

# Revisiting the use of dioxane as a reference compound for determination of the hydrodynamic radius of proteins by pulsed field gradient NMR spectroscopy

*Emil E. Tranchant<sup>1,2</sup>, Francesco Pesce<sup>1</sup>, Nina L. Jacobsen<sup>1,2</sup>, Catarina B. Fernandes<sup>1,2</sup>,  
Birthe B. Kragelund<sup>1,2\*</sup>, Kresten Lindorff-Larsen<sup>1\*</sup>*

<sup>1</sup>*Structural Biology and NMR Laboratory, The Linderstrøm-Lang Centre for Protein Science, and*

<sup>2</sup>*REPIN, Department of Biology, University of Copenhagen, Copenhagen, Denmark*

\*corresponding authors: [lindorff@bio.ku.dk](mailto:lindorff@bio.ku.dk); [bbk@bio.ku.dk](mailto:bbk@bio.ku.dk)

## ABSTRACT

Measuring the compaction of a protein or complex is key to understand the interactions within and between biomolecules. Experimentally, protein compaction is often probed either by estimating the radius of gyration ( $R_g$ ) obtained from small-angle X-ray scattering (SAXS) experiments or the hydrodynamic radius ( $R_h$ ) obtained for example by pulsed field gradient nuclear magnetic resonance (PFG NMR) spectroscopy. PFG NMR experiments generally report on the translational diffusion coefficient, which in turn can be used to estimate  $R_h$  using an internal standard. Here, we examine the use of 1,4-dioxane as an internal NMR standard to account for sample viscosity and uncertainty about the gradient strength. Specifically, we revisit the basis for the commonly used reference value for the  $R_h$  of dioxane (2.12 Å) that is used to convert measured diffusion coefficients into a hydrodynamic radius. We follow the same approach that was used to establish the current reference value for the  $R_h$  by measuring SAXS and PFG NMR data for a set of seven different proteins and using these as standards. Our analysis shows that the current  $R_h$  reference value for 1,4-dioxane  $R_h$  (2.12 Å) is underestimated, and we instead suggest a new value of  $2.27 \text{ \AA} \pm 0.04 \text{ \AA}$ . Using this updated reference value results in a ~7% increase in  $R_h$  values for proteins whose hydrodynamic radius have been measured by PFG NMR. We discuss the implications for ensemble descriptions of intrinsically disordered proteins and evaluation of effect resulting from for example ligand binding, posttranslational modifications, or changes to the environment.

**Keywords:** IDP, ensemble, diffusion coefficient, 1,4-dioxane, hydrodynamic radius, radius of gyration, SAXS, lysozyme

## INTRODUCTION

Proteins are dynamic entities that exist in ensembles of states whose average properties vary depending on their sequence properties, context, and post-translational modifications. Folded proteins typically have a narrow distribution of conformations, whereas the structures of so-called intrinsically disordered proteins (IDPs) vary substantially across the ensemble (1). When characterizing the structure and interactions of proteins it is often advantageous to be able to probe the expansion of the protein or the size of the assembly it forms. Experimentally, this can for example be assessed via probing the radius of hydration,  $R_h$ , and the radius of gyration,  $R_g$  (2, 3).  $R_h$  is often probed by pulsed field gradient nuclear magnetic resonance (PFG NMR) spectroscopy (4), fluorescence correlation spectroscopy, dynamic light scattering or, with a lower resolution, by size exclusion chromatography.  $R_g$  is most commonly probed by small-angle X-ray scattering (SAXS) (2, 5). As a robust and accurate determination of these two parameters is critical, and current estimates of  $R_h$  using PFG NMR rely on assumptions that for decades have mostly been left unexamined, we wanted to revisit one key step when determining the  $R_h$  of a protein by PFG NMR.

In PFG NMR, the location of the protein is “encoded” via a spatial field gradient and makes it possible to probe the translational diffusion coefficient ( $D_t$ ) and in turn estimate the  $R_h$ . As PFG NMR reports on the conformationally averaged  $D_t$ , and thus average  $R_h$ , the technique is especially useful when studying IDPs and their interactions, where conformational ensemble compactness can provide structural information and interaction affinities (6, 7). In a standard PFG NMR one records a series of NMR spectra at varying gradient strengths, where the decay of the NMR peak intensities with increasing gradient strengths can be fitted to the Stejskal-Tanner equation (Eq. 1) (8), that relates  $D_t$  to the measured peak intensities:

$$I = I_0 e^{-g^2 \gamma^2 \delta^2 (\Delta - \frac{\delta}{3}) D_t} \quad (\text{Eq. 1})$$

Here,  $g$  is the gradient strength,  $\gamma$  is the gyromagnetic ratio,  $\delta$  is the length of the gradient, and  $\Delta$  is the diffusion time. Assuming that the resulting translational diffusion of the nuclei is equal to that of the parent molecule, and the molecule diffuses as a spherical entity, the  $R_h$  of the selected peak can then be derived from the Stokes-Einstein relation (Eq. 2).

$$R_h = \frac{k_B T}{6\pi\eta D_t} \quad (\text{Eq. 2})$$

where  $k_B$  is the Boltzmann constant,  $T$  the temperature, and  $\eta$  the solvent viscosity; Eq. 2 can also be used as a definition of  $R_h$ . Complications may, however, arise when determining the  $R_h$  this way. First, estimating  $D_t$  on an absolute scale requires careful calibration of the field gradient. Second, the solvent viscosity is sensitive to type of solvent, solute and sample conditions. Thus, the buffer composition, temperature, and protein concentration may all affect viscosity and should be carefully controlled. Furthermore, any added  $D_2O$  used to lock the NMR frequency must be corrected for as there is a slight difference in viscosity between  $H_2O$  and  $D_2O$  (7). As the solvent viscosity is difficult to measure and control precisely, deriving reliable  $R_h$  of a protein from a selected peak just from its translational diffusion alone is difficult and may be imprecise. Instead, one often uses an internal reference compound that is added to the NMR sample. By knowing the

$R_h$  of the reference compound (9), the ratio of the  $D_t$  between the reference compound and the protein can be used to estimate the  $R_h$  of the protein according to Eq. 3.

$$R_{h,protein} = \frac{D_{t,reference}}{D_{t,protein}} * R_{h,reference} \quad (\text{Eq. 3})$$

Using an internal reference removes the need for accurate calibration of the gradient and measurement of the viscosity. Often, the reference compound of choice is 1,4-dioxane. 1,4-dioxane provides a single, easily discernible,  $^1\text{H}$  NMR peak at approx. 3.75 ppm and has been found not to interact substantially with several proteins at experimentally used concentrations (10). While other reference compounds such as  $\alpha$ - or  $\beta$ -cyclodextrin are sometimes suggested in place of 1,4-dioxane (11), cyclodextrins have also been shown to interact with proteins (12, 13), just as they contribute with many more signals in the NMR spectrum overlapping with those of the proteins.

Using 1,4-dioxane as a viscosity reference requires that its  $R_h$  is known and that it is insensitive to environmental changes. Early use of 1,4-dioxane as a reference in PFG NMR established the  $R_h$  of the molecule to be 2.12 Å, and this value has since been used as a reference when using PFG NMR to determine the  $R_h$  of proteins (14). The reference value for dioxane was determined as described above, but instead using a protein molecule as reference. Specifically, Wilkins et al performed PFG NMR experiments on a solution of 1,4-dioxane and hen egg white lysozyme (HEWL), where instead of an unknown  $R_h$  of the protein in Eq. 3, the  $R_h$  of dioxane was unknown and the  $R_h$  of the HEWL set to 19.8 Å. This  $R_h$  value originated from an earlier study where batch SAXS experiments on HEWL provided an experimental  $R_g$  for natively folded HEWL of  $15.3 \pm 0.2$  Å (15). By assuming the ratio between  $R_g$  and  $R_h$ ,  $\rho$ , for a globular protein such as HEWL to be that of a solid sphere (16), the  $R_h$  of HEWL was obtained with the  $R_g$  from SAXS using equation 4:

$$\rho = \frac{R_g}{R_h} = (3/5)^{1/2} \quad (\text{Eq. 4})$$

Using this approach and experimental batch SAXS data of HEWL, the authors (14) estimated the  $R_h$  of natively folded HEWL to be 19.8 Å. The ratio of the translational diffusion coefficients between HEWL and 1,4-dioxane was estimated from the PFG NMR data to be 9.33 (17), leading to an  $R_h$  of 1,4-dioxane of 2.12 Å.

Recently, a community study addressed the reproducibility of SAXS experiments between proteins and between instruments (18). Generally, for the studied set of proteins, experimental reproducibility and subsequent consensus curves were observed. However, batch SAXS experiments on HEWL specifically showed significant variability between experiments. Considering that the original derivation of the  $R_h$  of 1,4-dioxane was directly derived based on a batch SAXS measurement of HEWL, any uncertainty in the  $R_g$  of HEWL would result in uncertainty in the 1,4-dioxane  $R_h$  of 2.12 Å and hence impact  $R_h$  measurements of proteins using PFG NMR and dioxane referencing.

With this possible uncertainty on the  $R_h$  of 1,4-dioxane in mind, we decided to revisit the foundation of this reference value by exploring a larger set of proteins, using the same approach

as originally done in (14). We estimate the average and uncertainty of the  $R_h$  of 1,4-dioxane using a set of seven folded proteins for which we measured both translational diffusion coefficients by PFG NMR and determined the  $R_g$  by batch SAXS measurements. By considering the quality of the recorded data, we find that the established  $R_h$  value of 1,4-dioxane is slightly underestimated. Our data suggests that it should be increased by approximately 7% compared to the previous value, and we propose an updated standard  $R_h$  value of  $2.27 \pm 0.04 \text{ \AA}$  of 1,4-dioxane, which would also result in a 7% increase in derived protein  $R_h$  using Eq. 4.

## MATERIALS AND METHODS

### Protein purifications

Protein samples used in this work were either prepared from bought lyophilized powder stocks or from frozen, pre-purified stocks. Proteins from purchased powder-stocks were equine myoglobin (Sigma-Aldrich), bovine ribonuclease A (RNaseA) (GE Healthcare), and HEWL (Sigma-Aldrich). Pre-purified proteins include S100A13, prolactin, ACBP<sub>Y73F</sub>, and 14-3-3  $\zeta$ . Purification of prolactin was performed as described in (19). Purification of ACBP<sub>Y73F</sub> was performed as described in (20).

For S100A13, *E. coli* cells (BL21 DE3) (Biolabs) were transformed with pET-24a plasmid coding for His<sub>6</sub>-SUMO S100A13 (UniProt Q99584) and grown in high salt LB-broth medium (Sigma Aldrich). Cells were grown until an OD<sub>600</sub> of 0.6-0.8 and expression induced with 0.1 mM isopropyl- $\beta$ -D-1-thiogalactopyranoside (IPTG). After 4h growth at 37 °C, cells were harvested by centrifugation at 5000 x g for 15 min and stored at -20 °C. Cells were lysed in 50 mM Tris pH 8.0, 150 mM NaCl, 2 mM CaCl<sub>2</sub> through a French press cell disrupter at 25,000 psi (Constant Systems Ltd.) followed by centrifugation at 20,000 x g where the clear lysate was subsequently loaded onto a 5 mL Ni-NTA Sepharose column (GE Healthcare) equilibrated with 50 mM Tris, pH 8.0, 150 mM NaCl, 2 mM CaCl<sub>2</sub>. His<sub>6</sub>-SUMO S100A13 was eluted with 50 mM Tris pH 8.0, 150 mM NaCl, 500 mM Imidazole followed by overnight dialysis where 0.1 mg His-tagged ULP1 and 1 mM DTT was added to cleave the SUMO-tag. The sample was purified further using a reverse Ni-NTA step removing non-cleaved protein, ULP1 and SUMO followed by removal of DNA on a 1 mL Heparin column with 50 mM Tris, pH 7.4, 2 mM CaCl<sub>2</sub> and 50 mM Tris pH 7.4, 1 M NaCl, 2 mM CaCl<sub>2</sub>. For a final step, the sample was run on a Superdex 75 10/300 (GE Healthcare) in 50 mM Tris pH 7.4, 150 mM NaCl, fractions concentrated and stored at -20 °C.

The 14-3-3 $\zeta$  protein was expressed from a modified pET-24a vector designed to encode an N-terminal His<sub>6</sub>-SUMO tag, which was to be cleaved using ubiquitin-like protein protease 1 (ULP1). The plasmid was transformed into NiCo21(DE3) competent *E. coli* cells (New England BioLabs) grown in LB medium containing 50  $\mu$ g/mL kanamycin and the fusion protein expression was induced with 0.5 mM IPTG for 4-5 hours before harvesting cells by centrifugation at 5000xg, 15 min, 4 °C. The pellet was lysed in lysis/equilibration buffer (20 mM Bis-Tris pH 7.2, 10 mM Imidazole, 150 mM NaCl, 5 mM  $\beta$ -mercaptoethanol (bME)) using a French pressure cell disrupter (25 kpsi; Constant Systems Ltd, Daventry, UK), and the lysate was cleared by centrifugation at 20,000xg for 45 min at 4 °C. The His<sub>6</sub>-SUMO-14-3-3 $\zeta$  fusion protein was purified by immobilized metal affinity chromatography (IMAC) using Ni Sepharose 6 Fast Flow resin (5 mL; GE Healthcare) with standard IMAC procedures of sample application, high salt (1 M NaCl) washing step and imidazole elution. The eluted sample was dialyzed towards 2 L of buffer A (20 mM Bis-Tris pH 6.5, 5 mM bME) before the sample was applied to a 1 mL HiTrap Heparin HP column (Cytiva). The

column was washed with 15 mL of buffer A before the fusion protein was eluted with a linear two-step gradient of 0-30% over 3 mL and 30-100% over 20 mL of buffer B (20 mM Bis-Tris pH 6.5, 1 M NaCl, 5 mM bME). The His<sub>6</sub>-SUMO tag was cleaved off by supplementing the sample with 0.1 mg ULP1 and 2 mM DTT for at least 3 hours. This sample was re-applied to the IMAC column to remove the His<sub>6</sub>-SUMO-tag and ULP1, and the flow through containing pure 14-3-3 $\zeta$  was collected.

#### *NMR and SAXS sample preparations*

To prepare samples from purified proteins or lyophilized protein stocks, each protein was applied to a size exclusion chromatography (SEC) column (Superdex 75 10/300 GL, Cytiva) in a 20 mM sodium phosphate buffer, pH 7.4, 150 mM NaCl mounted on an Äkta Purifier system. Each collected sample was then analysed by SDS-PAGE to verify the purity of the protein. The samples for PFG NMR experiment were prepared with the following protein and concentrations: HEWL 200  $\mu$ M, RnaseA 300  $\mu$ M, myoglobin 400  $\mu$ M, S100A13 150  $\mu$ M, ACBP<sub>Y73F</sub> 200  $\mu$ M, prolactin 300  $\mu$ M, and 14-3-3 150  $\mu$ M. DSS was added to a final concentration of 25  $\mu$ M, D<sub>2</sub>O to a final concentration of 10% (v/v), and 1,4-dioxane to a final concentration between 0.04-0.06% (v/v). Sample volumes were either 100  $\mu$ L, 350  $\mu$ L, or 500  $\mu$ L, depending on the use of 3 mm Shigemi, 5 mm Shigemi, or 5 mm glass single-use NMR tubes (Bruker), respectively. Samples for SAXS experiments were prepared by concentrating the proteins after SEC to multiple samples of 1, 2, and 3 mg/ml in the 20 mM sodium phosphate buffer also used in the SEC. A pure buffer solution was also prepared for buffer subtraction of the background scattering.

#### *Pulsed field gradient NMR Spectroscopy*

PFG NMR experiments were performed on a Bruker Avance III HD 600 MHz spectrometer equipped with a Bruker proton-optimized quadrupole resonance NMR ‘inverse’ QCI cryoprobe. Each PFG NMR experiment was preceded by a 1D <sup>1</sup>H-spectrum used for referencing the spectra to the DSS peak at 0 ppm. Translational diffusion coefficients of proteins were determined by fitting peak intensity decays in the methyl and methylene region (2.5-0.5 ppm) to the Stejskal-Tanner equation (8, 9). The 1,4-dioxane translational diffusion coefficients were fitted to the intensity decay of the 1,4-dioxane peak at 3.75 ppm. Integration of selected peaks was performed by Bruker Dynamics Center v2.5.6. For every PFG NMR experiment, 64 scans were recorded with a gradient strength interval from 2% to 98% ( $\gamma=26752$  rad s<sup>-1</sup> Gauss<sup>-1</sup>) with a diffusion time ( $\Delta$ ) of 200 ms and a gradient length ( $\delta$ ) of 2 ms. Baseline correction and assessment of the Stejskal Tanner fitting intervals were performed in Bruker Topspin 3.6.2 and Dynamics Center v2.5.6, while final fitting of the translation diffusion was performed in GraphPad Prism 8.2.1.

#### *SAXS experiments*

SAXS experiments were performed at the CPHSAXS Facility, University of Copenhagen, on a Xenocs© BioXolver L with a wavelength of  $\lambda = 1.34$  Å. Primary data reduction was made in BIOXTAS RAW. For each sample protein the optimal protein concentration yielding a high signal to noise ratio while avoiding aggregation was chosen by inspecting the low- $q$  region of the scattering curves by Guinier analysis. Each scattering curve consisted of an accumulation of 10 measurements on each sample. Subsequent data analysis was performed in the ATSAS 3.0.5 Primus suite (21), with the “merge” function used on multiple scattering curves for each sample protein. Merged scattering curves of  $n$  SAXS experiments for each protein sample were then used for Guinier derivation of the protein  $R_g$ . The Primus “AutoRg” function was used to determine  $q$ -range

for  $R_g$  derivation. Pair distance distribution plots of the consensus curves were also calculated in the ATSAS 3.0.5 Primus suite with the  $D_{\max}$  value being set based on a qualitative assessment of the P®-function reaching 0.

#### *Radius of hydration and gyration calculation from atomic coordinates*

The anhydrous  $R_g$  from the atomic coordinates of proteins was calculated as the mass-weighted average distance of each atom from the protein's centre of mass. For a better comparison to  $R_g$  values obtained from experimental SAXS data, we also used the WAXSiS webserver (22, 23) to explicitly include a water layer around the protein, calculate the SAXS profile of the envelope and fit an  $R_g$  from this with the Guinier approach. The  $R_h$  was calculated with the HullRadSAS software (24). For PDB entries that are NMR ensembles, the  $R_g$  and  $R_h$  were calculated on all conformers of the ensembles and then averaged as  $\langle R_g \rangle = n^{-1} \sum_i^n R_{g,i}$  and  $\langle R_h \rangle = (n^{-1} \sum_i^n R_{h,i})^{-1}$ . Missing C-terminal residues in the 14-3-3  $\zeta$  structure were added with MODELLER (25, 26) prior to calculating  $R_h$  and  $R_g$ .

#### *$R_h$ determination of 1,4-dioxane*

To determine the  $R_h$  of 1,4-dioxane, the ratio of  $D_t$  between the sample protein and 1,4-dioxane was used in Eq. 3 alongside the assumed protein  $R_h$  values calculated from experimental SAXS  $R_g$  values and Eq. 4. To ensure an accurate estimate of the final  $R_h$  value of 1,4-dioxane, the data quality was factored in by weighing the data by its quality using a  $\chi^2$ -approach. The  $\chi^2$ -value was calculated by first calculating the observed ratios of diffusion coefficients from PFG NMR, and then by subtracting the expected ratios of the diffusion coefficients from SAXS  $R_g$  derived  $R_h$  values and an assumed 1,4-dioxane  $R_h$  value spanning an interval of 2.0 to 2.5 Å. The deviation in the diffusion ratio from the observed NMR-data and SAXS-derived expected data was then divided by the sum of the squared experimental standard error of fits from both PFG NMR and SAXS, leaving a  $\chi^2$ -value representing the data quality and fit to different possible  $R_h$  values for 1,4-dioxane spanning 2.0 to 2.5 Å. We thus calculated  $\chi^2$  using the following equation for different estimates of the  $R_h$  value of dioxane.

$$\chi^2 = \frac{(D_{ratio,NMR} - D_{ratio,expected})^2}{(SE_{fit,NMR}^2 + SE_{fit,SAXS}^2)} \quad (5)$$

where  $D_{ratio,NMR} = \frac{D_{t,1,4-dioxane}}{D_{t,protein}}$ , and  $D_{ratio,expected} = \frac{R_{h,SAXS}}{R_{h,1,4-dioxane}}$ .

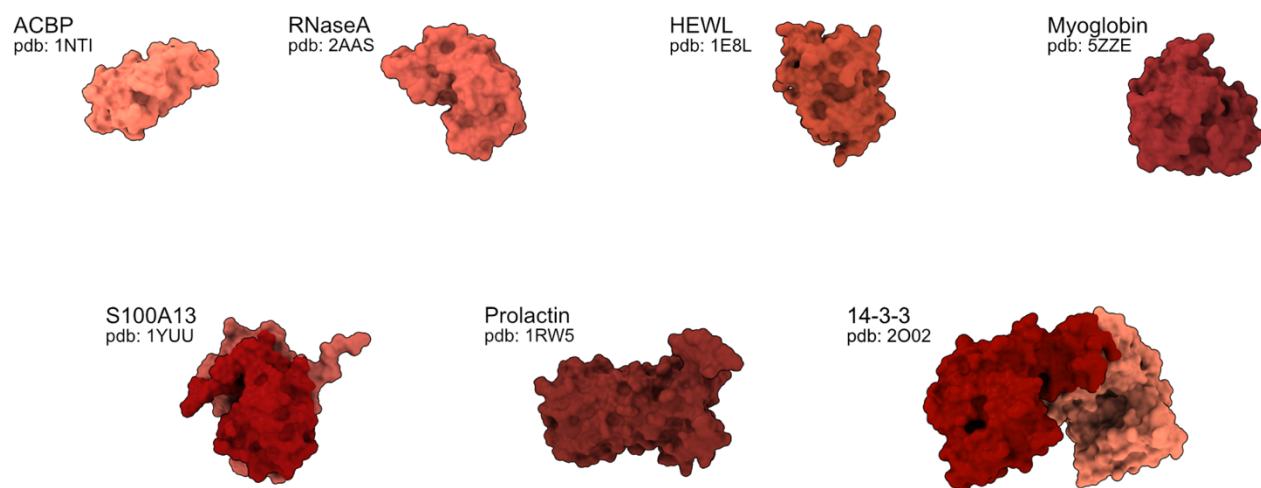
#### *Data availability*

Experimental data, processed data, and scripts to reproduce the content of this work are available at: [https://github.com/KULL-Centre/2023\\_dioxane-tranchant/](https://github.com/KULL-Centre/2023_dioxane-tranchant/).

## RESULTS & DISCUSSION

### Proteins and experimental measurements

To establish a foundation for determining the  $R_h$  of 1,4-dioxane we chose a subset of seven globular proteins varying in size, and recorded PFG NMR and batch SAXS experiments for each protein at different concentrations. The seven proteins were chosen based on availability, ease of experimental work (e.g., solubility and stability), size and roughly spherical shapes based on earlier experimental structures, as to adhere to the assumption made in Eq. 4. The seven proteins were HEWL, RNaseA, myoglobin, S100A13 (dimer), acyl-CoA-binding protein Y73F (ACBP<sub>Y73F</sub>), prolactin and the 14-3-3  $\zeta$  dimer (see Fig. 1 & Table 1). All proteins were checked for purity and homogeneity by an SDS-PAGE and size-exclusion chromatography prior to NMR and SAXS data acquisition (see supplemental Fig. 1).



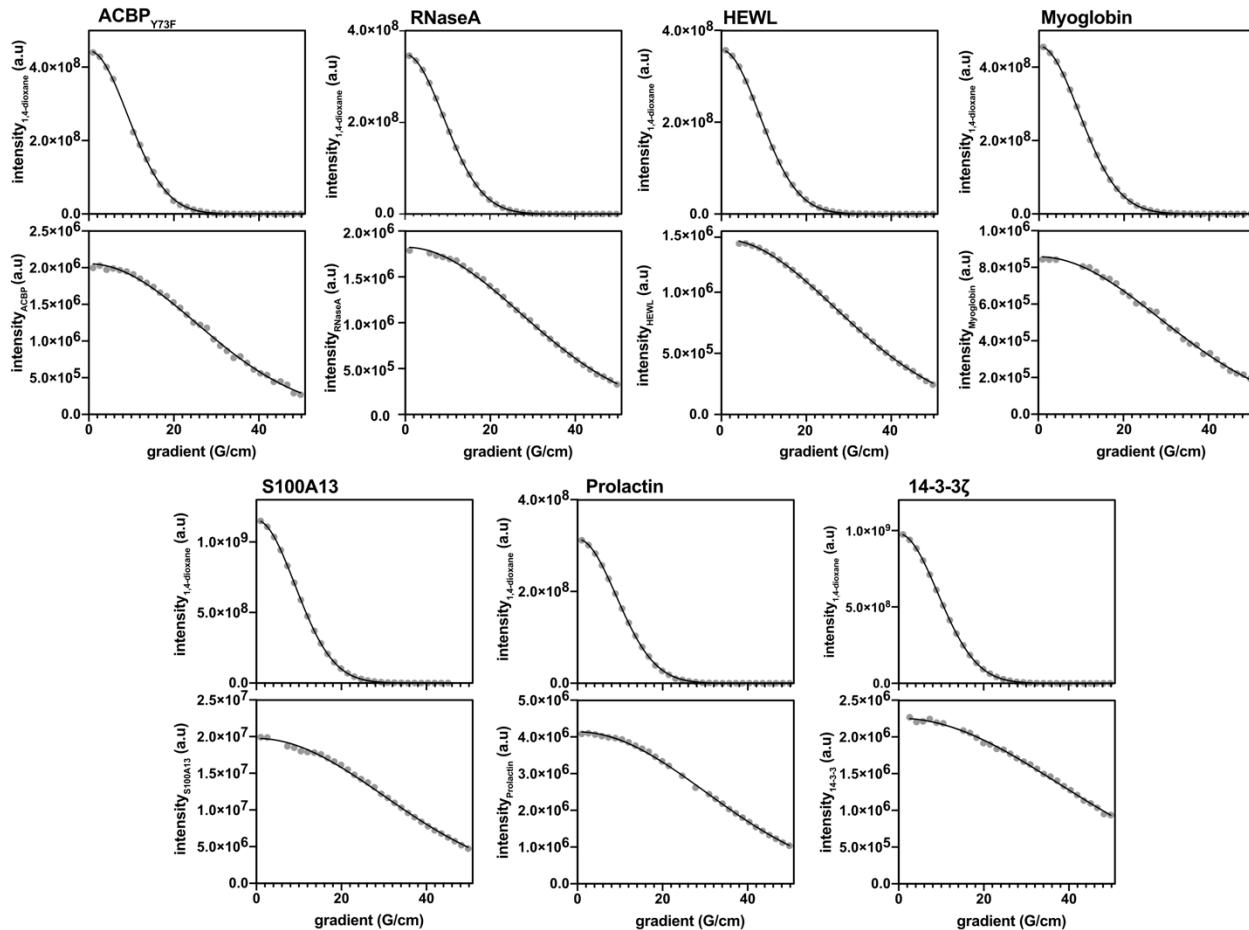
**Fig. 1: Surface contour representations of the seven sample proteins.** Model 1 of the ensemble in the respective PDB entries are shown. Proteins are arranged in order of molecular weight and shown on comparable scales. The structures used are from PDB ID 1NTI (20), 2AAS (27), 1E8L (28), 5ZZE (29), 1YUU (30), 1RW5 (31), 2002 (32).

Table 1. Overview of protein properties

PROTEIN	ACBP <sub>Y73F</sub>	RNASEA	HEWL	MYOGLOBIN	S100A13	PROLACTIN	14-3-3Z
ORGANISM	<i>B. taurus</i>	<i>B. taurus</i>	<i>G. gallus</i>	<i>E. caballus</i>	<i>H. sapiens</i>	<i>H. sapiens</i>	<i>H. sapiens</i>
SOURCE	<i>E. coli</i> expression	GE Healthcare	Sigma Aldrich	Sigma Aldrich	<i>E. coli</i> expression	<i>E. coli</i> expression	<i>E. coli</i> expression
UNIPROT ID (SEQUENCE)	P07108 (2–87, Y73F)	P61823 (27–150)	P00698 (19–147)	P68082 (2–154)	Q99584 (1–98)	P01236 (29–227)	P63104 (1–245)
PDB	1NTI	2AAS	1E8L	5ZZE	1YUU	1RW5	2002*
MW	9.9 kDa	13.7 kDa	14.3 kDa	17.1 kDa	22.9 kDa dimer	22.9 kDa	54 kDa dimer

\*lacks 15 C-terminal residues compared to the WT used in these experiments.

We first probed the diffusion coefficients of each protein using PFG NMR experiments with the addition of 1,4-dioxane as an internal standard. After picking peaks corresponding to either 1,4-dioxane or protein, we fitted the intensity decays as a function of the gradient strength to the Stejskal Tanner equation (Eq. 1) and the  $D_t$  of protein and dioxane was estimated (Fig. 2 and Table 2).



**Fig. 2: Peak intensity decays as a function of gradient strength in PFG NMR experiments.** Intensity decays for seven samples containing both 1,4-dioxane (top) and protein (bottom). Samples are shown in the following order: ACBP<sub>Y73F</sub>, RNaseA, HEWL, myoglobin, S100A13, prolactin, 14-3-3 $\zeta$ . Differences in peak intensity of 1,4-dioxane between samples are due to variations in concentration and differences in automatic selection of integration intervals performed in Bruker Dynamics Center.

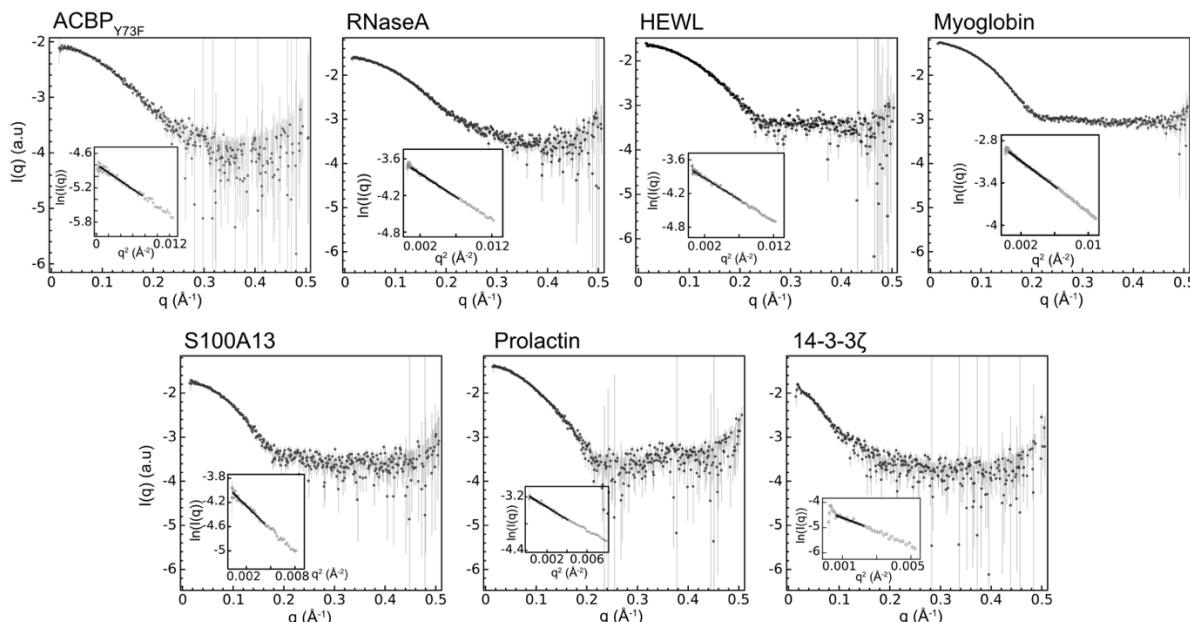
Table 2. Diffusion coefficients of 1,4-dioxane and seven proteins and their ratio

	ACBP	RNASEA	HEWL	MYOGLOBIN	S100A13	PROLACTIN	14-3-3
<b>1,4-DIOXANE</b>	$1.06 * 10^{-9}$ $\pm 5.65 * 10^{-12}$	$1.06 * 10^{-9}$ $\pm 7.45 * 10^{-13}$	$1.09 * 10^{-9}$ $\pm 3.09 * 10^{-12}$	$9.92 * 10^{-10}$ $\pm 7.67 * 10^{-13}$	$1.07 * 10^{-9}$ $\pm 8.02 * 10^{-13}$	$1.06 * 10^{-9}$ $\pm 2.86 * 10^{-12}$	$1.04 * 10^{-9}$ $\pm 6.79 * 10^{-13}$
<b>PROTEIN</b>	$1.37 * 10^{-10}$ $\pm 1.33 * 10^{-12}$	$1.20 * 10^{-10}$ $\pm 6.24 * 10^{-13}$	$1.22 * 10^{-10}$ $\pm 3.59 * 10^{-13}$	$1.07 * 10^{-10}$ $\pm 9.00 * 10^{-13}$	$9.84 * 10^{-11}$ $\pm 8.12 * 10^{-13}$	$9.74 * 10^{-11}$ $\pm 5.03 * 10^{-13}$	$6.22 * 10^{-11}$ $\pm 4.32 * 10^{-13}$
<b><math>D_T</math> RATIO</b>	$7.69 \pm 0.09$	$8.87 \pm 0.05$	$8.96 \pm 0.04$	$9.28 \pm 0.08$	$10.92 \pm 0.09$	$10.88 \pm 0.06$	$16.75 \pm 0.12$

While the 1,4-dioxane data fitted nicely to the Stejskal Tanner equation, the protein data showed greater variation. In general, the ratio of the diffusion coefficients between protein and 1,4-dioxane increased as the molecular weight of the protein increased, as would be expected. Differences in total intensity of 1,4-dioxane was observed, which was expected as the 1,4-dioxane concentration in some of the samples was slightly different (0.06% in S100A13 and 14-3-3, 0.04% in ACBP, HEWL, RNaseA, prolactin, and myoglobin). However, the observed difference in 1,4-dioxane intensity was also observed in experiments on triplicates of identical samples (supplemental figure 2), likely attributed to the automatic selection of integration area by the Dynamics Center software when peak picking.

Notably, the ratio of the diffusion coefficients between HEWL and 1,4-dioxane in our experiments was lower ( $8.96 \pm 0.05$ ) than in the original work where the  $R_h$  of 1,4-dioxane was estimated (9.33) (14, 17). This difference could be explained either by difference in pH or protein concentration between measurements, as the diffusion coefficient ratio of 9.33 was measured at pH 2, and with a HEWL concentration of ca. 1 mM, whereas our experiments were recorded at neutral pH and with 200  $\mu$ M of HEWL. To test whether the observed variation in diffusion ratio between our experiments and the original experiments could be explained by instrumental error, we recorded technical triplicate measurements of HEWL and RNaseA samples prepared in the same way from the same protein stock (supplemental Fig. 2 and Table 2). These results showed an approximately 1.4% variance in the ratio of the diffusion coefficients between technical replicas, smaller than the difference between 8.96 and 9.33.

Next, we recorded multiple SAXS curves for each protein at optimal protein concentrations and used the “merge” function in the ATSAS suite to derive a protein consensus curve (Fig. 3). We then performed a Guinier analysis of each consensus curve to estimate  $R_g$  for each protein to help minimize effects from aggregation on the scattering curves that could otherwise increase variance between measurements. The quality of the consensus scattering curves was evaluated by examining the pair distance distribution plots of each curve (supplemental Fig. 3).



**Fig. 3: Consensus scattering curves of experimental scattering curves of each of the seven proteins.** Total scattering curves used in each consensus curve: ACBP (from 2 measurements), RNaseA (4 measurements), HEWL (5 measurements), myoglobin (6 measurements), S100A13 (5 measurements), prolactin (3 measurements), and 14-3-3 (3 measurements). For each panel, the insert shows the low- $q$  region on a logarithmic scale and the linear fit used for the Guinier analysis.

We estimated the  $R_g$  values of the proteins by Guinier analysis and from the pair distance distribution plots (supplemental table 3); only values from Guinier analysis were used for estimating the 1,4-dioxane  $R_h$ . As expected, protein  $R_g$  increased as the molecular weight increased. Comparing the HEWL Guinier derived  $R_g$  of  $15.16 \pm 0.08 \text{ \AA}$  to the  $R_g$  used for the original 1,4-dioxane  $R_h$  derivation of  $15.3 \pm 0.2 \text{ \AA}$  (15), or to the mean batch-SAXS  $R_g$  of  $15.32 \pm 0.81 \text{ \AA}$  from (18), measured in the presence of 2 M urea, revealed no substantial differences in the mean value.

We then used the measured  $R_g$ -values for the seven proteins and the  $D_t$  ratios from PFG NMR to the  $R_h$  for 1,4-dioxane using Eq. 4 by estimating the protein  $R_h$ -values from Eq. 3. This approach yielded an estimated 1,4-dioxane  $R_h$  from each protein data set (Table 4).

**Table 4:** Experimentally derived diffusion coefficient ratios from PFG NMR, experimentally derived  $R_g$  values by Guinier analysis from batch SAXS, estimated protein  $R_h$  values from the SAXS data and resulting estimated  $R_h$  values from 1,4-dioxane. All reported errors are propagated standard errors of the original fits.

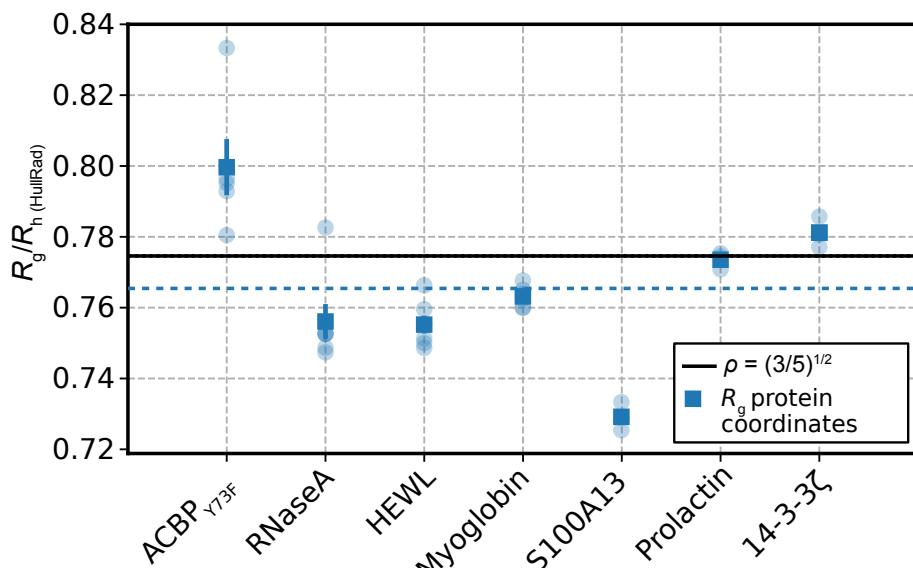
	ACBP	RNASEA	HEWL	MYOGLOBIN	S100A13	PROLACTIN	14-3-3
$D_t$ RATIO	$7.69 \pm 0.09$	$8.87 \pm 0.05$	$8.96 \pm 0.04$	$9.28 \pm 0.08$	$10.92 \pm 0.09$	$10.88 \pm 0.06$	$16.75 \pm 0.12$
PROTEIN $R_g(\text{\AA})$ (Guinier)	$14.79 \pm 0.18$	$15.20 \pm 0.08$	$15.16 \pm 0.08$	$16.41 \pm 0.04$	$19.77 \pm 0.24$	$20.29 \pm 0.16$	$27.47 \pm 1.00$
ESTIMATED PROTEIN $R_h(\text{\AA})$	$19.09 \pm 0.23$	$19.62 \pm 0.10$	$19.57 \pm 0.10$	$21.19 \pm 0.05$	$25.52 \pm 0.31$	$26.19 \pm 0.21$	$35.46 \pm 1.29$
DERIVED 1,4-DIOXANE $R_h(\text{\AA})$	$2.48 \pm 0.06$	$2.21 \pm 0.02$	$2.18 \pm 0.02$	$2.28 \pm 0.02$	$2.34 \pm 0.05$	$2.41 \pm 0.03$	$2.12 \pm 0.09$

Depending on the protein used as a standard, we arrived at different estimates of the  $R_h$  values for 1,4-dioxane ranging between  $2.12 \text{ \AA}$  to  $2.48 \text{ \AA}$ . The original estimate of  $R_h = 2.12 \text{ \AA}$  lies at the edge of this interval of  $R_h$  values but appears to be an underestimate as six of the seven protein data sets gave larger values. Taking the average 1,4-dioxane  $R_h$  from these data, and factoring in the errors for each protein set, we arrive at a weighted average of  $2.27 \text{ \AA}$ , with an error estimated by bootstrapping to be  $0.02 \text{ \AA}$ .

#### **Examining the relationship between $R_g$ and $R_h$ for proteins**

It is possible that differences in the  $R_h$  values estimated for 1,4-dioxane from the SAXS and NMR data across the seven proteins can in part be explained by deviation of the protein shapes from sphere of uniform mass distribution. As a consequence, the relationship between  $R_g$  and  $R_h$  may not be equal to  $\rho = (3/5)^{1/2}$ , as assumed in Eq. 4. To examine whether this assumption is reasonable

for the seven proteins, we used both the crystallographic and solution structures of each of the seven proteins to calculate  $R_g$  and  $R_h$ , and compared their ratio to the assumed value of  $\rho=(3/5)^{1/2}$  (Fig. 4) (23). Deviations from an ideal value of  $\rho$  would affect our analysis and the estimates of the  $R_h$  value for 1,4-dioxane presented above. We selected the seven proteins to have roughly spherical shapes and we expect small deviations from  $\rho=(3/5)^{1/2}$  to average out over the dataset. Overall, we observe the calculated  $R_g/R_h$  values to be close to  $\rho=(3/5)^{1/2}$ , with a small underestimation on average. As the  $R_g$  in Fig. 4 was calculated only for the protein coordinates, we subsequently included the contribution of the hydration layer to the calculated  $R_g$  to investigate if the apparent underestimation of  $R_g/R_h$  could be accounted for by considering the hydration of the proteins (supplemental Fig. 4) (24). This, however, led to a large overestimation of  $R_g/R_h$  compared to  $\rho=(3/5)^{1/2}$ . We examined whether the deviation from to  $\rho=(3/5)^{1/2}$  could be explained by deviations from a spherical shape by calculating the asphericity of the proteins (33); however, we find no strong correlation between asphericity and observed deviation (supplemental Fig. 5).

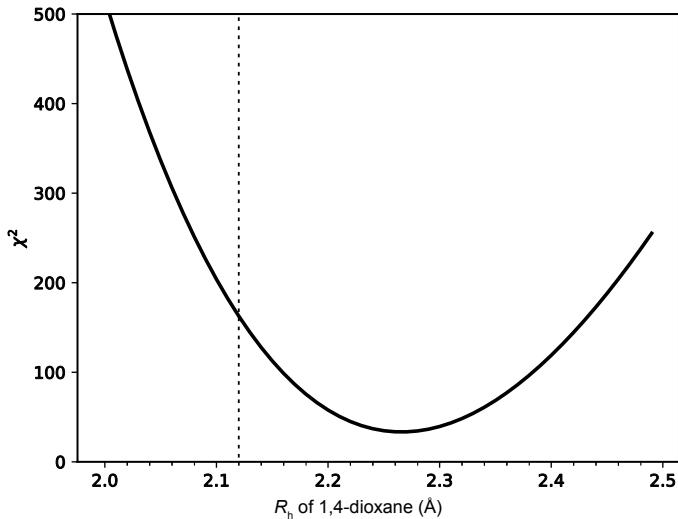


**Fig 4: Ratios of  $R_g/R_h$  obtained from 3D structures.** For each protein, three to five 3D structures were used, with the following PDB codes for each protein: HEWL (1e8l (28), 1dpx (34), 6abn (35), 5a3e (36)), RNaseA (2aas (27), 1fs3 (37), 1jvt (38), 4ooh(39), 1kf7 (40)), myoglobin (5zze (29), 1wla (41), 4dc8 (42), 5d5r (43), 5cn4 (43)), S100A13 (1yuu (30), 2h2k (44), 2egd (45)), ACBP (1nti (20), 1hb6 (46), 1hb8 (46), 2abd (47)), prolactin (1rw5 (31), 2q98 (48)), 14-3-3ζ (2o02 (32), 1qja (49), 1gjb (50)). AlphaFold structures (51, 52) were also included for each protein. Individual ratios are shown as partially transparent markers. For each protein we calculate the mean ratio over different structures and the standard error of the mean (solid markers). Dashed blue line represents the calculated average value of the  $R_g/R_h$ .

#### **Factoring in the data quality in calculating the expected $R_h$ of 1,4-dioxane**

As observed in the difference of the 1,4-dioxane  $R_h$  calculated from the different protein data sets, the choice of protein can affect the outcome. We therefore performed a global analysis of all seven protein data sets to help minimize protein-specific effects on the estimated value of  $R_h$  for 1,4-dioxane. For this, we calculated a  $\chi^2$  value (Eq. 5) between the measured ratio of diffusion coefficients and the value expected depending on (i) the estimated  $R_h$  for each protein and (ii) the  $R_h$  for 1,4-dioxane (Eq. 3). With this approach, we can assess an interval of likely 1,4-dioxane  $R_h$

values while taking into account the errors on the measured  $D_t$  ratios and the estimated value for the  $R_h$  for the seven proteins (from SAXS and  $\rho=(3/5)^{1/2}$ ) (16).



**Fig. 5:  $\chi^2$ -value for the seven protein data sets calculated as described in Eq. 5 for different possible  $R_h$  values for 1,4-dioxane.** The dashed vertical line highlights the  $R_h$  value of 1,4-dioxane of 2.12 Å determined in (14).

The resulting plot of  $\chi^2$  vs. the  $R_h$  for 1,4-dioxane (Fig. 5) shows a minimum around  $R_h$  of 2.27 Å and, as discussed above, suggests that the previously determined value of 2.12 Å is an underestimate (dashed line in Fig. 5). To estimate an error on the  $\chi^2$ -estimated 1,4-dioxane  $R_h$  of 2.27 Å, we performed a leave-one-out analysis which provided an error estimate of 0.02 Å, corresponding with the earlier weighted average 1,4-dioxane  $R_h$  and bootstrap error estimation. We also estimated an error using the technical replicate measurements of HEWL and RNaseA in both PFG NMR and SAXS (supplemental table 1 and table 2, respectively). From the technical replicates, we found a 1.4% and 1.1% error across samples measured by PFG NMR and SAXS Guinier analysis, respectively. Propagating these relative errors from the technical replicates to the  $\chi^2$ -estimated 1,4-dioxane  $R_h$ , we determine the  $R_h$  of 1,4-dioxane to be  $2.27 \pm 0.04$  Å.

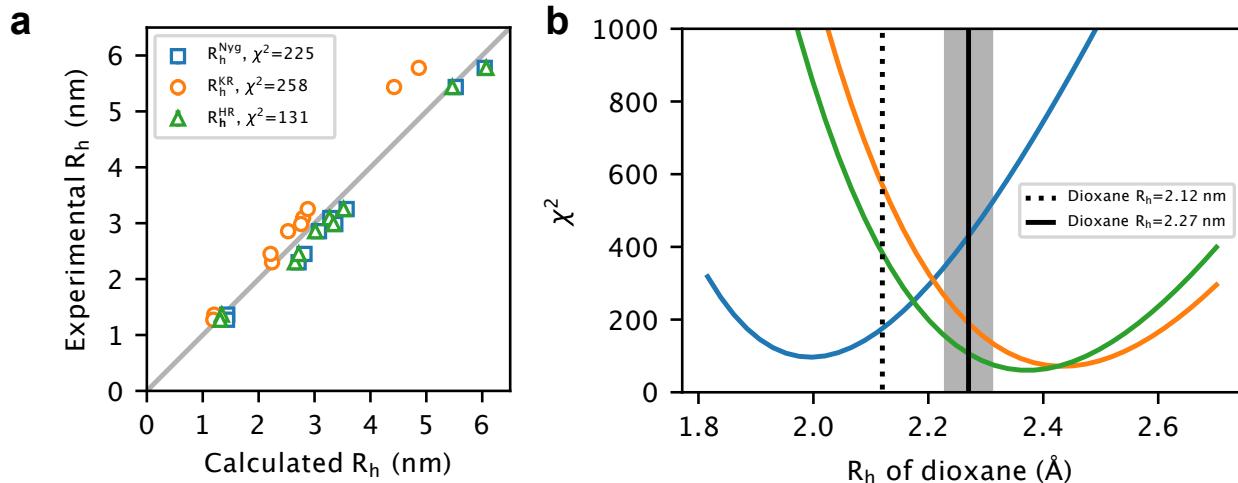
### Impact of an increase in the size of the hydrodynamic ratio of 1,4-dioxane

With an added uncertainty to the estimated 1,4-dioxane  $R_h$ , the above results suggest a ~7% ( $2.27 \text{ \AA}/2.12 \text{ \AA}=1.07$ ) increase in  $R_h$  compared to the commonly used reference. This change in reference value can be accounted for when examining previously published PFG NMR data that have used 1,4-dioxane as a reference by increasing the derived protein  $R_h$  by 7% as well. For example, we previously reported the  $R_h$  of prothymosin- $\alpha$  to be  $28.9 \pm 0.8$  Å using PFG NMR and with 1,4-dioxane as a reference with the  $R_h$  set as 2.12 Å (53). Using the updated reference value for dioxane and propagating the errors, the re-estimated  $R_h$  would be  $31 \pm 1$  Å. Often,  $R_h$  values from PFG NMR are used to track changes in protein dimensions following a selected perturbation. In these cases, the increase in 1,4-dioxane  $R_h$  would not be an issue, as the relative changes are unaffected by the increase in the absolute  $R_h$ . An example of this would be PFG NMR experiments studying a protein oligomeric state at different concentrations, by deriving the oligomeric state from the observed  $R_h$  at different concentrations (54).

### Consequences for estimating the conformational ensembles of IDPs

One important consequence of the results presented here is that they affect our understanding of how to compare conformational ensembles with experimental measurements of  $R_h$ . We recently compiled a list of the  $R_h$  of eleven IDPs as measured by PFG NMR, and used these together with SAXS experiments to evaluate different models to calculate  $R_h$  from conformational ensembles of IDPs (55). Our work led to the conclusion that, among the different approaches tested, the Kirkwood-Riseman equation (56) resulted in the best agreement between computational models and the  $R_h$  values measured by PFG NMR. We noted, however, how possible inaccuracies in the reference  $R_h$  of 1,4-dioxane would affect our results, by changing the experimentally determined  $R_h$  values and leading to a different conclusion.

As we here have shown that the  $R_h$  of 1,4-dioxane was previously underestimated, we re-examine the conclusion of our previous work considering the new reference value for the  $R_h$  of 1,4-dioxane. Therefore, we first increased the experimentally derived  $R_h$  values for the eleven IDPs by 7% (table S4) and applied an uncertainty of 2.1% (average relative error associated with the  $R_h$  of the IDPs) to the corrected  $R_h$  values, as we previously proposed to provide a uniform fidelity to the result of the PFG NMR experiments in our dataset. We used the same SAXS-optimized conformational ensembles used in our previous work (produced with the CALVADOS model (57)). In our previous work we used the Kirkwood-Riseman equation (56), two empirical relationships relating  $R_g$  and  $R_h$  (58) and HullRad (59) to calculate  $R_h$  from the conformational ensembles generated by CALVADOS. We here use the more recent HullRadSAS approach (24). When we compare the  $R_h$  values calculated from the SAXS-refined ensembles to the revised experimental values, the results are less clear than when we compared to the  $R_h$  values based on the original reference value for dioxane. Specifically, we now find that the Kirkwood-Riseman systematically underestimates the  $R_h$ , whereas the other approaches overestimate  $R_h$  (Fig. 6). Of the four methods examined, we find that HullRadSAS agrees more closely with the data. This observation is due to the finding that HullRadSAS giving rise to very good agreement with the experimental values for four of the IDPs; for the remaining the result is more complex as the experimental value lies in between the predictions from the different models (Fig. S6). In light of this result, we suggest the use of HullRadSAS when calculating the  $R_h$  of IDP conformations. Assuming that the SAXS-refined conformational ensembles that we have employed are accurate in reproducing the  $R_h$  of these IDPs, and therefore that the major source of uncertainty comes from the model used to calculate the  $R_h$ , we suggest that an uncertainty of 7% is taken into account on the calculated  $R_h$ .



**Figure 6: Assessment of models to calculate  $R_h$  from structural ensembles of intrinsically disordered proteins using the same conformational ensembles and approach as in our previous work (10).** We compare three approaches to calculate the  $R_h$  from conformational ensembles of several IDPs (Table S4): the Nygaard equation (in blue) (62), the Kirkwood-Riseman equation (in orange) (63) and HullRadSAS (in green) (24). (a) Agreement between calculated and experimentally derived  $R_h$  values. (b) As the value for the  $R_h$  of dioxane varies, so does our assessment of the models to calculate  $R_h$  from conformational ensembles of IDPs: when using a value of  $2.12 \text{ \AA}$ , the Kirkwood-Riseman equation leads to the best agreement with PFG NMR measurements. With the new proposed value of  $2.27 \pm 0.04 \text{ \AA}$ , HullRadSAS appears to be the model leading to the best agreement with PFG NMR measurements.

## CONCLUSIONS

From the PFG NMR and SAXS data on the seven proteins, and subsequent weighted average and  $\chi^2$  estimations of the 1,4-dioxane  $R_h$ , we find that the original  $2.12 \text{ \AA}$   $R_h$  is likely underestimated and suggest a value of  $2.27 \pm 0.04 \text{ \AA}$  to be used as the standard 1,4-dioxane  $R_h$  when performing PFG NMR to determine the hydrodynamic radius of a protein. The error on this final value is determined as the propagated relative error from technical replications measured by both PFG NMR and SAXS on HEWL and RNaseA; however, this uncertainty will rarely be the limiting factor in the accuracy of derived protein  $R_h$  values. Previously published PFG NMR protein  $R_h$  values can easily be re-referenced to our suggested 1,4-dioxane  $R_h$  by increasing the protein  $R_h$  with 7%, however this is only needed as long as absolute  $R_h$  values are reported. While our data suggest that the 1,4-dioxane  $R_h$  is greater than  $2.12 \text{ \AA}$ , one would ideally use an even greater protein dataset and/or other methods to analyse dioxane. Furthermore, the  $R_h$  of 1,4-dioxane might be both pressure- and temperature- dependent, and as such, the  $R_h$  of 1,4-dioxane at different experimental conditions should also be examined further (60, 61).

## Acknowledgements

We thank Signe A. Sjørup for skilled technical assistance and Johan G. Olsen for valuable discussions on water layers in relation to SAXS measurements. We thank Andreas Prestel, manager of the cOpenNMR facility <https://www1.bio.ku.dk/copennmr/> (grant no. NNF18OC0032996) for NMR assistance. We acknowledge access to the University of Copenhagen small-angle X-ray scattering facility, CPHSAXS, funded by the Novo Nordisk Foundation (grant no. NNF19OC0055857), and thank Pernille Sønderby Tuelung for assistance <https://drug.ku.dk/core-facilities/cphsaxs/>. We acknowledge access to computational resources from the Biocomputing Core Facility at the Department of Biology, University of Copenhagen.

## Funding

This work was supported by grants from the Novo Nordisk Foundation to the Challenge centres PRISM (NNF18OC0033950 to K.L.-L.) and REPIN (#NNF18OC0033926 to B.B.K.), the Lundbeck Foundation BRAINSTRUC initiative (R155-2015-2666 to B.K.K. and K.L.-L.) and the Danish Research Councils (#9040-00164B to B.B.K.).

## REFERENCES:

1. Mittag, T., and J.D. Forman-Kay. 2007. Atomic-level characterization of disordered protein ensembles. *Curr Opin Struct Biol.* 17:3–14.
2. Rieloff, E., and M. Slepö. 2020. Determining  $R_g$  of IDPs from SAXS Data. *Methods Mol Biol.* 2141:271–283.
3. Cordeiro, T.N., F. Herranz-Trillo, A. Urbanek, A. Estaña, J. Cortés, N. Sibille, and P. Bernadó. 2017. Small-angle scattering studies of intrinsically disordered proteins and their complexes. *Curr Opin Struct Biol.* 42:15–23.
4. Leeb, S., and J. Danielsson. 2020. Obtaining hydrodynamic radii of intrinsically disordered protein ensembles by pulsed field gradient NMR measurements. In: *Methods in Molecular Biology*. Humana Press Inc. pp. 285–302.
5. Sagar, A., D. Svergun, and P. Bernadó. 2020. Structural Analyses of Intrinsically Disordered Proteins by Small-Angle X-Ray Scattering. *Methods Mol Biol.* 2141:249–269.
6. Prestel, A., K. Bugge, L. Staby, R. Hendus-Altenburger, and B.B. Kragelund. 2018. Characterization of Dynamic IDP Complexes by NMR Spectroscopy. In: *Methods in Enzymology*. . pp. 193–226.
7. Danielsson, J., J. Jarvet, P. Damberg, and A. Gräslund. 2002. Translational diffusion measured by PFG-NMR on full length and fragments of the Alzheimer  $\text{A}\beta(1-40)$  peptide. Determination of hydrodynamic radii of random coil peptides of varying length. *Magnetic Resonance in Chemistry*. 40.
8. 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.
9. Wu, D.H., A. Chen, and C.S. Johnson. 1995. An Improved Diffusion-Ordered Spectroscopy Experiment Incorporating Bipolar-Gradient Pulses. *J Magn Reson A.* 115:260–264.
10. Pesce, F., E.A. Newcombe, P. Seiffert, E.E. Tranchant, J.G. Olsen, C.R. Grace, B.B. Kragelund, and K. Lindorff-Larsen. 2023. Assessment of models for calculating the hydrodynamic radius of intrinsically disordered proteins. *Biophys J.* 122:310–321.

11. Leeb, S., and J. Danielsson. 2020. Obtaining hydrodynamic radii of intrinsically disordered protein ensembles by pulsed field gradient NMR measurements. .
12. Aachmann, F.L., D.E. Otzen, K.L. Larsen, and R. Wimmer. 2003. Structural background of cyclodextrin-protein interactions. *Protein Eng.* 16:905–912.
13. Danielsson, J., J. Jarvet, P. Damberg, and A. Gräslund. 2004. Two-site binding of  $\beta$ -cyclodextrin to the Alzheimer A $\beta$ (1-40) peptide measured with combined PFG-NMR diffusion and induced chemical shifts. *Biochemistry*. 43:6261–6269.
14. 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.
15. Chen, L., K.O. Hodgson, and S. Doniach. 1996. A lysozyme folding intermediate revealed by solution X-ray scattering. *J Mol Biol.* 261:658–671.
16. Burchard, W., M. Schmidt, and W.H. Stockmayer. 1980. Influence of Hydrodynamic Preaveraging on Quasi-Elastic Scattering from Flexible Linear and Star-Branched Macromolecules. *Macromolecules*. 13:580–587.
17. Jones, J.A., D.K. Wilkins, L.J. Smith, and C.M. Dobson. 1997. Characterisation of protein unfolding by NMR diffusion measurements. *J Biomol NMR*. 10:199–203.
18. Trehewella, J., P. Vachette, J. Bierma, C. Blanchet, E. Brookes, S. Chakravarthy, L. Chatzimargas, T.E. Cleveland, N. Cowieson, B. Crossett, A.P. Duff, D. Franke, F. Gabel, R.E. Gillilan, M. Graewert, A. Grishaev, J.M. Guss, M. Hammel, J. Hopkins, Q. Huang, J.S. Hub, G.L. Hura, T.C. Irving, C.M. Jeffries, C. Jeong, N. Kirby, S. Krueger, A. Martel, T. Matsui, N. Li, J. Pérez, L. Porcar, T. Prangé, I. Rajkovic, M. Rocco, D.J. Rosenberg, T.M. Ryan, S. Seifert, H. Sekiguchi, D. Svergun, S. Teixeira, A. Thureau, T.M. Weiss, A.E. Whitten, K. Wood, and X. Zuo. 2022. A round-robin approach provides a detailed assessment of biomolecular small-angle scattering data reproducibility and yields consensus curves for benchmarking. *Acta Crystallogr D Struct Biol.* 78:1315–1336.
19. Teilum, K., J.C. Hoch, V. Goffin, S. Kinet, J.A. Martial, and B.B. Kragelund. 2005. Solution structure of human prolactin. *J Mol Biol.* 351:810–823.
20. Kragelund, B.B., K. Poulsen, K.V. Andersen, T. Baldursson, J.B. Krøll, T.B. Neergård, J. Jepsen, P. Roepstorff, K. Kristiansen, F.M. Poulsen, and J. Knudsen. 1999. Conserved residues and their role in the structure, function, and stability of acyl-coenzyme A binding protein. *Biochemistry*. 38:2386–2394.
21. 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. Sverguna, and D. Frankea. 2021. ATSAS 3.0: Expanded functionality and new tools for small-angle scattering data analysis. *J Appl Crystallogr.* 54:343–355.
22. Chen, P.C., and J.S. Hub. 2014. Validating solution ensembles from molecular dynamics simulation by wide-angle X-ray scattering data. *Biophys J.* 107:435–447.
23. Knight, C.J., and J.S. Hub. 2015. WAXSiS: A web server for the calculation of SAXS/WAXS curves based on explicit-solvent molecular dynamics. *Nucleic Acids Res.* 43:W225–W230.
24. Fleming, P.J., J.J. Correia, and K.G. Fleming. 2023. Revisiting macromolecular hydration with HullRadSAS. *European Biophysics Journal*.
25. Webb, B., and A. Sali. 2016. Comparative protein structure modeling using MODELLER. *Curr Protoc Protein Sci.* 2016:2.9.1-2.9.37.

26. Sali, A., and T.L. Blundell. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol.* 234:779–815.
27. Santoro, J., C. González, M. Bruix, J.L. Neira, J.L. Nieto, J. Herranz, and M. Rico. 1993. High-resolution three-dimensional structure of ribonuclease A in solution by nuclear magnetic resonance spectroscopy. *J Mol Biol.* 229:722–734.
28. Schwalbe, H., S.B. Grimshaw, A. Spencer, M. Buck, J. Boyd, C.M. Dobson, C. Redfield, and L.J. Smith. 2001. A refined solution structure of hen lysozyme determined using residual dipolar coupling data. *Protein Science.* 10:677–688.
29. Kitahara, M., S. Fudo, T. Yoneda, M. Nukaga, and T. Hoshino. 2019. Anisotropic Distribution of Ammonium Sulfate Ions in Protein Crystallization. *Cryst Growth Des.* 19:6136–6140.
30. Arnesano, F., L. Banci, I. Bertini, A. Fantoni, L. Tenori, and M.S. Viezzoli. 2005. Structural interplay between calcium(II) and copper(II) binding to S100A13 protein. *Angewandte Chemie - International Edition.* 44:6341–6344.
31. Teilum, K., J.C. Hoch, V. Goffin, S. Kinet, J.A. Martial, and B.B. Kragelund. 2005. Solution structure of human prolactin. *J Mol Biol.* 351:810–823.
32. Ottmann, C., L. Yasmin, M. Weyand, J.L. Veesenmeyer, M.H. Diaz, R.H. Palmer, M.S. Francis, A.R. Hauser, A. Wittinghofer, and B. Hallberg. 2007. Phosphorylation-independent interaction between 14-3-3 and exoenzyme S: From structure to pathogenesis. *EMBO Journal.* 26:902–913.
33. Steinhauser, M.O. 2005. A molecular dynamics study on universal properties of polymer chains in different solvent qualities. Part I. A review of linear chain properties. *Journal of Chemical Physics.* 122.
34. Weiss, M.S., G.J. Palm, and R. Hilgenfeld. 2000. Crystallization, structure solution and refinement of hen egg-white lysozyme at pH 8.0 in the presence of MPD. *Acta Crystallogr D Biol Crystallogr.* 56.
35. Seraj, Z., S. Ahmadian, M.R. Groves, and A. Seyedarabi. 2020. The aroma of TEMED as an activation and stabilizing signal for the antibacterial enzyme HEWL. *PLoS One.* 15.
36. Nannenga, B.L., D. Shi, A.G.W. Leslie, and T. Gonen. 2014. High-resolution structure determination by continuous-rotation data collection in MicroED. *Nat Methods.* 11.
37. Chatani, E. 2002. Conformational strictness required for maximum activity and stability of bovine pancreatic ribonuclease A as revealed by crystallographic study of three Phe120 mutants at 1.4 Å resolution. *Protein Science.* 11.
38. Vitagliano, L., A. Merlino, A. Zagari, and L. Mazzarella. 2002. Reversible substrate-induced domain motions in ribonuclease A. *Proteins: Structure, Function and Genetics.* 46.
39. Yadav, S.P.S., P. Sharma, and F.N.U. Ashish. Structure of Ribonuclease A at 40C - 4OOH. Structure available at <https://doi.org/10.2210/pdb4OOH/pdb> .
40. Berisio, R., F. Sica, V.S. Lamzin, K.S. Wilson, A. Zagari, and L. Mazzarella. 2002. Atomic resolution structures of ribonuclease A at six pH values. *Acta Crystallogr D Biol Crystallogr.* 58.
41. Maurus, R., C.M. Overall, R. Bogumil, Y. Luo, A.G. Mauk, M. Smith, and G.D. Brayer. 1997. A myoglobin variant with a polar substitution in a conserved hydrophobic cluster in the heme binding pocket. *Biochimica et Biophysica Acta - Protein Structure and Molecular Enzymology.* 1341.

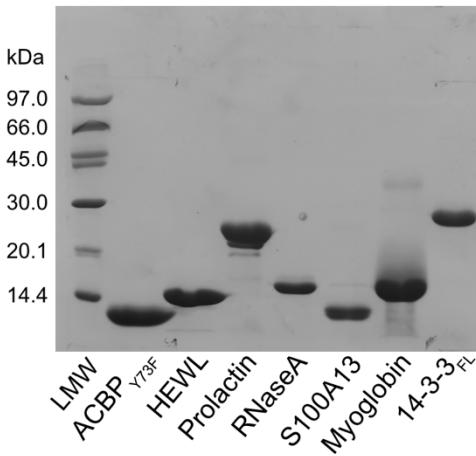
42. Kissick, D.J., C.M. Dettmar, M. Becker, A.M. Mulichak, V. Cherezov, S.L. Ginell, K.P. Battaile, L.J. Keefe, R.F. Fischetti, and G.J. Simpson. 2013. Towards protein-crystal centering using second-harmonic generation (SHG) microscopy. *Acta Crystallogr D Biol Crystallogr.* 69.
43. Barends, T.R.M., L. Foucar, A. Ardevol, K. Nass, A. Aquila, S. Botha, R.B. Doak, K. Falahati, E. Hartmann, M. Hilpert, M. Heinz, M.C. Hoffmann, J. Köfinger, J.E. Koglin, G. Kovacsova, M. Liang, D. Milathianaki, H.T. Lemke, J. Reinstein, C.M. Roome, R.L. Shoeman, G.J. Williams, I. Burghardt, G. Hummer, S. Boutet, and I. Schlichting. 2015. Direct observation of ultrafast collective motions in CO myoglobin upon ligand dissociation. *Science (1979).* 350.
44. Li, M., P.F. Zhang, X.W. Pan, and W.R. Chang. 2007. Crystal structure study on human S100A13 at 2.0 Å resolution. *Biochem Biophys Res Commun.* 356.
45. Imai, F.L., K. Nagata, N. Yonezawa, M. Nakano, and M. Tanokura. 2008. The crystal structure at 1.8 Å resolution of the calcium-bound human S100A13 at pH 7.5. *Acta Crystallogr A.* 64.
46. Van Aalten, D.M.F., K.G. Milne, J.Y. Zou, G.J. Kleywegt, T. Bergfors, M.A.J. Ferguson, J. Knudsen, and T.A. Jones. 2001. Binding site differences revealed by crystal structures of *Plasmodium falciparum* and bovine acyl-CoA binding protein. *J Mol Biol.* 309.
47. Andersen, K.V., and F.M. Poulsen. 1993. The three-dimensional structure of acyl-coenzyme A binding protein from bovine liver: Structural refinement using heteronuclear multidimensional NMR spectroscopy. *J Biomol NMR.* 3.
48. Jomain, J.B., E. Tallet, I. Broutin, S. Hoos, J. Van Agthoven, A. Ducruix, P.A. Kelly, B.B. Kragelund, P. England, and V. Goffin. 2007. Structural and thermodynamic bases for the design of pure prolactin receptor antagonists: X-ray structure of Del1-9-G129R-hPRL. *Journal of Biological Chemistry.* 282.
49. Rittinger, K., J. Budman, J. Xu, S. Volinia, L.C. Cantley, S.J. Smerdon, S.J. Gamblin, and M.B. Yaffe. 1999. Structural analysis of 14-3-3 phosphopeptide complexes identifies a dual role for the nuclear export signal of 14-3-3 in ligand binding. *Mol Cell.* 4.
50. Katz, B.A., P.A. Sprengeler, C. Luong, E. Verner, K. Elrod, M. Kirtley, J. Janc, J.R. Spencer, J.G. Breitenbucher, H. Hui, D. McGee, D. Allen, A. Martelli, and R.L. Mackman. 2001. Engineering inhibitors highly selective for the S1 sites of Ser190 trypsin-like serine protease drug targets. *Chem Biol.* 8.
51. Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature.* 596:583–589.
52. Varadi, M., S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, and S. Velankar. 2022. AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 50:D439–D444.
53. 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.

54. Hunter, H.N., D. Bruce Fulton, T. Ganz, and H.J. Vogel. 2002. The solution structure of human hepcidin, a peptide hormone with antimicrobial activity that is involved in iron uptake and hereditary hemochromatosis. *Journal of Biological Chemistry*. 277:37597–37603.
55. Pesce, F., E.A. Newcombe, P. Seiffert, E.E. Tranchant, J.G. Olsen, C.R. Grace, B.B. Kragelund, and K. Lindorff-Larsen. 2023. Assessment of models for calculating the hydrodynamic radius of intrinsically disordered proteins. *Biophys J.* 122:310–321.
56. Kirkwood, J.G., and J. Riseman. 1948. The Intrinsic Viscosities and Diffusion Constants of Flexible Macromolecules in Solution. *J Chem Phys.* 16:565–573.
57. 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. *Proc Natl Acad Sci U S A.* 118.
58. Nygaard, M., B.B. Kragelund, E. Papaleo, and K. Lindorff-Larsen. 2017. An Efficient Method for Estimating the Hydrodynamic Radius of Disordered Protein Conformations. *Biophys J.* 113.
59. 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.
60. Jensen, M.R., P.R.L. 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.
61. 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.
62. Nygaard, M., B.B. Kragelund, E. Papaleo, and K. Lindorff-Larsen. 2017. An Efficient Method for Estimating the Hydrodynamic Radius of Disordered Protein Conformations. *Biophys J.* 113:550–557.
63. Kirkwood, J.G., and J. Riseman. 1948. The intrinsic viscosities and diffusion constants of flexible macromolecules in solution. *J Chem Phys.* 16:565–573.

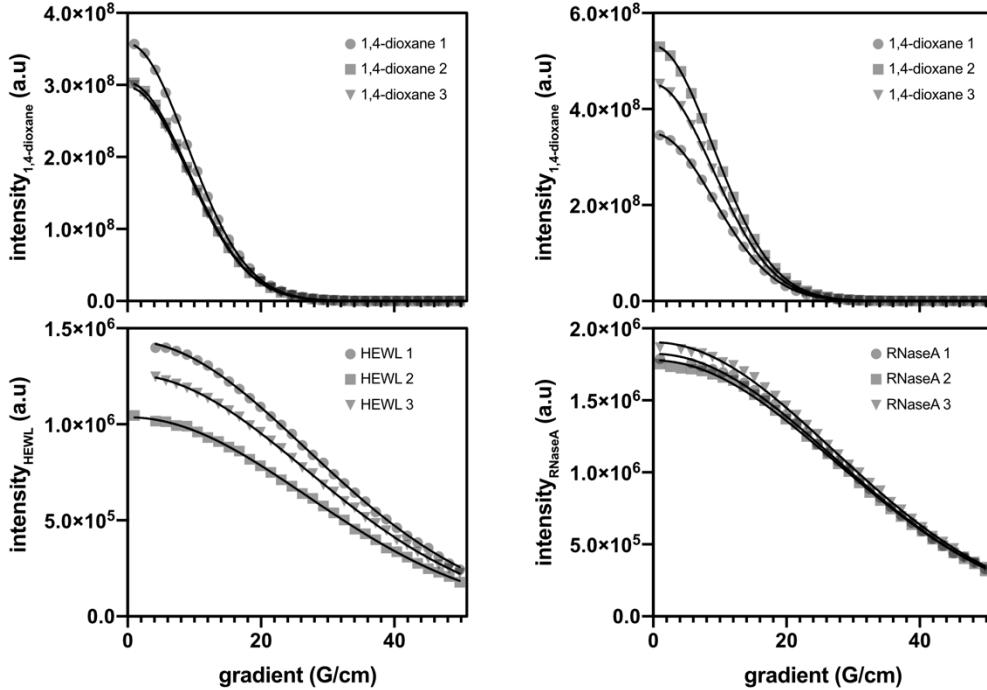
## Supplemental figures

### Revisiting the reference compound for determination of the hydrodynamic radius of proteins by pulsed field gradient NMR spectroscopy

Emil E. Tranchant<sup>1,2</sup>, Francesco Pesce<sup>1</sup>, Nina L. Jacobsen<sup>1,2</sup>, Catarina B. Fernandes<sup>1,2</sup>, Birthe B. Kragelund<sup>1,2\*</sup>, Kresten Lindorff-Larsen<sup>1\*</sup>



**Supplemental figure 1:** SDS-PAGE of the seven model proteins. The gel has been overloaded with protein to increase visibility of any impurities.



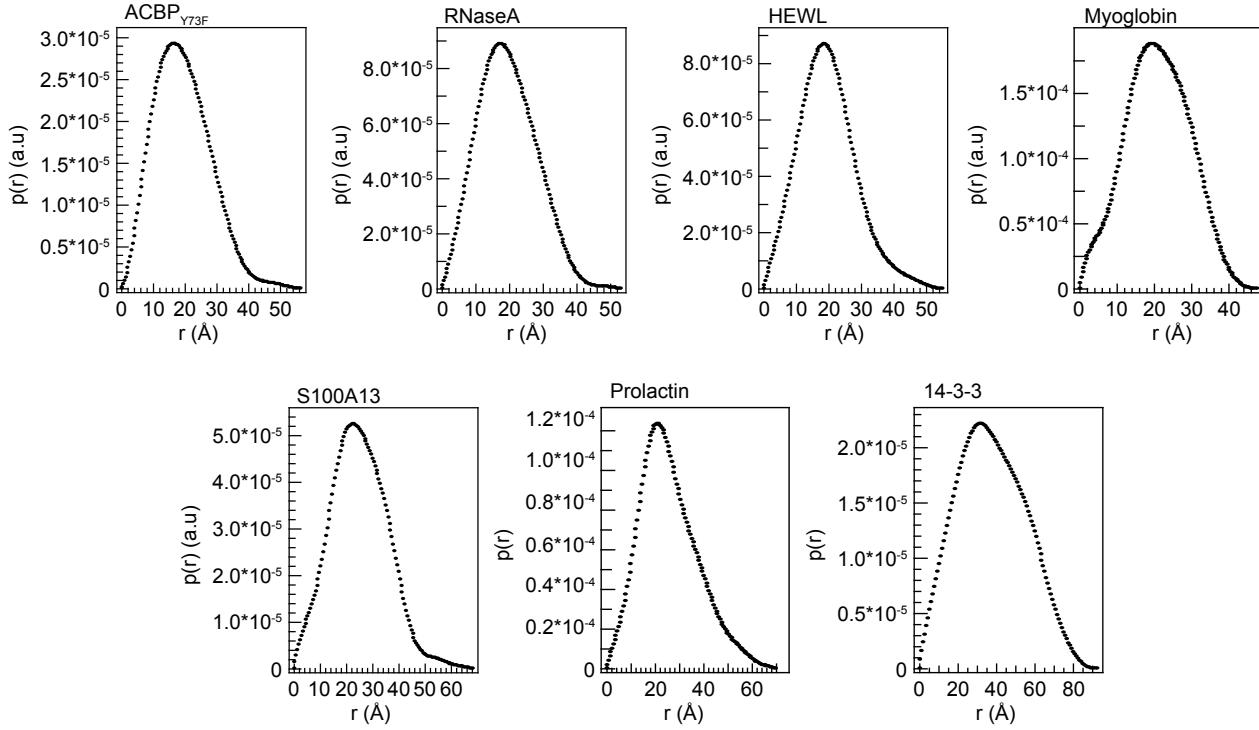
**Supplemental figure 2:** NMR peak intensity decays as a function of gradient strength. Intensity decays are shown for triplicate HEWL samples (left) and triplicate RNaseA samples (right) including decays of 1,4-dioxane (top) and proteins (bottom).

**Supplemental table 1:** Diffusion coefficients of triplicate HEWL and RNaseA and their averages, and the corresponding ratio of the diffusion coefficients (1,4-dioxane/protein) for each sample.

	HEWL 1	HEWL 2	HEWL 3	HEWL AVERAGE	RNASEA 1	RNASEA 2	RNASEA 3	RNASEA AVERAGE
<b>1,4-DIOXANE</b>	$1.09 \times 10^{-9}$ $\pm 3.09 \times 10^{-12}$	$1.08 \times 10^{-9}$ $\pm 1.86 \times 10^{-12}$	$1.08 \times 10^{-9}$ $\pm 4.52 \times 10^{-12}$	$1.08 \times 10^{-9}$ $\pm 9.46 \times 10^{-12}$	$1.06 \times 10^{-9}$ $\pm 7.45 \times 10^{-13}$	$1.09 \times 10^{-9}$ $\pm 2.72 \times 10^{-12}$	$1.08 \times 10^{-9}$ $\pm 1.82 \times 10^{-12}$	$1.08 \times 10^{-9}$ $\pm 5.26 \times 10^{-12}$
<b>PROTEIN</b>	$1.22 \times 10^{-10}$ $\pm 3.59 \times 10^{-13}$	$1.21 \times 10^{-10}$ $\pm 4.71 \times 10^{-13}$	$1.22 \times 10^{-10}$ $\pm 2.74 \times 10^{-13}$	$1.22 \times 10^{-10}$ $\pm 1.10 \times 10^{-12}$	$1.20 \times 10^{-10}$ $\pm 6.24 \times 10^{-13}$	$1.20 \times 10^{-10}$ $\pm 6.18 \times 10^{-13}$	$1.19 \times 10^{-10}$ $\pm 6.29 \times 10^{-13}$	$1.20 \times 10^{-10}$ $\pm 1.87 \times 10^{-12}$
<b>DIFFUSION RATIO (relative error)</b>	$8.96 \pm 0.05$	$8.88 \pm 0.05$	$8.84 \pm 0.06$	$8.89 \pm 0.11$ (1.3%)	$8.87 \pm 0.05$	$9.03 \pm 0.07$	$9.07 \pm 0.06$	$8.99 \pm 0.15$ (1.6%)

**Supplemental table 2:** Guinier derived  $R_g$  with standard error of the mean from triplicate SAXS measurements of HEWL (2mg/ml) and RNaseA (3mg/ml).

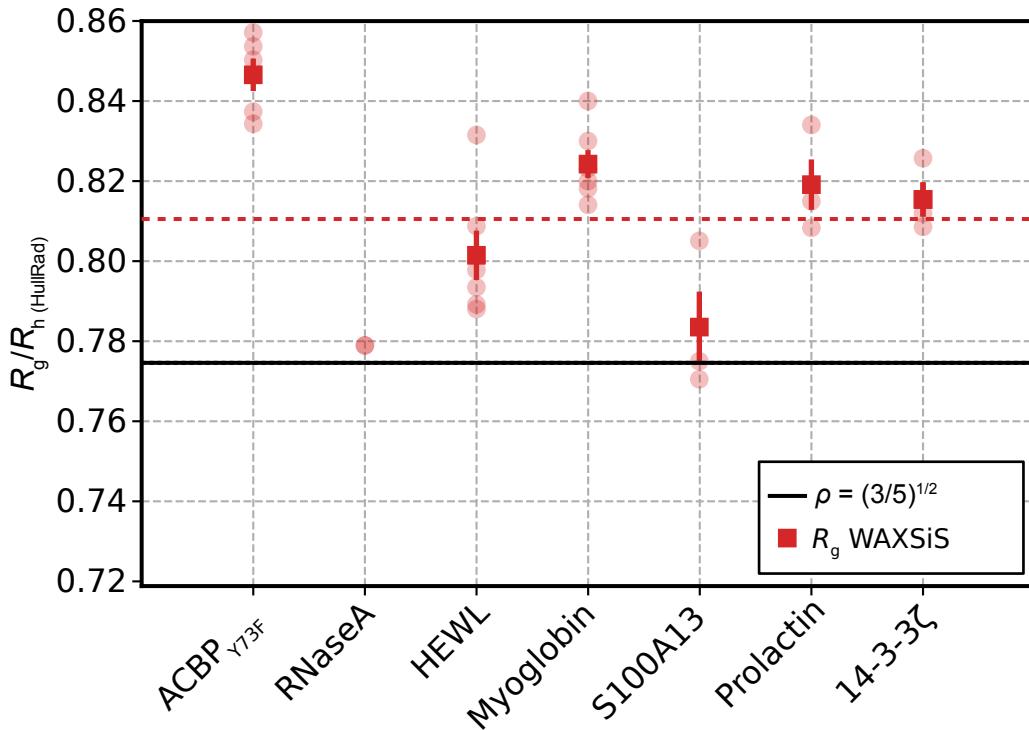
	HEWL 1	HEWL 2	HEWL 3	HEWL AVERAGE	RNASEA 1	RNASEA 2	RNASEA 3	RNASEA AVERAGE
<b>PROTEIN</b> $R_g$ (Å) (relative error)	$14.85 \pm 0.19$	$15.25 \pm 0.17$	$15.14 \pm 0.17$	$15.08 \pm 0.18$ (1.2%)	$15.44 \pm 0.15$	$15.35 \pm 0.15$	$15.34 \pm 0.16$	$15.38 \pm 0.15$ (1.0%)



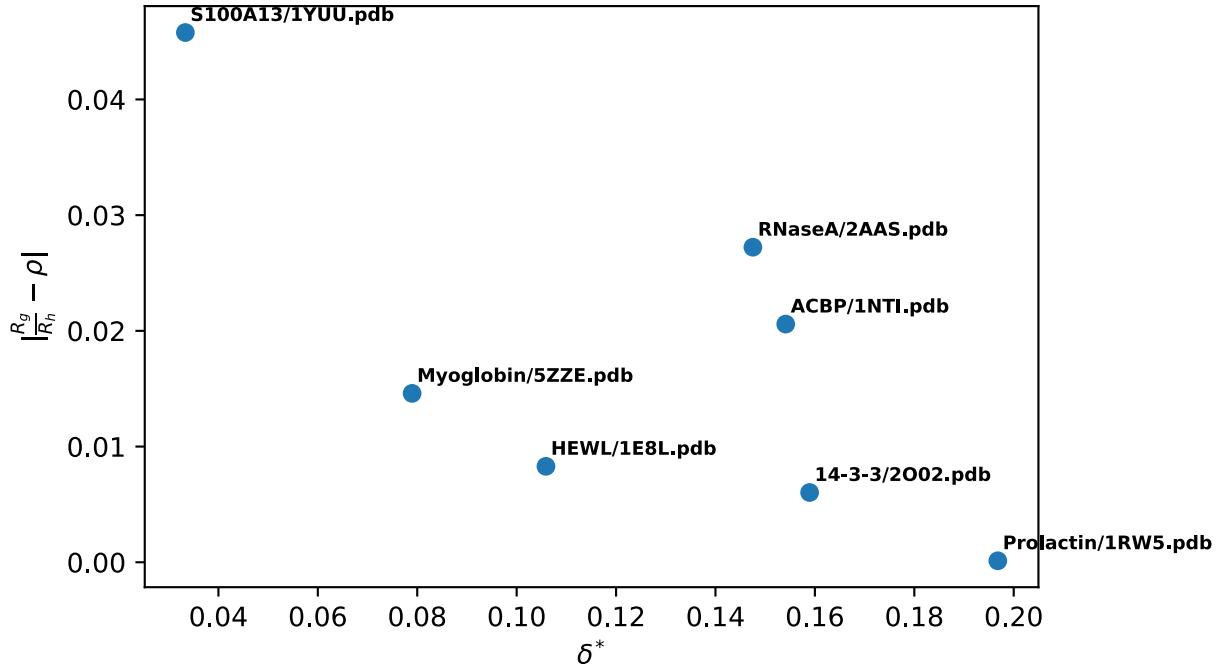
**Supplemental figure 3:** Pair-distance distribution plots ( $p(r)$ ) of the consensus scattering curve for each of the seven proteins.

**Supplemental table 3:** Experimentally derived  $R_g$  values by pair-distance distribution plots from batch SAXS. All values are reported with propagated standard errors of fit.

	ACBP	RNASEA	HEWL	MYOGLOBIN	S100A13	PROLACTIN	14-3-3
<b>PROTEIN</b> $R_g$ (Å) ( $p(r)$ )	$15.03 \pm 0.01$	$15.12 \pm 0.02$	$15.36 \pm 0.02$	$16.10 \pm 0.05$	$19.38 \pm 0.02$	$20.26 \pm 0.04$	$26.66 \pm 0.01$



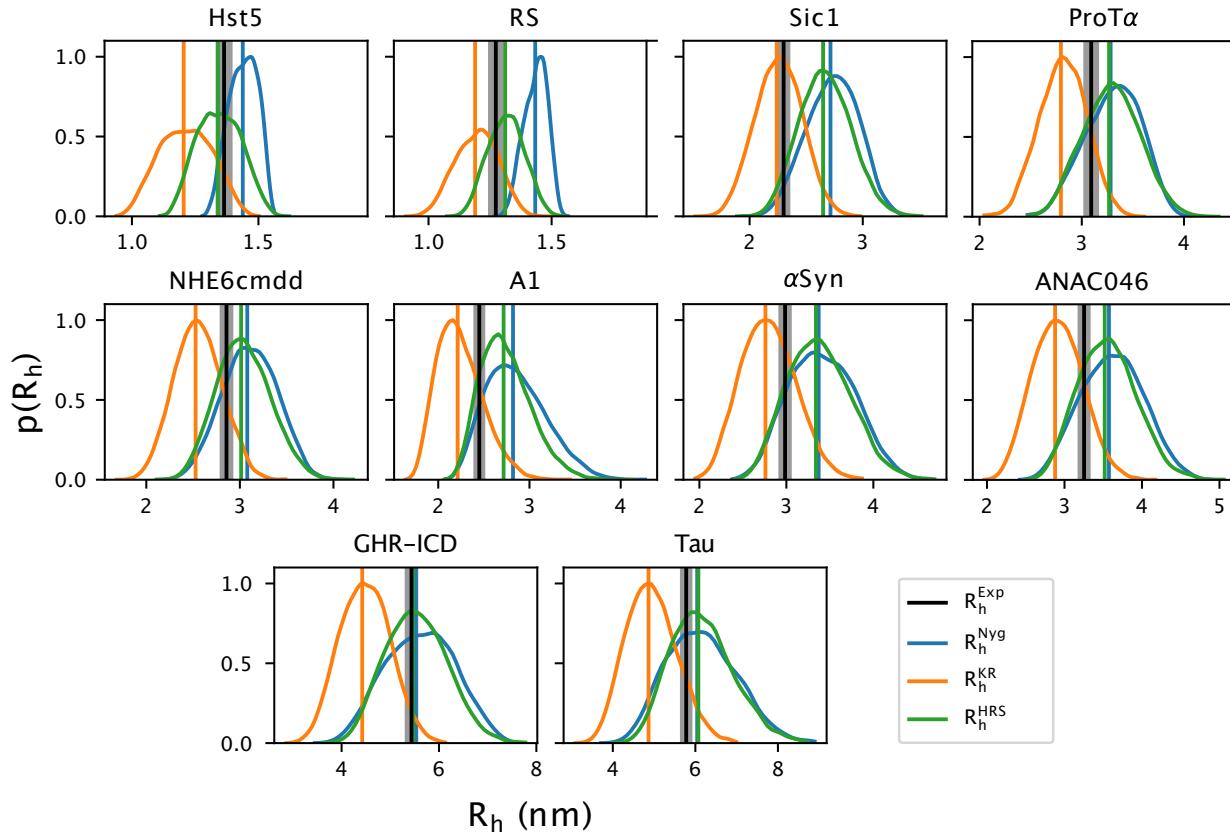
**Supplemental figure 4: Ratios of  $R_g/R_h$  obtained from 3D structures using an anhydrous  $R_g$  calculation and adding a water shell.** For each protein, three to five 3D structures were used, with the following PDB codes for each protein: HEWL (1e8l (28), 1dpx (34), 6abn (35), 5a3e (36)), RNaseA (2aas (27), 1fs3 (37), 1jvt (38), 4ooh(39), 1kf7 (40)), myoglobin (5zze (29), 1wla (41), 4dc8 (42), 5d5r (43), 5cn4 (43)), S100A13 (1yuu (30), 2h2k (44), 2egd (45)), ACBP (1nti (20), 1hb6 (46), 1hb8 (46), 2abd (47)), prolactin (1rw5 (31), 2q98 (48)), 14-3-3 $\zeta$  (2o02 (32), 1qja (49), 1gjb (50)). AlphaFold structures (51, 52) were also included for each protein. Individual ratios are shown as partially transparent markers. For each protein we calculated the mean ratio over different structures and the standard error of the mean (solid markers). Dashed red line represents the calculated average value of the  $R_g/R_h$ .



**Supplemental figure 5: Calculation of the asphericity ( $\delta^*$ , as defined in (33) for selected structures of the seven proteins and plot it against the absolute value of the difference between their  $R_g/R_h$  ratio to  $\rho$ .**

IDP	Hst5	RS	DSS1	Sic1	ProTa	NHE6-cmdd	A1	aSyn	ANAC-046	GHR-ICD	Tau
$R_h$ (Å) corrected	1.36	1.27	1.82	2.3	3.1	2.86	2.45	2.985	3.25	5.4	5.8
$R_h$ error (Å) corrected	0.04	0.02	0.07	0.1	0.1	0.05	0.08	0.006	0.06	0.1	0.2

**Supplemental table 4: PFG NMR derived  $R_h$  corrected according to a 1,4-dioxane  $R_h$  of  $2.27 \pm 0.04$  Å.**



**Supplemental figure 6: Probability distributions of the  $R_h$  and their ensemble averages calculated from 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 coordinates: the  $R_g$ -dependent Nygaard equation (Nyg, in blue), the Kirkwood-Riseman equation (KR, in orange), HullRadSAS (HRS, in green).**