Article

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], 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
[*]Correspondence: lindorff@bio.ku.dk, bbk@bio.ku.dk

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 ten 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 modelling 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 reports 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 modelling 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. We here 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.

1 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 modelling 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 (MD) 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).

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 modelling 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 non-uniform 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):

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

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:

$$D_t = \frac{k_b T}{6\pi\eta R_h} \tag{2}$$

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:

$$R_h(\text{protein}) = \frac{D_t(\text{protein})}{D_t(\text{reference})} R_h(\text{reference}) \tag{3}$$

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 (31–36). Despite this, differences of ca. 20% between the results provided by different forward models have been observed (30, 37).

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 modelling, 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 (35), we assume that these differences are small and would vary between different proteins. We thus used the SAXS-refined conformational ensembles as input to different forward models for R_h and compared the results with experiments. As a data set for benchmarking the forward models, we chose ten IDPs with varying lengths (24–441 residues) and amino acid composition 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.

2 MATERIALS AND METHODS

2.1 Protein purifications

2.1.1 The growth hormone receptor intracellular domain (GHR-ICD)

For SAXS experiments, the GHR-ICD (residues 270-620) was expressed and purified as described in (38). For NMR experiments, GHR-ICD was expressed as a His_6-SUMO fusion protein (His_6-SUMO-GHR-ICD) in E. coli BL21(DE3) cells, transformed by heat shock transformation. 1 L of LB-media supplemented with 50 μg μL^{-1} kanamycin was inoculated with a preculture, grown at 37 °C. At OD_{600}=0.6-0.8, expression was induced by addition of 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) and grown for four hours. 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 (PMSF), 1 tablet ethylenediaminetetraacetic acid (EDTA)-free protease inhibitor (Roche diagnostics, GmbH), pH

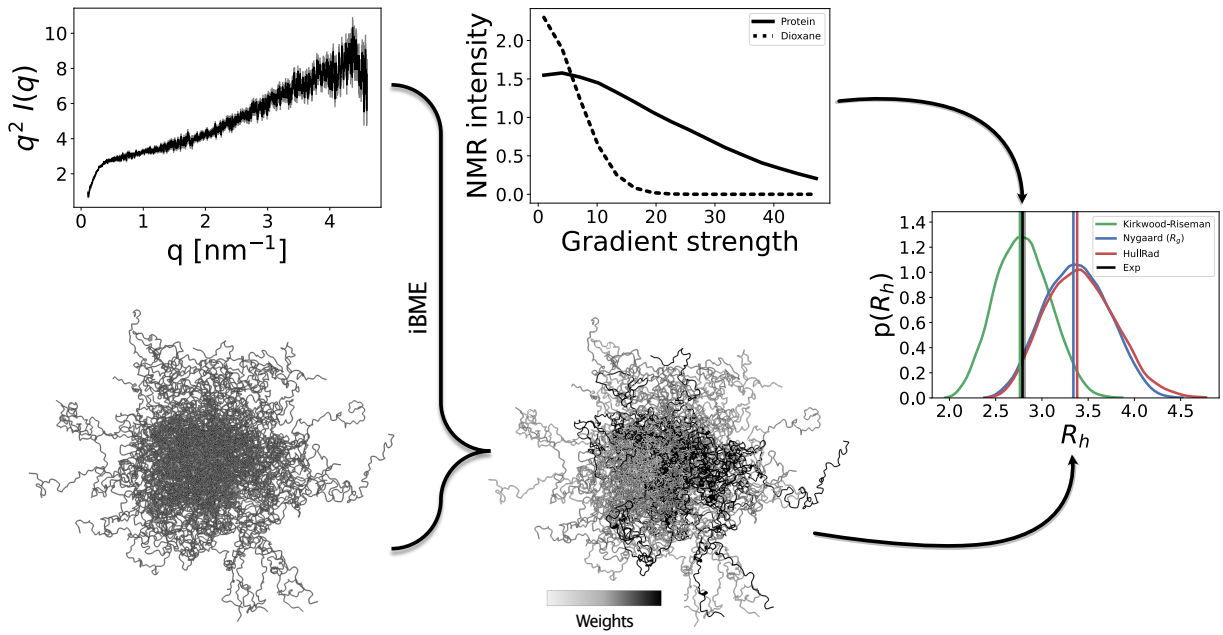

Figure 1: Overview of the approach. For each of the ten proteins, we computationally generate a conformational ensemble and optimize the weights of each conformer to get a reweighted ensemble consistent with SAXS data. Then we compute the R_h from the reweighted ensemble with different forward models and compare the values to the R_h determined experimentally by PFG NMR experiments.

8), and lysed by a French pressure cell disrupter (Constant Systems Ltd MC Cell Disrupter) at 20 kPsi. Lysate was cleared by centrifugation at 20000 g at 4 °C for 20 min. The supernatant was incubated with 2 mL Ni-NTA resin (GE Healthcare), equilibrated with buffer A (50 mM Tris-HCl, 150 mM NaCl, 10 mM imidazole, pH 8) for 1 hour 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 flow-through was re-incubated 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_6-SUMO tag was separated from GHR-ICD by incubating the sample with 2 mL Ni-NTA resin for 1 hour. The flow through was collected and used for further purification by reversed-phase chromatography using a Resource RPC column (GE Healthcare), equilibrated in MQ with 0.08% TFA (v/v),and eluted with a linear gradient from 0-100% of 70% acetonitrile (v/v), 0.1% TFA (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, 10x excess of DTT, 2% (v/v) glycerol as described, with protein concentration in the range from 1 to 6 mg mL^{-1} (39).

2.1.2 The human sodium-proton exchanger 6 intracellular distal domain (NHE6cmdd)

A modified pET-24b vector with an N-terminal His_6-SUMO tag was inserted with the 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, IPTG 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 Ltd MC Cell Disrupter). The lysate was centrifuged at 4 °C and 20000 g for 30 min and the supernatant applied to a gravity flow column with 4 mL pre-equilibrated 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 pre-soaked dialysis bag with a 3.5 kDa cut-off, 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 was collected. containing NHE6cmdd. The NHE6cmdd was concentrated to < 2 mL by centrifuging with 3 kDa cut-off 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, pre-equilibrated with 50 mM NH_4HCO_3, pH 7.8. A 0-100% linear gradient (20 column volumes) of 50 mM NH_5HCO_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.

2.1.3 Prothymosin-α

Prothymosin-α (ProTα) was produced and purified as described in (40), where also PFG NMR diffusion experiments are reported. SAXS intensities were recorded in 1x TBSK (10 mM Tris, 0.1 mM EDTA, 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 absorbs 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 1x 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 (41).

2.1.4 α-Synuclein

α-Synuclein (αSyn) was produced and purified as described in (42). 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 is from (43).

2.1.5 ANAC046$_{172\text{-}338}$

The disordered region (residues 172-338) of the arabidopsis NAC (no apical meristem [NAM], ATAF1/2, and cup-shaped cotyledon [CUC2]) transcription factor, ANAC046, was produced and purified as described in (42). 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}.

2.1.6 Dss1

Deleted in split hand/split foot 1 protein (Dss1) from S. pombe (44) was produced and purified as in (42) in the presence of 5 mM βME. NMR experiments were recorded in Tris buffer (20 mM Tris, 150 mM NaCl, 5 mM DTT, 10% D_2O, 0.25 mM DSS, 0.02% dioxane, 0.02% NaN_3) 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}.

2.2 Diffusion Ordered NMR Spectroscopy

Translational diffusion constants for each protein (50 μM-150 μM) and the internal reference were determined by fitting peak intensity decays within the methyl region centred around 0.85-0.9 ppm from diffusion ordered spectroscopy (DOSY) 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, 32 scans for αSyn) 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 = 26752 rad s^{-1} $Gauss^{-1}$) with a diffusion time (Δ) of 200 ms (299.9 ms for GHR-ICD) and gradient length (δ) of 3 ms, except for NHE6cmdd where this was 2 ms. Diffusion constants were fitted in Dynamics Center v2.5.6 (Bruker) 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.

2.3 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). 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 298K. 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.

2.4 Conformational ensembles

We generated conformational ensembles with two distinct methods, specifically Flexible-Meccano (here after FM) (49) and Langevin simulations with the CALVADOS (Coarse-graining Approach to Liquid-liquid phase separation Via an Automated Data-driven Optimisation Scheme) M1 parameters for a C_α-based coarse-grained model (50).

FM generates conformations for the backbone atoms of IDPs sampling from backbone dihedrals 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 20000 frames per simulation. Temperature 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 prior to SAXS calculations using PULCHRA (52) with default settings.

2.5 SAXS calculations

We calculated SAXS intensities using Pepsi-SAXS (53) as recently described (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 $1.025 \times r_m$, where r_m is the average atomic radius of the protein).

2.6 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, 55):

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

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 $\phi_{eff} = \exp(S_{rel})$ 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 artefacts in the reweighted ensemble (3). By scanning different values for θ and plotting χ_r^2 vs 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^2_{r}. In this approach, the χ^2_{r} in the $\mathcal{L}$ functional is:

$$\chi^2_{\mathrm{r}}(\omega_1 \cdots \omega_n, s, cst) = \frac{1}{m} \sum_{q}^{m} \frac{[(s \sum_j^n \omega_j I_q(x_j) + cst) - I_q^{\mathrm{EXP}}]^2}{\sigma^2_{\mathrm{BIFT},q}} \tag{5}$$

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 (56). 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 $\chi^2_{r,\mathrm{BIFT}}$ 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 $\sigma_{\mathrm{BIFT},q} = \sigma_q \sqrt{\chi^2_{r,\mathrm{BIFT}}}$. This procedure enabled a more direct comparison of χ^2_r values from different systems.

2.7 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_{g,C_α}) and the R_{h}:

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

   where the fitting paramters are α_1=0.216 Å^{-1}, α_2=4.06 Å and α_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.

2. The HULLRAD algorithm (29) that uses the convex hull method to predict hydrodynamic properties of proteins. The R_{h} computed with HULLRAD will here after be referred as $R_{\mathrm{h}}^{\mathrm{HR}}$.

3. The Kirkwood-Riseman equation (28, 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.

4. The linear fit proposed by Nygaard et al. (30) to approximate the R_{h} of HYDROPRO from $R_{\mathrm{h}}^{\mathrm{KR}}$: $R_{\mathrm{h}}^{\mathrm{Nyg\text{-}KR}} = 1.186 R_{\mathrm{h}}^{\mathrm{KR}} + 1.03$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 $\langle R_{\mathrm{h}} \rangle = \frac{1}{n^{-1} \sum_i^n (1/R_{\mathrm{h},i})}$ (31, 35). 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 $\exp(-R_{\mathrm{h}}^{-1})$ (Eq. 1 and 3). The exponential transformation can be omitted because it does not change the calculated average (31). For simplicity, we will here after 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 ten 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} (ca. 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.

3 RESULTS AND DISCUSSION

3.1 Proteins and experimental measurements

We collected a data set consisting of ten IDPs of different lengths (spanning from 24 to 441 residues) and sequence features (NCPR, 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 specify that, although different approaches exist to extract the R_{g} from a SAXS profile (58–60) (Table S2), we do not build our ensembles using the SAXS-derived R_{g}, rather we use the SAXS intensities themselves. To avoid discrepancies related to the dimensions of IDPs being influenced by experimental conditions (for 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). Also, buffers used for SAXS often contain glycerol to limit the radiation damage. Glycerol is not used for PFG NMR experiments. The average of the ratio R_g/R_h for the ten proteins is 1.2 (Table S2), in line with expectations from disordered Gaussian chains (61).

Table 1: Datasets used, consisting of ten IDPs with SAXS and PFG NMR measurements.

Name	Length	R_h [nm]	SAXS
Hst5	24	1.28 ± 0.02 (62)	(62)
RS	24	1.19 ± 0.01 (63)	(64)
Dss1	71	1.70 ± 0.06	This study
Sic1	90	2.15 ± 0.1 (65)	(66)
ProTα	111	2.89 ± 0.08 (40)	This study
NHE6cmdd	116	2.67 ± 0.02	This study
αSyn	140	2.79 ± 0.03	(43)
ANAC046	167	3.04 ± 0.01	This study
GHR-ICD	351	5.08 ± 0.02	(39)
Tau	441	5.40 ± 0.2 (67)	(68)

We measured SAXS data for Dss1, ProTα, NHE6cmdd and ANAC046 at different protein concentration (Fig. S1). We then inspected the resulting intensities as a function of the scattering angle to check for signs of aggregation or inter-particle 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).

3.2 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, 35), 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 ten proteins using both Flexible-Meccano (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, 69–71). 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, 72). Moreover, issues related to fitting free-parameters describing hydration layer and excluded volume in implicit-solvent based SAXS forward models have recently been addressed also in the context of IDPs (54, 73). We therefore calculated SAXS data from the conformational ensembles and compared these to experiments (Fig. 3, partially transparent bars). For both FM and CALVADOS we found very good agreements ($\chi^2_r \approx 1$) for nearly all proteins. The biggest outlier was the highly charged protein ProTα, where the FM ensemble did not provide as good a fit to the SAXS data (Fig. 3 and S3). 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 Baysian/Maximum Entropy (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 for 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 analysed the effect of the different priors (FM vs. CALVADOS) and reweighting by examining the distribution of the radius of gyration (Fig. 4). Overall, 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 did not observe any substantial change in the R_g distributions due to reweighting, in particular for the ensembles of ANAC046 and Tau generated with CALVADOS (Fig. 4). This is because the fit to the small-angle region of the SAXS profile was good in all cases for the CALVADOS prior, indicating that this prior is highly efficient in reproducing the average chain dimensions (Fig. S4). We note that this is likely explained—at least in part—by the fact that CALVADOS was parameterized to reproduce R_g for IDPs. Instead, in these cases we observed greater deviations at wider angles, that are responsible for the higher χ^2_r of ANAC046 and Tau. Nevertheless, these deviations were adjusted by reweighting

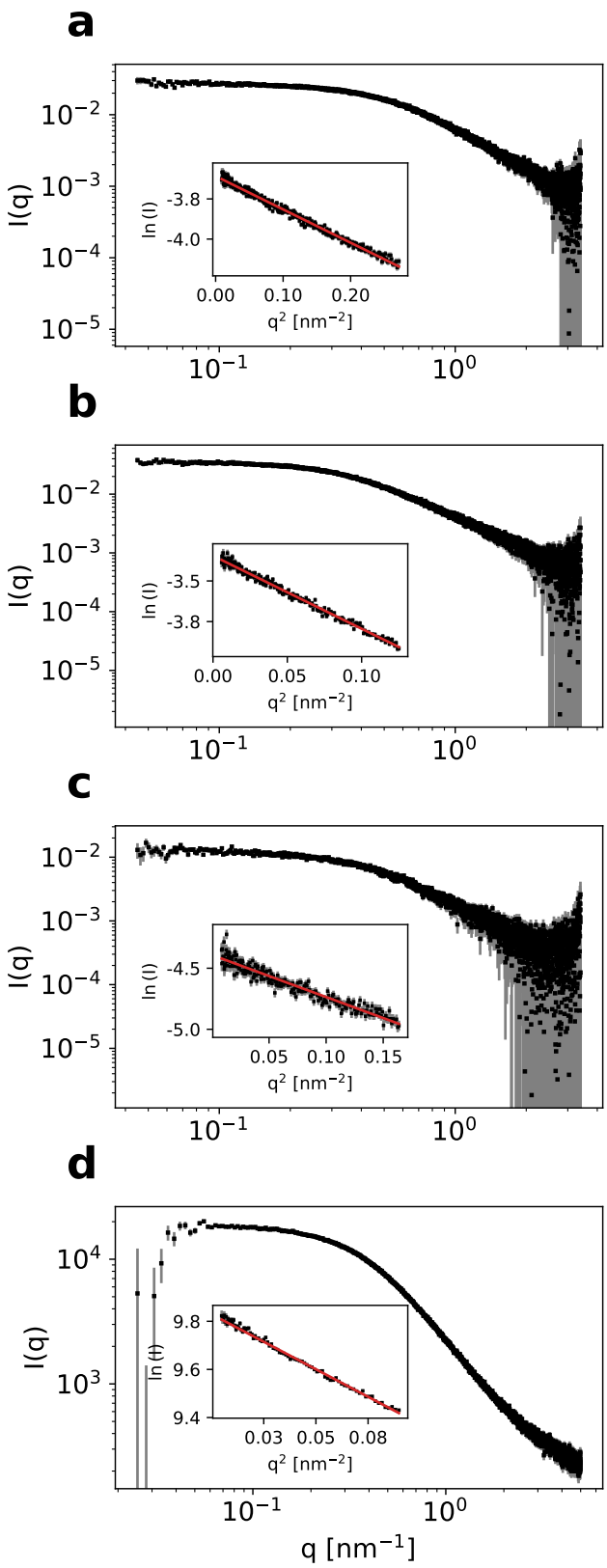

Figure 2: Experimental SAXS profiles for (a) Dss1, (b) ProTα, (c) NHE6cmdd, (d) ANAC046. Experimental intensities are shown as black squares with grey error bars. The inserts show the linear fit (red line) in the Guinier region (identified with the *autorg* tool of the ATSAS package (41)) for each profile.

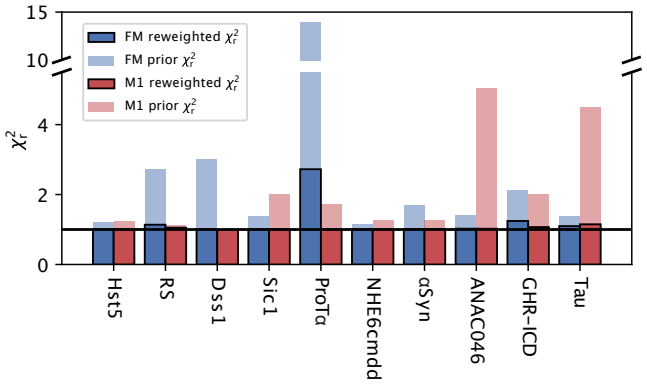

Figure 3: Results of reweighting the conformational ensembles generated with FM and CALVADOS against SAXS data. The χ_r^2 of the prior ensembles (before reweighting) are shown as partially transparent bars, while the χ_r^2 of the reweighted ensembles are shown as solid bars (blue for FM and red for CALVADOS). The horizontal black line delineates χ_r^2 =1, that, given the use of corrected errors for experimental SAXS intensities (see Materials and Methods), denotes reweighted ensembles in good agreement with the experimental data and devoid of overfitting.

the CALVADOS ensembles (Fig. S4).

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.

3.3 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 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 ten proteins, and find the Kirkwood-Riseman equation provided the best agreement with experiments (Fig. 5 and S5), with a χ^2 of 286. Indeed, for all but the two longest proteins (see below), the Kirkwood-Riseman equation gave very good agreement with experiments.

The other three forward models (the two equations described by Nygaard et al. and HullRad) gave χ^2 values that are three to four 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 (Fig. 5 and S5), 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 are 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 dataset and shows a sequence-length-dependent trend (Fig. 5 and S5).

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 larger than experiments. This picture, however, is different when looking at the two longest proteins, GHR-ICD (351 residues) and Tau (441 residues) (Fig. 5 and S5). For these two proteins the Kirkwood-Riseman equation apparently underestimates R_h, whereas the other three models give values closer to experiments. For NHE6cmdd (116 residues) and ANAC046 (167 residues) the R_h from PFG NMR lies in the middle of what the different forward models predict. The Nygaard equations, that are derived from HYDROPRO, and HullRad may be less precise because the models themselves are trained to predict the R_h of globular, folded proteins and, as we discuss in the Introduction, this is a common issue when using these forward models for IDPs. The Kirkwood-Riseman equation instead was developed in the context of theoretical studies on the hydrodynamic properties of disordered polymer chains. The behaviour of these chains is more similar to that of IDPs compared to 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

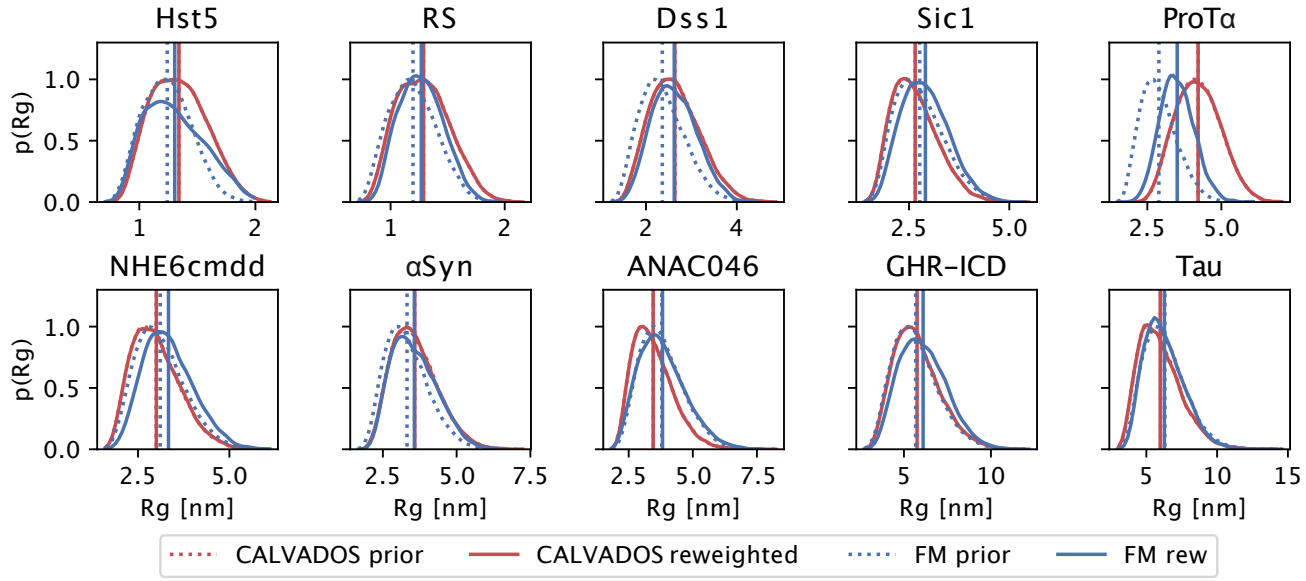

Figure 4: Probability distributions of the R_g calculated from the ensembles generated by CALVADOS (in red) and FM (in blue), both before (dotted lines) and after (solid lines) reweighting the ensembles against SAXS data.

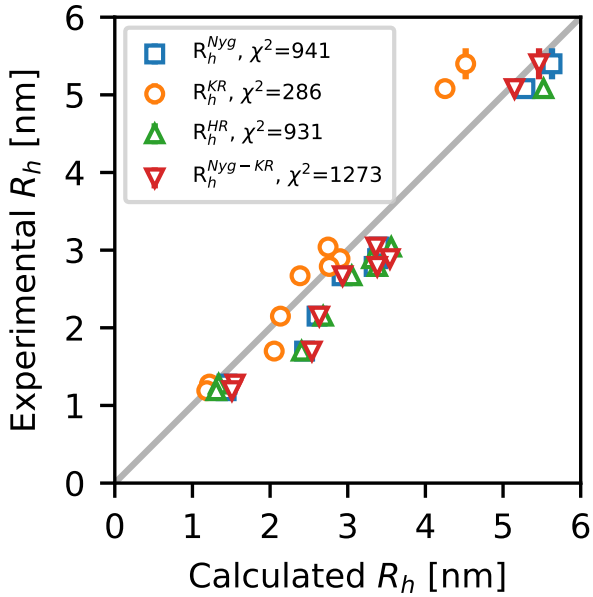

Figure 5: Ensemble-averaged R_h values calculated from the SAXS-reweighted CALVADOS ensembles, compared to the R_h determined by PFG NMR diffusion (in black). We tested four approaches to calculate the R_h from atomic coordinate: the R_g-dependent Nygaard equation (R_h^{Nyg}, in blue), the Kirkwood-Riseman 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).

disordered chains (30, 35, 61, 74), 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 to 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 proteins measured in this study) gave a similar picture as when comparing the R_h (Fig. S7).

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. S6) 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. S8).

Although the Kirkwood-Riseman equation provides a better overall agreement with PFG NMR diffusion experiments looking at the ten proteins, it appears that for the two longer chains the other three models perform better. The results for NHE6cmdd (116 residues) and ANAC046 (167 residues) suggest that the deviation between R_h^{KR} and experiments smoothly increases with chain length. In the SI we show that an empirical sequence-length-dependent correction can be fitted to the Kirkwood-Riseman equation to account for the observed discrepancies and determines a reduction of the χ^2 from 286 to 157 (Fig. S5 and S9). Nevertheless, we suggest that additional measurements should be made for long IDPs to probe the robustness of this correction.

CONCLUSIONS

Reliable forward models to compare conformational ensembles and biophysical measurements of IDPs are important both in integrative modelling 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 ten 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. We describe a sequence-length dependent correction to the Kirkwood-Riseman equation that fits the data better, but at this stage 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, 35, 37). 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; for this reason we have also focused only on measurements of R_h from PFG NMR. 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. 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 (75). 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 (62).

AUTHOR CONTRIBUTIONS

F.P, B.B.K. and K.L.-L. designed the study. E.A.N., P.S., E.T. 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 acknowledge the use of computational resources from the corefacility 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. 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 (#NNF18OC0033926 to B.B.K.). E.A.N has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 101023654.

REFERENCES

1. Wright, P. E., and H. J. Dyson, 2015. Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology* 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. *Biochemical Society Transactions* 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, New York, NY, 219–240. https://doi.org/10.1007/978-1-0716-0270-6_15.

4. Orioli, S., A. H. Larsen, S. Bottaro, and K. Lindorff-Larsen, 2020. Chapter Three - How to learn from inconsistencies: Integrating molecular simulations with experimental data. *In* B. Strodel, and B. Barz, editors, Computational Approaches for Understanding Dynamical Systems: Protein Folding and Assembly, Academic Press, volume 170 of *Progress in Molecular Biology and Translational Science*, 123–176. https://www.sciencedirect.com/science/article/pii/S1877117319302121.

5. Bonomi, M., C. Camilloni, A. Cavalli, and M. Vendruscolo, 2016. Metainference: A Bayesian inference method for heterogeneous systems. *Science Advances* 2:e1501177. https://www.science.org/doi/abs/10.1126/sciadv.1501177.

6. Hummer, G., and J. Köfinger, 2015. Bayesian ensemble refinement by replica simulations and reweighting. *The Journal of Chemical Physics* 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, P. Timmins, D. Marion, R. W. H. Ruigrok, and M. Blackledge, 2005. A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proceedings of the National Academy of Sciences* 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. *Proceedings of the National Academy of Sciences* 97:8868–8873. https://www.pnas.org/doi/abs/10.1073/pnas.160259697.

10. Martin, E. W., A. S. Holehouse, I. Peran, M. Farag, J. J. Incicco, A. Bremer, C. R. Grace, A. Soranno, R. V. Pappu, and T. Mittag, 2020. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science* 367:694–699. https://www.science.org/doi/abs/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. *Biophysical Journal* 112:2043–2046. https://www.sciencedirect.com/science/article/pii/S000634951730437X.

12. Thomasen, F. E., F. Pesce, M. A. Roesgaard, G. Tesei, and K. Lindorff-Larsen, 2022. Improving Martini 3 for Disordered and Multidomain Proteins. *Journal of Chemical Theory and Computation* 18:2033–2041. https://doi.org/10.1021/acs.jctc.1c01042, pMID: 35377637.

13. Henriques, J., C. Cragnell, and M. Skepö, 2015. Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment. *Journal of Chemical Theory and Computation* 11:3420–3431. https://doi.org/10.1021/ct501178z, pMID: 26575776.

14. Palazzesi, F., M. K. Prakash, M. Bonomi, and A. Barducci, 2015. Accuracy of Current All-Atom Force-Fields in Modeling Protein Disordered States. *Journal of Chemical Theory and Computation* 11:2–7. https://doi.org/10.1021/ct500718s, pMID: 26574197.

15. Rauscher, S., V. Gapsys, M. J. Gajda, M. Zweckstetter, B. L. de Groot, and H. Grubmüller, 2015. Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *Journal of Chemical Theory and Computation* 11:5513–5524. https://doi.org/10.1021/acs.jctc.5b00736, pMID: 26574339.

16. Piana, S., A. G. Donchev, P. Robustelli, and D. E. Shaw, 2015. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *The Journal of Physical Chemistry B* 119:5113–5123. https://doi.org/10.1021/jp508971m, pMID: 25764013.

17. Best, R. B., W. Zheng, and J. Mittal, 2014. Balanced Protein–Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *Journal of Chemical Theory and Computation* 10:5113–5124. https://doi.org/10.1021/ct500569b, pMID: 25400522.

18. Robustelli, P., S. Piana, and D. E. Shaw, 2018. Developing a molecular dynamics force field for both folded and disordered protein states. *Proceedings of the National Academy of Sciences* 115:E4758–E4766. https://www.pnas.org/doi/abs/10.1073/pnas.1800690115.

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

20. Stejskal, E. O., and J. E. Tanner, 1965. Spin Diffusion Measurements: Spin Echoes in the Presence of a Time-Dependent Field Gradient. *The Journal of Chemical Physics* 42:288–292. http://aip.scitation.org/doi/10.1063/1.1695690.

21. Rigler, R., Ü. Mets, J. Widengren, and P. Kask, 1993. Fluorescence correlation spectroscopy with high count rate and low background: analysis of translational diffusion. *European Biophysics Journal* 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. *Biophysical Reviews* 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. *Journal of Molecular Biology* 433:167196. https://www.sciencedirect.com/science/article/pii/S0022283621004290, from Protein Sequence to Structure at Warp Speed: How Alphafold Impacts Biology.

24. Wilkins, D. K., S. B. Grimshaw, V. Receveur, C. M. Dobson, J. A. Jones, and 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, pMID: 10600103.

25. Kärger, J., H. Pfeifer, and W. Heink, 1988. Principles and Application of Self-Diffusion Measurements by Nuclear Magnetic Resonance. Academic Press, volume 12 of *Advances in Magnetic and Optical Resonance*, 1–89. 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, New York, NY, 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. *Biophysical Journal* 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. *The Journal of Chemical Physics* 16:565–573. http://aip.scitation.org/doi/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. *Biophysical Journal* 114:856–869. https://www.sciencedirect.com/science/article/pii/S0006349518300651.

30. Nygaard, M., B. B. Kragelund, E. Papaleo, and K. Lindorff-Larsen, 2017. An Efficient Method for Estimating the Hydrodynamic Radius of Disordered Protein Conformations. *Biophysical Journal* 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, New York, NY, 429–445. https://doi.org/10.1007/978-1-0716-0524-0_21.

32. Naullage, P. M., M. Haghighatlari, A. Namini, J. M. C. Teixeira, J. Li, O. Zhang, C. C. Gradinaru, J. D. Forman-Kay, and T. Head-Gordon, 2022. Protein Dynamics to Define and Refine Disordered Protein Ensembles. *The Journal of Physical Chemistry B* 126:1885–1894. https://doi.org/10.1021/acs.jpcb.1c10925, pMID: 35213160.

33. Lincoff, J., M. Haghighatlari, M. Krzeminski, J. M. C. Teixeira, G.-N. W. Gomes, C. C. Gradinaru, J. D. Forman-Kay, and T. Head-Gordon, 2020. Extended experimental inferential structure determination method in determining the structural ensembles of disordered protein states. *Communications Chemistry* 3:74. http://www.nature.com/articles/s42004-020-0323-0.

34. Rauscher, S., V. Gapsys, M. J. Gajda, M. Zweckstetter, B. L. de Groot, and H. Grubmüller, 2015. Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *Journal of Chemical Theory and Computation* 11:5513–5524. https://doi.org/10.1021/acs.jctc.5b00736, pMID: 26574339.

35. Choy, W.-Y., F. A. Mulder, K. A. Crowhurst, D. Muhandiram, I. S. Millett, S. Doniach, J. D. Forman-Kay, and 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. *Journal of Molecular Biology* 316:101–112. https://www.sciencedirect.com/science/article/pii/S0022283601953288.

36. Lindorff-Larsen, K., S. Kristjansdottir, K. Teilum, W. Fieber, C. M. Dobson, F. M. Poulsen, and M. Vendruscolo, 2004. Determination of an Ensemble of Structures Representing the Denatured State of the Bovine Acyl-Coenzyme A Binding Protein. *Journal of the American Chemical Society* 126:3291–3299. https://doi.org/10.1021/ja039250g, pMID: 15012160.

37. Gomes, G.-N. W., M. Krzeminski, A. Namini, E. W. Martin, T. Mittag, T. Head-Gordon, J. D. Forman-Kay, and C. C. Gradinaru, 2020. Conformational Ensembles of an Intrinsically Disordered Protein Consistent with NMR, SAXS, and Single-Molecule FRET. *Journal of the American Chemical Society* 142:15697–15710. https://doi.org/10.1021/jacs.0c02088, pMID: 32840111.

38. Haxholm, G., L. Nikolajsen, J. Olsen, J. Fredsted, F. Larsen, V. Goffin, S. Pedersen, A. Brooks, M. Waters, and B. Kragelund, 2015. Intrinsically disordered cytoplasmic domains of two cytokine receptors mediate conserved interactions with membranes. *Biochemical Journal* 468:495–506. https://doi.org/10.1042/BJ20141243.

39. Seiffert, P., K. Bugge, M. Nygaard, G. W. Haxholm, J. H. Martinsen, M. N. Pedersen, L. Arleth, W. Boomsma, and B. B. Kragelund, 2020. Orchestration of signaling by structural disorder in class 1 cytokine receptors. *Cell Communication and Signaling* 18:132. https://biosignaling.biomedcentral.com/articles/10.1186/s12964-020-00626-6.

40. Borgia, A., M. B. Borgia, K. Bugge, V. M. Kissling, P. O. Heidarsson, C. B. Fernandes, A. Sottini, A. Soranno, K. J. Buholzer, D. Nettels, B. B. Kragelund, R. B. Best, and B. Schuler, 2018. Extreme disorder in an ultrahigh-affinity protein complex. *Nature* 555:61–66. http://www.nature.com/articles/nature25762.

41. Manalastas-Cantos, K., P. V. Konarev, N. R. Hajizadeh, A. G. Kikhney, M. V. Petoukhov, D. S. Molodenskiy, A. Panjkovich, H. D. T. Mertens, A. Gruzinov, C. Borges, C. M. Jeffries, D. I. Svergun, and D. Franke, 2021. *ATSAS 3.0*: expanded functionality and new tools for small-angle scattering data analysis. *Journal of Applied Crystallography* 54:343–355. https://doi.org/10.1107/S1600576720013412.

42. Newcombe, E. A., C. B. Fernandes, J. E. Lundsgaard, I. Brakti, K. Lindorff-Larsen, A. E. Langkilde, K. Skriver, and B. B. Kragelund, 2021. Insight into Calcium-Binding Motifs of Intrinsically Disordered Proteins. *Biomolecules* 11. https://www.mdpi.com/2218-273X/11/8/1173.

43. Ahmed, M. C., L. K. Skaanning, A. Jussupow, E. A. Newcombe, B. B. Kragelund, C. Camilloni, A. E. Langkilde, and K. Lindorff-Larsen, 2021. Refinement of α-Synuclein Ensembles Against SAXS Data: Comparison of Force Fields and Methods. *Frontiers in Molecular Biosciences* 8. https://www.frontiersin.org/article/10.3389/fmolb.2021.654333.

44. Crackower, M. A., S. W. Scherer, J. M. Rommens, C.-C. Hui, P. Poorkaj, S. Soder, J. M. Cobben, L. Hudgins, J. P. Evans, and 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. *Human molecular genetics* 5:571–579.

45. Wu, D., A. Chen, and C. Johnson, 1995. An Improved Diffusion-Ordered Spectroscopy Experiment Incorporating Bipolar-Gradient Pulses. *Journal of Magnetic Resonance, Series A* 115:260–264. https://www.sciencedirect.com/science/article/pii/S106418588571176X.

46. Prestel, A., K. Bugge, L. Staby, R. Hendus-Altenburger, and B. B. Kragelund, 2018. Chapter Eight - Characterization of Dynamic IDP Complexes by NMR Spectroscopy. *In* E. Rhoades, editor, Intrinsically Disordered Proteins, Academic Press, volume 611 of *Methods in Enzymology*, 193–226. https://www.sciencedirect.com/science/article/pii/S0076687918303057.

47. Blanchet, C. E., A. Spilotros, F. Schwemmer, M. A. Graewert, A. Kikhney, C. M. Jeffries, D. Franke, D. Mark, R. Zengerle, F. Cipriani, S. Fiedler, M. Roessle, and D. I. Svergun, 2015. Versatile sample environments and automation for biological solution X-ray scattering experiments at the P12 beamline (PETRA III, DESY). *Journal of Applied Crystallography* 48:431–443. http://scripts.iucr.org/cgi-bin/paper?S160057671500254X.

48. Franke, D., M. V. Petoukhov, P. V. Konarev, A. Panjkovich, A. Tuukkanen, H. D. T. Mertens, A. G. Kikhney, N. R. Hajizadeh, J. M. Franklin, C. M. Jeffries, and D. I. Svergun, 2017. *ATSAS 2.8* : a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *Journal of Applied Crystallography* 50:1212–1225. http://scripts.iucr.org/cgi-bin/paper?S1600576717007786.

49. Ozenne, V., F. Bauer, L. Salmon, J.-r. Huang, M. R. Jensen, S. Segard, P. Bernadó, C. Charavay, and 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, R. Crehuet, and K. Lindorff-Larsen, 2021. Accurate model of liquid-liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. *Proceedings of the National Academy of Sciences* 118:e2111696118. https://www.pnas.org/doi/abs/10.1073/pnas.2111696118.

51. Eastman, P., J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande, 2017. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology* 13:1–17. 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. *Journal of Computational Chemistry* 29:1460–1465. https://onlinelibrary.wiley.com/doi/abs/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 Crystallographica Section D* 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. *Biophysical Journal* 120:5124–5135. https://www.sciencedirect.com/science/article/pii/S0006349521008286.

55. Hummer, G., and J. Köfinger, 2015. Bayesian ensemble refinement by replica simulations and reweighting. *The Journal of Chemical Physics* 143:243150. https://doi.org/10.1063/1.4937786.

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

57. Kirkwood, J. G., 1954. The general theory of irreversible processes in solutions of macromolecules. *Journal of Polymer Science* 12:1–14. https://onlinelibrary.wiley.com/doi/10.1002/pol.1954.120120102.

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

59. Riback, J. A., M. A. Bowman, A. M. Zmyslowski, C. R. Knoverek, J. M. Jumper, J. R. Hinshaw, E. B. Kaye, K. F. Freed, P. L. Clark, and T. R. Sosnick, 2017. Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science* 358:238–241. https://www.science.org/doi/abs/10.1126/science.aan5774.

60. Zheng, W., and R. B. Best, 2018. An Extended Guinier Analysis for Intrinsically Disordered Proteins. *Journal of Molecular Biology* 430:2540–2553. https://www.sciencedirect.com/science/article/pii/S0022283618301359, intrinsically Disordered Proteins.

61. Oono, Y., and M. Kohmoto, 1983. Renormalization group theory of transport properties of polymer solutions. I. Dilute solutions. *The Journal of Chemical Physics* 78:520–528. https://doi.org/10.1063/1.444477.

62. Jephthah, S., L. Staby, B. B. Kragelund, and M. Skepö, 2019. Temperature Dependence of Intrinsically Disordered Proteins in Simulations: What are We Missing? *Journal of Chemical Theory and Computation* 15:2672–2683. https://doi.org/10.1021/acs.jctc.8b01281, pMID: 30865820.

63. Xiang, S., V. Gapsys, H.-Y. Kim, S. Bessonov, H.-H. Hsiao, S. Möhlmann, V. Klaukien, R. Ficner, S. Becker, H. Urlaub, R. Lührmann, B. de Groot, and 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.

64. Rauscher, S., V. Gapsys, M. J. Gajda, M. Zweckstetter, B. L. de Groot, and H. Grubmüller, 2015. Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *Journal of Chemical Theory and Computation* 11:5513–5524. https://doi.org/10.1021/acs.jctc.5b00736, pMID: 26574339.

65. Mittag, T., S. Orlicky, W.-Y. Choy, X. Tang, H. Lin, F. Sicheri, L. E. Kay, M. Tyers, and J. D. Forman-Kay, 2008. Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. *Proceedings of the National Academy of Sciences* 105:17772–17777. https://www.pnas.org/doi/abs/10.1073/pnas.0809222105.

66. Gomes, G.-N. W., M. Krzeminski, A. Namini, E. W. Martin, T. Mittag, T. Head-Gordon, J. D. Forman-Kay, and C. C. Gradinaru, 2020. Conformational Ensembles of an Intrinsically Disordered Protein Consistent with NMR, SAXS, and Single-Molecule FRET. *Journal of the American Chemical Society* 142:15697–15710. https://doi.org/10.1021/jacs.0c02088, pMID: 32840111.

67. Mukrasch, M. D., S. Bibow, J. Korukottu, S. Jeganathan, J. Biernat, C. Griesinger, E. Mandelkow, and M. Zweckstetter, 2009. Structural Polymorphism of 441-Residue Tau at Single Residue Resolution. *PLoS Biology* 7:e1000034. https://dx.plos.org/10.1371/journal.pbio.1000034.

68. Mylonas, E., A. Hascher, P. Bernadó, M. Blackledge, E. Mandelkow, and 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, pMID: 18771286.

69. Jensen, M. R., P. R. Markwick, S. Meier, C. Griesinger, M. Zweckstetter, S. Grzesiek, P. Bernadó, and 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.

70. Wells, M., H. Tidow, T. J. Rutherford, P. Markwick, M. R. Jensen, E. Mylonas, D. I. Svergun, M. Blackledge, and A. R. Fersht, 2008. Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proceedings of the National Academy of Sciences* 105:5762–5767. https://www.pnas.org/content/105/15/5762.

71. Mukrasch, M. D., P. Markwick, J. Biernat, M. von Bergen, P. Bernadó, C. Griesinger, E. Mandelkow, M. Zweckstetter, and M. Blackledge, 2007. Highly Populated Turn Conformations in Natively Unfolded Tau Protein Identified from Residual Dipolar Couplings and Molecular Simulation. *Journal of the American Chemical Society* 129:5235–5243. https://doi.org/10.1021/ja0690159, pMID: 17385861.

72. 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. *Journal of Applied Crystallography* 28:768–773. https://doi.org/10.1107/S0021889895007047.

73. Henriques, J., L. Arleth, K. Lindorff-Larsen, and M. Skepö, 2018. On the Calculation of SAXS Profiles of Folded and Intrinsically Disordered Proteins from Computer Simulations. *Journal of Molecular Biology* 430:2521–2539. https://www.sciencedirect.com/science/article/pii/S0022283618301232, intrinsically Disordered Proteins.

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

75. Ramanujam, V., T. R. Alderson, I. Pritišanac, J. Ying, and A. Bax, 2020. Protein structural changes characterized by high-pressure, pulsed field gradient diffusion NMR spectroscopy. *Journal of Magnetic Resonance* 312:106701. https://www.sciencedirect.com/science/article/pii/S1090780720300197.