bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Integrating smFRET, SAXS and NMR data to infer structural ensembles of an intrinsically-disordered protein

G-N. Gomes[1,2,*], M. Krzeminski[3,4], E. W. Martin[5], T. Mittag[5], T. Head-Gordon[6], J. D. Forman-Kay[3,4], and C. C. Gradinaru[1,2,*]

[1] *Department of Physics, University of Toronto, Toronto, Ontario M5G 1X8, Canada*

[2] *Department of Chemical and Physical Sciences, University of Toronto Mississauga, Mississauga, Ontario L5L 1C6, Canada*

[3] *Molecular Medicine Program, Hospital for Sick Children, Toronto, Ontario M5S 1A8*

[4] *Department of Biochemistry, University of Toronto, Toronto, Ontario M5G 1X8, Canada*

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

[6] *Departments of Chemistry, Bioengineering, Chemical and Biomolecular Engineering University of California, Berkeley, California 94720, United States*

[*] *To whom correspondence should be addressed. E-mail: gregory.gomes@mail.utoronto.ca, claudiu.gradinaru@utoronto.ca*

February 5, 2020

Classification: Biological Sciences, Biophysics and Computational Biology

Keywords: Intrinsically disordered protein | Conformational ensemble | Integrative modelling |Polymer physics

Author Contributions: G.G. and C.G. conceived the project; G.G. designed, performed and analysed smFRET experiments; E.M., G.G, and T.M. performed and analysed SAXS experiments; M.K., J.F.K., and T.H.G. guided the implementation of the ENSEMBLE analysis; G.G. performed simulations and calculated ensembles; G.G. and C.G. wrote the manuscript; G.G., T.M., T.H.G., J.F.K., and C.G. contributed insights, discussed results and interpretations, and edited the manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Intrinsically disordered proteins (IDPs) have fluctuating heterogeneous conformations, which makes structural characterization challenging. Transient long-range interactions in IDPs are known to have important functional implications. Thus, in order to calculate reliable structural ensembles of IDPs, the data used in their calculation must capture these important structural features. We use integrative modelling to understand and implement conformational restraints imposed by the most common structural techniques for IDPs: NMR spectroscopy, small-angle X-ray scattering (SAXS), and single-molecule Förster Resonance Energy Transfer (smFRET). Using the disordered N-terminal region of the Sic1 protein as a test case, we find that only Paramagnetic Relaxation Enhancement (PRE) and smFRET measurements are able to unambiguously report on transient long-range interactions. It is precisely these features which lead to deviations from homopolymer statistics and divergent structural inferences in non-integrative smFRET and SAXS analysis. Furthermore, we find that the sequence-specific deviations from homopolymer statistics are consistent with biophysical models of Sic1 function that are mediated by phospho-sensitive binding to its partner Cdc4. To our knowledge, these are the first conformational ensembles for an IDP in physiological conditions that are simultaneously consistent with smFRET, SAXS, and NMR data. Our results stress the importance of integrating the global and local structural information provided by SAXS and Chemical Shifts, respectively, with information on specific inter-residue distances from PRE and smFRET. Our integrative modelling approach and quantitative polymer-physics-based characterization of the experimentally-restrained ensembles could be used to implement a rigorous taxonomy for the description and classification of IDPs as heteropolymers.

Significance Statement

Intrinsically disordered proteins (IDPs) exhibit highly dynamic and heterogeneous conformations, which impedes rigorous structural characterization and understanding of their biological functions. Sic1 regulates the yeast cell cycle through phospho-sensitive binding to its partner Cdc4 and is paradigmatic of IDPs that bind tightly without partial/transient folding. In this paper, we integrated new and existing structural data from nuclear magnetic resonance, small-angle X-ray scattering and single-molecule fluorescence to calculate conformational ensembles for Sic1 and its phosphorylated state, pSic1. Data mining of these ensembles reveal unique features distinguishing Sic1/pSic1 from homopolymer statistics, such as overall compactness and large end-to-end distance fluctuations. Integrating experiments probing disparate scales, computational modelling, and polymer physics provides new and valuable insights into the conformation-to-function relationships in IDPs.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1 Introduction

Rather than encoding an energetically stable three-dimensional fold [1, 2], the primary amino-acid sequence of intrinsically disordered proteins (IDPs) encodes for a much flatter free-energy surface, allowing the protein to sample a large and heterogeneous set of conformations[3, 4]. These surfaces however, are not featureless, and IDPs may populate conformations with preferred local and long-range transient structure[5].

To better understand the relationship between primary sequence and the free-energy surface of IDPs, and to determine how the primary sequence encodes IDP functions, atomic-resolution descriptions of IDP conformational states have been developed[5, 6]. These structural ensembles are typically represented by a collection of conformations, each described by its atomic coordinates. The calculation or validation of these ensembles requires the input of multiple experimental observables, prompting the development of computational approaches for calculating structural ensembles consistent with a diverse set of experimental input data[5, 6].

One such approach is ENSEMBLE, which selects a subset of conformations from a starting pool of conformations to achieve agreement with Small Angle X-Ray Scattering (SAXS) and Nuclear Magnetic Resonance (NMR) data[5, 7, 8]. Because these ensembles have many free parameters, a polymer physics framework has also been used to concisely describe IDPs and unfolded proteins and to infer structural information from experimental data [9–12].

Given the importance of specific amino-acid motifs, sequence composition, and sequence patterning [13–16], it is not obvious that statistical laws derived for homopolymers in the limit of infinitely long chains will always satisfactorily describe finite-length heteropolymeric IDPs. This conformational diversity will likely not be captured by a single mean-field descriptor of polymer behaviour, such as a scaling exponent ν, or a single global dimension, such as the radius of gyration R_g. Though concise, homopolymer descriptions do not allow integration of different experimental observables, which is a problem especially if the IDP departs significantly from homopolymer statistics. However, the selection of sequences which deviate from simple statistical laws may be the result of a pressure to maintain a certain conformational ensemble for a certain biological function. Whether homopolymer or atomistically-detailed ensemble descriptions are used, the ultimate goal is to use structural information to generate hypotheses about protein function.

In yeast, the disordered protein Sic1 is eliminated via ubiquitination by the SCF^{Cdc4} ubiquitin ligase and subsequent degradation by the proteasome, allowing initiation of DNA replication[17, 18]. Sic1 binding to Cdc4 generally requires phosphorylation of a minimum of any six of the nine Cdc4 phosphodegron (CPD) sites on (full length) Sic1. This effectively sets a high threshold for the level of active G1 CDK required to initiate transition to S-phase. This ultrasensitivity with respect to G1 CDK activity ensures a coordinated onset of DNA synthesis and genomic stability[17]. The N-terminal 90 residues of Sic1 (henceforth Sic1) are sufficient for targeting to Cdc4 when highly phosphorylated (henceforth pSic1), making this region a valuable model for structural characterization[19].

Different biophysical models have been proposed to explain the ultrasensitive dependence of the Sic1-Cdc4 interaction on the number of phosphorylated CPDs. A kinetic model argued that the probability of Sic1 rebinding before diffusive exit could exhibit an ultrasensitive dependence on the number of phosphorylated CPDs, dependent on the timescales of diffusion and chain dynamics[20, 21]. A polyelectrostatic model suggested that long-range electrostatic interactions between the positively charged CPD binding pocket on Cdc4 and the constellation of unbound negatively charged phosphorylated CPDs on Sic1 could yield ultrasensitive binding[22, 23].

We sought to determine conformational ensembles of Sic1 and pSic1 with a combination of biophysical

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

methods that give insight into IDP conformations. We generated SAXS and single-molecule Förster Resonance Energy Transfer (smFRET) data on Sic1 and pSic1. These datasets initially appeared inconsistent when the datasets were analysed one at a time. Typical of reported smFRET-SAXS discrepancies, the smFRET dataset indicated more compact ensembles than did SAXS. We then used the ENSEMBLE method (Fig. 1) to understand the implications and advantages of combining multiple datasets, and to resolve the apparent discrepancy by joint refinement using the SAXS data and additional non-smFRET data. Specifically, we restrained conformational ensembles with SAXS data and with previously published NMR data[24, 25] to arrive at ensembles that are also consistent with the smFRET data.

We then contrasted the calculated conformational ensembles of Sic1 and pSic1 with the polymer-theoretic properties of infinitely long homopolymers to demonstrate the ways in which finite-length heteropolymers can deviate from these (often assumed) values. In contrast to infinitely long homopolymers which follow uniform power-law scaling of internal distances, Sic1 and pSic1 follow good-solvent scaling at short sequence separations and poor-solvent scaling at long sequence separations, leading to uncoupling between the root-mean-squared (rms) end-to-end distance (R_{ee}) and the rms radius of gyration (R_g). This heteropolymer effect has previously been hypothesized to explain apparently discrepant structural inferences between SAXS and smFRET [26–28].

We find through our approach of joint refinement by SAXS and NMR data, and validation by smFRET data, that the sequence-specific deviations from homopolymer statistics are consistent with biophysical models of Sic1 function, by encoding conformational states that give rise to ultrasensitive binding to its partner Cdc4. Our results provide a strong impetus for integrative modelling approaches over homopolymer approaches whenever possible.

2 Results

2.1 Measurements of R_{ee} and R_g inferred individually from smFRET or SAXS provide discrepant descriptions of Sic1 and pSic1 conformational ensembles

Fig. 2 A-C shows smFRET data measured on the Sic1 FRET construct, which is based on Sic1(1-90) and hereafter called Sic1. This construct was labelled stochastically at its termini with the FRET donor Alexa Fluor 488 and acceptor Alexa Fluor 647 (Förster radius $R_0 = 52.2 \pm 1.1$ Å, *SI Appendix 1.7*). The histogram is fit to a Gaussian function to extract the mean transfer efficiency $\langle E \rangle_{exp}$, which reports on the end-to-end distance distribution $P(r_{ee})$ (see *SI Appendix 1.10* for more details). Multisite phosphorylated Sic1 (pSic1) was generated via overnight incubation with Cyclin A/Cdk2 resulting in predominantly 6- and 7-fold phosphorylated Sic1, with a minor population of 5-fold phosphorylated Sic1 (determined by ESI mass spectrometery). Upon phosphorylation, $\langle E \rangle_{exp}$ decreases from 0.42 to 0.36 indicating chain expansion.

An estimate of the root-mean-squared end-to-end distance R_{ee} can be made from $\langle E \rangle$ by assuming $P(r_{ee})$ is described by a homopolymer model (*SI Appendix 1.10*). However, the smFRET data itself ($\langle E \rangle_{exp}$) does not suggest which (if any) homopolymer model is appropriate for a certain IDP. There is considerable flexibility in the choice of homopolymer model and in how to rescale the root-mean-squared inter-dye distance $R_{D,A}$ to R_{ee}, resulting in a range of R_{ee}. The inferred R_{ee} is 61-65 Å for Sic1 and 66-72 Å for pSic1, suggesting multisite phosphorylation results in an approximately 10% increase in R_{ee} (*SI Appendix, Table S2*). The smFRET data set, examined alone, suggests that Sic1 is 7-13% more compact than a self-avoiding random coil (RC) ensemble generated with the statistical coil generator TraDES for Sic1[32, 33].

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

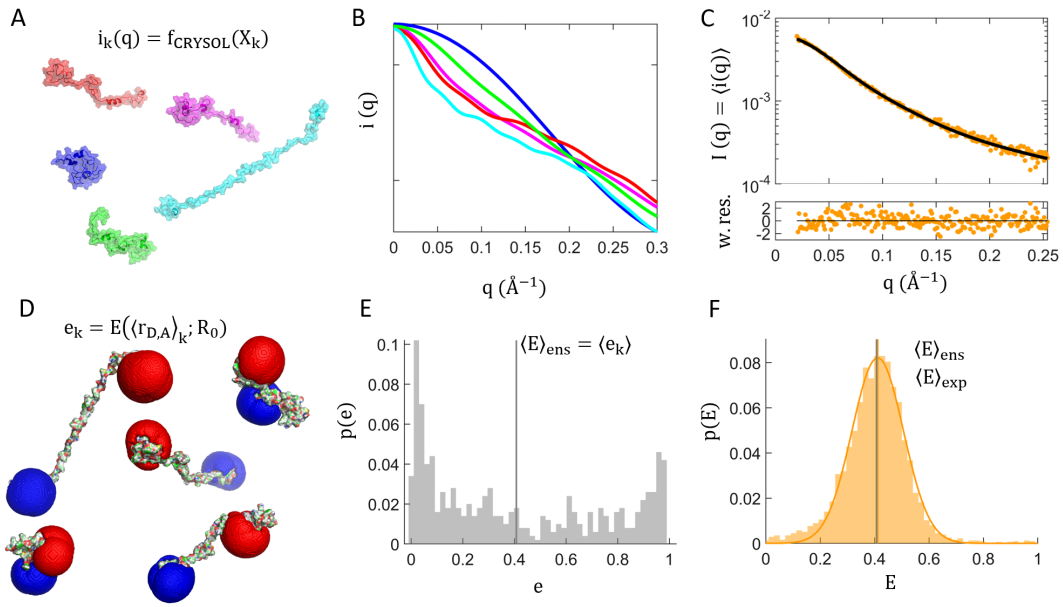

Figure 1: A schematic showing the ENSEMBLE approach for SAXS and smFRET data from an ensemble of structures $\mathbf{X} = [X_1, ..., X_k, ..., X_{N_{conf}}]$. (A-B) The SAXS intensity curve of each conformation $i_k(q)$ is back-calculated from the atomic coordinates using CRYSOL[29]. (C) The linear average of the CRYSOL-calculated SAXS profiles of individual conformers (black) is compared with the experimental SAXS profile (yellow). (D-E) Per-conformer FRET efficiencies e_k are calculated from the mean distance between dyes $\langle r_{DA}\rangle_k$ predicted by accessible volume simulations[30, 31]. (F) The ensemble-averaged transfer efficiency $\langle E\rangle_{ens} = \langle e_k\rangle$ (grey vertical line in E and F) is compared to the mean experimental transfer efficiency $\langle E\rangle_{exp}$ (yellow vertical line collinear with grey line in F).

.

To infer the root-mean-squared radius of gyration R_g from R_{ee} requires an additional assumption about the polymeric nature of system under study, namely the ratio $G = R_{ee}^2/R_g^2$. It has recently been shown that finite-length heteropolymeric chains can take on values of G that deviate from the values derived for infinitely long homopolymers in either the θ-state (Gaussian chains, $G = 6$) or excluded-volume (EV)-limit (self-avoiding walks, $G \approx 6.25$)[26–28]. Application of polymer-theoretic values of G to the smFRET inferred R_{ee} results in R_g 24-27 Å for Sic1 and 26-29 Å for pSic1 (*SI Appendix Table S3*). These inferred R_g values are systematically smaller than those inferred from the SAXS dataset (see below), or from integrated ensemble modelling (see below), similar to previously reported discrepancies between smFRET and SAXS[27, 34, 35].

Fig. 2 D-F shows SAXS data for Sic1 and pSic1. R_g was estimated to be approximately 30 Å for Sic1 and 32 Å for pSic1 using the Guinier approximation, and from the distance distribution function $P(r)$ obtained using the indirect Fourier transform of the regularized scattering curve (Fig. 2 E&F and *SI Appendix 2.1*). A model of chain statistics does not need to be specified, however, these methods are limited in describing IDPs and unfolded proteins[35, 36]. For example, the expanded and aspherical conformations of IDPs lead to a reduced range of scattering angles in which the Guinier approximation can be applied without systematic error[35]. The degree of underestimation of R_g increases as the maximum scattering angle q_{max} increases, while decreasing q_{max} reduces the number of points restraining the Guinier fit, which increases the uncertainty in R_g[35] (see also, *SI Appendix, Table S4*).

One solution to these limitations is to model the protein chain explicitly by generating ensembles of conformations. This is epitomized by the Ensemble Optimization Method (EOM) [37] and ENSEMBLE [7].

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

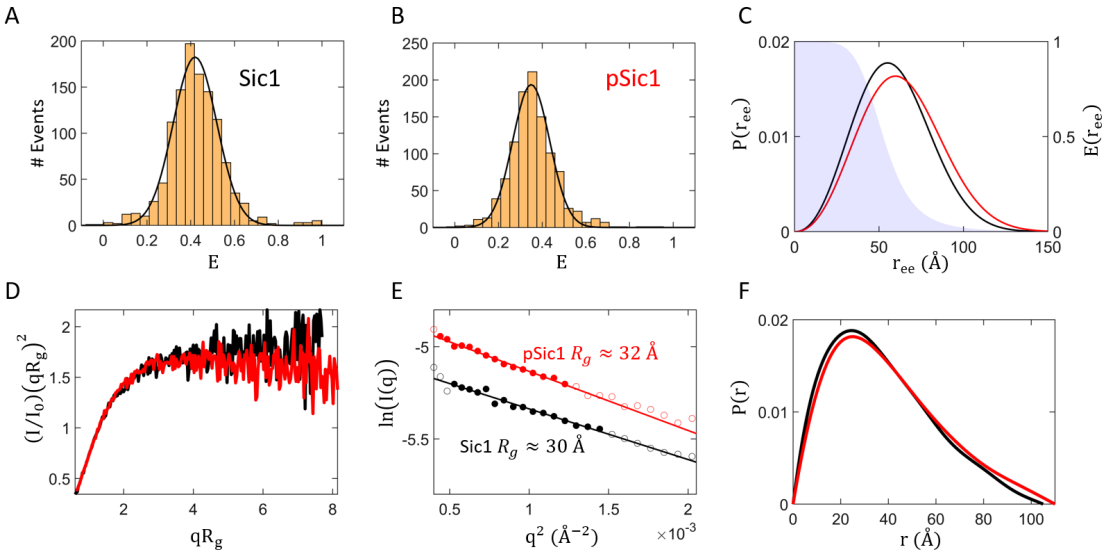

Figure 2: (A-B) smFRET efficiency (E) histograms of Sic1 (A) and pSic1 (B) labelled with Alexa Fluor 488 and Alexa Fluor 647 at positions -1C and T90C in TE buffer pH 7.5 150 mM NaCl. (C) Example SAW homopolymer $P(r_{ee})$ distributions (left vertical scale) for Sic1 (black, R_{ee} = 62.5 Å) and pSic1 (red, R_{ee} = 67.8 Å). The shaded underlying region shows the FRET distance dependence function $E(r_{ee})$ (right vertical scale). (D) Dimensionless Kratky plots of Sic1 (black) and pSic1 (red), normalized by initial intensity I_0 and the R_g estimated from the DATGNOM fit of the distance distribution function. (E) Guinier plots of Sic1 (black) and pSic1 (red). The solid circles are the data points selected for fitting ($q_{\max}R_g < 1.1$) and the solid lines show the Guinier fits using these data points. (F) The normalized distance distribution function $P(r)$ estimated by DATGNOM for Sic1 (black) and pSic1 (red).

Both approaches select a subset of conformations from an initial pool of conformations, such that the linear average of the CRYSOL-calculated SAXS profiles of individual conformers is in agreement with the full experimental SAXS profile (Fig. 1 A-C). However, the techniques differ in their generation of the initial pool of conformations and in the algorithm and cost-function used to minimize the disagreement with experiment (*SI Appendix 2.2*). Despite their differences, both ensemble-based approaches fit the SAXS data equally well, and resulted in nearly identical R_g values, which are similar to the "model-free" estimates (*SI Appendix, Table S5*). As was seen from the smFRET data, multisite phosphorylation results in chain expansion; this is in agreement with the SAXS data that indicates an approximately 6% increase in R_g.

Similarly, Riback and coworkers have recently introduced a procedure for fitting SAXS data by pre-generating ensembles of conformations with different properties (specifically, the strength and patterning of inter-residue attractions) and extracting dimensionless "molecular form factors" (MFFs)[34, 38]. The properties of interest are then inferred from the ensemble whose MFF best fits the data. Using the MFFs generated from homopolymer or heteropolymer simulations results in similar R_g to the aforementioned methods (*SI Appendix, Table S6*). Thus, R_g is strongly determined by the SAXS data, such that differences in the construction and refinement of models leads to minor differences in R_g.

Since conformations are explicitly represented in the EOM, ENSEMBLE, and MFF methods, the R_{ee} (and hence G) of the determined ensembles can be calculated. Although the various ensembles fit the SAXS data equally well, they have distinct values of R_{ee}, i.e., from 71-81 Å for Sic1 and from 71-87 Å for pSic1 depending on the method used (*SI Appendix, Tables S5&6*). These ensembles thus have G values from 5.1-7.7 and 4.6-7.9 for Sic1 and pSic1, respectively (*SI Appendix, Tables S5&6*). Unlike R_g, the SAXS data does not

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

uniquely determine R_{ee} and G, independent of modelling approach. Naturally, the accuracy of those aspects of the ensemble not strongly determined by the SAXS data will depend on the initial conformer generation and the optimization/selection algorithms. This suggests that integrating additional experimental data will improve structural inferences.

Similarly, for homopolymer-based smFRET inferences, modelling flexibility lead to a 4-6 Å range of inferred R_{ee} for Sic1, while the accuracy of $\langle E \rangle_{exp}$ (± 0.02), roughly corresponds to an uncertainty in the inferred R_{ee} of ± 2 Å. Likewise, G cannot be determined from the data itself, and must be assumed *a priori*. It would therefore be desirable to back calculate $\langle E \rangle_{ens}$ from a structural ensemble that is restrained by additional experimental data and to compare $\langle E \rangle_{ens}$ and $\langle E \rangle_{exp}$ directly.

2.2 Ensembles jointly restrained by SAXS and NMR data are consistent with measured FRET efficiencies

While attractive fluorophore interactions have been used to explain the smFRET technique's discrepant inferences for IDP ensembles relative to SAXS[38], we consider whether optimization using multiple solution data types might reduce or eliminate such discrepancies. One solution to reduce or eliminate the apparent smFRET and SAXS discrepancy is to jointly restrain ensembles with both data sets simultaneously[27, 35]. Instead, we hypothesized that jointly restraining ensembles with non-smFRET internal distance restraints and SAXS data could result in ensembles with back-calculated mean transfer efficiencies, $\langle E \rangle_{ens}$, in agreement with the experimental mean transfer efficiency $\langle E \rangle_{exp}$. This would provide complementary and much more compelling evidence that the smFRET and SAXS data sets are mutually consistent.

To provide non-smFRET information for joint refinement with SAXS data we used previously published NMR data on Sic1[24, 25]. Briefly, the NMR data consist of $^{13}C_{\alpha}$ and $^{13}C_{\beta}$ chemical shifts (CSs) from Sic1 and Paramagnetic Relaxation Enhancement (PRE) data from six single-cysteine Sic1 mutants using a nitroxide spin label (MTSL) coupled to cysteine residues in positions -1, 21, 38, 64, 83, and 90. We used the ENSEMBLE approach to calculate ensembles that are in agreement with the NMR and SAXS data (see *Materials and Methods* and *SI Appendix 3*). We used fluorophore accessible volume (AV) simulations[30] to back-calculate the mean transfer efficiency $\langle E \rangle_{exp}$ from the sterically accessible space of the dye attached to each conformation via its flexible linker (see *Materials and Methods* and *SI Appendix 3*).

The agreement of the experimental and back-calculated NMR and SAXS data was quantified using the reduced χ^2 metric to identifying statistically significant disagreement with experimental data ($\chi^2 >> 1$) (considering experimental and back-calculation errors, see *Materials and Methods* and *SI Appendix 3*). As a structureless null-hypothesis we also include a random coil (RC) ensemble generated by TraDES for Sic1. This RC ensemble is shown to be in very good agreement with excluded volume (EV) homopolymer statistics (see below). Table 1 summarizes the goodness of fit for Sic1 ensembles under various restraint combinations. *SI Appendix Fig. S3* shows typical examples of TraDES RC and SAXS-restrained fits to the experimental SAXS profiles.

For the TraDES RC ensemble, there is no statistically significant disagreement with the CS data ($\chi^2 < 1$). However, the agreement with the PRE, smFRET and SAXS data is poor. Internal distances between specific residues are generally larger in the RC ensemble than are expected from the PRE and smFRET data. On the other hand, the radius of gyration of this ensemble ($R_g \approx 28$ Å) is slightly smaller than SAXS-only estimates ($R_g \approx 30$ Å). Although specific internal distances in this ensemble are not reproduced, the mean-squared sum over all internal distances in this ensemble is only slightly less than indicated by the SAXS data, as

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1: Goodness of fit for Sic1 $N_{conf} = 500$ ensembles [a]

Restraints	χ^2 PRE	χ^2 $^{13}C_\alpha$ CS	χ^2 $^{13}C_\beta$ CS	χ^2 SAXS	$\langle E\rangle_{exp} - \langle E\rangle_{ens}$
TraDES RC (none)	1.51	0.479	0.466	1.85	0.12
SAXS	2.06	0.470	0.395	1.01	0.15
PRE	0.230	0.544	0.608	13.5	-0.22
SAXS+PRE	0.261	0.376	0.394	1.12	0.01
SAXS+PRE+CS+$\langle E\rangle$-filter	0.231	0.317	0.231	1.12	0.01

[a] Fit quality of $N_{conf} = 500$ ensembles derived by combining conformations from five independently calculated $N_{conf} = 100$ ensembles. Differences $|\langle E\rangle_{exp} - \langle E\rangle_{ens}| \leq \sqrt{\sigma^2_{E,exp} + \sigma^2_{E,ens}} \approx 0.02$ indicate no disagreement between back-calculated and experimental mean transfer efficiencies (see Materials and Methods).

$R_g = \sqrt{\frac{1}{2n^2}\sum_{ij}^{n}\langle r_{ij}^2\rangle}$.

When only the SAXS data are used as a restraint, the ensemble reproduces the SAXS curve very well. However, relative to the RC ensemble, the *overall* larger inter-residue distances in the SAXS-only ensemble further deteriorate the agreement with data reporting on specific inter-residue distances from PRE and smFRET.

When only the PRE data are used as a restraint, the agreement with the PRE data is achieved at the expense of not agreeing with all other observables. This ensemble reproduces specific inter-residue distances encoded by the PRE data, but not the overall distribution of inter-residue distances encoded by the SAXS data. A corollary of the r^{-6} PRE weighting is that the PRE ensemble average is dominated by contributions from compact conformations[39]. Consistent with this, the PRE-only ensemble is much more compact ($R_g \approx 22$ Å) than expected from the SAXS data. Similarly, the transfer efficiency calculated from the ensemble $\langle E\rangle_{ens}$ is larger than $\langle E\rangle_{exp}$ indicating either too short end-to-end distances overall, or some conformations with strongly underestimated end-to-end distances. Although for there is no disagreement with CS data ($\chi^2_\nu < 1$), the PRE-only ensemble is in worse agreement with the CS data than the TraDES RC or SAXS-only ensemble.

When the overall distribution of inter-residue distances from SAXS and the specific pattern of inter-residue distances from PRE are synthesized in one ensemble model, the transfer efficiency calculated from the ensemble, $\langle E\rangle_{ens}$, is in excellent agreement with the experimental transfer efficiency, $\langle E\rangle_{exp}$. The fit of the CS data (which were not used as a restraint for this ensemble) are also improved relative to the TraDES RC, the SAXS-only, and PRE-only ensembles. As was previously observed, generating ensembles by satisfying tertiary structure restraints seems to place some restraints on the backbone conformations[40].

Although the ensembles considered thus far are all consistent with CS data, we also calculated ensembles jointly restrained by SAXS, PRE, and CS data. However, introducing CSs (a non-distance restraint) decreases the relative weighting of distance-based restraints (SAXS and PRE) and causes a greater dispersion between $\langle E\rangle_{ens}$ for independently calculated ensembles. We therefore used a strategy of generating ensembles jointly restrained by SAXS, PRE, and CS data and filtered them against experimental transfer efficiencies (Table 1, SAXS+PRE+CS+$\langle E\rangle$-filter).

2.3 Integrative modelling provides a more comprehensive description of global dimensions than can be provided by SAXS or smFRET individually

To better understand the implications and advantages of combining multiple datasets we calculated global descriptions of Sic1 and pSic1 conformational ensemble dimensions (R_g, R_{ee}, and hydrodynamic radius R_h).

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SI Appendix Table S10 summarizes the global dimensions of five independently calculated ensembles with 100 conformations each ($N_{conf} = 100$).

The SAXS+PRE restrained ensembles have a mean $R_{ee} = 63.6 \pm 1.1$ Å for Sic1 and $R_{ee} = 64.7 \pm 0.7$ Å for pSic1. There is no significant discrepancy between R_{ee} inferred from smFRET using hompolymer models and the integrative approach ($\lessapprox$5% error). Inferences of R_{ee} using only the SAXS data overestimate the calculated mean R_{ee} by greater than 10% and depend highly on the initial conformer generation and the optimization/selection algorithm. Our approach of joint refinement/validation using SAXS, PRE and smFRET data addresses this issue. The Sic1 and pSic1 SAXS+PRE ensembles have back-calculated transfer efficiencies, $\langle E \rangle_{ens}$, which differ by five standard deviations, while their mean R_{ee} differ by less than one standard deviation. This is due to the increased sensitivity of $\langle E \rangle$ over R_{ee} due to the highly non-linear r^6 distance averaging. This demonstrates an additional advantage of integrative approaches, which use smFRET as explicit distance restraints or validation, rather than using derived values from the data via polymer theory assumptions.

The Sic1 and pSic1 SAXS+PRE ensembles' mean R_g ($R_g = 29.50 \pm 0.06$ Å for Sic1 and $R_g = 30.68 \pm 0.08$ for pSic1) are within 3% and 5% respectively of the model-free and the SAXS-only explicit chain estimates of R_g. In contrast, the determination of R_g from smFRET strongly depends on the model used. As shown below, the calculated ensembles have smaller values of G than do homopolymers resulting in systematically underestimated R_g when the polymer-theoretic values of G are used.

The calculated hydrodynamic radius, R_h, was found to be highly similar for all considered ensembles ($R_h \approx 21$ Å). The R_h of these ensembles is in excellent agreement with previously published pulsed-field gradient (PFG) diffusion NMR experiments, $R_h = 21.5 \pm 1.1$ Å and $R_h = 19.4 \pm 1.6$ Å for Sic1 and pSic1, respectively [25], and Fluorescence Correlation Spectroscopy (FCS) measurements $R_h = 22 \pm 2$ Å for Sic1[41].

2.4 Analysis of the conformational behaviour of calculated ensembles beyond global dimensions

We next sought to determine descriptions of the calculated conformational ensembles which go beyond global dimensions and would facilitate comparison with polymer theory reference states, and with IDPs and unfolded states of varying sequence and chain length, n. To this end, we used the fact that many aspects of homopolymer behaviour become universal, or independent of monomer identity, in the long chain (as $n \to \infty$) limit[42] (see below). This allowed us to clearly identify ways in which ensembles jointly restrained by SAXS and PRE data, and validated by smFRET data, deviate from homopolymer behaviour, and whether ensembles restrained only by SAXS data more resemble homopolymers, or the fully restrained ensembles.

For very long homopolymer chains, the scaling exponent ν tends to one of only three possible limits (1/3, 1/2, 0.588), describing the poor-solvent, θ-state, and excluded volume (EV)-limit respectively. Homopolymers in these limits have well-defined universal values for the size ratios $G = R_{ee}^2/R_g^2$ and $\rho = R_g/R_h$, the overall shape of the ensemble, as characterized by the average asphericity $\langle A \rangle$ ($A \sim 0$ for a sphere and $A \sim 1$ for a rod), the relative variance in the end-to-end distance distribution $\Delta R_{ee} = \sqrt{\langle r_{ee}^2 \rangle - \langle r_{ee} \rangle^2}/R_{ee}$, and the relative variance in the distribution of the shape of individual conformations $\Delta A = \sqrt{\langle A^2 \rangle - \langle A \rangle^2}/\langle A \rangle$. Table 2 summarizes the universal values expected for homopolymers in the θ-state or the EV-limit, in the case of very long chains (EV and θ-state $n \to \infty$) and for chains with similar length to Sic1 (EV $n = 90-100$).

As IDPs are finite-length heteropolymers, their apparent scaling exponents (ν_{app}, see below), can take on intermediate values to these three limits. Similarly, their behaviour can deviate from the universal values expected for homopolymers. Table 2 shows the nominally universal values calculated for the experimentally-

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

restrained ensembles. The TraDES RC, though not a homopolymer, is constructed with only excluded volume long-range interactions, and so is expected to have behaviour consistent with polymer theory predictions for an EV-limit polymers of similar chain-length (EV $n = 90 - 100$ Table 2).

Table 2: Nominally universal polymer properties of the TraDES RC ensemble, SAXS-only ensemble, and SAXS+PRE ensembles[a]

		G	ρ	$\langle A \rangle$	ΔA	ΔR_{ee}
Polymer Theory	EV ($n \to \infty$)	6.254	~1.59	0.431	0.442	0.374
	EV ($n = 90 - 100$)	6.32	1.27–1.39	0.4377	0.437	-
	θ-state ($n \to \infty$)	6	~ 1.5	0.396	-	0.422
Sic1	TraDES RC	6.37	1.33	0.438	0.438	0.352
	SAXS-only	6.34	1.35	0.447	0.430	0.363
	SAXS+PRE	4.99	1.33	0.349	0.461	0.417
pSic1	TraDES RC	6.35	1.33	0.438	0.432	0.366
	SAXS+PRE	4.83	1.31	0.361	0.440	0.388

[a] Reported values are the mean of 5 independently calculated $N_{conf} = 100$ ensembles. Table is reproduced in supplementary information with standard deviations of reported values and references for polymer theory values.

The values of G for the RC and SAXS-only ensembles are indistinguishable from the expected value for a homopolymer in the EV-limit ($G \approx 6.3$). Modelling the TraDES RC using phosphorylated residues at phosphorylation sites does not change G, consistent with the predicted universality. In contrast, for SAXS+PRE ensembles, G decreases from $G \approx 5.0$ for Sic1 to $G \approx 4.8$ for pSic1. Both values are outside the range $G_\theta = 6 \leq G \leq G_{EV} \approx 6.3$ despite the intermediate values of the apparent scaling exponents for Sic1 and pSic1 (see below).

For Sic1 and pSic1, ρ is not sensitive to deviations from homopolymer statistics at long sequence separations. The value of ρ remains ~1.3 for the RC, SAXS-only, and SAXS+PRE restricted ensembles, despite large changes in R_{ee} and G. The observed ρ are consistent with the range of polymer-theoretic values for a finite length EV homopolymer (EV $n = 90 - 100$ Table 2).

The Sic1 and pSic1 RC ensembles, have an average asphericity $\langle A \rangle$ very close to the polymer-theoretic value for a homopolymer in the EV-limit. The Sic1 SAXS-only ensembles are slightly more aspherical than the RC ensembles, consistent with the expected correlation between R_g and $\langle A \rangle$[27, 43]. Sic1 SAXS+PRE ensembles, however, are more spherical, with significantly lower $\langle A \rangle$, despite their larger-than-RC R_g. Similar to G, the values of $\langle A \rangle$ for the SAXS+PRE ensembles are not bound between the value predicted for the θ-state and EV-limit, despite these ensembles having intermediate values of the apparent scaling exponents (see below).

The relative variance in the end-to-end distance distribution, ΔR_{ee} is close to the EV-limit value ($\Delta R_{ee}^{EV} \approx 0.37$) for the TraDES RC and Sic1 SAXS-only restrained ensembles. In contrast, $\Delta R_{ee} \approx 0.42$ for the Sic1 SAXS+PRE ensembles, which is practically identical to the θ-state value. Sic1, although more compact than the RC, exhibits strong fluctuations in the end-to-end distance. Multisite phosphorylation appears to slightly reduce ΔR_{ee}, although it remains above the EV-limit values.

The RC and SAXS-only ensembles have a relative variance in the distribution of shapes, ΔA, similar to that of an EV-limit homopolymer, while that of the Sic1 SAXS+PRE ensemble is slightly larger. The broadness of the SAXS+PRE ensembles' A distribution stresses the fact that despite being more spherical than an EV polymer, the Sic1 ensemble contains conformations with a large distribution of shapes.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.5 Internal scaling profiles and apparent scaling exponents

To extract further insights regarding the effects of combining multiple solution data types on the statistics of internal distances in the ensembles, we calculated internal scaling profiles (ISPs, Fig. 3). ISPs quantify the mean internal distances ($R_{|i-j|} = \left\langle \langle r_{ij}^2 \rangle \right\rangle^{1/2}$) between all pairs of residues that are $|i-j|$ residues apart in the linear amino acid sequence (see *Materials and Methods*). The dependence of $R_{|i-j|}$ on sequence separation $|i-j|$ is often quantified by fitting to the power-law relation:

$$R_{|i-j|} = \sqrt{2 l_p b}\, |i-j|^{\nu_{app}} \tag{1}$$

where $b = 3.8$ Å is the distance between bonded C_α atoms and $l_p \approx 4$ Å is the persistence length. This persistence length was found to be applicable to a broad range of denatured and disordered states[27, 44, 45].

ISPs highlight important differences between ensembles. If the majority of internal distances are similar in the ISPs of two ensembles, their R_g values will be similar, as $R_g = \sqrt{\frac{1}{2n^2} \sum_{ij}^{n} \langle r_{ij}^2 \rangle}$ [27]. However, if their spatial separations start to diverge at long sequence separations, the ensembles will have dissimilar R_{ee} and $\langle E \rangle_{exp}$, when terminally labelled. This is illustrated by Fig. 3 A which shows the ISPs of the SAXS-only and SAXS+PRE Sic1 ensembles, which have similar R_g, but only the SAXS+PRE ensemble is consistent with the smFRET data.

Similarly, ISPs explain how R_{ee} and R_g can become decoupled for finite-length heteropolymers[27, 38]. Internal distances in very long homopolymers are expected to follow power-law scaling with a single global ν_{app} that defines the scaling behaviour at all sequence separations. We define the change in scaling behaviour at long sequence separations (ν_{app}^{long}) relative to intermediate sequence separations (ν_{app}^{int}) as $\Delta\nu_{app}^{ends} = \nu_{app}^{long} - \nu_{app}^{int}$. For long homopolymers we expect $\Delta\nu_{app}^{ends} \approx 0$. Negative (positive) values of $\Delta\nu_{app}^{ends}$ indicate ensembles with G less than (greater than) predicted from ν_{app}. The ISPs of SAXS-only (Fig. 3 A) and TraDES RC ensembles (Fig. 3 B&C) have $\Delta\nu_{app}^{ends} \approx 0$ consistent with the finding that these ensembles exhibit homopolymer behaviour (Table 2). In contrast, the SAXS+PRE ensembles, which are consistent with the smFRET data, have $\Delta\nu_{app}^{ends} << 0$, consistent with lower than expected G and $\langle A \rangle$[27].

To rigorously quantify deviations homopolymer statistics, we fit five independently calculated ensembles with 100 conformations ($N_{conf} = 100$, Table 3). In an intermediate regime ($15 \leq |i-j| \leq 51$), Sic1 and pSic1 SAXS+PRE ensembles have a scaling exponent $\nu_{app}^{int} \approx 0.53$, which suggests that at these scales, the physiological buffer is a marginally good solvent. At longer sequence separations ($51 < |i-j| \leq n_{res} - 5$), the ensembles show behaviour which is closer to the poor solvent scaling regime $\nu_{long} \approx 0.3$. We performed a paired t-test on the five $N_{conf} = 100$ ensembles to determine if the differences $\Delta\nu_{app}^{ends}$ come from a distribution with zero mean (Table 3). The deviations from homopolymer statistics (i.e., $\Delta\nu_{app}^{ends} \neq 0$) are statistically significant (p-value$<<$ 0.01) for the SAXS+PRE ensembles but not for the TraDES RC ensembles (p-value$\approx$ 0.14). The SAXS-only ensemble has weaker evidence for deviations from homopolymer statistics (p-value$\approx$ 0.03). Ensembles restrained only by SAXS data are more similar to EV homopolymers, than to the fully experimentally restrained ensembles.

2.6 Two dimensional scaling maps

To better describe the heteropolymeric nature of Sic1, a normalized two-dimensional (2D) scaling map was constructed (Fig. 4). In the first step, the ensemble-averaged distances between the C_α atoms of every unique pair of residues in the sequence is calculated for the experimentally-restrained ensemble ($\langle r_{ij}^2 \rangle_{ens}^{1/2}$), and for the respective TraDES RC ensemble ($\langle r_{ij}^2 \rangle_{\mathrm{RC}}^{1/2}$). Experimentally-restrained distances are normalized

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

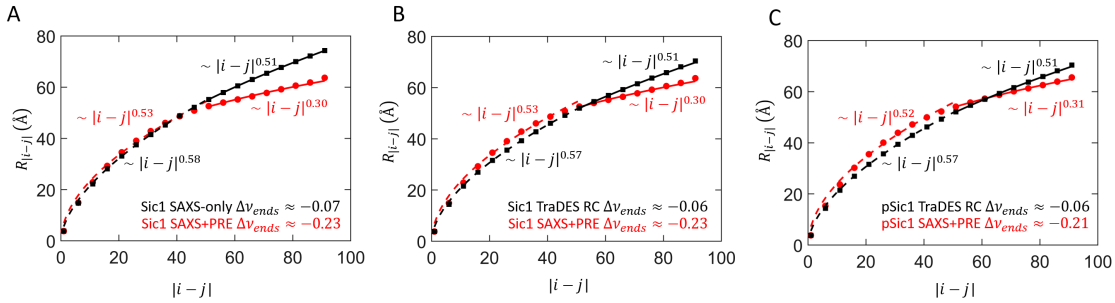

Figure 3: (A) $N_{conf} = 500$ Sic1 SAXS+PRE ensembles (circles) and Sic1 SAXS-only ensembles (squares) with fits to intermediate (dashed) and long (solid) sequence separations. (B) $N_{conf} = 500$ Sic1 SAXS+PRE ensembles (circles) and Sic1 TraDES RC (squares) with fits to intermediate (dashed) and long (solid) sequence separations. (C) $N_{conf} = 500$ pSic1 SAXS+PRE ensembles (black) and pSic1 TraDES RC (red) with fits to intermediate (dashed) and long (solid) sequence separations. For visualization, every fifth data point is shown.

Table 3: Fitting results for the TraDES RC ensemble, SAXS-only ensemble, and SAXS+PRE ensembles ISPs [a]

	TraDES RC[b]	Sic1 SAXS-only	Sic1 SAXS+PRE	pSic1 SAXS+PRE
ν_{app} (fixed $l_p = 4$ Å)	0.570	0.585	0.579	0.588
ν_{app}^{int}	0.566	0.583	0.527	0.524
ν_{app}^{long}	0.51	0.51	0.30	0.31
$\Delta\nu_{app}^{ends}$	-0.06	-0.07	-0.23	-0.21
Paired t-test p-value[c]	0.143	0.027	1.6×10^{-3}	5.1×10^{-4}

[a] Table results are the mean results from fitting 5 $N_{conf} = 100$ ensembles. Standard deviation of the mean for ν_{app} and ν_{app}^{int} is ≈ 0.005 and for ν_{app}^{long} and $\Delta\nu_{app}^{ends}$ is ≈ 0.03. See Materials and Methods for additional details.
[b] Sic1 TraDES RC and pSic1 TraDES RC result in nearly identical fits.
[c] Paired t-test for ν_{app}^{int} and ν_{app}^{long} differences.

by the RC distances and displayed as a 2D scaling map.

The normalized 2D scaling map for Sic1 (Fig. 4 A) displays regional biases for expansion ($\alpha_{ij} > 1$) and compaction ($\alpha_{ij} < 1$). Short internal distances $|i-j| \lessapprox 45$ show expansion relative to the RC, while $|i-j| \gtrapprox 60$ show compaction. The expansion, however, is heterogeneous. For example, the ~ 40 residue N-terminal region is significantly more expanded than the ~ 40 residue C-terminal region. Similar distinctions between the RC and pSic1 ensembles were observed (Fig. 4 B).

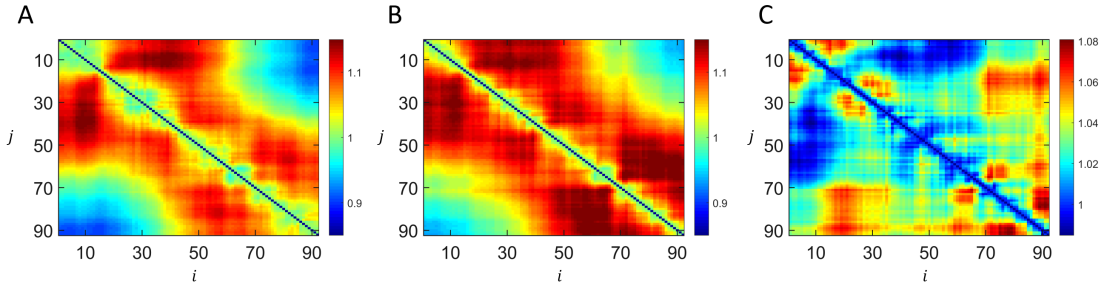

Figure 4: (A) Sic1 2D scaling map $\alpha_{ij} = \langle r_{ij}^2 \rangle_{ens}^{1/2} / \langle r_{ij}^2 \rangle_{RC}^{1/2}$ using the Sic1 (SAXS+PRE) $N_{conf} = 500$ and the Sic1 $N_{conf} = 500$ TraDES RC ensemble. (B) pSic1 2D scaling map $\alpha_{ij} = \langle r_{ij}^2 \rangle_{ens}^{1/2} / \langle r_{ij}^2 \rangle_{RC}^{1/2}$ using the pSic1 (SAXS+PRE) $N_{conf} = 500$ and the pSic1 $N_{conf} = 500$ TraDES RC ensemble. (C) pSic1 normalized by Sic1 dimensions.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

To compare Sic1 and pSic1 ensembles, the pSic1 ensemble was normalized by the Sic1 ensemble, (Fig. 4 C). This map describes the heterogeneous modulation of Sic1 upon multisite phosphorylation. Sic1 expansion upon phosphorylation has been attributed to transient tertiary contacts involving non-phosphorylated CPDs that are lost or weakened upon phosphorylation[24]. In our ensembles, expansion is also seen to be clustered around CPD sites, particularly those of the C-terminus. Expansion is also seen in the vicinity of Y14, previously implicated in tertiary interactions with CPDs[25] (see below).

2.7 Y14A mutation and phosphorylation disrupt tertiary contacts in Sic1

An intriguing possibility is that specific tertiary contacts, involving pi-pi[46] and cation-pi interactions[47], lead to compaction in Sic1. The Sic1 sequence is 23% residues with side chains containing pi-groups (F,N,Q,R, and Y), and 52% residues with relatively exposed pi-groups in peptide backbone amide groups (G,P,S, and T) [46]. In particular, PRE effects link CPDs with Y14 and ^{15}N relaxation experiments on Sic1 identified maxima in the R^2 rates near Y14[25]. Furthermore, the substitution Y14A led to an expansion in R_h of $\sim$20% in pSic1 [25]. We hypothesized that if Y14 engages in specific pi-pi and cation-pi interactions[1] throughout the chain, then removing its pi-character by mutation to alanine will disrupt these interactions, leading to larger R_{ee} and lower $\langle E \rangle_{exp}$. On the Kyte-Doolittle scale[48] this mutation increases the hydrophobicity ($H_Y = 0.36 \rightarrow H_A = 0.7$) suggesting that expansion would not result from reduced hydrophobicity.

We performed smFRET experiments for the Y14A mutants of Sic1 and pSic1 (Fig. 5 and *SI Appendix, Table S9*). Y14A mutation decreases Sic1 $\langle E \rangle_{exp}$ by approximately 7% (ca. 0.42 to 0.40, a small but reproducible shift). Similarly, Y14A mutation decreases pSic1 $\langle E \rangle_{exp}$ by approximately 8% (ca. 0.36 to 0.33). These results are consistent with chain expansion driven by the disruption of interaction with the pi-group of the tyrosine Y14. The cumulative effect of Y14A mutation and phosphorylation on Sic1 is to decrease $\langle E \rangle_{exp}$ by approximately 22%, such that $\langle E \rangle_{exp} = 0.33 \pm 0.02$ is similar to that of the Sic1 TraDES RC ($\langle E \rangle_{ens} = 0.30 \pm 0.01$).

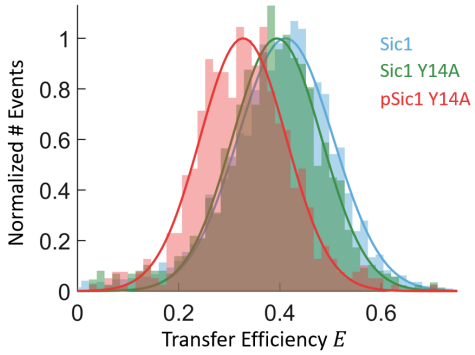

Figure 5: Y14A mutation and phosphorylation results in a shift to lower $\langle E \rangle_{exp}$ (more expanded conformations). Each histogram is normalized so that each Gaussian fit has a maximum of one.

[1]The Sic1 sequence has 6 K and 5 R residues. The predominantly 6- and 7-fold phosphorylated Sic1 studied here therefore has a net charge (excluding the dyes) of -1 to -3 assuming each phosphate group contribute -2 charge.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

3 Discussion

To better understand the implications and advantages of combining multiple datasets we generated SAXS and smFRET data on Sic1 and pSic1, and resolved their apparently discrepant inferences by joint refinement by the SAXS and PRE data. The ensembles restrained by SAXS and PRE data are, in addition, consistent with the smFRET data, chemical shift data, and hydrodynamic data (PFG-NMR and FCS).

We then explored the differences between ensembles restrained by different data types, and how they compare to homopolymer reference states, by calculating global descriptions of Sic1 and pSic1 conformational ensemble dimensions (R_g, R_{ee}, R_h), their size ratios (G and ρ), overall shape ($\langle A \rangle$) and the relative variances of their end-to-end distance and shape distributions (ΔR_{ee} and ΔA). To extract further insights we calculated internal scaling profiles and 2D scaling maps.

The picture that emerges when the entirety of the experimental data on Sic1 and pSic1 is considered, is that their conformational ensembles cannot be described by statistics derived for infinitely long homopolymers. This is unsurprising, given that Sic1 and pSic1 are finite-length heteropolymers. However, ensembles restrained only by the SAXS data are congruent with the set of homopolymer descriptions and scaling relationships for EV homopolymers. Neither the SAXS nor smFRET data, individually, suggest significant deviations from homopolymer statistics. Our results therefore provide a strong impetus for integrative modelling approaches over homopolymer approaches whenever multiple data types exist.

3.1 Experimental restraint contributions

Integrative ensemble modelling leverages the fact that different experimental restraints are sensitive to different aspects of disordered protein structure. Deviations from homopolymer statistics are encoded in the PRE and smFRET data, as these data types restrain distances between specific residues, rather than global averages. Deviations from homopolymer statistics at long sequence separations are encoded in the PRE data, such that joint restraint by PRE and SAXS data results in decoupled R_{ee} and R_g, and consistency with smFRET data.

We emphasize that the SAXS+PRE ensembles were not constructed by reweighting or selecting ensembles specifically to achieve agreement with $\langle E \rangle_{exp}$. In our approach, it was not guaranteed *a priori* that $\langle E \rangle_{ens}$ would match $\langle E \rangle_{exp}$, especially if either the introduction of PRE spin labels or smFRET fluorophores had perturbed the IDP ensemble. These deviations from homopolymer statistics are likely to be a general phenomenon for IDPs and unfolded proteins under refolding conditions, given their finite and heteropolymeric nature.

To fully understand the practical utility of different restraint types for characterizing types of structure in IDPs will require a more rigorous approach for scoring the probability of an ensemble on the basis of its agreement with diverse experimental data. For example, Lincoff and co-workers recently developed a Bayesian scoring formalism, the extended Experimental Inferential Structure Determination (X-EISD) method, to calculate the most probable ensembles for the drk SH3 unfolded state domain [49]. Using this method, they also found that FRET and PRE provide strong discriminatory power in determining the most probable ensemble.

3.2 Comparing smFRET- and SAXS-only estimates of ν_{app}

The Sic1 SAXS+PRE ISP in Fig. 3 A is consistent with the entirety of the experimental data on Sic1. We therefore consider this ISP as the benchmark in determining the scaling behaviour of the Sic1 ensemble and

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

compare it to recently published methods which infer scaling behaviour from only SAXS, or only smFRET data[34, 38, 50].

Analysing the SAXS data with the homopolymer MFF approach[34, 38] resulted in ν_{app} which is very similar to the ν_{app} obtained by globally fitting the ISPs with fixed l_p (*SI Appendix, Table S6*). Using a heterpolymer MFF analysis, which allows for deviations in power-law scaling at long sequence separations (MFF-het3, *SI Appendix, Table S6*), gave similar results to separately fitting ISPs at intermediate and long-sequence separations ($\nu_{app} = 0.56 \pm 0.01$ and $\Delta\nu_{app}^{ends} = -0.17 \pm 0.08$ for Sic1). However, in the absence of additional measurements, it is not clear when to prefer this model over the equally well-fit but more parsimonious homopolymer MFF models.

From the smFRET data, ν_{app} was estimated using the SAW-ν approach[50], which allows ν_{app} to vary in order to find a inter-dye distance distribution compatible with the observed $\langle E \rangle_{exp}$. The method assumes power law scaling of Eq. 1 with $|i-j|$ dictated by the segment length probed and with l_p and b fixed to the aforementioned values. Terminal labelling, therefore, results in a lower $\nu_{app} \sim 0.52$.

Part of the resolution to the recent controversy between ν_{app} determined by SAXS and smFRET experiments may be that negative values of $\Delta\nu_{app}^{ends}$ are common in IDPs and unfolded proteins under refolding conditions, causing lower estimates of ν_{app} by smFRET relative to SAXS. Since $\Delta\nu_{app}^{ends}$ may be undetectable without integrative modelling, the effect would be qualitatively similar to fluorophore-protein interactions, and like fluorophore-protein interactions it would disappear in high concentrations of denaturant where $\Delta\nu_{app}^{ends} \approx 0$[27, 38]. Deciding between fluorophore-interactions and heteropolymer effects requires an integrative modelling approach.

3.3 Conformation-to-function relationships

For soluble post-translationally modified IDPs, approximately good-solvent scaling may be unsurprising. The balance between chain-chain and chain-solvent interactions is a driving force for aggregation[54] and polymer theory predicts that proteins with overall good-solvent scaling in native-like conditions should remain soluble, while chains with poor-solvent scaling are expected to undergo aggregation. At short-to-intermediate sequence separations, good-solvent scaling provides read/write access of substrate motifs to modifying enzymes (e.g., phosphorylation and ubiquitination for Sic1).

Good-solvent scaling also confers advantages specifically to fuzzy or dynamic complexes as internal friction increases with increasing chain compaction[55]. Low internal friction and fast chain reconfigurations are therefore expected for short-to-intermediate separations. In a kinetic model of ultrasensitive binding, fast reconfiguration dynamics provides more opportunities for unbound CPDs to (re)bind before pSic1 diffuses out of proximity of Cdc4[20, 21, 23]. In the polyelectrostatic model, fast reconfiguration dynamics facilitates pSic1's dynamic interactions with Cdc4 through electrostatic averaging effects[22, 25].

The crossover to poor-solvent scaling at long sequence separations, $G < G_\theta < G_{EV}$, and $\langle A \rangle < \langle A \rangle_\theta < \langle A \rangle_{EV}$, imply that unbound CPDs that are sequence-distant from a bound CPD are on average closer to the WD40 binding pocket than they would be for an EV-chain, thus decreasing the solvent screening of electrostatic interactions. A prediction of the polyelectrostatic model is that decreasing the distance between the binding pocket and the overall mean-field charge distribution of Sic1, leads to sharper transitions in the fraction bound with respect to the number of phosphorylations[22]. In the kinetic model, these deviations from EV-statistics increase the effective concentration of CPDs in the vicinity of the binding pocket, which may increase the probability for any CPD to rebind before diffusive exit.

Although Sic1 ensembles are more spherical than EV-limit ensembles $\langle A \rangle < \langle A \rangle_\theta < \langle A \rangle_{EV}$, the relative

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

fluctuations in shape exceed those of the EV-limit ensembles (Table 2). Similarly, the relative fluctuations in R_{ee} (ΔR_{ee}) for these ensembles exceeds the expected fluctuations not only for an EV-limit polymer, but also those of the θ-state (Table 2). Large amplitude fluctuations in the shape and size of Sic1, effectively and rapidly sampling many different conformations, could allow CPDs in Sic1 to rapidly sample either the primary or secondary WD40 binding pocket, before the two proteins diffuse away. These fluctuations could also facilitate electrostatic averaging, permitting a mean-field treatment as assumed in the polyelectrostatic model.

4 Conclusions

Our work provides a high-resolution description of the conformational ensembles of Sic1 and pSic1 in physiological conditions. Our calculated ensembles are consistent with experimental data reporting on a wide range of spatial and sequence separation scales: local/secondary structure (CSs), non-local/tertiary structure (PRE and smFRET) and global/molecular size and shape (SAXS, PFG-NMR and FCS). To our knowledge, these are the first conformational ensembles for an IDP in native-like conditions (i.e., where heteropolymer effects are not abrogated by denaturant) which are consistent with smFRET, SAXS, and NMR data. Our results show that there are clear advantages of combining multiple datasets and that quantitative polymer-physics-based characterization of experimentally-restrained ensembles could be used to implement a rigorous taxonomy for the description and classification of IDPs as heteropolymers. The chain length independence of many of these properties facilitates comparison between different IDPs and unfolded states.

Our results suggest that for Sic1 and our dye pair, discrepant inferences between SAXS and smFRET cannot a priori be assumed to arise from "fluorophore-interactions." The impact of the fluorophores (or spin-labels) will of course depend on the physicochemical properties of the specific IDP sequence and the fluorophores (or spin-labels) used. Robustness to perturbation (e.g., labels or phosphorylation) may be built into Sic1's sequence via its patterning of charged and proline residues, as observed for other substrates of proline directed kinases (e.g., Ash1 [56]). Further ensemble modelling of IDPs will reveal to what extent robustness to labelling, and deviations from homopolymer statistics, are general to IDPs.

Ultimately, the prediction and elucidation of the structural details of IDPs and non-native states of proteins may prove to be more difficult than predicting the native structures of proteins with energetically stable three-dimensional folds. As such, an integrative use of multiple experiments probing disparate scales, computational modelling, and polymer physics, will provide valuable insights into IDPs and unfolded states and their biological functions.

5 Materials and Methods

5.1 Sic1 samples

The Sic1 N-terminal IDP region (1-90, henceforth Sic1) was expressed recombinantly as a Glutatione S-transferase (GST) fusion protein in *Escherichia coli* BL21 (DE3) codon plus cells and purified using glutathione-Sepharose affinity chromatography and cation-exchange chromatography. The correct molecular mass of the purified protein was verified by electrospray ionization mass spectrometry (ESI-MS).

A double cysteine variant of Sic1 (-1C-T90C) for smFRET experiments was generated via site directed mutagenesis from a single-cysteine mutant produced previously for PRE measurements[24, 25]. This construct was purified as above and the correct molecular mass of the purified protein was verified by ESI-MS. A Y14A mutant Sic1 (-1C-T90C-Y14A) was generated via site directed mutagenesis from the aforementioned double-cysteine mutant and was expressed, purified, and characterized using the same protocol.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The Sic1 smFRET construct was labelled stochastically with Alexa Fluor 488 C_5 Maleimide (ThermoFisher Scientific, Invitrogen, A10254) and Alexa Fluor 647 C_2 Maleimide (ThermoFisher Scientific, Invitrogen, A20347). After labelling with Alexa Fluor 647, cation-exchange chromatography was used to separate species with a single acceptor label, from doubly acceptor labelled and unlabelled species. The single-labelled species sample was then labelled with Alexa Fluor 488 and cation-exchange chromatography was used to separate doubly heterolabelled from acceptor only species. The correct mass of the doubly labelled sample was confirmed by mass spectrometry. The final FRET labelled sample was concentrated and buffer exchanged into PBS buffer pH 7.4 with 3 M GdmCl, 2 mM DTT and stored at -80 °C.

Phosphorylated samples were prepared by treatment of Sic1 with Cyclin A/Cdk2 (prepared according to Huang et al., [57])at a kinase:Sic1 ratio of 1:100 in the presence of 50 fold excess of ATP and 2.5 mM $MgCl_2$ overnight at 30 °C. The yield of phosphorylation reaction was determined by ESI-MS. Under these conditions the dominant species are 6- and 7-fold phosphorylated Sic1 (10195 Da and 10274 Da respectively) with a small fraction of 5-fold phosphorylated Sic1. After phosphorylation, the samples were buffer exchanged into PBS buffer pH 7.4 with 3 M GdmCl to prevent aggregation, denature kinase, and denature any phosphatases which may have inadvertently entered the solution. The samples were kept on ice in 4°C and measured within 24 hours.

Additional details regarding protein expression, purification and labelling are available in the supplementary information.

5.2 Single-molecule fluorescence

Single-molecule fluorescence experiments were performed on a custom-built multiparameter confocal microscope with microsecond alternating laser excitation. This instrumentation allows the simultaneous detection of the intensity, anisotropy, lifetime, and spectral properties of individual molecules and for the selection of fluorescence bursts in which both dyes are present and photophysically active.

Immediately prior to measurement samples were diluted to ∼50 pM in either (*i*) PBS buffer: 10 mM sodium phosphate and 140 mM NaCl pH 7.0, 1 mM EDTA (to replicate NMR measurement buffer of Ref [24]) or (*ii*) Tris buffer: 50 mM Tris and 150 mM NaCl, pH 7.5. (to replicate SAXS measurement buffer). No difference in $\langle E \rangle_{exp}$ was detected when comparing buffer conditions and results are shown for Tris buffer conditions.

The acquired data were subjected to multiparameter fluorescence analysis[58, 59] and ALEX filtering[60]. The burst search was performed using an All Photon Burst Search (APBS)[61, 62] with $M = 10$, $T = 500$ μs and $L = 50$. Transfer efficiencies were determined burst-wise and corrected for differences in the quantum yields of the dyes and detection efficiencies, as described in further detail in the *SI Appendix*.

The Förster radius R_0 was calculated assuming a relative dipole orientation factor $\kappa^2 = 2/3$ and the refractive index of water $n = 1.33$. The assumption of $\kappa^2 = 2/3$ is supported by subpopulation-specific steady-state anisotropies for the donor in the presence of the acceptor r_{DA} (*SI Appendix, Table S1*). The overlap integral J was measured for each sample and found not to change upon phosphorylation or Y14A mutation. The minimal variation in donor-only lifetimes τ_{D0} suggested minimal variation in the donor-quantum yield ϕ_D. R_0 was therefore calculated to be $R_0 = 52.2 \pm 1.1$ Å for all samples, and variation between samples within this uncertainty.

We estimate the precision for $\langle E \rangle_{exp}$ to be ca. 0.005 (for measurements performed on the same day, with approximately equal sample dependent calibration factors). We estimate the accuracy of $\langle E \rangle_{exp}$, $\sigma_{E,exp}$, to be ca. 0.02 (due to uncertainty in the instrumental and sample dependent calibration factors). Further details about the instrumentation, photoprotection, laser excitations, burst detection, filtering and multiparameter fluorescence analysis can be found in the *SI Appendix*.

5.3 Small-angle X-ray scattering

Small angle X-ray scattering data were collected at beamline 12-ID-B at the Argonne National Laboratory Advanced Photon Source. Protein samples were freshly prepared using size exclusion chromatography (GE Life Sciences, Superdex 75 10/300 GL) in a buffer containing 50 mM Tris pH 7.5, 150 mM NaCl, 5 mM DTT, and 2 mM TCEP. Fractions were loaded immediately after elution without further manipulation. Buffer collected one column volume after protein elution from the column was used to record buffer data before and after each protein sample. SAXS data were acquired manually; protein samples were loaded, then gently refreshed with a syringe pump to prevent x-ray damage. A Pilatus 2M detector provided q-range coverage from 0.015 $Å^{-1}$ to 1.0 $Å^{-1}$. Wide-angle x-ray scattering data were acquired with a Pilatus 300k detector and had a q range of 0.93 – 2.9 $Å^{-1}$. Calibration of the q-range calibration was performed with a silver behenate sample. Twenty sequential images were collected with 1 sec exposure time per image with each detector. Data were inspected for anomalous exposures and mean buffer data were subtracted from sample data using the WAXS water peak at q∼1.9 $Å^{-1}$ as a subtraction control. Details about the SAXS data analysis can be found in the *SI Appendix*.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

5.4 ENSEMBLE

ENSEMBLE 2.1 [7] was used to determine a subset of conformations from an initial pool of conformers created by the statistical coil generator TraDES[32, 33]. For restraining with SAXS and PRE data, both modules were given rank 1. When SAXS, PRE and CS modules were used, PRE and SAXS were given rank 1 and CSs were given rank 2. All other ENSEMBLE parameters were left at their default values.

To achieve a balance between the concerns of over-fitting (under-restraining) and under-fitting (over-restraining) we performed multiple independent ENSEMBLE calculations with 100 conformers, $N_{conf} = 100$, as suggested by Ref [40], and averaged the results from independent ensemble calculation or combined them to form ensembles with larger numbers of conformers (e.g., $N_{conf} = 500$). Structural features resulting from over-fitting (fitting the "noise" in the experimental data) should be averaged out in independent ENSEMBLE calculations, while rare states which are conserved in independently calculated ensembles (and thus have evidence from the data) should accumulate in weight when the ensembles are combined[40]. The agreement with experimental data and polymer-theory based ensemble descriptions were highly similar for repeated independently calculated ensembles (see *SI Appendix Table S10-11*). This supports their averaging or consolidation into one larger ensemble. Similarly, Lincoff and coworkers demonstrated overall convergence for $N_{conf} = 100$ ensembles of drk SH3 unfolded state domain[49].

NMR data was obtained from BMRB accession numbers 16657 (Sic1) and 16659 (pSic1)[24]. A total of 413 PRE restraints were used with a typical conservative upper- and lower-bound on PRE distance restraints of ±5 Å[39, 63]. This tolerance was used in computing the χ^2 metric for the PRE data. CSs were back-calculated using the SHIFTX calculator[64] and a total of 61 C_α CSs and 56 C_β CSs were used. The CS χ^2 metric was computed using the experimental uncertainty $\sigma_{res} = \pm 0.4$ ppm and the uncertainty in the SHIFTX calculator ($\sigma_{SHIFTX} = 0.98$ ppm for C_α CSs and $\sigma_{SHIFTX} = 1.10$ ppm for C_β CSs[64]).

CRYSOL[29] with default solvation parameters was used to predict the solution scattering from individual structures from their atomic coordinates. A total of 235 data points from $q = 0.02$ to $q = 0.254$ Å^{-1} were used in SAXS restrained ensembles. The SAXS χ^2 metric was computed using the experimental uncertainty in each data point. In principle, the calculation of χ^2 should also include an uncertainty in the SAXS back-calculation (in particular due to uncertainty in modelling the solvation shells of IDPs[65]).

Accessible volume (AV) simulations[30, 31] were used to predict the sterically accessible space of the dye attached to each conformation via its flexible linker. These calculations were performed using the "FRET-restrained positions and screening" (FPS) software [30] provided by the Seidel group. The mean distance between dyes for conformer k, $\langle r_{DA}\rangle_k$, is used to calculate the per-conformer FRET efficiency $e_k(\langle r_{DA}\rangle_k; R_0)$ (Figure 1D). As chain reconfigurations are much faster than the averaging time in smFRET, the smFRET experiment measures $P(E)$, the probability of observing a *burst* with efficiency E, rather than the probability distribution of per-conformer FRET efficiencies $p(e)$ (Figure 1E-F). Ensembles are therefore evaluated by the discrepancy $\langle E\rangle_{exp} - \langle E\rangle_{ens}$ (Figure 1F). The uncertainty in $\langle E\rangle_{ens}$, $\sigma_{E,ens}$, is ca. 0.01, primarily due to uncertainty in R_0; a similar value was obtained by Lincoff and coworkers[49]. Differences $|\langle E\rangle_{exp} - \langle E\rangle_{ens}| \leq \sqrt{\sigma^2_{E,exp} + \sigma^2_{E,ens}} \approx 0.02$ indicate no disagreement between back-calculated and experimental mean transfer efficiencies. A comprehensive description of the ENSEMBLE calculations, restraints and back-calculations can be found in the *SI Appendix*.

5.5 Polymer scaling analysis

The distance $R^2_{|i-j|} = \left\langle \langle r^2_{ij}\rangle \right\rangle_{ens}$ between C_α atoms is an average first over all pairs of residues that are separated by $|i-j|$ residues, and then over all conformations in the ensemble. The apparent scaling exponent ν_{app} was estimated by fitting an ISP calculated for each $N_{conf} = 100$ ensemble to the following expression:

$$\ln\left(R_{|i-j|}\right) = \nu_{app} \ln\left(|i-j|\right) + A_0 \tag{2}$$

Eq. 2 is derived for homopolymers in the infinitely long chain limit. Following Peran and coworkers[28], for finite-length chains, a lower bound of $|i-j| > 15$ was used to exclude deviations from infinitely long chain scaling behaviour at short sequence-separations and an upper bound of $|i-j < |n_{res} - 5$ was used to exclude deviations due to "dangling ends." With these restrictions, finite-length homopolymers are expected to be well fit by Eq. 2. Evenly spaced points in log-log space were used during fitting. Fitting the entire $15 < |i-j < |n_{res} - 5$ range was used to obtain ν_{app}. A_0 was either fixed at log(5.51) (l_p=4 Å) or left as a free fitting parameter.

To test for differences in scaling behaviour at intermediate and long sequence separations, the $15 < |i-j < |n_{res} - 5$ range was evenly divided into intermediate ν^{int}_{app} ($15 \leq |i-j| \leq 51$) and long ν^{long}_{app} regimes ($51 < |i-j| \leq n_{res} - 5$). Homopolymers are expected to have $\nu^{long}_{app} \approx \nu^{int}_{app}$. A paired t-test was performed in MATLAB R2018b using five $N_{conf} = 100$ ν^{long}_{app} and ν^{int}_{app} estimates, to test whether the set of $\Delta\nu_{ends} = \nu^{long}_{app} - \nu^{int}_{app}$ come from a normal distribution with mean equal to zero

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and unknown variance. Tests for normality (Jarque–Bera, Anderson-Darling, Lilliefors) were performed in MATLAB 2018b ($p > 0.4$). Table 3 reports the mean difference for the five ensembles.

6 Acknowledgements

G.N.G., M.K.,J.F-K., and T.H.G thank the National Institutes of Health for support under Grant 5R01GM127627-03. J.F-K. thanks the Natural Sciences and Engineering Research Council of Canada for support under RGPIN-2016-06718 Fund 490974. C.C. G. thanks the Natural Sciences and Engineering Research Council of Canada for support under RGPIN 2017–06030.

References

[1] Hue Sun Chan and Ken A. Dill. Energy Landscapes and the Collapse Dynamics of Homopolymers. *The Journal of Chemical Physics*, 99(3):2116–2127, August 1993. ISSN 0021-9606. doi: 10.1063/1.465277.

[2] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai. Navigating the Folding Routes. *Science*, 267(5204): 1619–1620, March 1995. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.7886447.

[3] Charles K Fisher and Collin M Stultz. Constructing Ensembles for Intrinsically Disordered Proteins. *Current Opinion in Structural Biology*, 21(3):426–431, June 2011. ISSN 0959-440X. doi: 10.1016/j.sbi.2011.04.001.

[4] Vladimir N. Uversky, Christopher J. Oldfield, and A. Keith Dunker. Intrinsically Disordered Proteins in Human Diseases: Introducing the D2 Concept. *Annual Review of Biophysics*, 37(1):215–246, 2008. doi: 10.1146/annurev.biophys.37.032807.125924.

[5] Joseph A. Marsh and Julie D. Forman-Kay. Ensemble modeling of protein disordered states: Experimental restraint contributions and validation. *Proteins: Structure, Function, and Bioinformatics*, 80 (2):556–572, 2012. ISSN 1097-0134. doi: 10.1002/prot.23220.

[6] Malene Ringkjøbing Jensen, Markus Zweckstetter, Jie-rong Huang, and Martin Blackledge. Exploring Free-Energy Landscapes of Intrinsically Disordered Proteins at Atomic Resolution Using NMR Spectroscopy. *Chemical Reviews*, 114(13):6632–6660, July 2014. ISSN 0009-2665. doi: 10.1021/cr400688u.

[7] Mickaël Krzeminski, Joseph A. Marsh, Chris Neale, Wing-Yiu Choy, and Julie D. Forman-Kay. Characterization of Disordered Proteins with ENSEMBLE. *Bioinformatics*, 29(3):398–399, February 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts701.

[8] Joseph A. Marsh and Julie D. Forman-Kay. Structure and Disorder in an Unfolded State under Nondenaturing Conditions from Ensemble Models Consistent with a Large Number of Experimental Restraints. *Journal of Molecular Biology*, 391(2):359–374, August 2009. ISSN 0022-2836. doi: 10.1016/j.jmb.2009.06.001.

[9] H S Chan and K A Dill. Polymer Principles in Protein Structure and Stability. *Annual Review of Biophysics and Biophysical Chemistry*, 20(1):447–490, 1991. doi: 10.1146/annurev.bb.20.060191.002311.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[10] Gregory-Neal Gomes and Claudiu C. Gradinaru. Insights into the Conformations and Dynamics of Intrinsically Disordered Proteins Using Single-Molecule Fluorescence. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1865(11, Part B):1696–1706, November 2017. ISSN 1570-9639. doi: 10.1016/j.bbapap.2017.06.008.

[11] Benjamin Schuler, Andrea Soranno, Hagen Hofmann, and Daniel Nettels. Single-Molecule FRET Spectroscopy and the Polymer Physics of Unfolded and Intrinsically Disordered Proteins. *Annual Review of Biophysics*, 45:207–231, May 2016. ISSN 1936-1238. doi: 10.1146/annurev-biophys-062215-010915.

[12] Alex S. Holehouse and Rohit V. Pappu. Collapse Transitions of Proteins and the Interplay Among Backbone, Sidechain, and Solvent Interactions. *Annual Review of Biophysics*, 47(1):19–39, 2018. doi: 10.1146/annurev-biophys-070317-032838.

[13] Celeste J Brown, Audra K Johnson, A Keith Dunker, and Gary W Daughdrill. Evolution and Disorder. *Current Opinion in Structural Biology*, 21(3):441–446, June 2011. ISSN 0959-440X. doi: 10.1016/j.sbi.2011.02.005.

[14] Alex S. Holehouse, Rahul K. Das, James N. Ahad, Mary O. G. Richardson, and Rohit V. Pappu. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophysical Journal*, 112(1):16–21, January 2017. ISSN 0006-3495. doi: 10.1016/j.bpj.2016.11.3200.

[15] Harry Amri Moesa, Shunichi Wakabayashi, Kenta Nakai, and Ashwini Patil. Chemical Composition Is Maintained in Poorly Conserved Intrinsically Disordered Regions and Suggests a Means for Their Classification. *Molecular BioSystems*, 8(12):3262–3273, October 2012. ISSN 1742-2051. doi: 10.1039/C2MB25202C.

[16] Taraneh Zarin, Caressa N. Tsai, Alex N. Nguyen Ba, and Alan M. Moses. Selection Maintains Signaling Function of a Highly Diverged Intrinsically Disordered Region. *Proceedings of the National Academy of Sciences*, 114(8):E1450–E1459, February 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1614787114.

[17] Piers Nash, Xiaojing Tang, Stephen Orlicky, Qinghua Chen, Frank B. Gertler, Michael D. Mendenhall, Frank Sicheri, Tony Pawson, and Mike Tyers. Multisite Phosphorylation of a CDK Inhibitor Sets a Threshold for the Onset of DNA Replication. *Nature*, 414(6863):514–521, November 2001. ISSN 1476-4687. doi: 10.1038/35107009.

[18] R. Verma, H. McDonald, J. R. Yates, and R. J. Deshaies. Selective degradation of ubiquitinated Sic1 by purified 26S proteasome yields active S phase cyclin-Cdk. *Molecular Cell*, 8(2):439–448, August 2001. ISSN 1097-2765. doi: 10.1016/s1097-2765(01)00308-2.

[19] R. Verma, R. M. Feldman, and R. J. Deshaies. SIC1 Is Ubiquitinated in Vitro by a Pathway That Requires CDC4, CDC34, and Cyclin/CDK Activities. *Molecular Biology of the Cell*, 8(8):1427–1437, August 1997. ISSN 1059-1524.

[20] Peter Klein, Tony Pawson, and Mike Tyers. Mathematical Modeling Suggests Cooperative Interactions between a Disordered Polyvalent Ligand and a Single Receptor Site. *Current Biology*, 13(19):1669–1678, September 2003. ISSN 0960-9822. doi: 10.1016/j.cub.2003.09.027.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[21] Jason W. Locasale. Allovalency Revisited: An Analysis of Multisite Phosphorylation and Substrate Rebinding. *The Journal of Chemical Physics*, 128(11):115106, March 2008. ISSN 0021-9606. doi: 10.1063/1.2841124.

[22] Mikael Borg, Tanja Mittag, Tony Pawson, Mike Tyers, Julie D. Forman-Kay, and Hue Sun Chan. Polyelectrostatic Interactions of Disordered Ligands Suggest a Physical Basis for Ultrasensitivity. *Proceedings of the National Academy of Sciences*, 104(23):9650–9655, June 2007. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0702580104.

[23] Veronika Csizmok, Stephen Orlicky, Jing Cheng, Jianhui Song, Alaji Bah, Neda Delgoshaie, Hong Lin, Tanja Mittag, Frank Sicheri, Hue Sun Chan, Mike Tyers, and Julie D. Forman-Kay. An Allosteric Conduit Facilitates Dynamic Multisite Substrate Recognition by the SCF\textsuperscriptCdc4 Ubiquitin Ligase. *Nature Communications*, 8:13943, January 2017. ISSN 2041-1723. doi: 10.1038/ncomms13943.

[24] Tanja Mittag, Joseph Marsh, Alexander Grishaev, Stephen Orlicky, Hong Lin, Frank Sicheri, Mike Tyers, and Julie D. Forman-Kay. Structure/Function Implications in a Dynamic Complex of the Intrinsically Disordered Sic1 with the Cdc4 Subunit of an SCF Ubiquitin Ligase. *Structure*, 18(4):494–506, March 2010. ISSN 0969-2126. doi: 10.1016/j.str.2010.01.020.

[25] Tanja Mittag, Stephen Orlicky, Wing-Yiu Choy, Xiaojing Tang, Hong Lin, Frank Sicheri, Lewis E. Kay, Mike Tyers, and Julie D. Forman-Kay. Dynamic Equilibrium Engagement of a Polyvalent Ligand with a Single-Site Receptor. *Proceedings of the National Academy of Sciences*, 105(46):17772–17777, November 2008. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0809222105.

[26] Jianhui Song, Gregory-Neal Gomes, Tongfei Shi, Claudiu C. Gradinaru, and Hue Sun Chan. Conformational Heterogeneity and FRET Data Interpretation for Dimensions of Unfolded Proteins. *Biophysical Journal*, 113(5):1012–1024, September 2017. ISSN 1542-0086. doi: 10.1016/j.bpj.2017.07.023.

[27] Gustavo Fuertes, Niccolò Banterle, Kiersten M. Ruff, Aritra Chowdhury, Davide Mercadante, Christine Koehler, Michael Kachala, Gemma Estrada Girona, Sigrid Milles, Ankur Mishra, Patrick R. Onck, Frauke Gräter, Santiago Esteban-Martín, Rohit V. Pappu, Dmitri I. Svergun, and Edward A. Lemke. Decoupling of Size and Shape Fluctuations in Heteropolymeric Sequences Reconciles Discrepancies in SAXS vs. FRET Measurements. *Proceedings of the National Academy of Sciences*, 114(31):E6342–E6351, August 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1704692114.

[28] Ivan Peran, Alex S. Holehouse, Isaac S. Carrico, Rohit V. Pappu, Osman Bilsel, and Daniel P. Raleigh. Unfolded states under folding conditions accommodate sequence-specific conformational preferences with random coil-like dimensions. *Proceedings of the National Academy of Sciences*, page 201818206, June 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1818206116.

[29] D. Svergun, C. Barberato, and M. H. J. Koch. CRYSOL – a Program to Evaluate X-Ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *Journal of Applied Crystallography*, 28(6):768–773, December 1995. ISSN 0021-8898. doi: 10.1107/S0021889895007047.

[30] Stanislav Kalinin, Thomas Peulen, Simon Sindbert, Paul J. Rothwell, Sylvia Berger, Tobias Restle, Roger S. Goody, Holger Gohlke, and Claus A. M. Seidel. A Toolkit and Benchmark Study for FRET-Restrained High-Precision Structural Modeling. *Nature Methods*, 9(12):1218–1225, December 2012. ISSN 1548-7105. doi: 10.1038/nmeth.2222.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[31] Simon Sindbert, Stanislav Kalinin, Hien Nguyen, Andrea Kienzler, Lilia Clima, Willi Bannwarth, Bettina Appel, Sabine Müller, and Claus A. M. Seidel. Accurate Distance Determination of Nucleic Acids via Förster Resonance Energy Transfer: Implications of Dye Linker Length and Rigidity. *Journal of the American Chemical Society*, 133(8):2463–2480, March 2011. ISSN 1520-5126. doi: 10.1021/ja105725e.

[32] Howard J. Feldman and Christopher W. V. Hogue. A fast method to sample real protein conformational space. *Proteins: Structure, Function, and Bioinformatics*, 39(2):112–131, May 2000. ISSN 1097-0134. doi: 10.1002/(SICI)1097-0134(20000501)39:2⟨112::AID-PROT2⟩3.0.CO;2-B.

[33] Howard J. Feldman and Christopher W. V. Hogue. Probabilistic Sampling of Protein Conformations: New Hope for Brute Force? *Proteins: Structure, Function, and Bioinformatics*, 46(1):8–23, January 2002. ISSN 1097-0134. doi: 10.1002/prot.1163.

[34] Joshua A. Riback, Micayla A. Bowman, Adam M. Zmyslowski, Catherine R. Knoverek, John M. Jumper, James R. Hinshaw, Emily B. Kaye, Karl F. Freed, Patricia L. Clark, and Tobin R. Sosnick. Innovative Scattering Analysis Shows That Hydrophobic Disordered Proteins Are Expanded in Water. *Science*, 358(6360):238–241, October 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aan5774.

[35] Alessandro Borgia, Wenwei Zheng, Karin Buholzer, Madeleine B. Borgia, Anja Schüler, Hagen Hofmann, Andrea Soranno, Daniel Nettels, Klaus Gast, Alexander Grishaev, Robert B. Best, and Benjamin Schuler. Consistent View of Polypeptide Chain Expansion in Chemical Denaturants from Multiple Experimental Methods. *Journal of the American Chemical Society*, 138(36):11714–11726, September 2016. ISSN 0002-7863. doi: 10.1021/jacs.6b05917.

[36] Alexey G. Kikhney and Dmitri I. Svergun. A Practical Guide to Small Angle X-Ray Scattering (SAXS) of Flexible and Intrinsically Disordered Proteins. *FEBS letters*, 589(19 Pt A):2570–2577, September 2015. ISSN 1873-3468. doi: 10.1016/j.febslet.2015.08.027.

[37] G. Tria, H. D. T. Mertens, M. Kachala, and D. I. Svergun. Advanced Ensemble Modelling of Flexible Macromolecules Using X-Ray Solution Scattering. *IUCrJ*, 2(2):207–217, March 2015. ISSN 2052-2525. doi: 10.1107/S205225251500202X.

[38] Joshua A. Riback, Micayla A. Bowman, Adam M. Zmyslowski, Kevin W. Plaxco, Patricia L. Clark, and Tobin R. Sosnick. Commonly Used FRET Fluorophores Promote Collapse of an Otherwise Disordered Protein. *Proceedings of the National Academy of Sciences*, page 201813038, April 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1813038116.

[39] Debabani Ganguly and Jianhan Chen. Structural Interpretation of Paramagnetic Relaxation Enhancement-Derived Distances for Disordered Protein States. *Journal of Molecular Biology*, 390(3): 467–477, July 2009. ISSN 0022-2836. doi: 10.1016/j.jmb.2009.05.019.

[40] Joseph A. Marsh and Julie D. Forman-Kay. Ensemble Modeling of Protein Disordered States: Experimental Restraint Contributions and Validation. *Proteins: Structure, Function, and Bioinformatics*, 80 (2):556–572, October 2011. ISSN 0887-3585. doi: 10.1002/prot.23220.

[41] Baoxu Liu, Darius Chia, Veronika Csizmok, Patrick Farber, Julie D. Forman-Kay, and Claudiu C. Gradinaru. The Effect of Intrachain Electrostatic Repulsion on Conformational Disorder and Dynamics of the Sic1 Protein. *The Journal of Physical Chemistry B*, 118(15):4088–4097, April 2014. ISSN 1520-6106. doi: 10.1021/jp500776v.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[42] L Schäfer. *Excluded Volume Effects in Polymer Solutions as Explained by the Renormalization Group*. Springer, Berlin, 1999.

[43] Jianhui Song, Gregory-Neal Gomes, Claudiu C. Gradinaru, and Hue Sun Chan. An Adequate Account of Excluded Volume Is Necessary To Infer Compactness and Asphericity of Disordered Proteins by Förster Resonance Energy Transfer. *The Journal of Physical Chemistry B*, 119(49):15191–15202, December 2015. ISSN 1520-6106. doi: 10.1021/acs.jpcb.5b09133.

[44] Hagen Hofmann, Andrea Soranno, Alessandro Borgia, Klaus Gast, Daniel Nettels, and Benjamin Schuler. Polymer Scaling Laws of Unfolded and Intrinsically Disordered Proteins Quantified with Single-Molecule Spectroscopy. *Proceedings of the National Academy of Sciences*, 109(40):16155–16160, October 2012. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1207719109.

[45] Huan-Xiang Zhou. Polymer Models of Protein Stability, Folding, and Interactions. *Biochemistry*, 43 (8):2141–2154, March 2004. ISSN 0006-2960. doi: 10.1021/bi036269n.

[46] Robert McCoy Vernon, Paul Andrew Chong, Brian Tsang, Tae Hun Kim, Alaji Bah, Patrick Farber, Hong Lin, and Julie Deborah Forman-Kay. Pi-Pi contacts are an overlooked protein feature relevant to phase separation. https://elifesciences.org/articles/31486, February 2018.

[47] Timothy J. Nott, Evangelia Petsalaki, Patrick Farber, Dylan Jervis, Eden Fussner, Anne Plochowietz, Timothy D. Craggs, David P. Bazett-Jones, Tony Pawson, Julie D. Forman-Kay, and Andrew J. Baldwin. Phase Transition of a Disordered Nuage Protein Generates Environmentally Responsive Membraneless Organelles. *Molecular Cell*, 57(5):936–947, March 2015. ISSN 1097-2765. doi: 10.1016/j.molcel.2015.01.013.

[48] Jack Kyte and Russell F. Doolittle. A Simple Method for Displaying the Hydropathic Character of a Protein. *Journal of Molecular Biology*, 157(1):105–132, May 1982. ISSN 0022-2836. doi: 10.1016/0022-2836(82)90515-0.

[49] James Lincoff, Mickael Krzeminski, Mojtaba Haghighatlari, João M. C. Teixeira, Gregory-Neal W. Gomes, Claudiu C. Gradinaru, Julie D. Forman-Kay, and Teresa Head-Gordon. Extended Experimental Inferential Structure Determination Method for Evaluating the Structural Ensembles of Disordered Protein States. *arXiv:1912.12582 [physics, q-bio] (submitted)*, December 2019.

[50] Wenwei Zheng, Gül H. Zerze, Alessandro Borgia, Jeetain Mittal, Benjamin Schuler, and Robert B. Best. Inferring Properties of Disordered Chains from FRET Transfer Efficiencies. *The Journal of Chemical Physics*, 148(12):123329, February 2018. ISSN 0021-9606. doi: 10.1063/1.5006954.

[51] Rahul K. Das and Rohit V. Pappu. Conformations of Intrinsically Disordered Proteins Are Influenced by Linear Sequence Distributions of Oppositely Charged Residues. *Proceedings of the National Academy of Sciences*, 110(33):13392–13397, August 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1304749110.

[52] Abhinav Nath, Maria Sammalkorpi, David C. DeWitt, Adam J. Trexler, Shana Elbaum-Garfinkle, Corey S. O'Hern, and Elizabeth Rhoades. The Conformational Ensembles of α-Synuclein and Tau: Combining Single-Molecule FRET and Simulations. *Biophysical Journal*, 103(9):1940–1949, November 2012. ISSN 0006-3495. doi: 10.1016/j.bpj.2012.09.032.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[53] Lucas Sawle and Kingshuk Ghosh. A Theoretical Method to Compute Sequence Dependent Configurational Properties in Charged Polymers and Proteins. *The Journal of Chemical Physics*, 143(8):085101, August 2015. ISSN 0021-9606. doi: 10.1063/1.4929391.

[54] Rohit V. Pappu, Xiaoling Wang, Andreas Vitalis, and Scott L. Crick. A Polymer Physics Perspective on Driving Forces and Mechanisms for Protein Aggregation. *Archives of biochemistry and biophysics*, 469(1):132–141, January 2008. ISSN 0003-9861. doi: 10.1016/j.abb.2007.08.033.

[55] Wenwei Zheng, Hagen Hofmann, Benjamin Schuler, and Robert B. Best. Origin of Internal Friction in Disordered Proteins Depends on Solvent Quality. *The Journal of Physical Chemistry B*, 122(49): 11478–11487, December 2018. ISSN 1520-6106. doi: 10.1021/acs.jpcb.8b07425.

[56] Erik W. Martin, Alex S. Holehouse, Christy R. Grace, Alex Hughes, Rohit V. Pappu, and Tanja Mittag. Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation. *Journal of the American Chemical Society*, 138(47):15323–15335, November 2016. ISSN 1520-5126. doi: 10.1021/jacs.6b10272.

[57] Yongqi Huang, Mi-Kyung Yoon, Steve Otieno, Moreno Lelli, and Richard W. Kriwacki. The Activity and Stability of the Intrinsically Disordered Cip/Kip Protein Family Are Regulated by Non-Receptor Tyrosine Kinases. *Journal of Molecular Biology*, 427(2):371–386, January 2015. ISSN 1089-8638. doi: 10.1016/j.jmb.2014.11.011.

[58] Evangelos Sisamakis, Alessandro Valeri, Stanislav Kalinin, Paul J. Rothwell, and Claus A. M. Seidel. Chapter 18 - Accurate Single-Molecule FRET Studies Using Multiparameter Fluorescence Detection. In Nils G. Walter, editor, *Methods in Enzymology*, volume 475 of *Single Molecule Tools, Part B:Super-Resolution, Particle Tracking, Multiparameter, and Force Based Methods*, pages 455–514. Academic Press, January 2010. doi: 10.1016/S0076-6879(10)75018-7.

[59] Volodymyr Kudryavtsev, Martin Sikor, Stanislav Kalinin, Dejana Mokranjac, Claus A. M. Seidel, and Don C. Lamb. Combining MFD and PIE for Accurate Single-Pair Förster Resonance Energy Transfer Measurements. *Chemphyschem: A European Journal of Chemical Physics and Physical Chemistry*, 13 (4):1060–1078, March 2012. ISSN 1439-7641. doi: 10.1002/cphc.201100822.

[60] Achillefs N. Kapanidis, Ted A. Laurence, Nam Ki Lee, Emmanuel Margeat, Xiangxu Kong, and Shimon Weiss. Alternating-laser excitation of single molecules. *Accounts of chemical research*, 38(7):523–533, 2005. ISSN 0001-4842.

[61] C. Eggeling, S. Berger, L. Brand, J. R. Fries, J. Schaffer, A. Volkmer, and C. A. Seidel. Data Registration and Selective Single-Molecule Analysis Using Multi-Parameter Fluorescence Detection. *Journal of Biotechnology*, 86(3):163–180, April 2001. ISSN 0168-1656.

[62] Eyal Nir, Xavier Michalet, Kambiz M. Hamadani, Ted A. Laurence, Daniel Neuhauser, Yevgeniy Kovchegov, and Shimon Weiss. Shot-Noise Limited Single-Molecule FRET Histograms:\, Comparison between Theory and Experiments. *The Journal of Physical Chemistry B*, 110(44):22103–22124, November 2006. ISSN 1520-6106. doi: 10.1021/jp063483n.

[63] Joel R. Gillespie and David Shortle. Characterization of Long-Range Structure in the Denatured State of Staphylococcal Nuclease. I. Paramagnetic Relaxation Enhancement by Nitroxide Spin labels11Edited

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.05.935890; this version posted February 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

by P. E. Wright. *Journal of Molecular Biology*, 268(1):158–169, April 1997. ISSN 0022-2836. doi: 10.1006/jmbi.1997.0954.

[64] Stephen Neal, Alex M. Nip, Haiyan Zhang, and David S. Wishart. Rapid and Accurate Calculation of Protein 1H, 13C and 15N Chemical Shifts. *Journal of Biomolecular NMR*, 26(3):215–240, July 2003. ISSN 1573-5001. doi: 10.1023/A:1023812930288.

[65] João Henriques, Lise Arleth, Kresten Lindorff-Larsen, and Marie Skepö. On the Calculation of SAXS Profiles of Folded and Intrinsically Disordered Proteins from Computer Simulations. *Journal of Molecular Biology*, 430(16):2521–2539, August 2018. ISSN 0022-2836. doi: 10.1016/j.jmb.2018.03.002.