Universal nature of collapsibility in the context of protein folding and evolution

D. Thirumalai,[1] Himadri S. Samanta,[1] Hiranmay Maity,[2] and Govardhan Reddy[2]

[1]Department of Chemistry, University of Texas at Austin, TX 78712

[2]Solid State and Structural Chemistry Unit,

Indian Institute of Science, Bangalore, Karnataka, India 560012

(Dated: October 29, 2018)

Abstract

Theory and simulations predicted sometime ago that the sizes of unfolded states of globular proteins should decrease continuously as the denaturant concentration is shifted from a high to a low value. However, small angle X-ray scattering (SAXS) data were used to assert the opposite, while interpretation of single molecule Forster resonance energy transfer experiments (FRET) supported the theoretical predictions. The disagreement between the two experiments is the SAXS-FRET controversy. By harnessing recent advances in SAXS and FRET experiments and setting these findings in the context of a general theory and simulations, we establish that compaction of unfolded states is universal. The theory also predicts that proteins rich in β-sheets are more collapsible than α-helical proteins. Because the extent of compaction is small, experiments have to be accurate and their interpretations should be as model free as possible. Theory also suggests that collapsibility itself could be a physical restriction on the evolution of foldable sequences, and provides a physical basis for the origin of multi-domain proteins.

1. PROTEIN COLLAPSIBILITY - WHAT IS THE PROBLEM?

The number of protein sequences with N amino acids that can be synthesized from twenty amino acids is 20^N, which is approximately 10^{130} for $N = 100$. On the other hand, the number of folds in natural globular proteins, which may be associated with low energy compact structures, is only on the order of a few thousands [1, 2]. Clearly, the sequence space is dense in contrast to the sparse structure space. The dramatic reduction that occurs from the dense sequence space to the countable number of folds may be rationalized by merely imposing the restriction that folded globular proteins be Minimum Energy Compact Structures (MECS) [3]. Precise calculations using lattice models [4, 5] for proteins show that the number of compact structures grows exponentially with N [3], as predicted by polymer theory [6]. Remarkably, the number of MECS likely grows only as $\ln N$, which has remained a surprising but under appreciated result [3]. The implication of this finding is that for a vast number of sequences the MECS must be topologically similar. In other words, the basins of attraction in the structure space are so rare that a vast number of sequences map on to precisely one structure with a specific topology. This plausibility, also established using lattice models [7], tidily explains the emergence of vastly limited number of structures from the astronomically large number of sequences. Thus, the propensity to form MECS is the distinguishing feature of biologically foldable sequences. Similar arguments could be made for RNA, which are made from four nucleotides. Indeed, it has been suggested that the requirement of compactness may be the key constraint for single stranded viral RNA evolution [8]. It is likely that compaction as a selection mechanism holds more generally for ribozymes as well.

The structures and dynamics of MECS as well the unfolded states are usually investigated by varying the concentration of denaturants, such as Urea or Guanidinium Chloride (GdmCl). A fundamental question that goes to the heart of protein collapsibility problem is: how do the shapes of the folded and unfolded states change as a function of the concentration of denaturants? In order to unpack the answer to this seemingly simple question, let us consider the folding of simple two-state proteins in which only the folded ($\boldsymbol{F}$) and unfolded ($\boldsymbol{U_D}$) states are appreciably populated. The balance between the population of these $\boldsymbol{F}$ and $\boldsymbol{U_D}$ states in experiments are altered by changing denaturant concentrations. The radii of gyration of the folded states, R_g^Fs, change imperceptibly [9] as the concentration of denaturants, $[C]$, is altered. However, as $[C]$ decreases below the mid-point $[C_m]$, the concentration at which the folded and unfolded protein populations are equal, whether or not the size of the unfolded states, $R_g^{U_D}$, decreases becoming more compact has, until recently, remained in dispute [10]. In a nutshell, does the $\boldsymbol{U_D}$ become compact forming the $\boldsymbol{U_C}$ state, with $R_g^{U_C} < R_g^{U_D}$ below $[C_m]$?

Twenty five years ago, the answer to this question posed above was given in the affirmative [11] using theory. Subsequently by taking into account consequences of the finite size of globular proteins it was shown that folding cooperativity increases universally as $\sim N^{2.2}$, where N is the number of amino acids [12]. In the process, we argued that water is only a moderately good solvent, and is more likely to be closer to the θ solvent [12] (Box 1 gives a background of polymer physics terminology commonly used to analyze experimental data). However, analyses of the data using two different experimental methods have arrived at contradictory conclusions. Based on a number of Small Angle X-Ray Scattering (SAXS) experiments, it had been asserted emphatically for nearly two decades [10, 13, 14] that the dimension of the $\boldsymbol{U_D}$ state does not change as $[C]$ decreases. It, therefore, follows that $R_g^{U_D} = R_g^{U_C}$ at all $[C]$. In sharp contrast, using single molecule Forster resonance energy transfer experiments (FRET), it was concluded that $R_g^{U_D} > R_g^{U_C}$ at low $[C]$ [15–17], in accord with our theoretical predictions [3]. In light of experimental and theoretical advances in the last two years, we survey the current status and come to the conclusion that denatured state ensemble (DSE) collapse of single domain globular proteins is universal. As a corollary, we also suggest that, as a rule, Intrinsically Disordered Proteins (IDPs) must expand as the denaturant concentration increases.

Box 1. Polymer physics language for describing states of proteins: Despite substantial differences between proteins and homopolymers, the language and concepts to describe the latter, principally developed by Flory [18], have been adopted to characterize $\boldsymbol{U_D}$ states and IDPs. Folded states of globular proteins are roughly spherical and are nearly maximally compact with high packing densities [19–21]. The radius of gyration (R_g^N) of folded proteins is well described by the Flory law with $R_g^N \approx a_N N^{\frac{1}{3}}$, with $a_N \approx 3.3$ Å [22]. At high denaturant concentrations, proteins swell adopting expanded conformations. In unfolded $\boldsymbol{U_D}$ states $R_g^{U_D} \approx r_0 N^{\nu}$ where $\nu \approx 0.6$

is the Flory exponent with $r_0 \approx 2.0$ Å [23]. Estimates of a_D vary greatly and is one of the difficulties in assessing solvent quality (Box 2). Thus, viewed from this perspective, we could surmise that proteins must undergo a coil-to-globule transition [24–26], a process that is reminiscent of the equilibrium collapse transition in homopolymers with $N \gg 1$ [27, 28]. The latter is driven by a balance between conformational entropy and intra-polymer interaction energy. By analogy, we surmise that the swollen state is realized in good solvents (interaction between proteins and solvents is favorable) whereas in the collapsed state intra protein interactions are preferred. It is tempting to identify high (low) denaturant concentrations with good (poor) solvent for proteins.

The simple physical picture given above is not wholly accurate because two additional states need to be considered in order to understand the collapsibility problem in polypeptide chains. First, upon increasing the denaturant concentration from zero, the side chains, which are densely packed in the $\boldsymbol{F}$ state, could become disordered while preserving the overall fold. Such a state is referred to as the dry globule ($\boldsymbol{DG}$). Experiments [29, 30], theory [31] and simulations [32, 33] have provided evidence for the $\boldsymbol{DG}$ state, which we postulated to be a universal intermediate [34] in the folding landscape (Figure 1). As the denaturant concentration decreases, the $\boldsymbol{U_D}$ state becomes compact, forming the $\boldsymbol{U_C}$ state. These states ($\boldsymbol{F}$, $\boldsymbol{U_D}$, $\boldsymbol{DG}$, and $\boldsymbol{U_C}$) can be distinguished using two order parameters. One is the density, $\rho = \frac{N}{R_g^3}$ and the other is χ, which measures how similar a given conformation is to the $\boldsymbol{F}$ state [35]. If the protein is folded then $\rho \sim \mathcal{O}(1)$ and $\chi \sim 0$ whereas in the $\boldsymbol{U_C}$ state the value of ρ are small and $\chi \sim \mathcal{O}(1)$. Similarly, the$\boldsymbol{DG}$ state is characterized by $\rho \sim \mathcal{O}(1)$ and $\chi \neq 0$, but not too large. Finally, the value of ρ in $\boldsymbol{U_C}$ is greater than in the $\boldsymbol{U_D}$ state whereas χ is smaller. Figure 1 illustrates the states of a globular protein as a function of ρ and χ.

The theory intended for describing the coil-globule transition in homopolymers as a function of the solvent quality is altered is strictly valid only when $N \gg 1$. However, single domain proteins are small (typically contain less than about 200 residues). As a result, the transitions between the states are rounded when studied using ρ or χ. As a result the values of the order parameters do not change precipitously. In particular, the difference between the radii of gyration between the $\boldsymbol{U_D}$ and $\boldsymbol{U_C}$ states are small [36], requiring accurate measurements over a wide range of $[C]$. The absence of such experiments, until recently, had created a robust and useful controversy. An unambiguous answer to the collapsibility problem, which is important in protein folding and has ramifications for IDPs as well, had therefore remained elusive from an experimental perspective although it has been under theoretical control for twenty five years.

Simulations and SAXS experiments show that $\boldsymbol{U_D}$ of Ubiquitin undergoes modest compaction: As a case study that illustrates succinctly the many nuances in the collapsibility of the $\boldsymbol{U_D}$ states, we consider the protein Ubiquitin (UB), whose folding has been investigated by changing denaturants [37, 38], mechanical forces [39], and more recently pressure [40]. UB is a 76-residue protein with complicated topology. The crystal structure[41] (PDB ID: 1UBQ) shows that in the folded state it has 5 β-strands and 2 α-helices (Figure 2A). Experiments [42] and simulations [36, 43] show that the GdmCl midpoint for UB is $\sim$ 3.8 M at neutral pH (Figure 2B). Simulations [36] in which the effects of GdmCl is modeled using the Molecular Transfer Model [9] show that R_g of the protein increases from $\sim$ 13 Å to $\sim$ 25.6 Å as the UB unfolds (Figure 2B). The predicted unfolding transition, as monitored by GdmCl induced swelling, is in excellent agreement with the SAXS experiments [42]. The radius of gyration, $R_g^{U_D}$, of the $\boldsymbol{U_D}$ state of UB, decreases continuously from $\sim$ 25.6 Å to $\sim$ 23 Å as $[GdmCl]$ is diluted from 6 to 0.75 M (Figure 2B). The predictions using simulations and the most recent SAXS measurements [44], which show that UB does become compact by 2.2 Å as $[GdmCl]$ is diluted from 6 to 0.7 M are in good agreement. Thus, the unfolded states of UB do become compact, albeit only modestly so, as GdmCl concentration is decreased from a high to a low value.

Is UB in high Urea concentration a random coil?: Simulations [36] and FRET experiments [16, 38] showed that compaction of the $\boldsymbol{U_D}$ state of UB, upon denaturant dilution, is driven by the changes in the solvent quality. To infer the nature of the solvent quality, we calculated the probability distribution of the normalized end-to-end distance (R_{ee}) of UB, defined as $x \equiv R_{ee}/\langle R_{ee}^2\rangle^{1/2}$ ($\langle R_{ee}\rangle = \langle R_{ee}^2\rangle^{1/2}$ is the mean R_{ee}). In good solvents, $P(x)$, should be described by the universal shape corresponding to the self-avoiding walk (SAW) provided $N \gg 1$. The universal shape of $P(x)$ for the SAW is given by [45–48]

$$P(x) = 4\pi A x^{2+g} \exp[-\alpha x^{\delta}], \tag{1}$$

where ν is the Flory scaling exponent, $g = (\gamma - 1)/\nu$, $\delta = 1/(1-\nu)$, and $\gamma \approx 1.1619$ for 3 dimensional SAW [49]. The constants, A and α, are evaluated using the conditions, $\int\limits_0^\infty P(x)dx = \int\limits_0^\infty x^2 P(x)dx = 1$.

In acidic pH and high denaturant conditions, $[urea] = 8$ M, $P(x)$ for UB is well fit using Eq. 1. However, the value of ν extracted from the fit is 0.75, which is greater than 0.60 expected for long SAWs (Figure 3A). The discrepancy is attributable to the finite size of UB ($N = 76$). In $[urea] = 2$ M, $P(x)$ shows a bimodal distribution as the C and N termini β-strands (β_1 and β_5) (Figure 3A) make contacts with a non-negligible probability leading to a peak in $P(x)$ at $x \approx 0.5$ (Figure 3B). This shows that in low denaturant concentrations, the topology of the folded state plays a critical role in determining the extent of compaction of the collapsed states [50].

Although the fit of the calculated $P(x)$ has the form given by Eq. 1, the extracted ν value from experiments and simulations should be viewed as an effective exponent, ν_{eff}, because there are finite size corrections to the Flory exponent ν (Box 2). Therefore, based solely on the value of ν_{eff}, we should not conclude that 8 M urea is a good solvent for unfolded UB.

Solvent quality: The solvent quality for a protein may also be inferred by measuring intramolecular distances between labelled residues, and fitting the results to predictions based on polymer theory. Recently, FRET experiments were performed on UB by positioning the donor and acceptor dyes at different positions to extract the intra chain root mean square distance, $\sqrt{\langle r^2 \rangle}$ between the dyes[38]. The solvent quality at a particular denaturant concentration is inferred from the effective scaling exponent using the relation, $\sqrt{\langle r^2 \rangle} = r_o N_{aa}^{\nu_{eff}}$, where N_{aa} is the number of amino acids separating the donor and acceptor dyes (Figure 3C), and r_o is an unknown parameter. The exponent, ν_{eff}, computed using this procedure from both experiments[38] and simulations[36] for ubiquitin shows that ν_{eff} varies continuously from ~ 0.6 to ~ 0.5 as $[urea]$ is varied from 8 M to 1 M. From this perspective, the solvent quality changes from good solvent like conditions to poor as $[urea]$ is diluted (Figure 3C). However, the expectation that for $\nu = 0.6$, $P(x)$ must obey the universal shape in eq. 1 is not satisfied because the fit of the simulation data yield $\nu_{eff} = 0.75$. The values of the effective exponent have also been estimated using SAXS data [44] by calculating

$R(|i-j| \approx |i-j|^{\nu_{off}}$ ($R(|i-j|$ is the mean distance between residues i and j) from the DSE generated by a new computational method to analyze SAXS data. As outlined in Box 2, this way of estimating ν_{off} is also not without difficulties. Thus, different ways of analyzing the data may not be consistent with each other, casting doubts on the assessment of the solvent quality based on estimates of ν_{off} from FRET or SAXS experiments.

Box 2. SAXS and FRET experiments. The two commonly used methods to measure the radii of gyration (R_gs) of polypeptide chains are SAXS and single molecule FRET experiments. The R_g can be directly calculated using the Guinier approximation to the scattering intensity, $I(q)$ (Figure 4), which for small q is given by $I(q) \approx I(0)\exp(-\frac{q^2R_g^2}{3})$ where $I(0) \propto$ to the molecular weight. Thus, the slope of the plot of $\ln I(q)$ versus q^2 yields $\frac{R_g^2}{3}$. A practical difficulty in determining R_g is that $I(q)$ has to be measured accurately for values of $qR_g \ll 1$. This is particularly important for polypeptide chains for which the changes in R_g are not large. Apparently, the problem is exacerbated at low denaturant concentrations [51], which might contribute to large errors in measuring $R_g^{U_C}$.

SAXS also provides information about conformations of the DSE through the distance distribution function, $P(r)$ given by,

$$I(q) = 4\pi \int_0^{D_{max}} dr\ p(r)\ \frac{\sin(qr)}{qr}, \qquad (2)$$

where because of the finite size of the polypeptide chain the upper limit is D_{max}, which is related to q_{min}, the smallest wave vector accessible in SAXS experiments. The average value of the square of the radius of gyration is the second moment of $P(r)$. There are uncertainties in the estimate of D_{max}, which has to be chosen with care in order to ensure consistency between Guinier approximation and R_g calculated from $P(r)$. These problems have to be taken into account when SAXS data for the DSE are analyzed.

In smFRET experiments, donor and acceptor dyes are attached to two positions (Figure 4), typically but not always, to the ends of the polypeptide chain. If the dyes are at the ends, then the mean FRET efficiency $\langle E \rangle$ is given by,

$$\langle E \rangle = \int_0^\infty \frac{P(R_{ee})}{1 + (\frac{R_{ee}}{R_0})^6} dR_{ee}, \tag{3}$$

where $P(R_{ee})$ is the normalized distribution of the end-to-end distance, R_{ee}, and R_0 is the dye-dependent Forster radius at which $\langle E \rangle = 0.5$. There are limitations in inferring R_g from the measured values of $\langle E \rangle$. (i) Calculating $P(R_{ee})$ using the above equation is a non-trivial inverse problem. A commonly used assumption is that $P(R_{ee})$ is a Gaussian,

$$P(R_{ee}) = 4\pi R_{ee}^2 \left(\frac{3}{2\pi \langle R_{ee}^2 \rangle} \right)^{3/2} \exp\left(-\frac{3R_{ee}^2}{2\langle R_{ee}^2 \rangle} \right). \tag{4}$$

Using eq. 3 and 4, the average radius of gyration of the protein, $\langle R_g \rangle$ is computed using the relation [52], $\langle R_g \rangle = \sqrt{\langle R_{ee}^2 \rangle / 6}$, which holds for a Gaussian polymer chain. (ii) The assumption that the unfolded states of polypeptide chains behave as ideal polymers is not accurate, which matters in resolving the apparent SAXS-FRET controversy [53] because in the transition from $\boldsymbol{U_D} \to \boldsymbol{U_C}$ the changes in the radius of gyration are small. A consequence is that $\langle R_g \rangle$ of proteins in the denatured ensemble inferred from FRET experiments do not agree with the values obtained from SAXS experiments, and the disagreement is significant for proteins like ubiquitin and protein L [10, 17, 24, 42]. (iii) It is important to point out that for the standard polymer models (Gaussian, Self-Avoiding walks, and Worm-like Chain Model), for which analytic expressions for $P(R_{ee})$ are available [53], it can be shown that the values of $\langle R_{ee} \rangle$ exceeds $\langle R_g \rangle$. This implies that $[\delta R_g]_{FRET} > [\delta R_g]_{SAXS}$ where $\delta R_g = R_g^{U_D} - R_g^{U_C}$. Consequently, FRET experiments exaggerate the extent of compaction of polypeptide chains [9, 36, 43] as the denaturant concentration is decreased. (iv) Finally, the attached dyes could have an effect on the conformations of the polypeptide chains [44], although practitioners of FRET experiments insist that this is not the case [54, 55].

The use of the expression in Eq. 1, with ν as an adjustable parameter to solve Eq. 3 is also unsatisfactory from a theoretical perspective. The lack of theory, connecting the distance distribution function $P(r)$ (eq. 2) to $P(R_{ee})$ for any polymer model other than the Gaussian chain makes it difficult to compare data from the SAXS and FRET techniques in a straightforward manner.

The quality of the solvent (discussed in Box 1) is assessed using ν_{eff} extracted from experimental data. In FRET experiments, ν_{eff} is calculated using $[C]$-dependent $R_g = r_o N^{\nu_{eff}}$ where the prefactor r_0 is assumed to be independent of $[C]$. In the most recent SAXS experiments [56] the MFF method is used to generate DSE from which the value of ν_{eff} is extracted using $\langle R(|i-j|)\rangle \approx |i-j|^{\nu_{eff}}$ ($\langle R(|i-j|)\rangle$ is the mean distance between residues i and j). The need to know r_o in FRET data analysis is not satisfactory as is the reliance on the accuracy of the DSE conformations generated by the MFF method [44]. If Eq. 1 holds then $\langle R(|i-j|)\rangle \approx |i-j|^{\nu_{eff}}$ calculated for a particular value of i and j for one protein ought to be identical to the value for another protein as long as $|i-j|$ is the same. In addition, $P(R(|i-j|))$ should also have the same universal form given by eq. 1 at least when $|i-j|$ is large.

In addition, there are finite size corrections [57] to the exponent in the relation, $\langle R_{ee}^2\rangle \sim N^{2\nu}$, given by,

$$\langle R_{ee}^2\rangle = AN^{2\nu}\left(1 + \frac{B}{N^{\Delta}} + \frac{C}{N} + \cdots\right), \qquad (5)$$

where ν and the correction to scaling Δ are universal, while A, B and C are system-specific constants. In order to accurately infer the solvent quality using ν, the corrections to the scaling relation should also be extracted carefully along with ν, to account for finite N [58–60]. Some of these shortcomings together with the broad range in the denaturant concentration range over which the solvent quality changes, render such analyses to be of limited value.

Extent of compaction is small: The relative compaction in the protein dimensions upon denaturant dilution is small unlike in long synthetic polymers that undergo coil-globule transition, which is akin to a genuine phase transition [61]. The finite-size of proteins is one contributing factor [11, 12]. The relative compaction, Δ, defined as

$$\Delta = \frac{R_g^{U_D}([C_h]) - R_g^{U_D}([C_l])}{R_g^{U_D}([C_h])} \qquad (6)$$

is typically less than 0.2 for a majority of proteins for which reliable simulation and experimental data are available (Table I)[36]. Table I shows that the length variation of the proteins used in experiments to study collapse in the denatured ensemble is small. As a result the change in $R_g^{U_D}([C])$ as a function of $[C]$ for most of the proteins is not large but is measurable. From Table I it is also clear that there is only a weak correlation between the size of the protein alone and relative compaction. The maximum value of relative compaction is observed for the cold shock protein, which predominantly has a β-sheet secondary structure. Thus, besides N the topology of the folded protein should also dictate the extent of compaction in the protein unfolded ensemble[50].

Status of the SAXS-FRET controversy: The details in Box 2 give a glimpse of the difficulties in probing collapsibility of proteins using experiments, which in part explains the SAXS-FRET controversy. Until last year, the persistent claim based on analyses of SAXS data was that the dimensions of the $\boldsymbol{U_D}$ state remains unchanged at all denaturant concentrations including $[C] = 0$ [10, 14]. In an important development, a new analysis method, referred to as Molecular Form Factor (MFF), was used to generate the ensemble of conformations of the unfolded states that are consistent with the measured SAXS profiles. Using MFF and new data for the denatured state ensembles, it was concluded that [44] the radii of gyration of two IDPs decreases between (20-28)% as the polypeptide chains are transferred from 6 M aqueous GdmCl solution to water. For Ubiquitin, $R_g^{U_D}$ decreases by about 2.2 Å upon a similar change in the GdmCl concentration. In addition, the effective Flory exponent decreases from 0.6 to about 0.5. The analyses of smFRET data for a number of proteins and IDPs consistently indicate that the $\boldsymbol{U_D}$ states become compact. It is gratifying that both camps are in qualitative agreement that the radius of gyration of the $\boldsymbol{U_C}$ states are less than in the $\boldsymbol{U_D}$ states. A perusal of the two commentaries [54, 55] and the response [56] to the article [44] shows that the debate now focusses on what is the quality of aqueous denaturant solution for denatured states of globular proteins and IDPs. We contend based on general theoretical grounds that it is difficult to answer this question using SAXS or FRET. Some of the reasons are outlined in Box 2.

For the subtle issue of collapsibility of the $\boldsymbol{U_D}$ state resolution usually requires theory and accurate simulations that do not require inputs from SAXS or FRET experiments. We believe that the theoretical and simulation results, summarized in Figures 2-5 go a long away in solving the SAXS-FRET controversy. Based on the advances in recent experimental [44, 51] and simulations [36] and theory [50], we summarize the current status of polypeptide chain collapse.

- The propensity of unfolded states of globular proteins to collapse is universal [50]. This conclusions is in harmony with analyses of FRET data. Similarly, simulations [36] and SAXS experiments [56] agree regarding the extent of collapse of the $\boldsymbol{U_D}$ state of UB. For example, on decreasing $[GdmCl]$ from 4.7 M to 0.9 M, SAXS experiments suggest that $R_g^{U_D}$ of the $\boldsymbol{U_D}$ state of UB decreases by about 7% whereas our simulations show that, at neutral pH, the decrease is ≈13% as $[GdmCl]$ decreases from 7 M to 0.25 M (see Table I). The predicted changes are in qualitative agreement with SAXS experiments. We pass the baton to experimentalists to measure changes in $R_g^{U_D}$ at acidic pH, which we predict is larger (Table I).

- Extraction of the effective Flory exponent from SAXS or FRET data is not straightforward, thus making claims about solvent quality dubious. Strictly speaking, the estimates of the Flory exponent is only meaningful for $N \gg 1$, which is not satisfied in the studies of single domain globular proteins or even the larger IDPs [56]. Thus, a careful finite size corrections (Eq. 5) must be considered in analyzing data. Nevertheless, theory [12] predicts that for most unfolded states of globular proteins water behaves as a θ-solvent, which means that intra protein and water-protein interactions nearly (but not perfectly) cancel with each other. The difficulties in extracting ν_{eff} not withstanding (discussed in Box 2), the cross over in the values of ν_{eff} from good ($\nu_{eff} \approx 0.6$) to θ-solvent ($\nu_{eff} = 0.5$) condition occurs over a broad range of $[C]$.

- From Table I, containing the extent of compaction for the denatured states of several proteins, leads to the unexpected conclusion that the length of the protein is not the determining factor in the extent of compaction. It is statistically more closely related to the topology of the folded state, as shown in Figure. 5.

Box 3. Theory for Collapsibility: The theoretical basis for assessing collapsibility of a given sequence starts with the Hamiltonian:

$$\mathcal{H} = \frac{3k_BT}{2a_0^2}\int_0^N \left(\frac{\partial \vec{r}}{\partial s}\right)^2 ds + k_BT\mathcal{V}(\vec{r}(s)), \tag{7}$$

where $\vec{r}(s)$ is the position of monomer s, and a_0 is the monomer size. The first term accounts for the polymeric nature of polypeptide chains. The interactions between the residues in the above equation are contained in $\mathcal{V}(\vec{r}(s))$ as follows:

$$\mathcal{V}(\vec{r}(s)) = \frac{v}{(2\pi a_0^2)^{3/2}} \sum_{s=0}^{N} \sum_{s'=0}^{N} e^{-\frac{(\vec{r}(s)-\vec{r}(s'))^2}{2a_0^2}} - \frac{\kappa}{(2\pi\sigma^2)^{3/2}} \sum_{\{s_i,s_j\}} e^{-\frac{(\vec{r}(s_i)-\vec{r}(s_j))^2}{2\sigma^2}}. \qquad (8)$$

The excluded volume interactions between the residues are represented by the $v > 0$ term. The attraction (term $\propto \kappa$) exists between specific residues, where the sum is over the set of specific interactions between residue pair $\{s_i, s_j\}$. We use the proteins contact maps, computed using the protein data bank (PDB) structures in order to assign the specific interactions [50]. A contact is assigned to any two residues s_i and s_j, if the distance between their C_α atoms is less than $R_c = 8$ Å and $|s_i - s_j| > 2$. For the excluded volume repulsion, the range is the size of the monomer $a_0 = 3.8$ Å, and for the specific attraction, the range is equal to the average distance between C_α atoms involved in contact formation, which averaged across a selection of proteins from PBD, is $\sigma = 6.3$ Å.

Changing the value of κ, and hence the strength of attraction, results in the transition between the extended and compact states. Decreasing κ is analogous to alteration of the concentration of the denaturants ($[C]$). At high $[C]$ ($\kappa \approx 0$, good solvent) the excluded volume repulsion dominates, while at low denaturant (high κ, poor solvent) the attractive interactions are important. The point where attraction balances repulsion is the θ-point, and the strength of attraction is κ_θ at the θ-point. At the θ-point, which is well defined if $N \gg 1$, the polypeptide chain behaves like an ideal polymer.

We define collapsibility using a measure of how easy it is to reach the θ-point. In others words, the smaller the value of κ_θ is easier it is for the polypeptide chain to undergo compaction as $[C]$ is decreased. Determination of κ_θ is complicated [50] but the final expression may be written in a manner that can be evaluated numerically for any globular protein. The theory makes two major predictions. (1) Polypeptide chain compaction is universal. However, the values of κ_θ depend on the number of residues, N, in the polypeptide chains. We showed that [50] κ_θ scales as N^β with $\beta > 0$. This implies that larger proteins tend to be less collapsible. (2) There is a caveat in the conclusion stated above. The κ_θ values also depend on the folded structures of the globular proteins, with β-sheet proteins being more collapsible than α-helical proteins.

β-sheet proteins are more collapsible than α-helical proteins: The theory outlined in Box 3 allows us to calculate κ_θ for globular proteins. The values of κ_θ (small values imply ease of collapse), plotted in Figure 5 for 2306 proteins, shows that for predominantly α-helix proteins ($> 90\%$), κ_θ increases rapidly with N with most of the proteins lying closer to the minimum collapsibility line. In contrast, the κ_θ values for proteins with high content of β-sheet ($> 70\%$) are closer to the minimum collapsible line. The values of κ_θ for the two sets are very distinct with minimal overlap. These results show that the extent of collapse of proteins that are mostly α-helical is much less than those with predominantly β-sheet structures. As the denaturation concentration is lowered below the midpoint, the $\boldsymbol{U_D}$ states reach one of the MECS, which are stabilized predominantly by native-like interactions. Consequently, the extent of compaction at low $[C]$ is determined by the topology of the native state, which partly explains the results in Figure 5. Thus, both N and even more importantly the topology of the $\boldsymbol{F}$ state determine κ_θ.

The reason for ease of collapsibility of proteins that are rich in β sheets is that the folded states of these proteins are stabilized by larger number of non-local contacts compared α-helical proteins. Indeed, there is a clear separation in the distribution, $P(\frac{N_{nc}^{NL}}{N_{nc}})$ where (N_{nc}^{NL} is the number of non-local contacts and N_{nc} is the total number contacts in a folded protein) between these two classes of proteins. By surveying 2306 proteins, we find that a remarkable separation in $P(\frac{N_{nc}^{NL}}{N_{nc}})$ between α-rich and β-rich proteins, explaining the ease of compaction in the latter compared to the former.

Compaction as a sequence selection mechanism? There are two reasons that support the assertion that natural sequences of foldable proteins have evolved to be collapsible. First, precise studies using lattice models show that the constraint of compaction and low energy reduces the number of viable protein structures [3, 7]. Second, the results of our recent theory show that the protein collapse is encoded in the structure [50] to which globular proteins are biased at low denaturant concentrations. Therefore, we conclude that simple biophysical constraint of compaction serves as a plausible selection mechanism in the evolution of foldable sequences. Interestingly, a similar conclusion has been reached for the evolution of viral single stranded RNA molecules [8]. In this context, it has been shown that viral ssRNAs are more compact than random RNA sequences with the same length and similar chemical composition. Thus, based on the principle of parsimony we would suggest that compaction alone might explain evolvability of protein and RNA sequences.

If compaction is a selection mechanism in the evolution of protein sequences then it follows that if the κ_θ values are large, as is the case when N increases ($\kappa_\theta \approx N^\beta$ with β greater than unity), then such proteins might split into multi-domains even at the expense of creating an interface. Each of the individual domains would have κ_θ in a reasonable range, on the order of few k_BT (Fig. 5).

What about IDPs?: It is estimated that between (30 - 40)% of eukaryotic proteome are either are intrinsically disordered or contain intrinsically disordered regions [62, 63]. Their roles in phase separation resulting in membraneless organelles containing droplets of IDPs [64, 65] has raised the need to understand their statistical properties in isolation. Because they do not form ordered structures they are ideal model systems for obtaining insights into $\boldsymbol{U_D}$ states of globular proteins. Typically, IDPs are low-complexity sequences containing more than usual fraction of charged and polar residues. Statistically they could exhibit properties that are reminiscent of synthetic polyampholytes (PAs) [66] or polyelectrolytes (PEs) [67]. Indeed, FRET experiments [68] on a highly charge IDP have been interpreted using PA theory. Because of the charged nature of certain IDPs salt effects, which are almost always present in the buffer, makes it difficult to analyze SAXS and FRET data. Indeed, the phase diagrams of IDPs in terms of the denaturants and salt concentration are only recently being revealed using theoretical calculations [69, 70]. At a fixed salt concentration, SAXS experiments have shown that the radius of gyration of the 334-residue PNt (an IDP) decreases by about 17% as GdmCl concentration is lowered from about 4M to a low value [44]. Similarly, FRET experiments on both the C and N termini highly charged prothymosin α apparently swell as GdmCl concentration is increased[68]. These findings are in line with the effect of denaturants on the unfolded states of globular proteins. There are exceptions to the general of denaturant-induced expansion of IDPs (see the last two entries in Table I). A complete picture for IDPs will emerge only by simultaneously changing both the denaturant and salt concentrations.

Concluding Remarks and Remaining problems:

For nearly two decades there was no consensus among experimentalists [10, 15], despite early theoretical predictions [11], on whether the sizes of the unfolded states decrease as the concentration of denaturants decreases. Here, we have made a compelling case using the recent advances in theory, simulations, and new SAXS and FRET data that the dimensions of the unfolded states of globular proteins must decrease as the denaturant concentration is lowered, thus putting to bed the long standing SAXS-FRET controversy. The apparent controversy did bring to sharper focus the issue of compaction of the $\boldsymbol{U_D}$ states, which is fundamentally important in understanding the assembly mechanism of proteins. Thanks to advances, just in the last three years [44, 51], we can assert that the compaction of the unfolded states of proteins is universal [50]. However, the changes in the denaturant dependent dimensions in the $\boldsymbol{U_D} \to \boldsymbol{U_C}$ is small, thus requiring precise measurements. The debate now seems to have shifted to determining whether water is a good or a θ-solvent [54, 56]. For reasons given here, this question can only be answered by measuring the second viral coefficient of the $\boldsymbol{U_D}$ state as function of denaturant concentration.

Despite the overall consensus that collapse is encoded in the evolved protein sequences, there are a few issues that require quantitative answers. Some of these are: (i) There is an urgent need to develop reliable force-fields, which includes the effects of commonly used denaturants, for use in atomically detailed molecular dynamics simulations so that simulations that are independent of experiments can be performed. (ii) In the same vein, a completely model-independent way of analyzing the integral equation connecting FRET efficiency and the distance distribution is needed. (iii) The accuracy of the commonly used Guinier approximation to extract the radius of gyration has been criticized [51, 54] because of substantial errors at low values of the scattering vector (see the figure in Box 2), which is exacerbated at small denaturant concentrations. (iv) The perpetual question of whether dyes attached to polypeptide chains affect their conformations [56] may be partially resolved by doing SAXS and FRET experiments on many other proteins, as has been done recently [71].

There are also exciting questions that deserve scrutiny. First, are aqueous denaturant solutions good solvents for Intrinsically Disordered Proteins? Although there are tentative answers to this question, the analyses methods could be criticized for reasons alluded to in this article. Second, can one quantitatively explain the emergence of multi-domain proteins using collapsibility as an essential physical constraint? We have given a preliminary answer to this question [50] but it remains a speculation. Finally, do these ideas and concepts carry over to ribozymes many of which are also compact? In the RNA field the importance of being compact, driven by cations, has long been accepted especially for ribozymes [72]. However, quantitative description of basic issues such as the dependence of persistence length of RNA and DNA, which is related to flexibility and the propensity of these biological molecules to collapse, is still lacking.

Acknowledgements: We are grateful to Ben Schuler and Tobin Sosnick for useful discussions and for sharing their data in tabular form. This work was supported by a grant from the National Science Foundation (CHE 16-36424). Additional support from the Collie-Welch Chair through the Welch Foundation (F-0019) is greatly appreciated. GR acknowledges support from the Department of Science and Technology through the grant EMR/2016/001356.

[1] S Govindarajan, R Recabarren, and RK Goldstein. Estimating the total number of protein folds. *Proteins*, 35(4):408–414, 1999.

[2] YI Wolf, NV Grishin, and EV Koonin. Estimating the number of protein folds and families from complete genome data. *J. Mol. Biol.*, 299(4):897–905, 2000.

[3] C. J. Camacho and D. Thirumalai. Minimum energy compact structures of random sequences of heteropolymers. *Phys. Rev. Lett.*, 71:2505–2508, 1993.

[4] HS Chan and KA Dill. The effects of internal constraints on the configurations od chain molecules. *J. Chem. Phys.*, 92(5):3118–3135, 1990.

[5] KA Dill, S Bromberg, KZ Yue, KM Fiebig, DP Yee, PD Thomas, and HS Chan. Principles of protein-folding - A perspective from simple exact models. *Protein Sci.*, 4(4):561–602, 1995.

[6] H Orland, C Itzykson, and C De Dominicis. An Evaluation of the Number of Hamiltonian Paths. *J. de Physique Lett.*, 46(8):L353–L357, 1985.

[7] H Li, R Helling, C Tang, and N Wingreen. Emergence of preferred structures in a simple model of protein folding. *Science*, 273(5275):666–669, 1996.

[8] Luca Tubiana, Anze Losdorfer Bozic, Cristian Micheletti, and Rudolf Podgornik. Synonymous Mutations Reduce Genome Compactness in Icosahedral ssRNA Viruses. *Biophys. J.*, 108(1):194–202, 2015.

[9] E. P. O'Brien, G. Ziv, G. Haran, B. R. Brooks, and D. Thirumalai. Effects of denaturants and osmolytes on proteins are accurately predicted by the molecular transfer model. *Proc. Natl. Acad. Sci. USA*, 105:13403–13408, 2008.

[10] Tae Yeon Yoo, Steve P. Meisburger, James Hinshaw, Lois Pollack, Gilad Haran, Tobin R. Sosnick, and Kevin Plaxco. Small-angle X-ray scattering and single-molecule FRET spectroscopy produce highly divergent views of the low-denaturant unfolded state. *J. Mol. Biol.*, 418(3-4):226–236, 2012.

[11] C. J. Camacho and D. Thirumalai. Kinetics and thermodynamics of folding in model proteins. *Proc. Natl Acad Sci USA*, 90(13):6369–6372, 1993.

[12] MS Li, DK Klimov, and D Thirumalai. Finite size effects on thermal denaturation of globular proteins. *Phys. Rev. Lett.*, 93(26):268107, 2004.

[13] KW Plaxco, IS Millett, DJ Segel, S Doniach, and D Baker. Chain collapse can occur con-

comitantly with the rate-limiting step in protein folding. *Nat. Struct. Biol.*, 6(6):554–556, 1999.

[14] Herschel M. Watkins, Anna J. Simon, Tobin R. Sosnick, Everett A. Lipman, Rex P. Hjelm, and Kevin W. Plaxco. Random coil negative control reproduces the discrepancy between scattering and FRET measurements of denatured protein dimensions. *Proc. Natl. Acad. Sci.*, 112(21):6631–6636, 2015.

[15] Benjamin Schuler and William A. Eaton. Protein folding studied by single-molecule FRET. *Curr. Opin. Struct. Biol.*, 18(1):16–26, 2008.

[16] H Hofmann, A Soranno, A Borgia, K Gast, D Nettels, and B Schuler. Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single molecule spectroscopy. *Proc. Natl. Acad. Sci. USA*, 109:16155–16160, 2012.

[17] K. A. Merchant, R. B. Best, J. M. Louis, I. V. Gopich, and W. A. Eaton. Characterizing the unfolded states of proteins using single-molecule FRET spectroscopy and molecular simulations. *Proc. Natl. Acad. Sci. USA*, 104:1528–1533, 2007.

[18] Paul J Flory. *Statistical Mechanics of Chain Molecules.* Interscience, New York, 1980.

[19] FM Richards. Interpretation of protein structures - Total volume, group volume distributions and packing density. *J. Mol. Biol.*, 82(1):1–14, 1974.

[20] J Liang and KA Dill. Are proteins well-packed? *Biophys. J.*, 81(2):751–766, 2001.

[21] S Bromberg and KA Dill. Side-chain entropy and packing in proteins. *Protein Sci.*, 3(7):997–1009, 1994.

[22] R.I. Dima and D. Thirumalai. Asymmetry in the shapes of folded and denatured states of proteins. *J. Phys. Chem. B*, 108:6564–6570, 2004.

[23] J.E. Kohn, I.S. Millett, J. Jacob, B. Zagrovic, T.M. Dillon, N. Cingel, R.S. Dothager, S. Seifert, P. Thiyagarajan, T.R. Sosnick, M.Z. Hasan, V.S. Pande, I. Ruczinski, S. Doniach, and K.W. Plaxco. Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc. Natl Acad Sci USA*, 101(34):12491–12496, 2004.

[24] E. Sherman and G. Haran. Coil-globule transition in the denatured state of a small protein. *Proc. Natl. Acad. Sci. USA*, 103:11539–11543, 2006.

[25] Gilad Haran. How, when and why proteins collapse: the relation to folding. *Curr. Opin. Struct. Biol.*, 22(1):14–20, 2012.

[26] G. Ziv, D. Thirumalai, and G. Haran. Collapse transition in proteins. *Phys. Chem. Chem.*

Phys., 11:83–93, 2009.

[27] IM Lifshitz, AY Grosberg, and AR Khokhlov. Some problems of statistical physics od poymer-chains with volume interactions. *Rev. Mod. Phys.*, 50(3):683–713, 1978.

[28] A Yu Grosberg and AR Khokhlov. *Statistical Physics of Macromolecules.* American Institute of Physics: New York, 1994.

[29] T Kiefhaber, AM Labhardt, and RL Baldwin. Direct nmr evidence for an intermediate preceding the rate-limiting step in the unfolding of ribonuclease-a. *Nature*, 375(6531):513–515, 1995.

[30] Robert L. Baldwin, Carl Frieden, and George D. Rose. Dry molten globule intermediates and the mechanism of protein unfolding. *Proteins*, 78(13):2725–2737, 2010.

[31] AV Finkelstein and EI Shakhnovich. Theory of cooperative transitions in protein molecules .2. phase-diagram for a protein molecule in solution. *Biopolymers*, 28(10):1681–1694, 1989.

[32] RD Mountain and D Thirumalai. Molecular dynamics simulations of end-to-end contact formation in hydrocarbon chains in water and aqueous urea solution. *J. Am. Chem. Soc.*, 125(7):1950–1957, 2003.

[33] Lan Hua, Ruhong Zhou, D. Thirumalai, and B. J. Berne. Urea denaturation by stronger dispersion interactions with proteins than water implies a 2-stage unfolding. *Proc. Natl. Acad. Sci. U. S. A.*, 105(44):16928–16933, 2008.

[34] D Thirumalai, Zhenxing Liu, Edward P O'Brien, and Govardhan Reddy. Protein folding: From theory to practice. *Curr. Opin. Struct. Biol.*, 23(1):22–29, 2013.

[35] D Thirumalai and DK Klimov. Deciphering the timescales and mechanisms of protein folding using minimal off-lattice models. *Curr. Opin. Struct. Biol.*, 9(2):197–207, 1999.

[36] Govardhan Reddy and D. Thirumalai. Collapse precedes folding in denaturant-dependent assembly of ubiquitin. *J. Phys. Chem. B*, 121(5):995–1009, 2017.

[37] Benjamin Schuler, Andrea Soranno, Hagen Hofmann, and Daniel Nettels. Single-molecule FRET spectroscopy and the polymer physics of unfolded and intrinsically disordered proteins. *Annu. Rev. Biophys.*, 45:207–231, 2016.

[38] Mikayel Aznauryan, Leonildo Delgado, Andrea Soranno, Daniel Nettels, Jie-rong Huang, Alexander M. Labhardt, Stephan Grzesiek, and Benjamin Schuler. Comprehensive structural and dynamical view of an unfolded protein from the combination of single-molecule FRET, NMR, and SAXS. *Proc. Natl. Acad. Sci. U. S. A.*, 113(37):E5389–E5398, 2016.

[39] Sergi Garcia-Manyes, Lorna Dougan, Carmen L. Badilla, Jasna Brujic, and Julio M. Fernandez. Direct observation of an ensemble of stable collapsed states in the mechanical folding of ubiquitin. *Proc. Natl. Acad. Sci.*, 106(26):10534–10539, 2009.

[40] Cyril Charlier, T. Reid Alderson, Joseph M. Courtney, Jinfa Ying, Philip Anfinrud, and Adriaan Bax. Study of protein folding under native conditions by rapidly switching the hydrostatic pressure inside an NMR sample cell. *Proc. Natl. Acad. Sci.*, 115(18):E4169–E4178, 2018.

[41] S Vjaykumar, CE Bugg, and WJ Cook. Structure of ubiquitin refined at 1.8 A resolution. *J. Mol. Biol.*, 194(3):531–544, 1987.

[42] J Jacob, B Krantz, RS Dothager, P Thiyagarajan, and TR Sosnick. Early collapse is not an obligate step in protein folding. *J. Mol. Biol.*, 338(2):369–382, 2004.

[43] Govardhan Reddy and D Thirumalai. Dissecting ubiquitin folding using the self-organized polymer model. *J. Phys. Chem. B*, 119(34):11358–11370, 2015.

[44] 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, 2017.

[45] ME Fisher. Shape of a self-avoiding walk or polymer chain. *J. Chem. Phys.*, 44(2):616–622, 1966.

[46] DS McKenzie and MA Moore. Shape of a self-avoiding walk or polymer chain. *J. Phys. A-Math. Gen.*, 4(5):L82–L86, 1971.

[47] JD Cloizeau. Lagrangian theory for a self-avoiding random chain. *Phys. Rev. A*, 10(5):1665–1669, 1974.

[48] S Redner. Distribution-functions in the interior of polymer-chains. *J. Phys. A-Math. Gen.*, 13(11):3525–3541, 1980.

[49] D Macdonald, DL Hunter, K Kelly, and N Jan. Self-avoiding walks in 2 to 5 dimensions - exact enumerations and series study. *J. Phys. A-Math. Gen.*, 25(6):1429–1440, 1992.

[50] Himadri S. Samanta, Pavel I. Zhuravlev, Michael Hinczewski, Naoto Hori, Shaon Chakrabarti, and D. Thirumalai. Protein collapse is encoded in the folded state architecture. *Soft Matter*, 13(19):3622–3638, 2017.

[51] Alessandro Borgia, Wenwei Zheng, Karin Buholzer, Madeleine B. Borgia, Anja SchÃŒler,

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. *J. Am. Chem. Soc.*, 138(36):11714–11726, 2016.

[52] Michael Rubinstein and Ralph H Colby. *Polymer Physics*. OUP Oxford, 2003.

[53] E. P. O'Brien, G. Morrison, B. R. Brooks, and D. Thirumalai. How accurate are polymer models in the analysis of Forster resonance energy transfer experiments on proteins? *J. Chem. Phys.*, 130:124903, 2009.

[54] Robert B. Best, Wenwei Zheng, Alessandro Borgia, Karin Buholzer, Madeleine B. Borgia, Hagen Hofmann, Andrea Soranno, Daniel Nettels, Klaus Gast, Alexander Grishaev, and Benjamin Schuler. Comment on "Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water". *Science*, 361(6405), 2018.

[55] Gustavo Fuertes, Niccolo Banterle, Kiersten M. Ruff, Aritra Chowdhury, Rohit V. Pappu, Dmitri I. Svergun, and Edward A. Lemke. Comment on "Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water". *Science*, 361(6405), 2018.

[56] Joshua A. Riback, Micayla A. Bowman, Adam Zmyslowski, Catherine R. Knoverek, John Jumper, Emily B. Kaye, Karl F. Freed, Patricia L. Clark, and TobinR. Sosnick. Response to Comment on "Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water". *Science*, 361(6405), 2018.

[57] ZV Djordjevic, I Majid, HE Stanley, and RJ Dossantos. Correction-to-scaling exponents and amplitudes for the correlation length of liner-polymers in 2 dimensions. *J. Phys. A-Math. Gen.*, 16(14):L519–L523, 1983.

[58] S Havlin and D Benavraham. Corrections to scaling in self-avoiding walks. *Phys. Rev. A*, 27(5):2759–2762, 1983.

[59] JW Lyklema and K Kremer. Correction to scaling exponent for the 2-dimensional self-avoiding walk. *Phys. Rev. B*, 31(5):3182–3184, 1985.

[60] DC Rapaport. On 3-dimensional self-avoiding walks. *J. Phys. A-Math. Gen.*, 18(1):113–126, 1985.

[61] Pierre-Gilles De Gennes. *Scaling concepts in polymer physics*. Cornell university press, 1979.

[62] M. Madan Babu, Richard W. Kriwacki, and Rohit V. Pappu. Versatility from Protein Disorder. *Science*, 337(6101):1460–1461, 2012.

[63] Zachary A. Levine and Joan-Emma Shea. Simulations disordered proteins and systems with conformational heterogeneity. *Curr. Opin. Struct. Biol.*, 43:95–103, APR 2017.

[64] Salman F. Banani, Hyun O. Lee, Anthony A. Hyman, and Michael K. Rosen. Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.*, 18(5):285–298, 2017.

[65] Yi-Hsuan Lin, Julie D. Forman-Kay, and Hue Sun Chan. Sequence-Specific Polyampholyte Phase Separation in Membraneless Organelles. *Phys. Rev. Lett.*, 117(17), 2016.

[66] P. Higgs and J.-F. Joanny. Theory of polyampholyte solutions. *J. Chem. Phys.*, 94:1543, 1991.

[67] B.-Y. Ha and D. Thirumalai. Conformations of a polyelectrolyte chain. *Phys. Rev. A*, 46:R3012–R3015, 1992.

[68] Sonja Mueller-Spaeth, Andrea Soranno, Verena Hirschfeld, Hagen Hofmann, Stefan Rueegger, Luc Reymond, Daniel Nettels, and Benjamin Schuler. Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 107(33):14609–14614, 2010.

[69] Taylor Firman and Kingshuk Ghosh. Sequence charge decoration dictates coil-globule transition in intrinsically disordered proteins. *J. Chem. Phys.*, 148(12), 2018.

[70] Himadri S Samanta, Debayan Chakraborty, and D Thirumalai. Charge fluctuation effects on the shape of flexible polyampholytes with applications to intrinsically disordered proteins. *J. Chem. Phys.*, 149(16):163323, 2018.

[71] Gustavo Fuertes, Niccolò Banterle, Kiersten M Ruff, Aritra Chowdhury, Davide Mercadante, Christine Koehler, Michael Kachala, Gemma Estrada Girona, Sigrid Milles, Ankur Mishra, et al. Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements. *Proc. Natl. Acad. Sci. USA*, pages E6342–E6351, 2017.

[72] Sarvin Moghaddam, Gokhan Caliskan, Seema Chauhan, Changbong Hyeon, R. M. Briber, D. Thirumalai, and Sarah A. Woodson. Metal Ion Dependence of Cooperative Collapse Transitions in RNA. *J. Mol. Biol.* , 393(3):753–764, 2009.

[73] Hiranmay Maity and Govardhan Reddy. Folding of protein L with implications for collapse in the denatured state ensemble. *J. Am. Chem. Soc.*, 138(8):2609–2616, 2016.

[74] Hiranmay Maity and Govardhan Reddy. Thermodynamics and Kinetics of Single-Chain Monellin Folding with Structural Insights into Specific Collapse in the Denatured State Ensemble.

J. Mol. Biol., 430(4, SI):465–478, 2018.

[75] Zhenxing Liu, Govardhan Reddy, and D. Thirumalai. Folding PDZ2 domain using the molecular transfer model. *J. Phys. Chem. B*, 120(33):8090–8101, 2016.

[76] Zhenxing Liu, Govardhan Reddy, Edward P O'Brien, and D Thirumalai. Collapse kinetics and chevron plots from simulations of denaturant-dependent folding of globular proteins. *Proc. Natl. Acad. Sci. USA*, 108(19):7787–7792, 2011.

TABLE I: Relative changes in the radius of gyration of various proteins in the unfolded states between high and low denaturant concentrations

Protein	$R_g^{U_D}([C_h])$[c] ([GdmCl] = 4-7 M)	$R_g^{U_D}([C_l])$[d] ([GdmCl] = 0-1 M)	$\Delta = \frac{R_g^{U_D}([C_h])-R_g^{U_D}([C_l])}{R_g^{U_D}([C_h])}$
Protein L[73] (N_{res}=64)	26.3 Å ([C] = 7 M)	22.5 Å ([C] = 0.25 M)	14.4%
Monellin (N_{res}=96; Neutral pH)[74]	27.8 Å ([C] = 7 M)	25.5 Å ([C] = 0.25 M)	8.3%
PDZ2 Domain[75] (N_{res}=94)[a]	32.2 Å ([C] = 7 M)	29.8 Å ([C] = 0.25 M)	7.5%
Ubiquitin (N_{res}=76; Neutral pH)	25.9 Å ([C] = 7 M)	22.5 Å ([C] = 0.25 M)	13.1%
Ubiquitin (N_{res}=76; Low pH)	30.4 Å ([C] = 7 M)	23.3 Å ([C] = 0.25 M)	23.4%
SH3[76] (N_{res}=56)	23.7 Å ([C] = 7 M)	20.3 Å ([C] = 0.2 M)	14.3%
Cold Shock[9] (N_{res}=70)	26.4 Å ([C] = 7 M)	17.8 Å ([C] = 1 M)	32.6%
ACTR[51] (N_{res}=73)[b]	29.7 Å ([C] = 7 M)	24.7 Å ([C] = 0.3 M)	16.8%
R17d[51] (N_{res}=116)[b]	40.3 Å ([C] = 7 M)	33.2 Å ([C] = 0.6 M)	17.6%
N49[71] (N_{res}=38)[a,b,e]	16.9 Å ([C] = 6 M)	15.9 Å ([C] = 0 M)	5.9%
NUS[71] (N_{res}=82)[a,b,e]	31.3 Å ([C] = 6 M)	24.9 Å ([C] = 0 M)	20.4%
NUL[71] (N_{res}=114)[a,b,e]	35 Å ([C] = 6 M)	30 Å ([C] = 0 M)	14.3%
PNt[44] (N_{res}=334)[b]	62 Å ([C] = 4 M)	51.3 Å ([C] = 0.15 M)	17.3%
IBB[71] (N_{res}=99)[a,b,e]	31.2 Å ([C] = 6 M)	32 Å ([C] = 0 M)	≈ 0%
NLS[71] (N_{res}=46)[a,b,e,f]	23.3 Å ([C] = 6 M)	24 Å ([C] = 0 M)	≈ 0%

[a] Denaturant used is urea.

[b] Data is from experiments.

[c] $R_g^{U_D}([C_h])$ is the R_g in the $\boldsymbol{U_D}$ state at high denaturant concentrations.

[d] $R_g^{U_D}([C_l])$ is the R_g in the $\boldsymbol{U_C}$ state at low denaturant concentrations.

[e] Sequence details of the IDPs are in the supplementary information of Ref. [71].

[f] The IDP NLS has high mean charge and low mean hydrophobicity. A different buffer, compared to other IDPs, was used in the measurements [71].

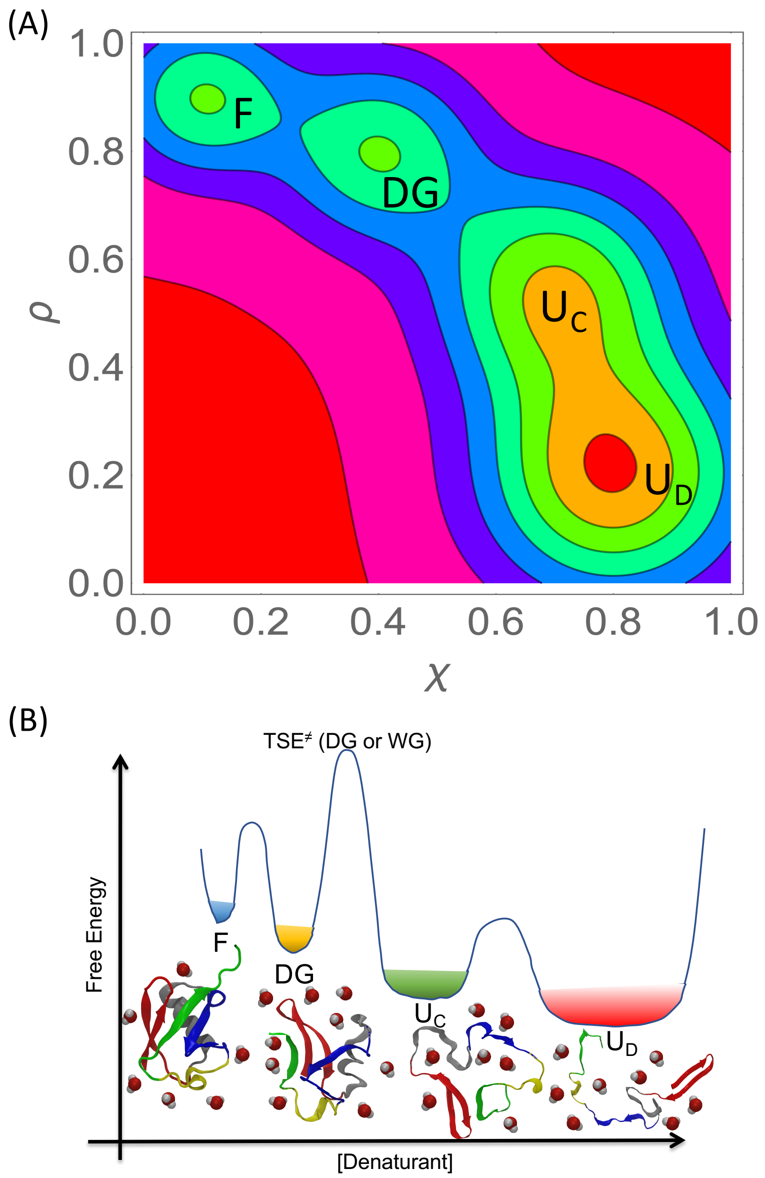

FIG. 1: **Box 1 Figure**. (A) Folding landscape of a globular protein projected onto the order parameters, packing density (ρ) and the structural overlap function (χ [11]). The folding landscape shows the dry globule state ($\boldsymbol{DG}$) and unfolded collapsed state ($\boldsymbol{U_C}$) in between the protein folded ($\boldsymbol{F}$) and unfolded states ($\boldsymbol{U_D}$). As the protein folds from the unfolded state, ρ increases and χ decreases. The order parameters are described in Box 1. (B) A schematic illustrating the effect of denaturants on the folding landscape. In the folded state, the protein the amino acids are tightly packed and protein hydrophobic core is devoid of water molecules. As the denaturant concentration increases, the folded core loosens forming the molten globule state, which can be wet containing water molecules or dry ($\boldsymbol{DG}$ state). As the denaturant concentration increases, the protein unfolds to compact unfolded states ($\boldsymbol{U_C}$). The barrier between $\boldsymbol{U_C}$ and $\boldsymbol{U_D}$ states is shown only for visual purposes.

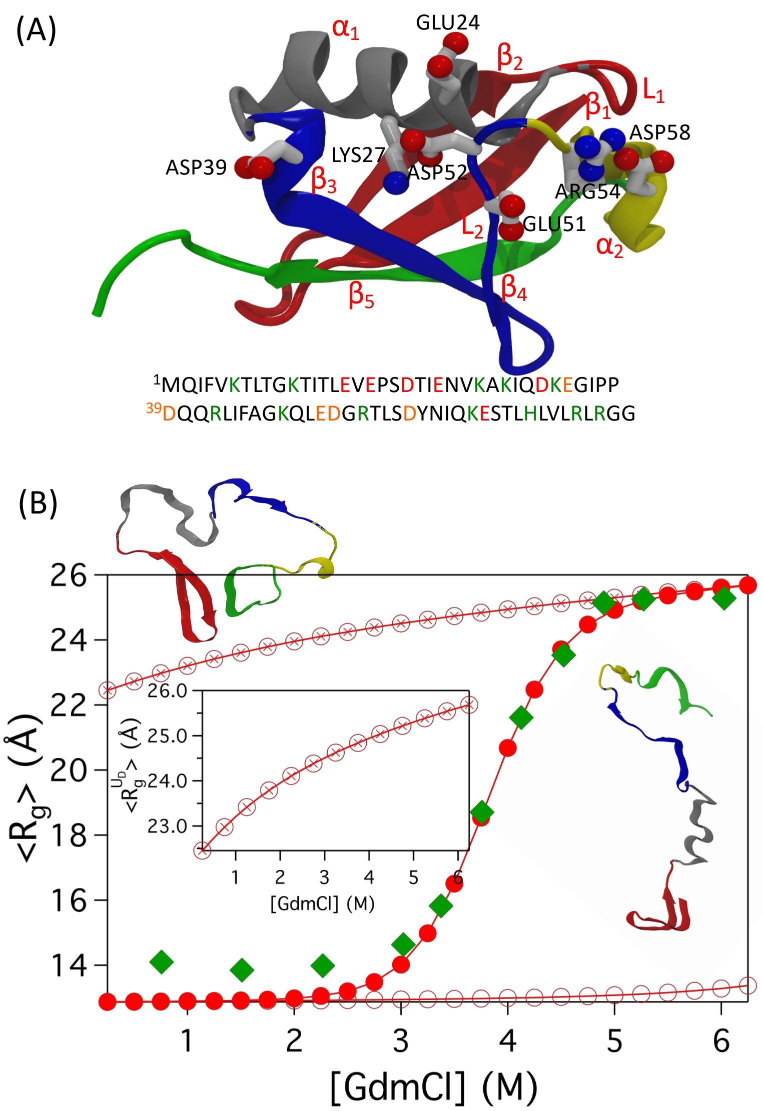

FIG. 2: (A) Crystal structure of UB in the folded state (PDB ID: 1UBQ)[41]. UB has 5 β-strands (β_1-β_5) shown in red, blue and green, 2 α-helices (α_1 - α_2) shown in grey and yellow, and it has 2 disordered loops with contacts labeled L_1 and L_2. UB sequence, in a single letter amino acid code, is shown below the structure. Amino acids in green and red letters are positively and negatively charged, respectively. Some of the charged residues are highlighted in the structure. (B) Radius of gyration, R_g, as a function of $[GdmCl]$ in neutral pH. Simulation data [36], shown in red solid circles, empty circles and circles with crosses correspond to R_g, R_g^F and $R_g^{U_D}$, respectively. The inset highlights the continuous decrease in $R_g^{U_D}$ as a function of $[GdmCl]$. Green squares give R_g from SAXS experiments [42]. Simulation snapshots of unfolded UB in the extended and compact form are shown in the figure.

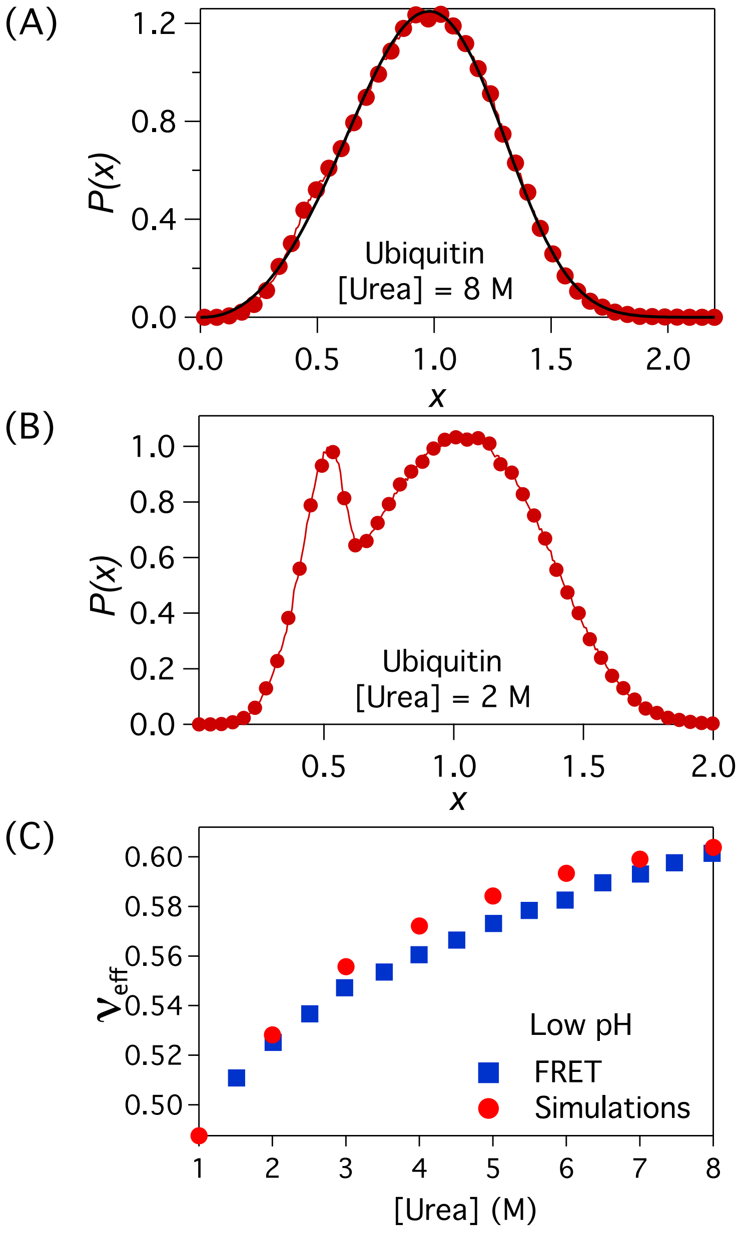

FIG. 3: (A) The normalized end-to-end distance, $x(= R_{ee}/\sqrt{\langle R^2_{ee}\rangle})$, distribution in low pH and $[urea] = 8.0$ M is shown in red circles. The black line is a fit to eq. 1 with $\nu_{eff} = 0.75$. (B) $P(x)$ in low pH and $[urea] = 2.0$ M. (C) The effective exponent, ν_{eff}, calculated using the relation $\sqrt{\langle r^2\rangle} \sim N_{aa}^{\nu_{eff}}$, as a function of $[urea]$. Here, $\sqrt{\langle r^2\rangle}$ is the intra molecular distance in UB between two residues separated by N_{aa} residues. The data in blue squares and red circles are from FRET experiments [38] and simulations [36], respectively.

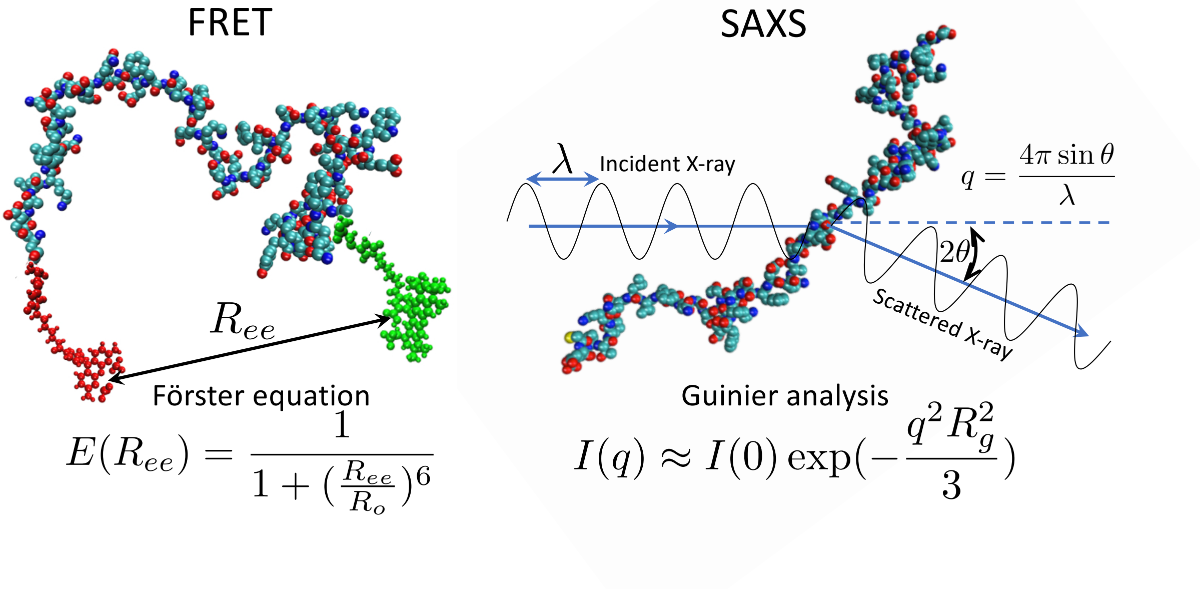
FRET
R_{ee}
Förster equation
$E(R_{ee}) = \frac{1}{1+(\frac{R_{ee}}{R_o})^6}$
SAXS
λ
Incident X-ray
$q = \frac{4\pi \sin\theta}{\lambda}$
2θ
Scattered X-ray
Guinier analysis
$I(q) \approx I(0)\exp(-\frac{q^2R_g^2}{3})$

FIG. 4: **Box 2 Figure**. Schematic of the FRET and SAXS experiments. In the FRET experiments, the donor and acceptor dyes, shown in red and green, respectively are attached to the protein termini. The efficiency of energy transfer, E, between the dyes, which depends on the denaturant concentration is measured. From the measured E, the distance between the dyes and the size of the protein is inferred using the Förster relation. In the SAXS experiments, R_g is extracted using the Gunier approximation, which depends on the ratio of the intensities of the incident, $I(0)$, and scattered, $I(q)$, of X-rays at small scattering angles, θ.

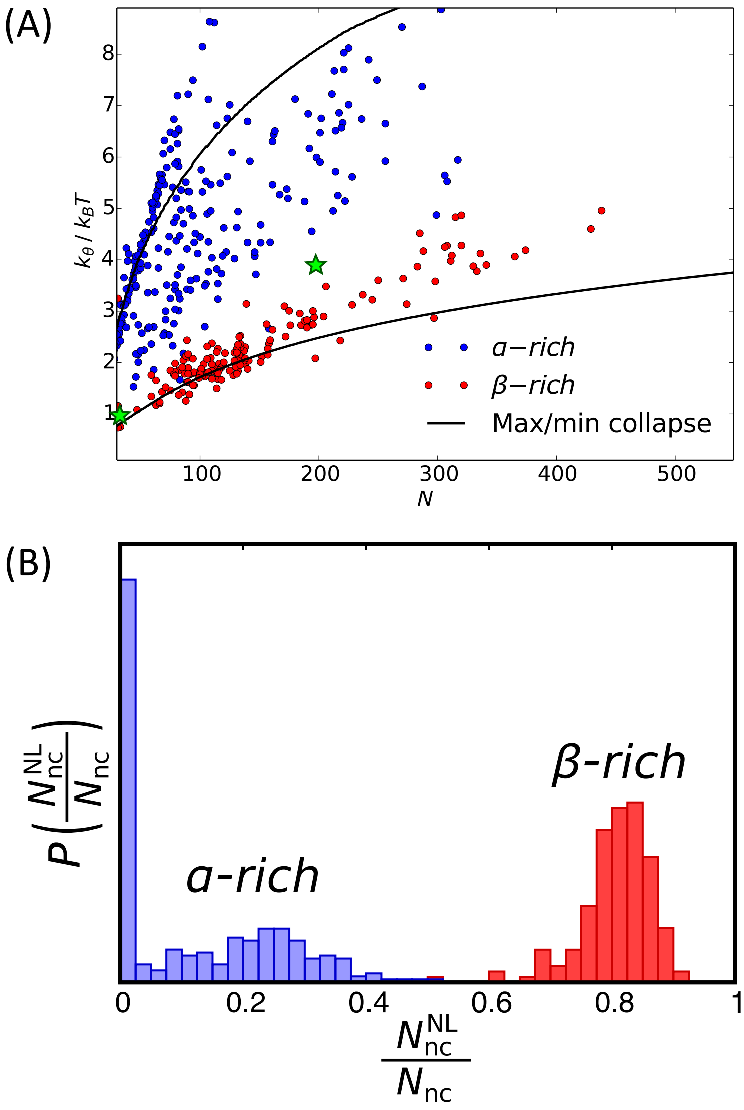

FIG. 5: (A) Plot of κ_θ, specifying the average interaction strength between two residues at the θ-point in units of k_BT, as a function of the number of residues, N. For identical values of N, proteins rich in β-sheet are more collapsible (have smaller κ_θ values) compared to those rich in α-helices. The green star on the left is for a RNA psueudoknot and the other is for the *Azoarcus* ribozyme. (B) $P(\frac{N_{nc}^{NL}}{N_{nc}})$ is the distribution of $\frac{N_{nc}^{NL}}{N_{nc}}$, where N_{nc} is the total number of native contacts in a protein, and N_{nc}^{NL} is the number of long range native contacts. Proteins rich in β-sheets have more long-range contacts compared to proteins rich in α-helices.