. Indeed, for all but the two longest proteins and Dss1, the Kirkwood-Riseman equation resulted in very good agreement with experiments.

We generally use distances between the C_α atoms (or beads) when applying the Kirkwood-Riseman equation to calculate R_h. We explored the effect of instead using the center of mass to represent the position of each amino acid. Overall, we find very similar results from the two different approaches for all but the shortest IDPs (Fig. S6). The two approaches also give very similar agreement with experiments (χ^2 = 188 for C_α atoms and 232 for centers of mass); additional work is needed to examine which method is more accurate for shorter proteins.

The other three forward models (the two equations described by Nygaard et al. and HullRad) gave χ^2 values that are five to eight times greater than that for the Kirkwood-Riseman equation. For all three models, this higher χ^2 is due to an apparently overestimated R_h (Figs. 5 and S7), and we see also that the three models are very similar to one another. The exception is for the shortest proteins, Histatin 5 (Hst5) and the RS repeat peptide (RS). This may be because the length of Hst5 and RS is at the limit of the chain length range the Nygaard equations were parametrized for. Despite the overall better agreement when using the Kirkwood-Riseman equation, the agreement with the R_h from PFG NMR experiment is not uniform across the data set and shows a sequence-length-dependent trend (Figs. 5 and S7). When looking at the two longest proteins, GHR-ICD (351 residues) and Tau (441 residues) (Figs. 5 and S7), the Kirkwood-Riseman equation apparently underestimates R_h, whereas the other three models give values closer to experiments.

The general picture for the eight shortest proteins (24–167 residues) is thus that the Kirkwood-Riseman equation provides an accurate model for R_h, and that the other three models provide relatively similar values that are generally greater than the experimental values. For most proteins, there is thus good agreement between the SAXS-refined ensembles and the R_h value calculated from the ensembles using the Kirkwood-Riseman equation without the need for any further refinement of the ensembles. In contrast, for the other three methods, the experimental R_h values lie in the tail of the R_h distributions (Fig. S7). Thus, while it would be possible to construct ensembles that simultaneously agree with both the SAXS and PFG NMR data (31,34), this would require a greater level of reweighting.

We suspect that the Nygaard equations, which are derived from HYDROPRO, and HullRad may be less precise for disordered proteins because the models themselves are derived to predict the R_h of globular, folded proteins. The Kirkwood-Riseman equation instead was developed in the context of theoretical studies on the hydrodynamic properties of disordered polymer chains. The behavior of these chains is more similar to that of IDPs compared with folded proteins, as they exist in an extreme disordered state governed only by self-avoidance of the component particles. This observation is supported by calculating the ratio R_g/R_h from the experiments. For globular proteins, this ratio is expected to be around 0.78 and between 1.2 and 1.5 for disordered chains (30,34,66,75), and indeed we find the average to be 1.2 with some variation across proteins (Table S3).

In the analyses above, we compared the ensembles with the estimated values of R_h; however, this value is derived from fitting the intensity profiles in the PFG NMR experiments. To examine whether a more direct comparison would give a different picture, we used the predicted values of R_h to derive the diffusion profiles using Eqs. 1 and 3. We found that comparing the calculated and experimental diffusion profiles (for the PFG NMR measurements reported in this study) gave a similar picture as when comparing the R_h (Fig. S8).

We also tested the forward models on the ensembles generated by FM. Since SAXS-refined FM and CALVADOS ensembles are similar in terms of the level of compaction (except ProTα), the results were similar. In particular, we saw an overall better agreement with experiments using the Kirkwood-Riseman equation (Fig. S9) and a sequence-length-dependent discrepancy in this agreement. We also note that Dss1 appears to be an outlier, since both ensembles seem to be more expanded than what the PFG NMR diffusion experiment detects for all forward models for the R_h used.

To find possible reasons for this apparent sequence length dependency, we looked at different conformational properties of the ensembles produced with CALVADOS and their relation to sequence length. We computed the sequence-length-normalized asphericity (the degree to which a molecule deviates from a fully spherical shape) and the relative shape anisotropy from the ensembles produced with CALVADOS to highlight potential differences in the shapes adopted by short and long chains. Nevertheless, we did not find properties for which the two longest proteins stood out (Fig. S10).

Expanding the data set

The analyses above were made possible by collecting a set of proteins for which both SAXS and PFG NMR data had been measured on the same protein and under comparable conditions. While the set of 11 proteins covers a wide range of lengths and sequence properties (Table S1), there are two areas that are not covered well. First, there are no proteins of length between 167 and 351 residues (Table S1). Second, most of the proteins are relatively expanded IDPs, with 9 of the 11 proteins having SAXS-derived scaling exponents $\nu \geq 0.55$ (Table S2). To complement our analysis described above we therefore collected data from the literature for an additional 11 proteins for which R_h had been measured using PFG NMR, although not in all cases measured using internal referencing by dioxane (Table S5). We used CALVADOS to generate ensembles for these 11 proteins and calculated R_h using the 4 different models (Fig. S11 *a*). As expected from the fact that the ensembles were not refined using SAXS data and that the data are more heterogeneous, the agreement is more noisy. We find that, within this set of proteins, the four different approaches perform comparably well, with χ^2 values in the range of 606 (R_h^{HR}) to 669 (R_h^{Nyg-KR}) (compared with the span of 188–1406 for the proteins with both SAXS and PFG NMR data).

Comparing the distribution of R_h calculated with the different methods for these 11 proteins with experiments shows that the largest discrepancies between experiments and calculations using the Kirkwood-Riseman model are for A2, FUS, and SBD (Fig. S12). Both A2 and FUS are known to form relatively compact ensembles and so we examined whether there was a relationship between the compaction of the protein—evaluated by the scaling exponent calculated from the conformational ensembles—and the accuracy of the values calculated using the Kirkwood-Riseman model (Fig. S11 *b*). Overall we find a correlation between compaction (ν) and the error in the calculated values of R_h. We note, however, that we obtain accurate results for the two compact disordered proteins Ddx4 and A1, and note that the experimental measurements of A2 and FUS did not use internal referencing with dioxane. Finally, we analyzed simulations of 200-residue-long homopolymeric peptides to examine whether the calculations of R_h using the Kirkwood-Riseman model capture the expected relationship between R_g/R_h and compaction (30,34,66,75). Indeed, we find a high correlation between the calculated scaling exponent, ν, and the R_g/R_h ratio, so that the most compact peptides have a ratio < 1 and the most expanded peptides have ratios approaching 1.4 (Fig. S13).

CONCLUSIONS

Reliable forward models to compare conformational ensembles and biophysical measurements of IDPs are important both in integrative modeling and for benchmarking and optimizing molecular mechanics force fields. Here, we have explored the accuracy of forward models to calculate the R_h from structural ensembles of IDPs. To do so, we first constructed conformational ensembles for 11 IDPs, ranging in length from 24 to 441 residues and diverse in sequence composition. We then determined and optimized the agreement of these ensembles with SAXS data to reproduce the average chain dimensions in solution encoded in SAXS data. Finally, we used four different models to calculate R_h from the refined ensembles and assessed their accuracy by comparison to measurements by PFG NMR measurements. Of the four models that we tested, the Kirkwood-Riseman equation gives the best overall agreement with experiments. Nevertheless, we also found that the accuracy of this model appears to drop in a sequence-length-dependent fashion, which is evident for GHR-ICD and Tau. It is not clear if the source of the sequence-length-dependent discrepancy is due to inaccuracies in the forward model or in the ensembles, or if it is a property of long IDPs, and further studies are needed to clarify this.

In addition to collecting additional data for long IDPs, one approach to get some insight might be to refine the ensembles of GHR-ICD, Tau (and other IDPs of similar length) simultaneously against the R_h and SAXS data. Indeed, previous studies have demonstrated that it is possible to combine SAXS and PFG NMR measurements to refine the distribution of conformations in a disordered ensemble (31,34,36). Examining whether such a refinement is possible and which conformations are retained might give insights into whether the problems are with the ensembles or the forward model. Nevertheless, a more detailed analysis would require measurements on several more proteins.

Another issue to consider relates to how the R_h values have been obtained. In particular, the values depend on the R_h value for dioxane (2.12 Å from Wilkins et al. (24)) that is used as a reference in the PFG NMR diffusion experiments. This value, however, comes with some uncertainty. Specifically, it is based on the assumption that for a globular protein $R_g = \sqrt{3/5}R_h$ and the use of a SAXS-derived value for R_g for a single protein (24). If the reference R_h used for dioxane is not exact, the forward models may implicitly absorb such a scale factor. Given that our calculations predict

R_h relatively accurately, our results suggest that the estimate of the R_h for dioxane is probably quite accurate. To examine to what extent our conclusions depend on the reference value for dioxane, we analyzed how the χ^2 of the R_h obtained with the different models from the CALVADOS simulation would change if different values for the R_h of dioxane had been used to obtain the R_h of the proteins from PFG NMR (Fig. S14). We find that 2.12 Å lies very close to the χ^2 minimum for the Kirkwood-Riseman equation, and that a R_h for dioxane greater than ca. 2.25 Å would be needed to change the conclusion that the Kirkwood-Riseman model provides the better fit to the data (Fig. S14). Additional measurements on folded proteins as well as measurements using other references, such as cyclodextrin (26), might help understand these issues better. Finally, it has recently been observed that the R_h of dioxane and other reference compounds may be pressure dependent (76). This suggests that the value could also be temperature dependent, and this should be studied in more detail to help interpret better PFG NMR diffusion experiments at different temperatures (58).

We also analyzed an additional set of 11 proteins with PFG NMR measurements. These proteins do not have measured SAXS data and so we instead rely on the overall accuracy of the CALVADOS model to capture the expansion of these proteins. We selected these proteins to include more proteins with intermediate lengths and to represent proteins with a wider range of properties. While the results confirmed that all models perform relatively well, the results are less clear than for the proteins with consistent SAXS and PFG NMR data. The results hint at a dependency on the accuracy of the Kirkwood-Riseman model on the level of compaction in line with the expectation that this model is expected to work best for disordered expanded polymers. A more detailed analysis, however, would ideally be based on a set of proteins that have been referenced in a consistent way and for which both SAXS and PFG NMR data have been recorded at near identical conditions.

In summary, we present an analysis of 11 proteins for which we have collected SAXS, PFG NMR, and simulation data to generate conformational ensembles. We have used these data to compare different methods to calculate the hydrodynamic radius from conformational ensembles of disordered proteins. Overall we find good agreement from all models, and that the Kirkwood-Riseman model gives the best overall agreement.

SUPPORTING MATERIAL

Supporting material can be found online at https://doi.org/10.1016/j.bpj.2022.12.013.

AUTHOR CONTRIBUTIONS

F.P., B.B.K., and K.L.-L. designed the study. E.A.N., P.S., E.E.T., F.P., and C.R.G. purified proteins for NMR and SAXS measurements and recorded and analyzed the NMR data. J.G.O. and F.P. analyzed the SAXS data. F.P. produced and analyzed the ensembles. F.P. and K.L.-L. analyzed the data and wrote the paper with input from all authors.

ACKNOWLEDGMENTS

We thank our colleagues who measured the data that have made this work possible. We acknowledge the use of computational resources from the core facility for biocomputing at the Department of Biology and support for NMR infrastructure from Villumfonden. We thank the beamline scientists Cy Jeffries at EMBL DESY P12 and Nathan Cowieson at DIAMOND B21 for technical support and data acquisition, and Jacob H. Martinsen for assistance with protein purification and Eric M. Morrow and Stine F. Pedersen for input on NHE6cmdd. We thank Amanda D. Due for preparation of the ANAC046 samples. We thank Anne Bremer and Tanja Mittag for support and assistance in purification and studies of the A1 LCD, and acknowledge access to the St. Jude Biomolecular NMR Spectroscopy Center. We thank Giulio Tesei for fruitful discussion on how to calculate the Kirkwood-Riseman equation. We thank David de Sancho for comments on our work. This research was supported by the Lundbeck Foundation BRAINSTRUC initiative (R155-2015-2666 to B.K.K. and K.L.-L.) and the Novo Nordisk Foundation Challenge grant REPIN (no. NNF18OC0033926 to B.B.K.). E.A.N. has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement no. 101023654.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Wright, P. E., and H. J. Dyson. 2015. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* 16:18–29. http://www.nature.com/articles/nrm3920.
2. Babu, M. M. 2016. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem. Soc. Trans.* 44:1185–1200. https://doi.org/10.1042/BST20160172.
3. Bottaro, S., T. Bengtsen, and K. Lindorff-Larsen. 2020. Integrating Molecular Simulation and Experimental Data: A Bayesian/Maximum Entropy Reweighting Approach. Springer US, pp. 219–240. https://doi.org/10.1007/978-1-0716-0270-6_15.
4. Orioli, S., A. H. Larsen, …, K. Lindorff-Larsen. 2020. Chapter Three - how to learn from inconsistencies: integrating molecular simulations with experimental data. *In* Computational Approaches for Understanding Dynamical Systems: Protein Folding and Assembly. B. Strodel and B. Barz, eds Academic Press, pp. 123–176, Volume 170 of Progress in Molecular Biology and Translational Science. https://www.sciencedirect.com/science/article/pii/S1877117319302121.
5. Bonomi, M., C. Camilloni, …, M. Vendruscolo. 2016. Metainference: a Bayesian inference method for heterogeneous systems. *Sci. Adv.* 2:e1501177. https://doi.org/10.1126/sciadv.1501177.
6. Hummer, G., and J. Köfinger. 2015. Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.* 143:243150. https://doi.org/10.1063/1.4937786.
7. Różycki, B., Y. C. Kim, and G. Hummer. 2011. SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Structure.* 19:109–116. http://www.sciencedirect.com/science/article/pii/S0969212610003953.
8. Bernadó, P., L. Blanchard, …, M. Blackledge. 2005. A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proc. Natl. Acad. Sci. USA.* 102:17002–17007. https://www.pnas.org/content/102/47/17002.

9. Shoemaker, B. A., J. J. Portman, and P. G. Wolynes. 2000. Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc. Natl. Acad. Sci. USA*. 97:8868–8873. https://doi.org/10.1073/pnas.160259697.

10. Martin, E. W., A. S. Holehouse, …, T. Mittag. 2020. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science*. 367:694–699. https://doi.org/10.1126/science.aaw8653.

11. Lin, Y.-H., and H. S. Chan. 2017. Phase separation and single-chain compactness of charged disordered proteins are strongly correlated. *Biophys. J.* 112:2043–2046. https://www.sciencedirect.com/science/article/pii/S000634951730437X.

12. Thomasen, F. E., F. Pesce, …, K. Lindorff-Larsen. 2022. Improving martini 3 for disordered and multidomain proteins. *J. Chem. Theor. Comput.* 18:2033–2041. https://doi.org/10.1021/acs.jctc.1c01042.

13. Henriques, J., C. Cragnell, and M. Skepö. 2015. Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment. *J. Chem. Theor. Comput.* 11:3420–3431. https://doi.org/10.1021/ct501178z.

14. Palazzesi, F., M. K. Prakash, …, A. Barducci. 2015. Accuracy of current all-atom force-fields in modeling protein disordered states. *J. Chem. Theor. Comput.* 11:2–7. https://doi.org/10.1021/ct500718s.

15. Rauscher, S., V. Gapsys, …, H. Grubmüller. 2015. Structural ensembles of intrinsically disordered proteins depend strongly on force field: a comparison to experiment. *J. Chem. Theor. Comput.* 11:5513–5524. https://doi.org/10.1021/acs.jctc.5b00736.

16. Piana, S., A. G. Donchev, …, D. E. Shaw. 2015. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J. Phys. Chem. B*. 119:5113–5123. https://doi.org/10.1021/jp508971m.

17. Best, R. B., W. Zheng, and J. Mittal. 2014. Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association. *J. Chem. Theor. Comput.* 10:5113–5124. https://doi.org/10.1021/ct500569b.

18. Robustelli, P., S. Piana, and D. E. Shaw. 2018. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. USA*. 115:E4758–E4766. https://doi.org/10.1073/pnas.1800690115.

19. Mertens, H. D. T., and D. I. Svergun. 2010. Structural characterization of proteins and complexes using small-angle X-ray solution scattering. *J. Struct. Biol.* 172:128–141, new Trends in Protein Expression. https://www.sciencedirect.com/science/article/pii/S1047847710001905.

20. Stejskal, E. O., and J. E. Tanner. 1965. Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient. *J. Chem. Phys.* 42:288–292. https://doi.org/10.1063/1.1695690.

21. Rigler, R., U. Mets, …, P. Kask. 1993. Fluorescence correlation spectroscopy with high count rate and low background: analysis of translational diffusion. *Eur. Biophys. J.* 22. http://link.springer.com/10.1007/BF00185777.

22. Stetefeld, J., S. A. McKenna, and T. R. Patel. 2016. Dynamic light scattering: a practical guide and applications in biomedical sciences. *Biophys. Rev.* 8:409–427. http://link.springer.com/10.1007/s12551-016-0218-6.

23. Lindorff-Larsen, K., and B. B. Kragelund. 2021. On the potential of machine learning to examine the relationship between sequence, structure, dynamics and function of intrinsically disordered proteins. *J. Mol. Biol.* 433:167196, from Protein Sequence to Structure at Warp Speed: How Alphafold Impacts Biology. https://www.sciencedirect.com/science/article/pii/S0022283621004290.

24. Wilkins, D. K., S. B. Grimshaw, …, L. J. Smith. 1999. Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques. *Biochemistry*. 38:16424–16431. https://doi.org/10.1021/bi991765q.

25. Kärger, J., H. Pfeifer, and W. Heink. 1988. Principles and Application of Self-Diffusion Measurements by Nuclear Magnetic Resonance. Academic Press, pp. 1–89, volume 12 of Advances in Magnetic and Optical Resonance. https://www.sciencedirect.com/science/article/pii/B978012025512250004X.

26. Leeb, S., and J. Danielsson. 2020. Obtaining Hydrodynamic Radii of Intrinsically Disordered Protein Ensembles by Pulsed Field Gradient NMR Measurements. Springer US, pp. 285–302. https://doi.org/10.1007/978-1-0716-0524-0_14.

27. Ortega, A., D. Amorós, and J. García de la Torre. 2011. Prediction of hydrodynamic and other solution properties of rigid proteins from atomic- and residue-level models. *Biophys. J.* 101:892–898. https://www.sciencedirect.com/science/article/pii/S0006349511007764.

28. Kirkwood, J. G., and J. Riseman. 1948. The intrinsic viscosities and diffusion constants of flexible macromolecules in solution. *J. Chem. Phys.* 16:565–573. https://doi.org/10.1063/1.1746947.

29. Fleming, P. J., and K. G. Fleming. 2018. HullRad: fast calculations of folded and disordered protein and nucleic acid hydrodynamic properties. *Biophys. J.* 114:856–869. https://www.sciencedirect.com/science/article/pii/S0006349518300651.

30. Nygaard, M., B. B. Kragelund, …, K. Lindorff-Larsen. 2017. An efficient method for estimating the hydrodynamic radius of disordered protein conformations. *Biophys. J.* 113:550–557. https://www.sciencedirect.com/science/article/pii/S0006349517306926.

31. Ahmed, M. C., R. Crehuet, and K. Lindorff-Larsen. 2020. Computing, Analyzing, and Comparing the Radius of Gyration and Hydrodynamic Radius in Conformational Ensembles of Intrinsically Disordered Proteins. Springer US, pp. 429–445. https://doi.org/10.1007/978-1-0716-0524-0_21.

32. Naullage, P. M., M. Haghighatlari, …, T. Head-Gordon. 2022. Protein dynamics to define and refine disordered protein ensembles. *J. Phys. Chem. B*. 126:1885–1894. https://doi.org/10.1021/acs.jpcb.1c10925.

33. Lincoff, J., M. Haghighatlari, …, T. Head-Gordon. 2020. Extended experimental inferential structure determination method in determining the structural ensembles of disordered protein states. *Commun. Chem.* 3:74. http://www.nature.com/articles/s42004-020-0323-0.

34. Choy, W.-Y., F. A. A. Mulder, …, L. E. Kay. 2002. Distribution of molecular size within an unfolded state ensemble using small-angle X-ray scattering and pulse field gradient NMR techniques. *J. Mol. Biol.* 316:101–112. https://www.sciencedirect.com/science/article/pii/S0022283601953288.

35. Lindorff-Larsen, K., S. Kristjansdottir, …, M. Vendruscolo. 2004. Determination of an ensemble of structures representing the denatured state of the bovine acyl-coenzyme A binding protein. *J. Am. Chem. Soc.* 126:3291–3299. https://doi.org/10.1021/ja039250g.

36. Gomes, G.-N. W., M. Krzeminski, …, C. C. Gradinaru. 2020. Conformational ensembles of an intrinsically disordered protein consistent with NMR, SAXS, and single-molecule FRET. *J. Am. Chem. Soc.* 142:15697–15710. https://doi.org/10.1021/jacs.0c02088.

37. Haxholm, G. W., L. F. Nikolajsen, …, B. B. Kragelund. 2015. Intrinsically disordered cytoplasmic domains of two cytokine receptors mediate conserved interactions with membranes. *Biochem. J.* 468:495–506. https://doi.org/10.1042/BJ20141243.

38. Seiffert, P., K. Bugge, …, B. B. Kragelund. 2020. Orchestration of signaling by structural disorder in class 1 cytokine receptors. *Cell Commun. Signal.* 18:132. https://biosignaling.biomedcentral.com/articles/10.1186/s12964-020-00626-6.

39. Borgia, A., M. B. Borgia, …, B. Schuler. 2018. Extreme disorder in an ultrahigh-affinity protein complex. *Nature*. 555:61–66. http://www.nature.com/articles/nature25762.

40. Manalastas-Cantos, K., P. V. Konarev, …, D. Franke. 2021. *Atsas 3.0*: expanded functionality and new tools for small-angle scattering data analysis. *J. Appl. Crystallogr.* 54:343–355. https://doi.org/10.1107/S1600576720013412.

41. Newcombe, E. A., C. B. Fernandes, …, B. B. Kragelund. 2021. Insight into calcium-binding motifs of intrinsically disordered proteins. *Biomolecules*. 11:1173. https://www.mdpi.com/2218-273X/11/8/1173.

42. Ahmed, M. C., L. K. Skaanning, …, K. Lindorff-Larsen. 2021. Refinement of α-synuclein ensembles against SAXS data: comparison of

force fields and methods. *Front. Mol. Biosci.* 8:654333. https://www.frontiersin.org/article/10.3389/fmolb.2021.654333.

43. Crackower, M. A., S. W. Scherer, …, L.-C. Tsui. 1996. Characterization of the split hand/split foot malformation locus SHFM1 at 7q21. 3–q22. 1 and analysis of a candidate gene for its expression during limb development. *Hum. Mol. Genet.* 5:571–579.

44. Bremer, A., M. Farag, …, T. Mittag. 2022. Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. *Nat. Chem.* 14:196–207. https://www.nature.com/articles/s41557-021-00840-w.

45. Wu, D., A. Chen, and C. Johnson. 1995. An improved diffusion-ordered spectroscopy experiment incorporating bipolar-gradient pulses. *J. Magn. Reson., Ser. A.* 115:260–264. https://www.sciencedirect.com/science/article/pii/S106418588571176X.

46. Prestel, A., K. Bugge, …, B. B. Kragelund. 2018. Chapter eight - characterization of dynamic IDP complexes by NMR spectroscopy. *In* Intrinsically Disordered Proteins. E. Rhoades, ed Academic Press, pp. 193–226, volume 611 of Methods in Enzymology. https://www.sciencedirect.com/science/article/pii/S0076687918303057.

47. Blanchet, C. E., A. Spilotros, …, D. I. Svergun. 2015. Versatile sample environments and automation for biological solution X-ray scattering experiments at the P12 beamline (PETRA III, DESY). *J. Appl. Crystallogr.* 48:431–443. http://scripts.iucr.org/cgi-bin/paper?S160057671500254X.

48. Franke, D., M. V. Petoukhov, …, D. I. Svergun. 2017. *Atsas 2.8* : a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J. Appl. Crystallogr.* 50:1212–1225. http://scripts.iucr.org/cgi-bin/paper?S1600576717007786.

49. Ozenne, V., F. Bauer, …, M. Blackledge. 2012. Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics.* 28:1463–1470. https://doi.org/10.1093/bioinformatics/bts172.

50. Tesei, G., T. K. Schulze, …, K. Lindorff-Larsen. 2021. Accurate model of liquid-liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. *Proc. Natl. Acad. Sci. USA.* 118. e2111696118. https://doi.org/10.1073/pnas.2111696118.

51. Eastman, P., J. Swails, …, V. S. Pande. 2017. OpenMM 7: rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* 13:e1005659. https://doi.org/10.1371/journal.pcbi.1005659.

52. Rotkiewicz, P., and J. Skolnick. 2008. Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem.* 29:1460–1465. https://doi.org/10.1002/jcc.20906.

53. Grudinin, S., M. Garkavenko, and A. Kazennov. 2017. *Pepsi-SAXS*: an adaptive method for rapid and accurate computation of small-angle X-ray scattering profiles. *Acta Crystallogr. D Struct. Biol.* 73:449–464. https://doi.org/10.1107/S2059798317005745.

54. Pesce, F., and K. Lindorff-Larsen. 2021. Refining conformational ensembles of flexible proteins against small-angle x-ray scattering data. *Biophys. J.* 120:5124–5135. https://www.sciencedirect.com/science/article/pii/S0006349521008286.

55. Larsen, A. H., and M. C. Pedersen. 2021. Experimental noise in small-angle scattering can be assessed using the Bayesian indirect Fourier transformation. *J. Appl. Crystallogr.* 54:1281–1289. https://doi.org/10.1107/S1600576721006877.

56. Kirkwood, J. G. 1954. The general theory of irreversible processes in solutions of macromolecules. *J. Polym. Sci.* 12:1–14. https://doi.org/10.1002/pol.1954.120120102.

57. Clisby, N., and B. Dünweg. 2016. High-precision estimate of the hydrodynamic radius for self-avoiding walks. *Phys. Rev. E.* 94:052102. https://doi.org/10.1103/PhysRevE.94.052102.

58. Jephthah, S., L. Staby, …, M. Skepö. 2019. Temperature dependence of intrinsically disordered proteins in simulations: what are we missing? *J. Chem. Theor. Comput.* 15:2672–2683. https://doi.org/10.1021/acs.jctc.8b01281.

59. Xiang, S., V. Gapsys, …, M. Zweckstetter. 2013. Phosphorylation drives a dynamic switch in serine/arginine-rich proteins. *Structure.* 21:2162–2174. https://www.sciencedirect.com/science/article/pii/S0969212613003651.

60. Mittag, T., S. Orlicky, …, J. D. Forman-Kay. 2008. Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. *Proc. Natl. Acad. Sci. USA.* 105:17772–17777. https://doi.org/10.1073/pnas.0809222105.

61. Mukrasch, M. D., S. Bibow, …, M. Zweckstetter. 2009. Structural polymorphism of 441-residue Tau at single residue resolution. *PLoS Biol.* 7:e1000034. https://dx.plos.org/10.1371/journal.pbio.1000034.

62. Mylonas, E., A. Hascher, …, D. I. Svergun. 2008. Domain conformation of Tau protein studied by solution small-angle X-ray scattering. *Biochemistry.* 47:10345–10353. https://doi.org/10.1021/bi800900d.

63. Guinier, A. 1939. La diffraction des rayons X aux très petits angles : application à l'étude de phénomènes ultramicroscopiques. *Ann. Phys.* 11:161–237.

64. Riback, J. A., M. A. Bowman, …, T. R. Sosnick. 2017. Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science.* 358:238–241. https://doi.org/10.1126/science.aan5774.

65. Zheng, W., and R. B. Best. 2018. An extended guinier analysis for intrinsically disordered proteins. *J. Mol. Biol.* 430:2540–2553, intrinsically Disordered Proteins. https://www.sciencedirect.com/science/article/pii/S0022283618301359.

66. Oono, Y., and M. Kohmoto. 1983. Renormalization group theory of transport properties of polymer solutions. I. Dilute solutions. *J. Chem. Phys.* 78:520–528. https://doi.org/10.1063/1.444477.

67. Soranno, A., B. Buchli, …, B. Schuler. 2012. Quantifying internal friction in unfolded and intrinsically disordered proteins with single-molecule spectroscopy. *Proc. Natl. Acad. Sci. USA.* 109:17800–17806. https://doi.org/10.1073/pnas.1117368109.

68. Moses, D., F. Yu, …, S. Sukenik. 2020. Revealing the hidden sensitivity of intrinsically disordered proteins to their chemical environment. *J. Phys. Chem. Lett.* 11:10131–10136. https://doi.org/10.1021/acs.jpclett.0c02822.

69. Jones, J. A., D. K. Wilkins, …, C. M. Dobson. 1997. Characterisation of protein unfolding by NMR diffusion measurements. *J. Biomol. NMR.* 10:199–203. http://link.springer.com/10.1023/A:1018304117895.

70. Jensen, M. R., P. R. Markwick, …, M. Blackledge. 2009. Quantitative determination of the conformational properties of partially folded and intrinsically disordered proteins using NMR dipolar couplings. *Structure.* 17:1169–1185. https://www.sciencedirect.com/science/article/pii/S0969212609002986.

71. Wells, M., H. Tidow, …, A. R. Fersht. 2008. Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proc. Natl. Acad. Sci. USA.* 105:5762–5767. https://www.pnas.org/content/105/15/5762.

72. Mukrasch, M. D., P. Markwick, …, M. Blackledge. 2007. Highly populated turn conformations in natively unfolded Tau protein identified from residual dipolar couplings and molecular simulation. *J. Am. Chem. Soc.* 129:5235–5243. https://doi.org/10.1021/ja0690159.

73. Svergun, D., C. Barberato, and M. H. J. Koch. 1995. *Crysol* – a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.* 28:768–773. https://doi.org/10.1107/S0021889895007047.

74. Henriques, J., L. Arleth, …, M. Skepö. 2018. On the calculation of SAXS profiles of folded and intrinsically disordered proteins from computer simulations. *J. Mol. Biol.* 430:2521–2539, intrinsically Disordered Proteins. https://www.sciencedirect.com/science/article/pii/S0022283618301232.

75. Burchard, W., M. Schmidt, and W. H. Stockmayer. 1980. Information on polydispersity and branching from combined quasi-elastic and intergrated scattering. *Macromolecules.* 13:1265–1272.

76. Ramanujam, V., T. R. Alderson, …, A. Bax. 2020. Protein structural changes characterized by high-pressure, pulsed field gradient diffusion NMR spectroscopy. *J. Magn. Reson.* 312:106701. https://www.sciencedirect.com/science/article/pii/S1090780720300197.

Biophysical Journal, Volume *122*

Supplemental information

Assessment of models for calculating the hydrodynamic radius of intrinsically disordered proteins

Francesco Pesce, Estella A. Newcombe, Pernille Seiffert, Emil E. Tranchant, Johan G. Olsen, Christy R. Grace, Birthe B. Kragelund, and Kresten Lindorff-Larsen

Supplementary material: Assessment of models for calculating the hydrodynamic radius of intrinsically disordered proteins

Francesco Pesce,[†] Estella A. Newcombe,[†] Pernille Seiffert,[†] Emil E. Tranchant,[†] Johan G. Olsen,[†] Christy R. Grace,[‡] Birthe B. Kragelund,[*,†] and Kresten Lindorff-Larsen[*,†]

[†]*Structural Biology and NMR Laboratory, The Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Copenhagen, Denmark*

[‡]*Department of Structural Biology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA*

E-mail: bbk@bio.ku.dk; lindorff@bio.ku.dk

Supplementary figures and tables

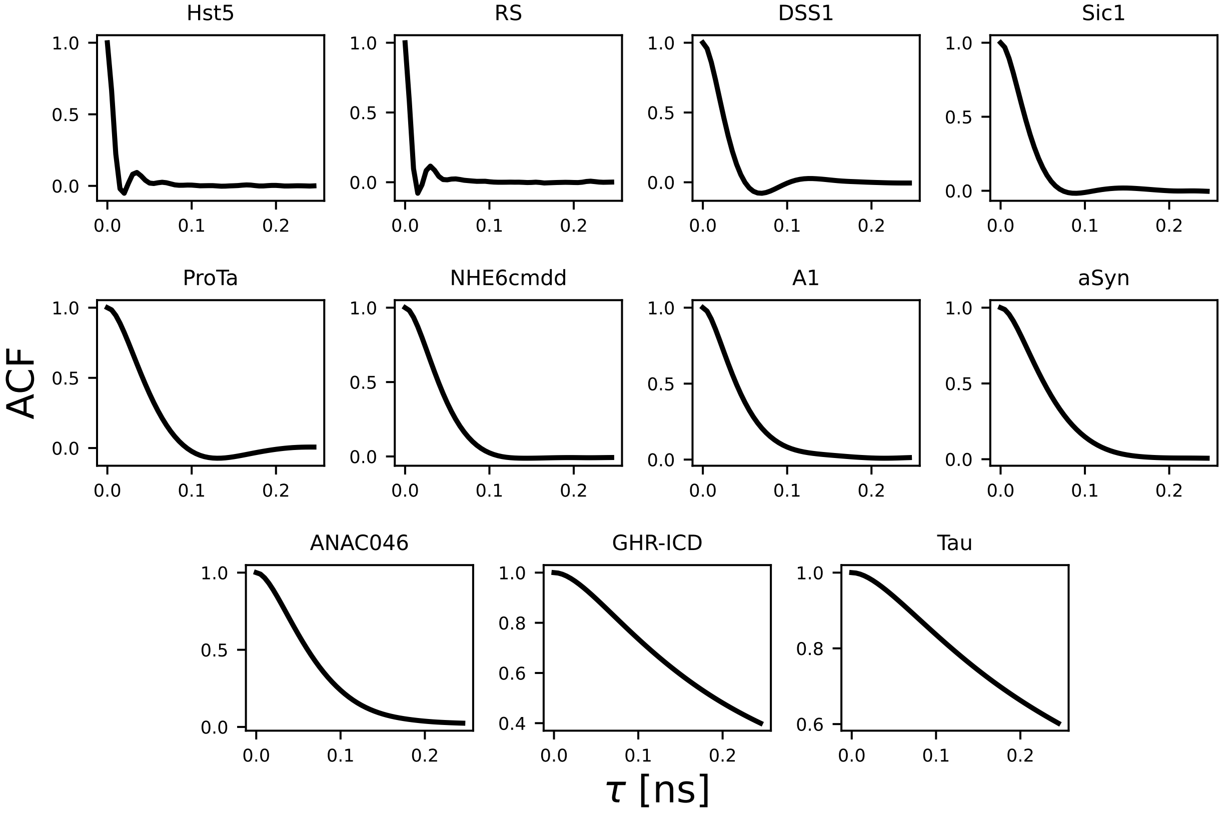

Figure S1: Autocorrelation function of the radius of gyration from the CALVADOS simulations shown up to a lag-time (τ) of 0.25 ns.

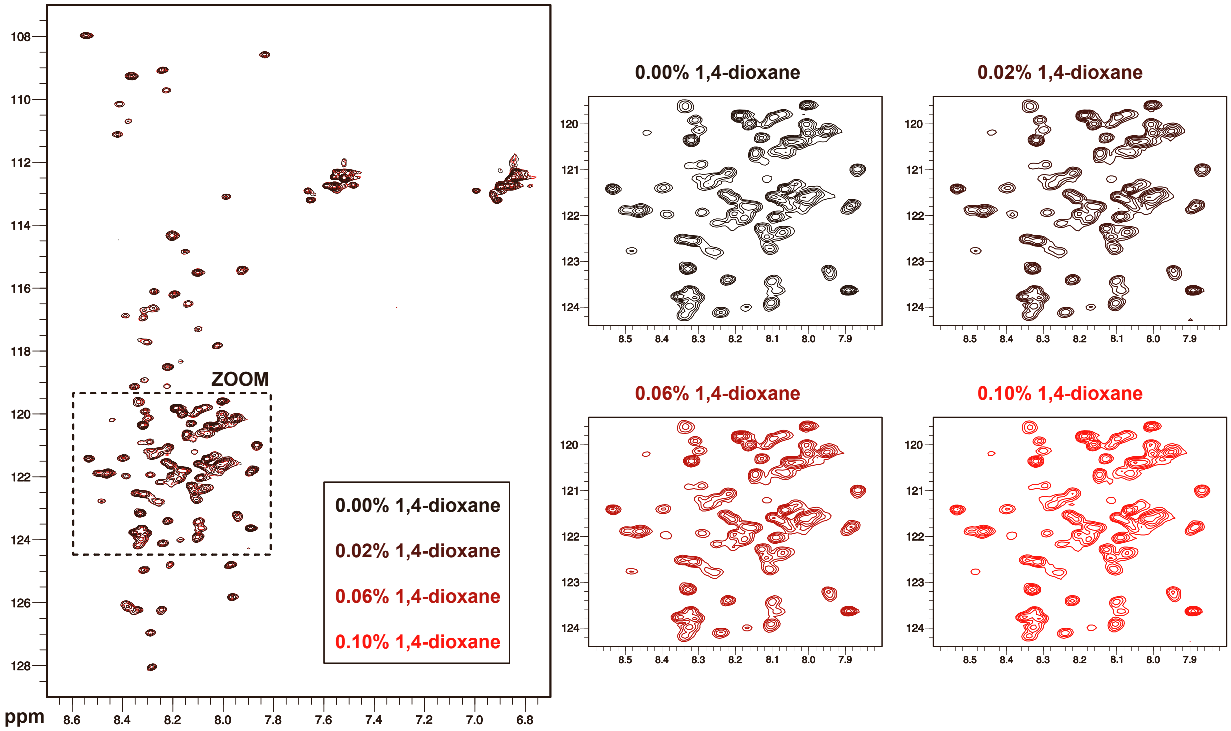

Figure S2: Titration of ANAC046 with dioxane. The figure shows ^{1}H-^{15}N HSQC NMR spectra of ^{15}N-labeled ANAC046 alone and in presence of 0.02%, 0.06% and 0.1% of dioxane. Spectra were recorded in 20 mM sodium phosphate (pH 7.0), 100 mM NaCl, 2 mM TCEP, 25 μM DSS, 10% D_2O at 25°C.

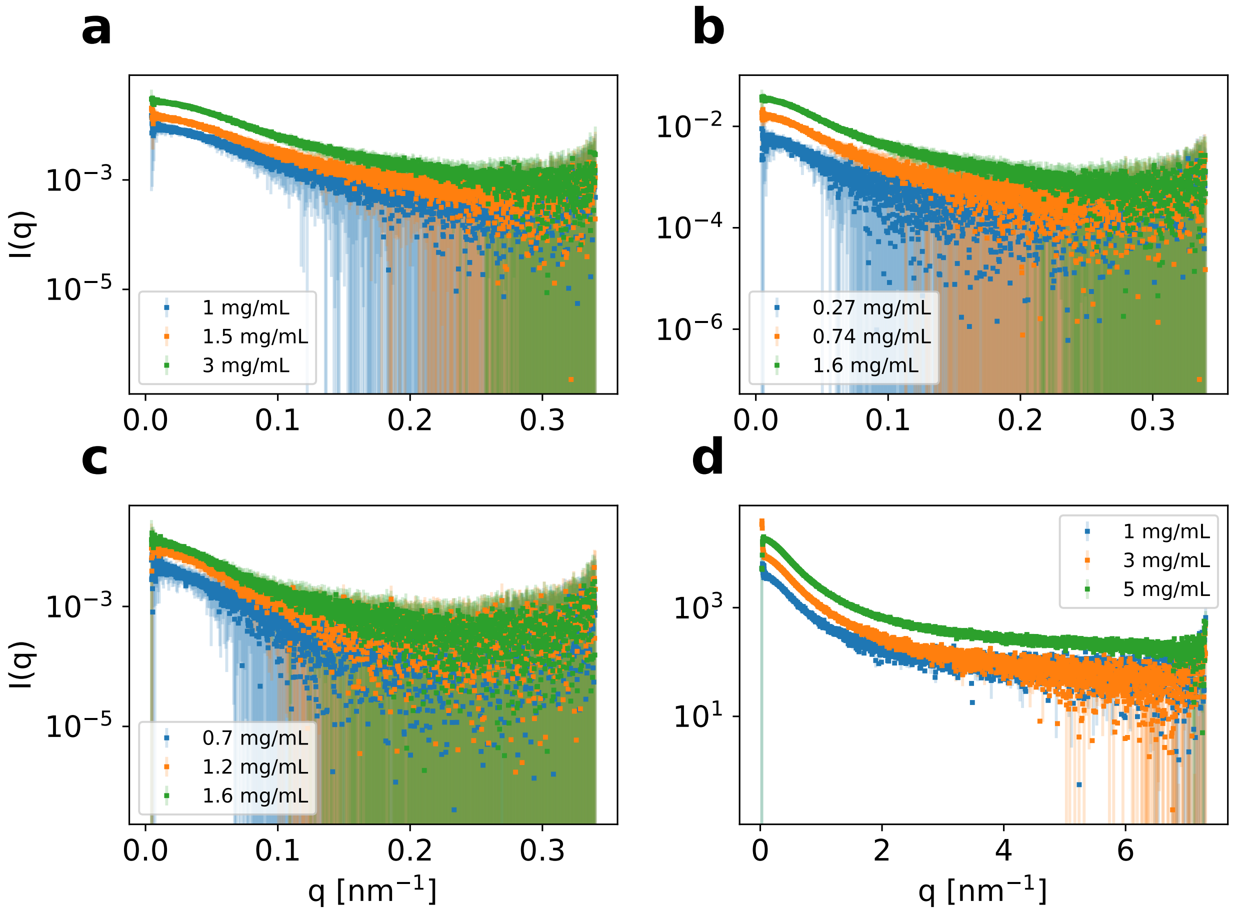

Figure S3: Experimental SAXS profiles for (a) Dss1, (b) ProTα, (c) NHE6cmdd, (d) ANAC046. SAXS from samples at different protein concentrations are shown in different colours.

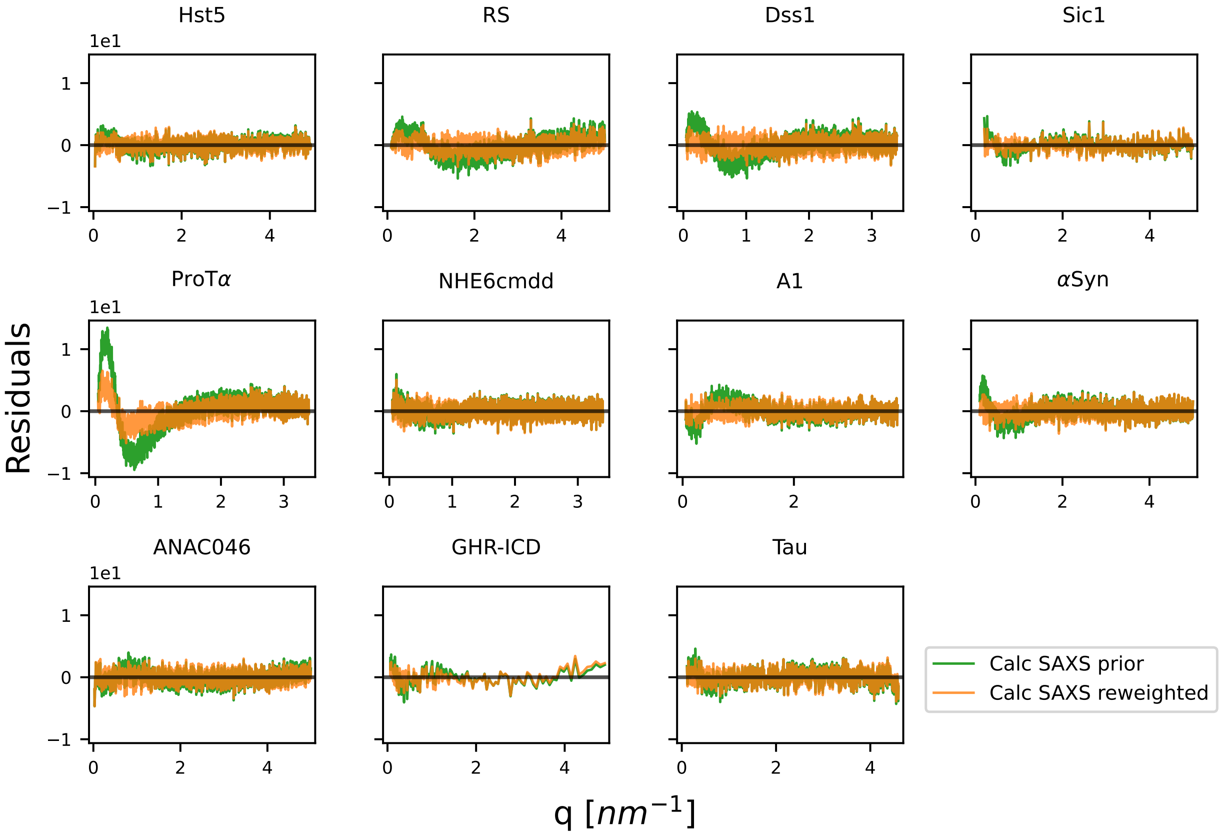

Figure S4: Residuals of the SAXS intensities calculated from the FM ensembles before (orange) and after (green) reweighting.

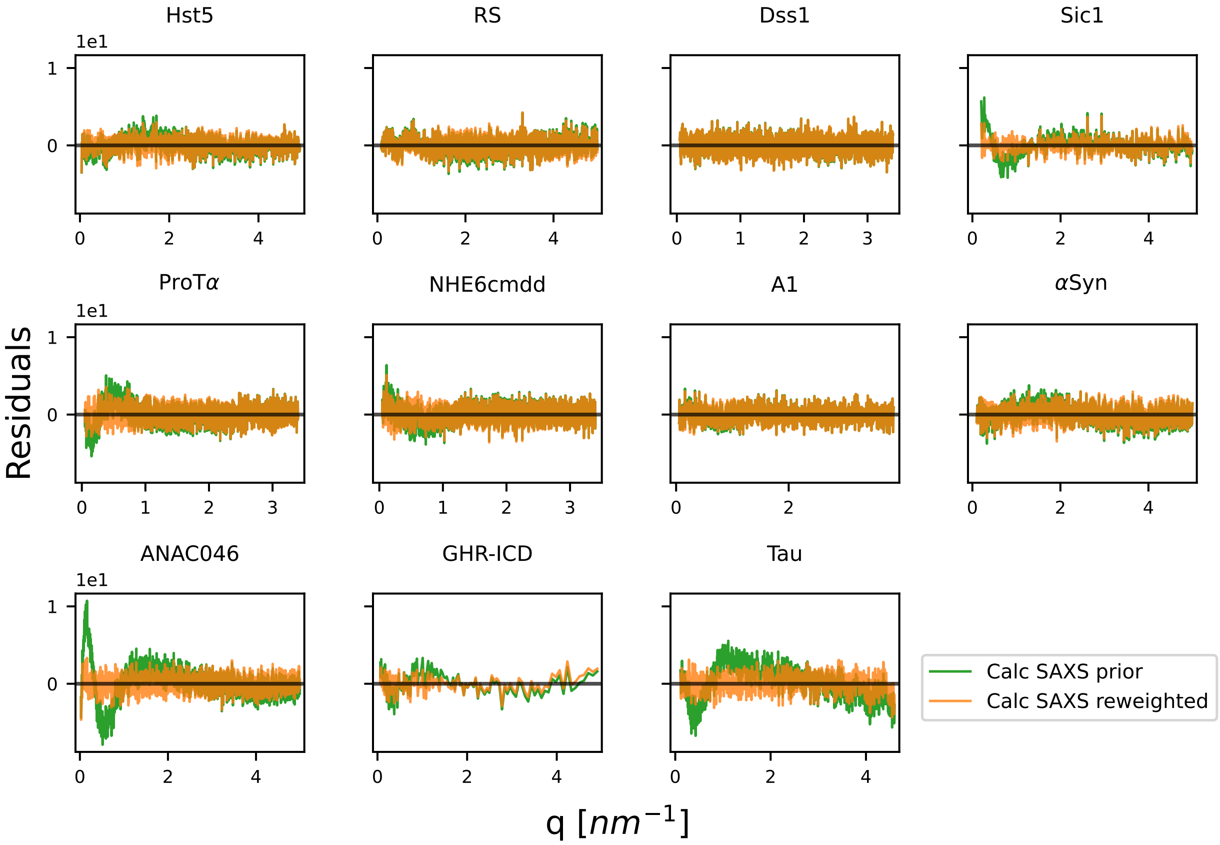

Figure S5: Residuals of the SAXS intensities calculated from the CALVADOS ensembles before (orange) and after (green) reweighting.

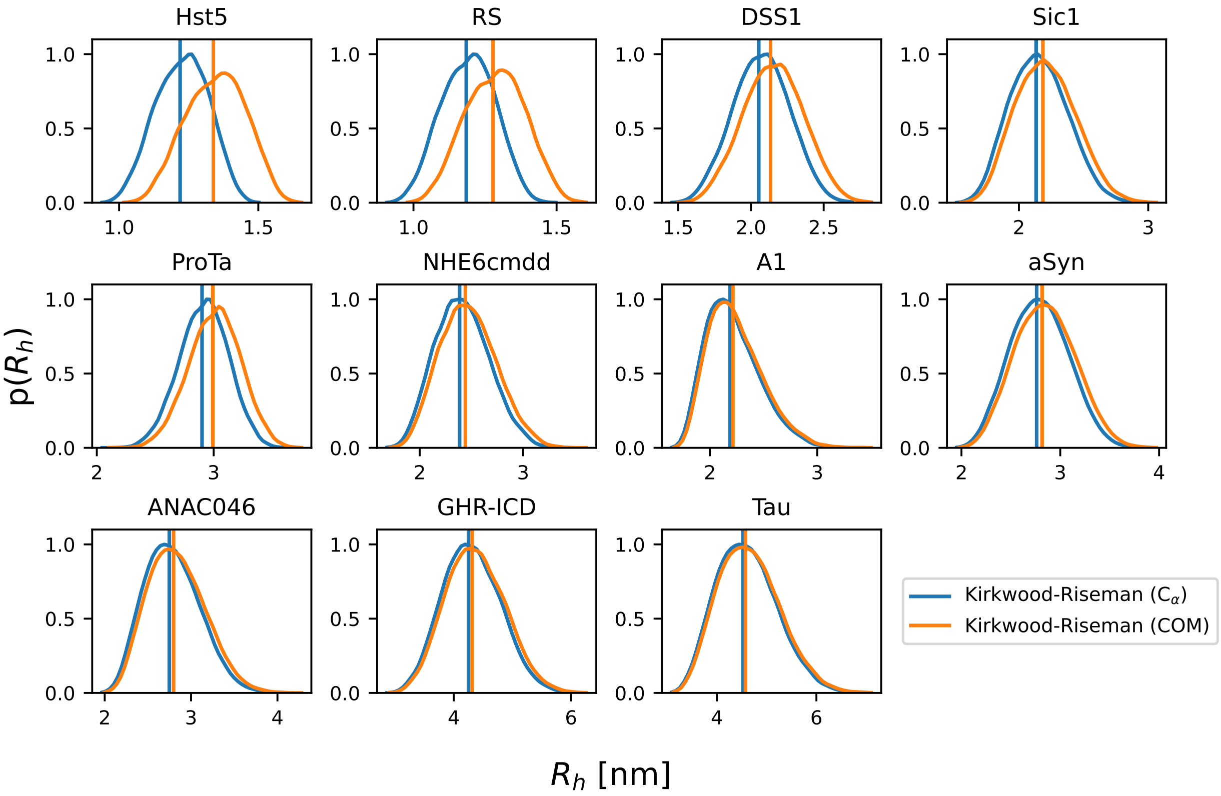

Figure S6: Distributions of the R_h^KR calculated from the CALVADOS ensembles using either the C_α coordinates or from the center of mass of the residues after converting the ensembles to all-atom structures.

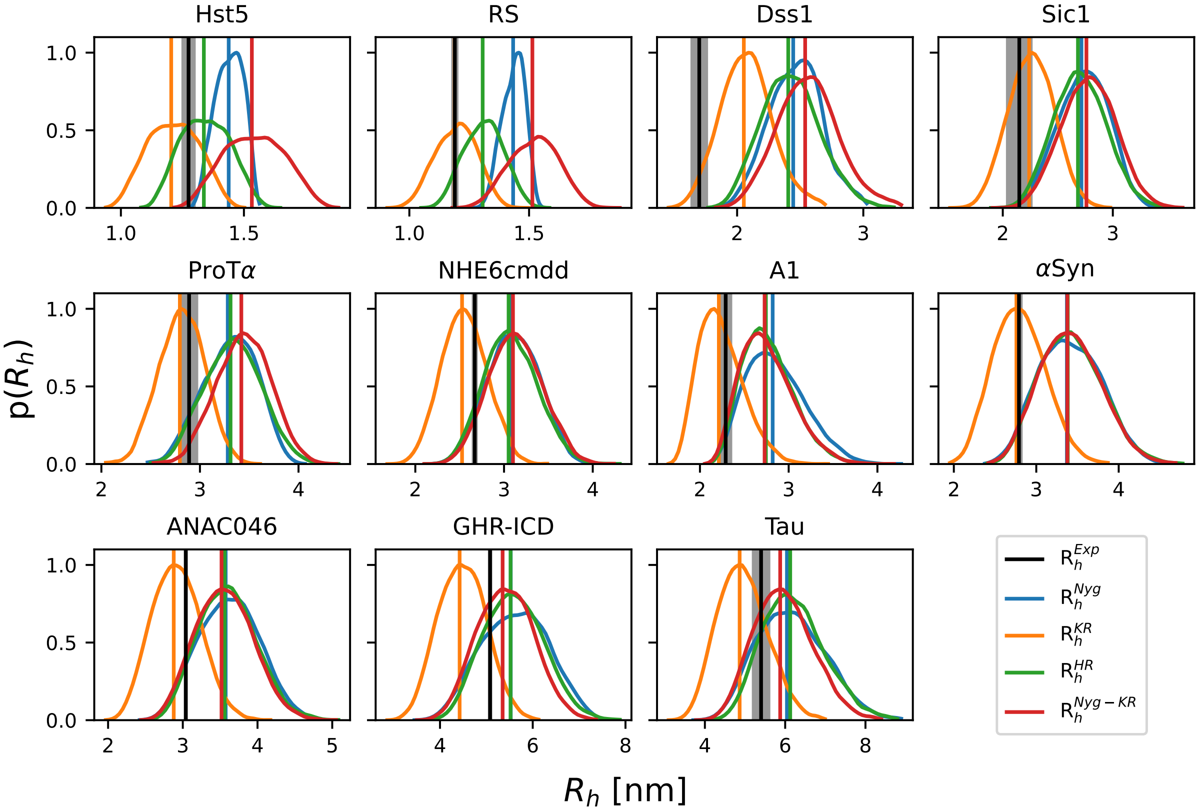

Figure S7: Probability distributions of the R_h and their ensemble averages calculated from the SAXS-reweighted CALVADOS ensembles, compared with the R_h determined by PFG NMR diffusion (in black). We tested four approaches to calculate the R_h from atomic coordinate: the R_g-dependent Nygaard equation ($R_\mathrm{h}^\mathrm{Nyg}$, in blue), the Kirkwood-Riseman equation (R_h^KR, in orange), HullRad (R_h^HR, in green), and the Nygaard correction to the Kirkwood-Riseman equation ($R_\mathrm{h}^\mathrm{Nyg\text{-}KR}$, in red).

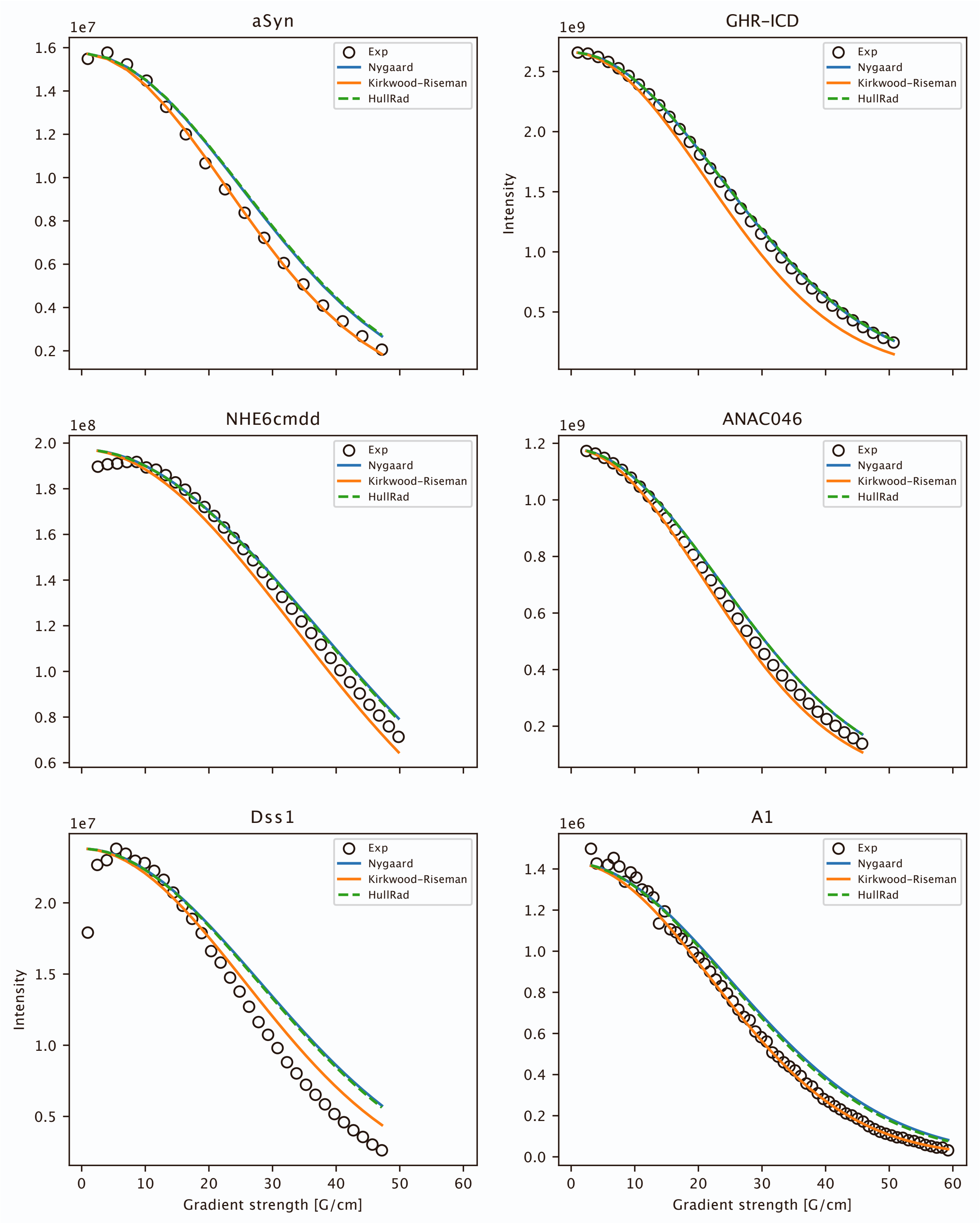

Figure S8: We use the average calculated R_h with three different forward models to derive the diffusion profiles using the Stejskal-Tanner equation. We show this for the PFG NMR experiments performed in this study.

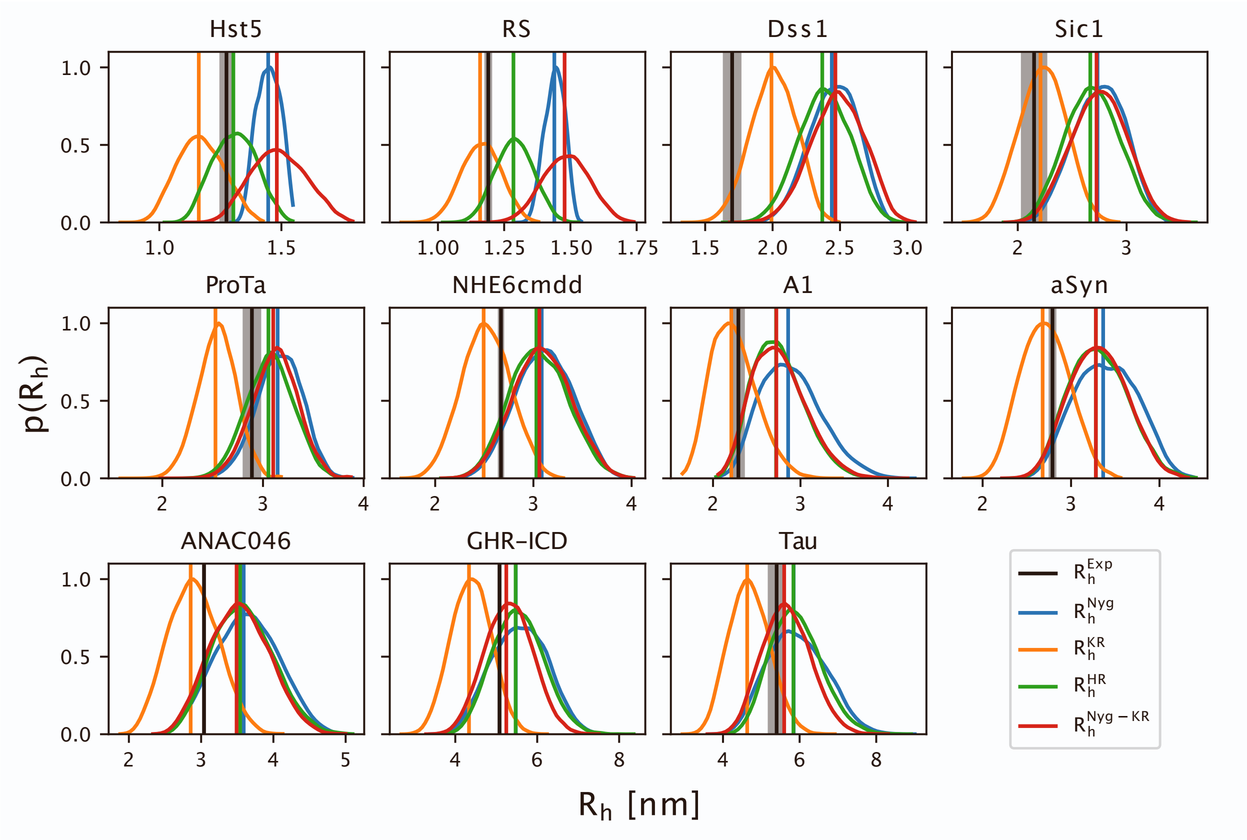

Figure S9: Probability distributions of the R_h and their ensemble averages calculated from the SAXS-reweighted FM ensembles, compared with the R_h determined by PFG NMR diffusion (in black). We tested four approaches to calculate the R_h from atomic coordinate: the R_g-dependent Nygaard equation ($R_\mathrm{h}^\mathrm{Nyg}$, in blue), the Kirkwood-Riseman equation (R_h^KR, in orange), HullRad (R_h^HR, in green), and the Nygaard correction to the Kirkwood-Riseman equation ($R_\mathrm{h}^\text{Nyg-KR}$, in red).

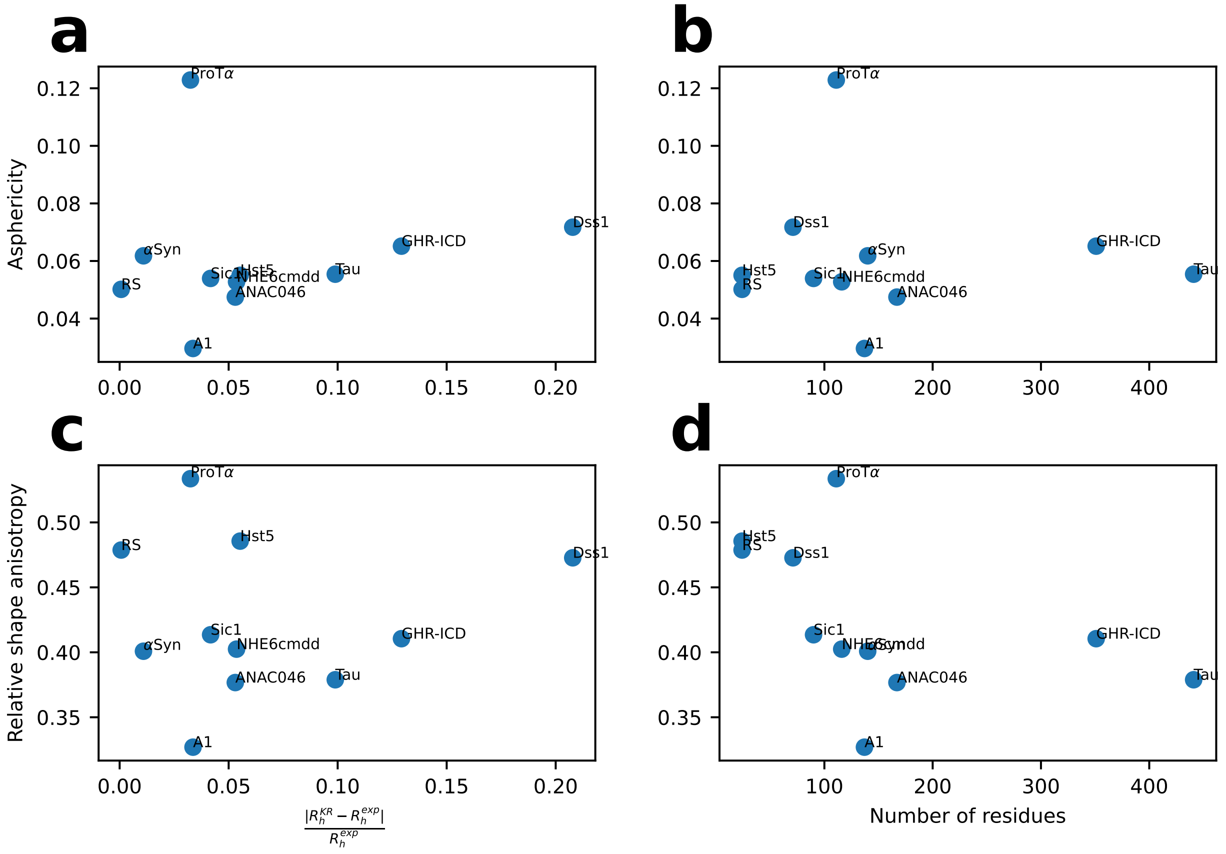

Figure S10: We calculate (a, b) the average asphericity and (c, d) the average relative shape anisotropy of the CALVADOS ensembles, and plot them against (a, c) the relative difference of the R_h calculated with the Kirkwood-Riseman equation from the experimental R_h and (b, d) the number of residues.

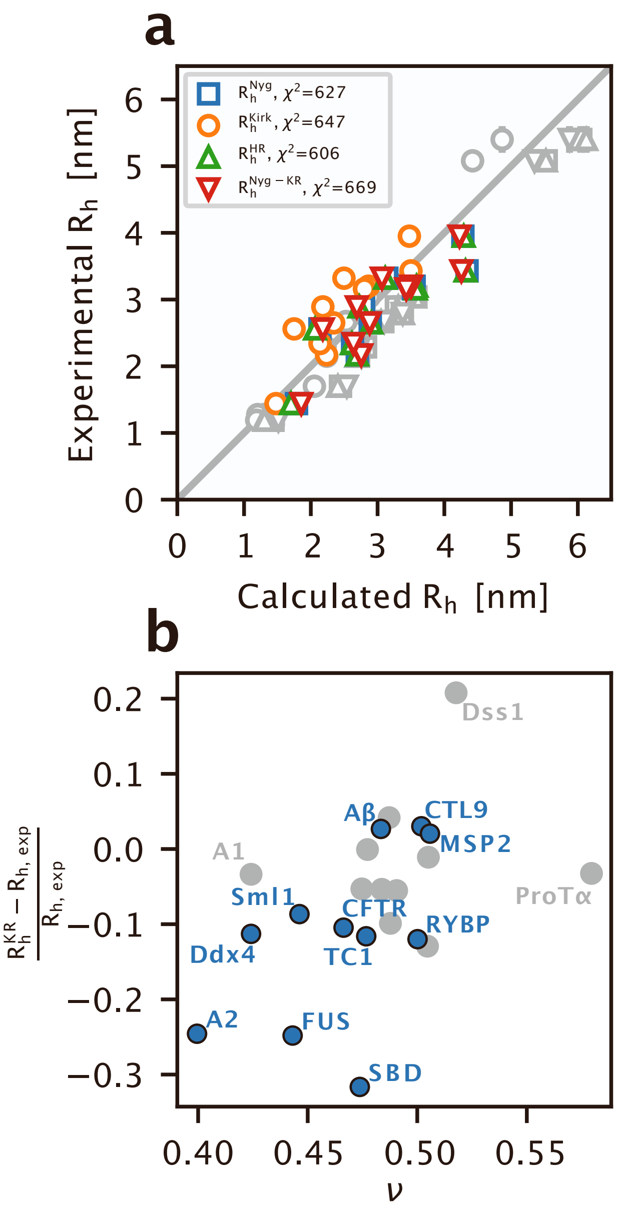

Figure S11: Comparison of the R_h values from PFG NMR measurements with predictions from the CALVADOS ensembles for the eleven proteins in Table S5. (a) R_h calculated from the CALVADOS ensembles using the Nygaard equation (R_h^{Nyg}, in blue), the Kirkwood-Riseman equation (R_h^{KR}, in orange), HullRad (R_h^{HR}, in green), and the Nygaard correction to the Kirkwood-Riseman equation ($R_h^{Nyg\text{-}KR}$, in red) are compared to the experimental R_h values. The legend reports the χ^2 over these 11 proteins. (b) Plot of the scaling exponent (ν) calculated from the CALVADOS ensembles vs. the relative difference between the calculated R_h^{KR} and the experimental R_h values. Points in grey refer to those proteins for which we have SAXS data (Table 1). The Pearson correlation coefficient across the entire set of proteins is 0.56.

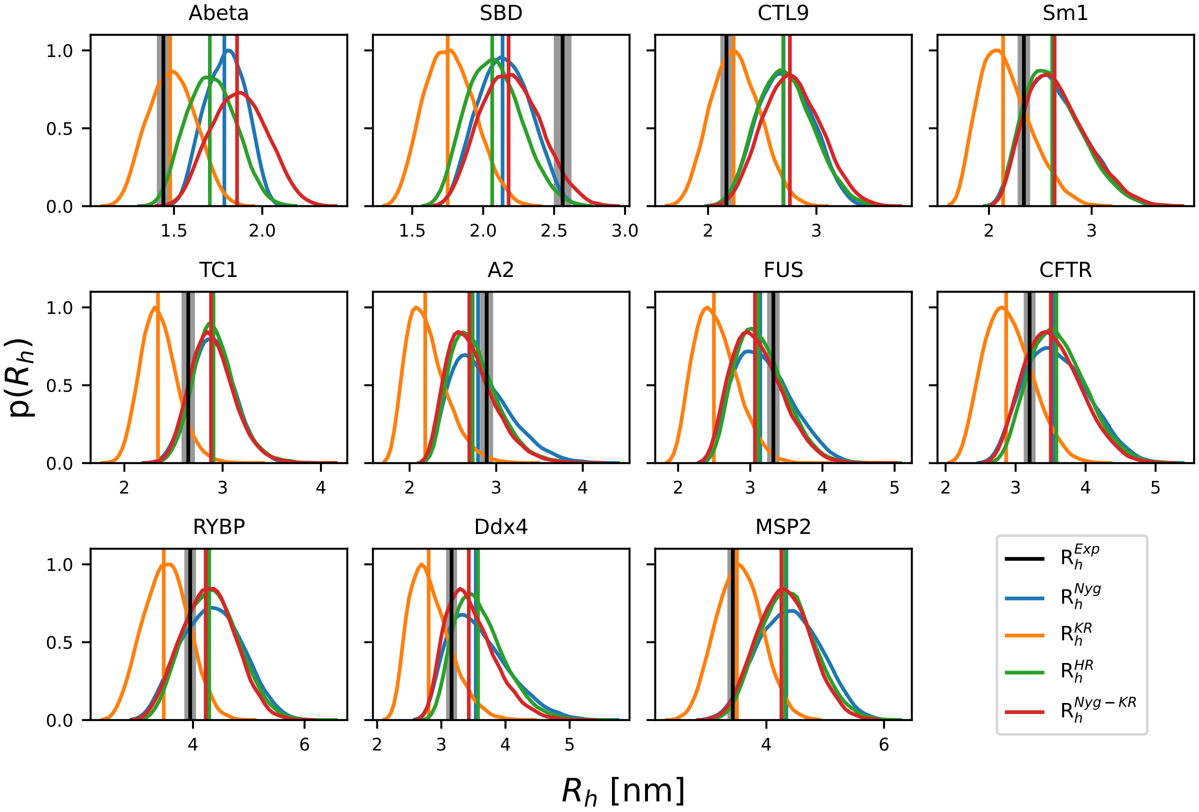

Figure S12: Probability distributions of the R_h and their ensemble averages calculated from the CALVADOS ensembles of the eleven proteins in Table S5, compared to the R_h determined by PFG NMR diffusion (in black). The results are shown for the four models to calculate the R_h from the coordinates: the R_g-dependent Nygaard equation ($R_\mathrm{h}^\mathrm{Nyg}$, in blue), the Kirkwood-Riseman equation (R_h^KR, in orange), HullRad (R_h^HR, in green), and the Nygaard correction to the Kirkwood-Riseman equation ($R_\mathrm{h}^\mathrm{Nyg\text{-}KR}$, in red).

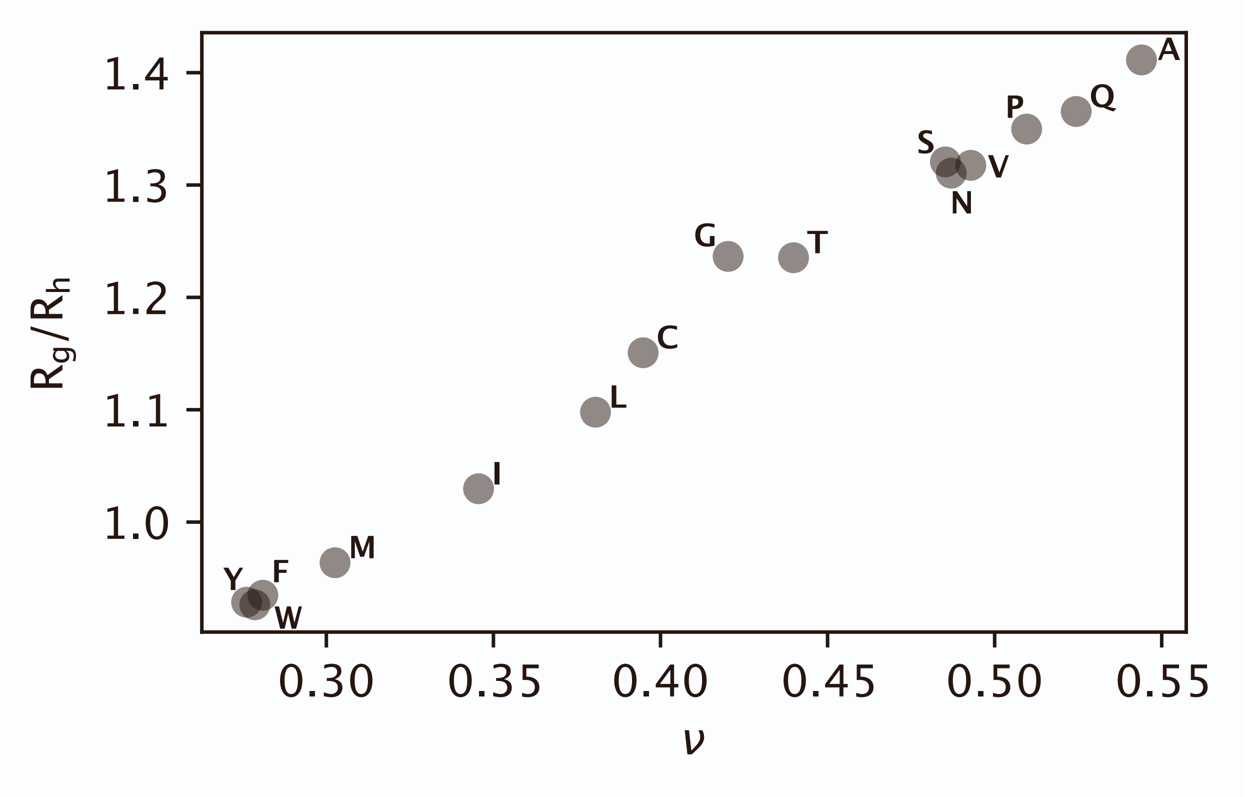

Figure S13: We simulated 200-residue-long homopolymers of the 15 non-ionic (i.e. excluding Asp, Glu, His, Lys and Arg) amino acids with CALVADOS. CALVADOS was not trained or tested to give accurate results for homopolymers, and we instead use the simulations to create 15 gradually expanded conformational ensembles. We calculated R_{g} and R_{h} (using the Kirkwood-Riseman equation) from each of these ensembles to explore whether the calculations recover the expected increase in the ratio (calculated as $\langle R_g^2\rangle^{1/2}/\langle R_h^{-1}\rangle^{-1}$) as a function of the expansion of the chain (quantified using the calculated scaling exponent, ν).

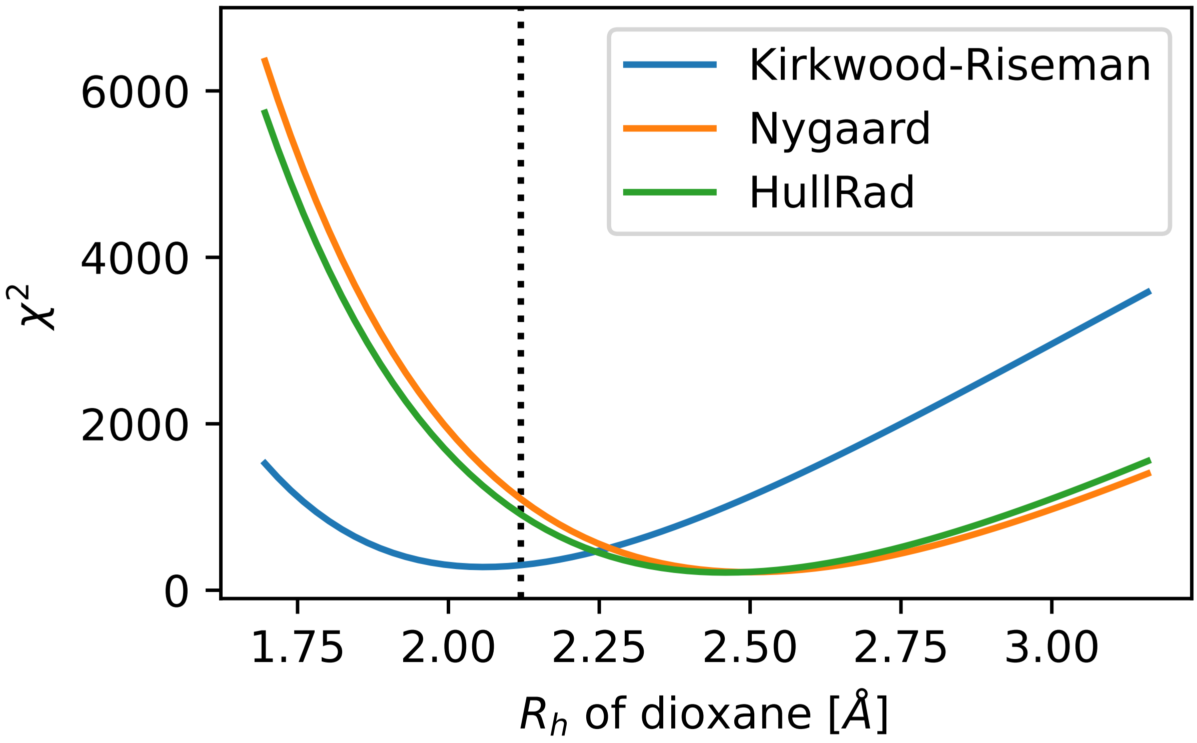

Figure S14: We use different values for the R_h of dioxane and derive the resulting experimental R_h values for the eleven proteins in this study. We then calculate the χ^2 of the R_h obtained with either the Kirkwood-Riseman equation, the Nygaard equation or HullRad from the CALVADOS ensembles against the new sets of experimental R_h. Plot of χ^2 vs. the R_h of dioxane shows that the Nygaard equation and HullRad would give rise to a better agreement with the data only if the R_h of dioxane were greater than 2.25 Å.

Table S1: Sequence features and number of conformers generated with Flexible-Meccano (FM) for each protein.

name	Length	SCD*	NCPR**	Fraction positive	Fraction negative	Fraction proline	FM conformers number
Hst5	24	0.8	0.2	29.2	8.3	0.0	10000
RS	24	2.7	0.3	33.3	0.0	4.2	10000
DSS1	71	6.9	-0.3	7.0	32.4	2.8	15000
Sic1	90	3.5	0.1	12.2	0.0	16.7	15000
ProTα	111	39.5	-0.4	9.0	47.7	1.8	15000
NHE6cmdd	116	2.6	-0.1	6.9	17.2	9.5	15000
A1	137	1.3	0.1	8.8	2.9	1.5	20000
aSyn	140	-1.2	-0.1	10.7	17.1	3.6	20000
ANAC046	167	2.2	-0.1	4.2	10.8	8.4	20000
GHR-ICD	351	10.5	-0.1	7.4	16.0	8.5	25000
Tau	441	-8.1	0.0	13.2	12.7	9.8	30000

* The sequence charge decoration (SCD)[1] is a measure of charge patterning in the sequence. High values mean that charges are uniformly mixed, while low values indicates a separation of positive and negative charges.

** Net charge per residue.

Table S2: Radius of gyration calculated from the experimental SAXS profiles using the Guinier analysis (with the ATSAS package),[2] the Extended Guinier analysis (EGA)[3] and the Molecular Form Factor (MFF)[4] approach. Additionally we show the scaling exponent ν estimated by both the EGA and MFF analysis and the ratio between R_{g} (from ATSAS) and the R_{h} from PFG NMR. This ratio takes values around 0.78 for globular proteins and 1.2 for ideal Gaussian chains[5,6]

.

	R_{g} (ATSAS)	R_{g} (EGA)	R_{g} (MMF)	$\nu(EGA)$	ν (MFF)	$\frac{R_{\mathrm{g}}^{\mathrm{ATSAS}}}{R_{\mathrm{h}}}$
name						
Hst5	1.34 ± 0.05	1.38	1.39 ± 0.01	0.58	0.5 ± 0.1	1.05
RS	1.26 ± 0.08	1.34	1.36 ± 0.01	0.57	0.60 ± 0.04	1.06
DSS1	2.5 ± 0.1	2.6	2.636 ± 0.004	0.58	0.567 ± 0.005	1.46
Sic1	2.9 ± 0.1	3.03	3.06 ± 0.02	0.58	0.58 ± 0.01	1.33
ProTα	3.7 ± 0.2	3.95	3.94 ± 0.01	0.62	0.595 ± 0.003	1.27
NHE6cmdd	3.2 ± 0.2	3.36	3.40 ± 0.01	0.57	0.55 ± 0.01	1.21
A1	2.5 ± 0.1	2.6	2.73 ± 0.02	0.49	0.45 ± 0.01	1.11
aSyn	3.56 ± 0.04	3.62	3.68 ± 0.01	0.56	0.591 ± 0.003	1.28
ANAC046	3.6 ± 0.3	3.73	3.768 ± 0.004	0.55	0.576 ± 0.001	1.19
GHR-ICD	6.0 ± 0.5	6.09	5.96 ± 0.04	0.56	0.557 ± 0.003	1.19
Tau	6.4 ± 0.5	6.24	6.66 ± 0.04	0.54	0.588 ± 0.001	1.18

Table S3: Experimental conditions used in the SAXS and PFG NMR measurements.

Protein	SAXS buffer	SAXS T [K]	PFG NMR buffer	PFG NMR T [K]
Hst5	20 mM Tris (pH 7.5), 150 mM NaCl	293	20 mM Na_2HPO_4/NaH_2PO_4 (pH 7.0), 10% D_2O, 0.25 mM DSS, and 0.25% 1,4-dioxane	293
RS	50 mM Na_2HPO_4/NaH_2PO_4 (pH 7.0), 100 mM NaCl	298	50 mM Na_2HPO_4/NaH_2PO_4 (pH 7.0), 100 mM NaCl	298
Dss1	20 mM Tris (pH 7.4), 150 mM NaCl, 2% glycerol, 5 mM DTT	288	20 mM Tris, 150 mM NaCl, 5 mM DTT, 2% glycerol, 10% D_2O, 0.25 mM DSS, 0.02% 1,4-dioxane, 0.02% NaN_3	288
Sic1	50 mM Tris (pH 7.5), 150 mM NaCl, 5 mM DTT, and 2 mM TCEP	ND	10 mM Na_2HPO_4/NaH_2PO_4 (pH 7.0), 140 mM NaCl, 1 mM EDTA, 0.2% NaN_3, 10% D_2O	278
ProTα	10 mM Tris (pH 7.4), 0.1 mM EDTA, 155 mM KCl, 2% glycerol	288	10 mM Tris (pH 7.4), 0.1 mM EDTA, 155 mM KCl	288
NHE6cmdd	20 mM Tris-HCl (pH 7.4), 150 mM NaCl, 2% glycerol, 5 mM DTT	288	20 mM Tris-HCl (pH 7.4), 150 mM NaCl, 5 mM DTT, 0.1% 1,4-dioxane, 25 μM DSS, 10% D_2O	288
A1	20 mM HEPES (pH 7.0), 150 mM NaCl	298	20 mM HEPES (pH 7.0), 150 mM NaCl, 1,4-dioxane, 0.02% dioxane, 10% D_2O	298
αSyn	20 mM Na_2HPO_4/NaH_2PO_4 (pH 7.4), 150 mM NaCl, 2% glycerol	293	20 mM Na_2HPO_4/NaH_2PO_4 (pH 7.4), 150 mM NaCl, 2% glycerol 10% D_2O, 0.25 mM DSS, 0.02% 1,4-dioxane, 0.02% NaN_3	293
ANAC046	20 mM Na_2HPO_4/NaH_2PO_4 (pH 7.0), 100 mM NaCl, 5 mM DTT	298	20 mM Na_2HPO_4/NaH_2PO_4 (pH 7.0), 100 mM NaCl, 1 mM DTT, 10% D_2O, 0.25 mM DSS, 0.04% 1,4-dioxane, 0.02% NaN_3	298
GHR-ICD	20 mM Na_2HPO_4/NaH_2PO_4 (pH 7.3), 300 mM NaCl, 10x excess of DTT, 2% glycerol	298	20 mM Na_2HPO_4/NaH_2PO_4 (pH 7.3), 150 mM NaCl, 10 mM B-ME, 10% D_2O, 0.25 mM DSS, 0.05% 1,4-dioxane, 0.02% NaN_3	298
Tau	137 mM NaCl, 3 mM KCl, 10 mM Na_2HPO_4/NaH_2PO_4 (pH 7.4), 2 mM KH_2PO_4, and 1 mM DTT	288	99.9% D_2O, 50 mM Na_2HPO_4/NaH_2PO_4 (pH 7.0) (pH 6.9), 2% 1,4-dioxane	ND

Table S4: Parameters of the BME reweighting of ensembles against SAXS data. The χ_r^2 values refer to the values after reweighting.

Protein	Flexible meccano			CALVADOS		
	χ_r^2	ϕ_{eff}	θ	χ_r^2	ϕ_{eff}	θ
Hst5	1.00	0.85	50	1.00	0.97	250
RS	1.14	0.73	500	1.05	0.95	150
Dss1	1.00	0.82	500	0.98	0.94	25
Sic1	1.00	0.95	200	1.00	0.88	200
ProTα	2.72	0.51	5000	1.00	0.87	500
NHE6cmdd	1.00	0.92	150	1.00	0.86	150
A1	1.00	0.74	100	1.00	0.99	750
αSyn	1.01	0.86	150	1.02	0.77	100
ANAC046	1.02	0.82	150	1.02	0.91	1000
GHR-ICD	1.24	0.90	75	1.07	0.90	50
Tau	1.10	0.86	150	1.15	0.81	750

Table S5: Additional dataset of eleven IDPs with only PFG NMR measurements.

Name	Length	R_{h} [nm]
Aβ*	40	1.44[7]
SBD	61	2.56 ± 0.07[8]
CTL9-I98A**	92	2.17[9]
Sml1*	105	2.34 ± 0.1[10]
TC1	112	2.65 ± 0.05[11]
A2*	155	2.89 ± 0.03[12]
FUS*	163	3.32 ± 0.04[12]
CFTR R region	189	3.2 ± 0.1[13,14]
RYBP	234	3.95 ± 0.02[15]
Ddx4*	236	3.16[16]
3D7-6H MSP2	237	3.43[17]

* Not derived with dioxane as reference molecule.

** Cold denatured state.

References

(1) Sawle, L.; Ghosh, K. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *The Journal of Chemical Physics* **2015**, *143*, 085101.

(2) Manalastas-Cantos, K.; Konarev, P. V.; Hajizadeh, N. R.; Kikhney, A. G.; Petoukhov, M. V.; Molodenskiy, D. S.; Panjkovich, A.; Mertens, H. D. T.; Gruzinov, A.; Borges, C.; Jeffries, C. M.; Svergun, D. I.; Franke, D. *ATSAS 3.0*: expanded functionality and new tools for small-angle scattering data analysis. *Journal of Applied Crystallography* **2021**, *54*, 343–355.

(3) Zheng, W.; Best, R. B. An Extended Guinier Analysis for Intrinsically Disordered Proteins. *Journal of Molecular Biology* **2018**, *430*, 2540–2553, Intrinsically Disordered Proteins.

(4) Riback, J. A.; Bowman, M. A.; Zmyslowski, A. M.; Knoverek, C. R.; Jumper, J. M.; Hinshaw, J. R.; Kaye, E. B.; Freed, K. F.; Clark, P. L.; Sosnick, T. R. Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science* **2017**, *358*, 238–241.

(5) Choy, W.-Y.; Mulder, F. A.; Crowhurst, K. A.; Muhandiram, D.; Millett, I. S.; Doniach, S.; Forman-Kay, J. D.; Kay, L. E. Distribution of molecular size within an unfolded state ensemble using small-angle X-ray scattering and pulse field gradient NMR techniques. *Journal of Molecular Biology* **2002**, *316*, 101–112.

(6) Oono, Y.; Kohmoto, M. Renormalization group theory of transport properties of polymer solutions. I. Dilute solutions. *The Journal of Chemical Physics* **1983**, *78*, 520–528.

(7) Danielsson, J.; Jarvet, J.; Damberg, P.; Gräslund, A. Translational diffusion measured by PFG-NMR on full length and fragments of the Alzheimer Aβ(1–40) peptide. Deter-

mination of hydrodynamic radii of random coil peptides of varying length. *Magnetic Resonance in Chemistry* **2002**, *40*, S89–S97.

(8) Chong, P. A.; Ozdamar, B.; Wrana, J. L.; Forman-Kay, J. D. Disorder in a Target for the Smad2 Mad Homology 2 Domain and Its Implications for Binding and Specificity*. *Journal of Biological Chemistry* **2004**, *279*, 40707–40714.

(9) Shan, B.; McClendon, S.; Rospigliosi, C.; Eliezer, D.; Raleigh, D. P. The Cold Denatured State of the C-terminal Domain of Protein L9 Is Compact and Contains Both Native and Non-native Structure. *Journal of the American Chemical Society* **2010**, *132*, 4669–4677, PMID: 20225821.

(10) Danielsson, J.; Liljedahl, L.; Bárány-Wallje, E.; Sønderby, P.; Kristensen, L. H.; Martinez-Yamout, M. A.; Dyson, H. J.; Wright, P. E.; Poulsen, F. M.; Mäler, L.; Gräslund, A.; Kragelund, B. B. The Intrinsically Disordered RNR Inhibitor Sml1 Is a Dynamic Dimer. *Biochemistry* **2008**, *47*, 13428–13437.

(11) Gall, C.; Xu, H.; Brickenden, A.; Ai, X.; Choy, W. Y. The intrinsically disordered TC-1 interacts with Chibby via regions with high helical propensity. *Protein Science* **2007**, *16*, 2510–2518.

(12) Ryan, V. H.; Dignon, G. L.; Zerze, G. H.; Chabata, C. V.; Silva, R.; Conicella, A. E.; Amaya, J.; Burke, K. A.; Mittal, J.; Fawzi, N. L. Mechanistic View of hnRNPA2 Low-Complexity Domain Structure, Interactions, and Phase Separation Altered by Mutation and Arginine Methylation. *Molecular Cell* **2018**, *69*, 465–479.e7.

(13) Baker, J. M. R. Structural Characterization and Interactons of the CFTR Regulatory Region. *PhD Thesis). Department of Biochemistry. University of Toronto, Toronto.* **2009**,

(14) Marsh, J. A.; Forman-Kay, J. D. Sequence determinants of compaction in intrinsically disordered proteins. *Biophysical Journal* **2010**, *98*, 2383–2390.

(15) Neira, J. L.; Román-Trufero, M.; Contreras, L. M.; Prieto, J.; Singh, G.; Barrera, F. N.; Renart, M. L.; Vidal, M. The Transcriptional Repressor RYBP Is a Natively Unfolded Protein Which Folds upon Binding to DNA. *Biochemistry* **2009**, *48*, 1348–1360, PMID: 19170609.

(16) Brady, J. P.; Farber, P. J.; Sekhar, A.; Lin, Y.-H.; Huang, R.; Bah, A.; Nott, T. J.; Chan, H. S.; Baldwin, A. J.; Forman-Kay, J. D.; Kay, L. E. Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation. *Proceedings of the National Academy of Sciences* **2017**, *114*, E8194–E8203.

(17) Zhang, X.; Perugini, M. A.; Yao, S.; Adda, C. G.; Murphy, V. J.; Low, A.; Anders, R. F.; Norton, R. S. Solution Conformation, Backbone Dynamics and Lipid Interactions of the Intrinsically Unstructured Malaria Surface Protein MSP2. *Journal of Molecular Biology* **2008**, *379*, 105–121.