Phase Phenomena of Proteins in Living Matter

by
Andrei Gabriel Gasic

A dissertation submitted to the Department of Physics,
College of Natural Sciences and Mathematics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in Physics

Chair of Committee: Prof. Margaret S. Cheung, Advisor
Committee Member: Prof. Kevin E. Bassler
Committee Member: Prof. Gemunu H. Gunaratne
Committee Member: Prof. Anatoly B. Kolomeisky, Rice University
Committee Member: Prof. Vassiliy Lubchenko
Committee Member: Prof. Greg Morrison

University of Houston
December 2020

I would like to dedicate this dissertation to my loving parents and wife.

Acknowledgements

First, a special thank you goes to my parents, my sister, my uncle Greg, and my wife for their loving support.

Although this thesis is very personal, the research could not have be done in isolation. I would like to acknowledge the members of the Cheung group and the Center for Theoretical Biological Physics (CTBP) at Rice University. Thanks to Fabio Zagarra, my first publication was realized. Conversations with Dirar Homouz, Pengzhi Zhang, Victor Tsai, Michele Di Pierro, Ryan Cheng, and Nick Schafer were always enlightening. I am thankful for my mentees, Caleb Daugherty, Lucas Babel, and Atrayee Sarkar. I also want to acknowledge my PhD committee, whose critiques and advise has greatly influenced me.

Theory is usually more fun in collaboration with experimentalists. Most of the investigations in this thesis have an experimental validation part. I thank Mayank Boob, Max Prigozhin, Kapil Dave, and Martin Gruebele for performing the experiments.

Additionally, I would like to acknowledge my training fellowship on the Houston Area Molecular Biophysics Program (T32 GM008280) and funding by the National Science Foundation (MCB-1412532, PHY-1427654, OAC-1531814). Of course, I thank computing resources from the Center for Advanced Computing and Data Systems at UH.

Last but certainly not least, I would like to thank my advisor, Margaret S. Cheung. She gave me a chance to work in her lab when I was not even sure if I could do research. Through her guidance, I have gained the skills and wisdom to become a successful principal

investigator. It has been privilege to work with her, and I am forever grateful.

Abstract

Proteins must fold and function in the immensely complex environment of a cell—this is far from the ideal test-tube setting. This thesis answers an important question at the interface of physics and biology: how does the crowded cellular environment influence the dynamics of proteins and their folding phases?

This thesis is presented in two parts. First, we review protein folding and investigate the effects of hydrostatic pressure using coarse-grained molecular simulations. There are two common forms of pressure-dependent potentials of mean force (PMFs) for coarse-grained molecular simulations of protein folding and unfolding under hydrostatic pressure. We investigated the two different pressure-dependencies on the desolvation potential in a structure-based protein model using coarse-grained molecular simulations. We showed that the protein's folding transition curve on the pressure–temperature phase diagram depends on the relationship between the potential well minima and pressure.

In the second part, we move toward understanding the effects of crowded environment of the cell. Certain proteins' properties are exhibited by systems near a critical point, where distinct phases merge. This concept goes beyond previous studies that propose proteins have a well-defined folded and unfolded phase boundary in the pressure-temperature plane. Here, by modeling the protein phosphoglycerate kinase (PGK) on the temperature (T), pressure (P), and crowding volume-fraction (ϕ) phase diagram, we demonstrate a critical transition where phases merge, and PGK exhibits large structural fluctuations. Above the critical

temperature (T_c), the difference between the intermediate and unfolded phases disappears. When ϕ increases, the T_c moves to a lower T. Crowding shifts PGK closer to a critical line in its parameter space, where conformational changes can occur without costly free-energy barriers.

To understand the "quniary" interaction between cells, we examine the interplay between folding and inter-domain interactions of engineered FiP35 WW domain repeat proteins with $n = 1$ through 5 repeats using a coarse-grained simulated annealing with the AWSEM Hamiltonian. We show misfolded structures become increasingly prevalent as one proceeds from monomer to pentamer.

Finally, we discuss the implications and connection to the organization and dynamics of the cytoplasm, unifying the single protein scale with the many-protein architectures at the subcellular scale.

Table of Contents

List of Figures

List of Tables

List of Symbols & Acronyms

Roman Symbols

C_p	heat capacity
$\boldsymbol{F}_i$	force vector on particle i
F	free energy
G	Gibbs free energy
$\mathcal{H}$	Hamiltonian of the system
k_B	Boltzmann constant $\approx 8.314 \times 10^{-3}\,\mathrm{kJ\,mol^{-1}\,K^{-1}}$
$\mathbb{M}$	renormalization operator
m	mass
$\mathcal{N}(0,1)$	Gaussian (or normal) distribution with zero mean and unit variance
N, n	number of particles
$\mathcal{O}(x)$	terms "on the order of" or "infinitesimally asymptotic to" x
P	pressure
$\boldsymbol{r}_i$,$\boldsymbol{x}_i$	position vector of particle i
R_{ee}	end-to-end distance
r_{ij}	distance between particles i and j
R_g	radius of gyration
S	entropy
T	temperature

t	time
$\boldsymbol{v}_i$	velocity vector of particle i
V	volume

Greek Symbols

α	thermal expansion coefficient
β	inverse temperature $1/(k_B T)$
Γ	the set of variable for a configuration, specifically $\{\{r_{ij}\}, \{\theta_i\}, \{\phi_i\}\}$
Δ	difference symbol: is often used as a prefix signifying a finite change
$\delta(x)$	Dirac delta function
δ_{ij}	Kronecker delta function
ϵ	unit energy of system
η	viscosity
$\Theta(x)$	Heaviside step function
θ_i	the angle between three consecutive residues
κ	compressibility
λ_m	mean tryptophan fluorescence wavelength
ξ	random kick
σ	unit length of system
σ_{ij}	direct contact distance between particles i and j
ϕ	volume fraction
ϕ_i	the dihedral angle defined over four sequential residues
χ	overlap parameter $\in [0, 1]$, where the crystal state has a value of 0
Ψ	Landau-Ginsberg free energy order parameter
ω	excluded-volume parameter

Superscripts

0	signifying original, initial, or crystal
γ	crowding-induced polymer collapse scaling exponent
ν	Flory scaling exponent

Subscripts

c	critical point value
i, j	index variables
tot	denoting 'total'

Other Symbols

$\approx$	"is approximately equal to"
$\sim$	"is asymptotic equivalent to"
$\equiv$	"is defined as"
$\propto$	"is proportional to"
$\langle \cdots \rangle$	ensemble average
$\{\cdots\}$	a set of elements
$\xrightarrow{f}$	a map f from one set to another
$\leftarrow$	assignment operator sets the value of a variable in an algorithm
$\in$	"is an element of"
$\lim_{x \to a^-}$	the limit as x increases in value approaching a ("from the left" or "from below")
∇	gradient operator $\hat{\mathbf{e}}_x \frac{\partial}{\partial x} + \hat{\mathbf{e}}_y \frac{\partial}{\partial y} + \hat{\mathbf{e}}_z \frac{\partial}{\partial z}$
$\int \mathcal{D}x$	path (or functional) integral denotes integration over all paths
δf^2	variance of f, which is $\langle f^2 \rangle - \langle f \rangle^2$

Acronyms

3D	three-dimensional

C	Crystal state
CC	Collapsed Crystal state
CG	coarse grain
FRET	Förster Resonance Energy Transfer
GCMC	grand canonical Monte Carlo
I	Intermediate state
LD	Langevin dynamics
MD	molecular dynamics
PGK	phosphoglycerate kinase
PMF	potential of mean force
REM	replica exchange method
Sph	Spherical state
SU	Swollen Unfolded state
U	Unfolded state

One can best appreciate, from a study of living things, how primitive physics still is.

—Albert Einstein

1

Introduction: From Protein Folding in the Test-tube to Folding in the Cell

The field of protein folding and dynamics is expanding towards cellular biological physics. An understanding of the physical principles of macromolecules inside the cell is needed to bridge the scale of single proteins to the scales of subcellular organizations, such as supramolecular assemblies [1], aggregation [2], or phase separation and cellular compartmentalization [3]. As macromolecules and cytosolic scaffolds crowd the interior of a cell, robust networks of smart matter form that are capable of making collective decisions for cellular survival. The structure and dynamics of these crowded spatial networks of macromolecules manifest from a multitude of weak enthalpic interactions, called 'quinary' interactions [4], and collective entropic forces. Since subcellular systems are held together through heterogenous weak interactions and entropic effects, the macroscopic degrees of freedom are often interwoven with the microscopic. One such example is the non-equilibrium dynamics of the cytoskeletal network [5]. This is strikingly different from conventional condensed matter physics systems

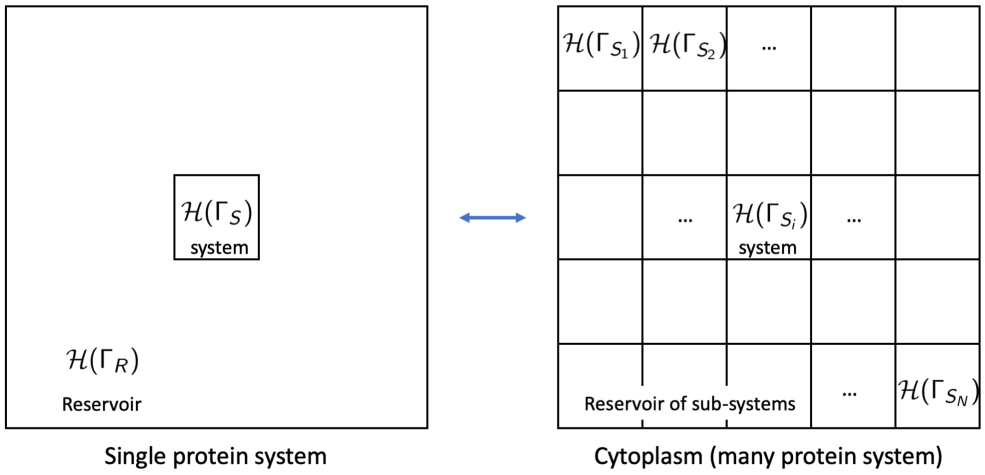

Figure 1.1 Moving from single protein to many protein systems and multi-protein assemblies in the cytoplasm.

where a key triumph of the past century came from averaging out the weak microscopic interactions (i.e. renormalization group [6]). Moreover, the cytoplasm allows for signals (mechanical or chemical) to propagate over many lengths of a single protein. The challenge in understanding signal transduction lies in the unknown complexity of molecular interactions in a jammed-packed space. It is our group's long term research goal to understand how signals from meaningful external stimuli passes through networks of proteins across the noisy environment of the cell. In order to reach that goal, both theory and experiment are needed to link the relationship between the cellular environment and protein structure and function.

In summary, the combination of crowding, solvent fluctuations, quinary interaction and chemical changes can greatly change the dynamics and stability of a protein, and in turn, its function. Our understanding of relationships between the cellular environment and protein

folding and dynamics will unify the single molecule scale to the subcellular scale. More theory and experiment will continue to rely on each other to push this goal forward. As we rebuild the complexity back into the ideal solution (i.e. in vitro) we can begin to develop in vivo principles. At the other end of the complexity spectrum, in vivo experiments and all-atom simulations of cellular cytoplasm will give us a phenomenological understanding. The challenges in understanding protein folding and dynamics in the cell are daunting, but the drive to discover is greater. Moving forward to mesoscopic assembles and supramolecular machine dynamics, further bridging the length and time scales, we will elucidate the stunning complexity of subcellular organization and dynamics.

1.1 Protein Structure and History of the Folding Problem

Proteins are weakly branched heteropolymers with monomers composed of 20 possible amino acids[7]. The sequence of these amino acids is termed "primary structure" [Fig. 1.2(a)]. This is the first of five levels of protein structure hierarchy. Since the average protein is approximately 200 amino acids long, there are 20^{200} ($\approx 1.6 \times 10^{260}$) possible primary structures[8]. However, real proteins occupy a minute fraction of the total sequence space [9, 10].

This tiny slice of sequence space are primary structures that will fold into a 3D compact form and carry out various biological functions. Within some flexibility, proteins fold into unique structures. This unlike a collapsed polymer, which may have a large variability in possible compact structures. Pauling and Corey proposed in the 1950s that the folded amino acid chain could form a periodic array of hydrogen bonds, leading the the suggestion of helix and sheet structures [11, 12]. These structures became to be known as the "secondary structures", α-helices and β-sheets [Fig. 1.2(b)], and are arguably the building blocks of folding.

Next in the hierarchy of protein structures is "tertiary stucture", which is the folding of α-helices and β-sheets into a compact globular structure shown in Fig. 1.2(c). Stably folding into a globule is primarily driven by the hydrophobic amino acids being buried in the core away from the solvent. Depending on the solvent conditions or external stresses, protein can lose their 3D folded structure. Throughout this thesis, this process is interchangeably referred to as denaturation or unfolding. In the 1960s, Anfinsen experimental realized that denatured proteins may be re-natured (folded again) into a globule by returning the solvent or other external perturbations to normal physiological conditions [13]. This observations lead to the hypothesis that the folded protein is the state with the lowest Gibbs free energy.

Shortly after Anfinsen's famous experiments, Levinthal proposed the argument that there are too many possible confirmations for proteins to "find the needle in the haystack by random searching" (needle meaning folded state and haystack meaning conformational space) [15, 16]. His argument became known as the Levinthal paradox. This paradox was at the center of what was termed "the protein folding problem"*. The main battle was between thermodynamic or kinetic control for reaching a protein's free energy minimum. Thermodynamic control meant that reaching a global minimum was path independent and required an extensive search. Whereas, kinetic control was path dependent allowing quick folding. For kinetic control, the physical system would need to be non-ergodic. Examples of physical systems displaying kinetic control are glasses [18] and metastable phase transitions[19, 20]. The solution to this problem will be discussed in the following section (1.2).

The same interactions that fold a protein also give rise to assembly of multiple tertiary structures. In the hierarchy, this is called "quaternary structure" shown in Fig. 1.2(d) and is also referred as protein complexes. These assemblies are stable and permanent.

Closely related to quaternary structure is the final level of structure hierarchy: "quinary

*The "problem" is actually a group of related problems. See ref.[17] for more details.

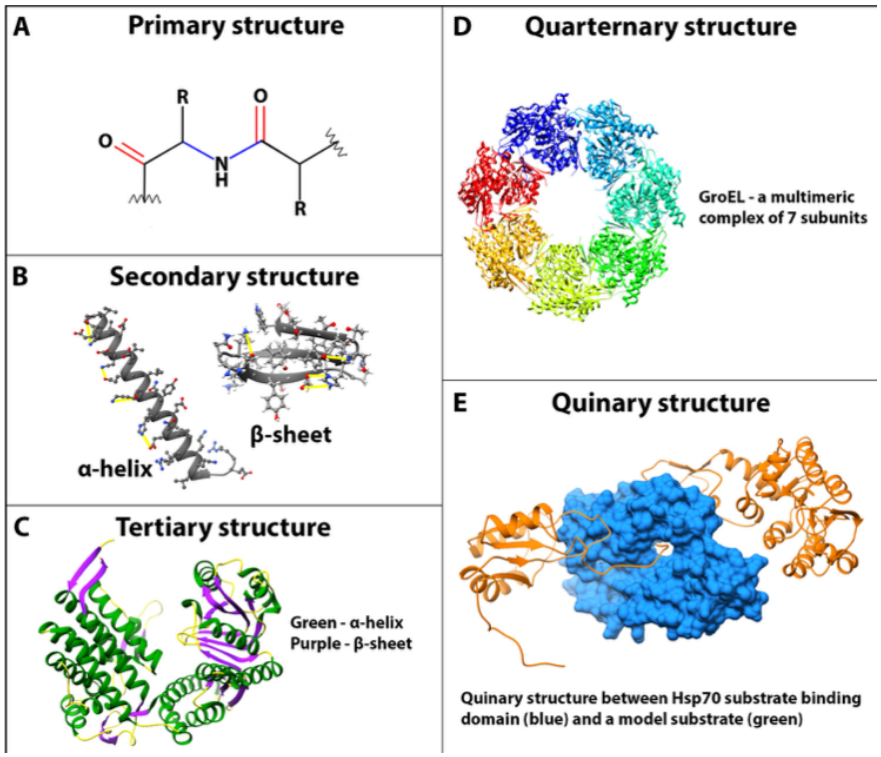

Figure 1.2 Hierarchy of protein structures. (A) Primary structure is a 1D sequence of amino acids held together by amide bonds. (B) Examples of secondary structure are α-helices and β-sheets. (C) An example of tertiary structure is a fully folded GroEL (PDB ID: 1SS8) consisting of both α-helices and β-sheets. (D)A complex made of seven GroEL (PDB ID: 1SS8) proteins is an example of quaternary structure. (E) Quinary structure (weak binding) between Hsp70 substrate binding domain (blue, PDB ID: 2KHO) and a model substrate phosphoglycerate kinase (orange, PDB ID: 3PGK). From ref.[14]

structure". At this level of orgainization, proteins interact weakly and form transient (low thermodynamic stability) functional complexes in the cell. Additionally, the interactions that underlie quinary structure formation are referred to as "quinary interactions". This quinary interaction network may span the entire proteome the organism [14].

1.2 Proteins as Heteropolymers and Energy Landscape Theory

The solution to the protein folding problem came from two the unification of two fields, polymers physics[21, 22] and spin glasses[23, 24, 25]. Since proteins are heteropolymers, certain control variables that dictated the behavior of polymers should also dictate protein behavior. One universal property is that protein size depends on its length or number of amino acids (N). When denatured, proteins act as a random coil, and when folded, proteins are maximally compact. From Flory theory, the radius of gyration (R_g, defined as $R_g^2 \equiv \frac{1}{2N^2}\sum_{i,j}\left(\boldsymbol{r}_i - \boldsymbol{r}_j\right)^2$ where $\boldsymbol{r}_i$ is the position of the i^{th} residue) of proteins should scale as,

$$R_g \sim N^{\nu}, \tag{1.1}$$

where the Flory exponent $\nu = 0.6$ for a random coil and $\nu = 1/3$ for a collapsed polymer. Indeed, experimentally derived data confirms this size scaling behavior, which is shown in Fig. 1.3(a & b).

Depending on the solvent conditions or temperature, polymers (including proteins) will be either in a random coil or collapsed phase. Meaning that at a certain point, a phase transition occurs from one phase to the other.[†] However, proteins exhibit remarkable co-operativity and "all-or-nothing" folding transitions (first-order), unlike homopolymers or

[†]Strictly speaking, phase transitions are only exhibited by infinite systems; nevertheless, it is usefully characterize collapse and folding transitions as such.

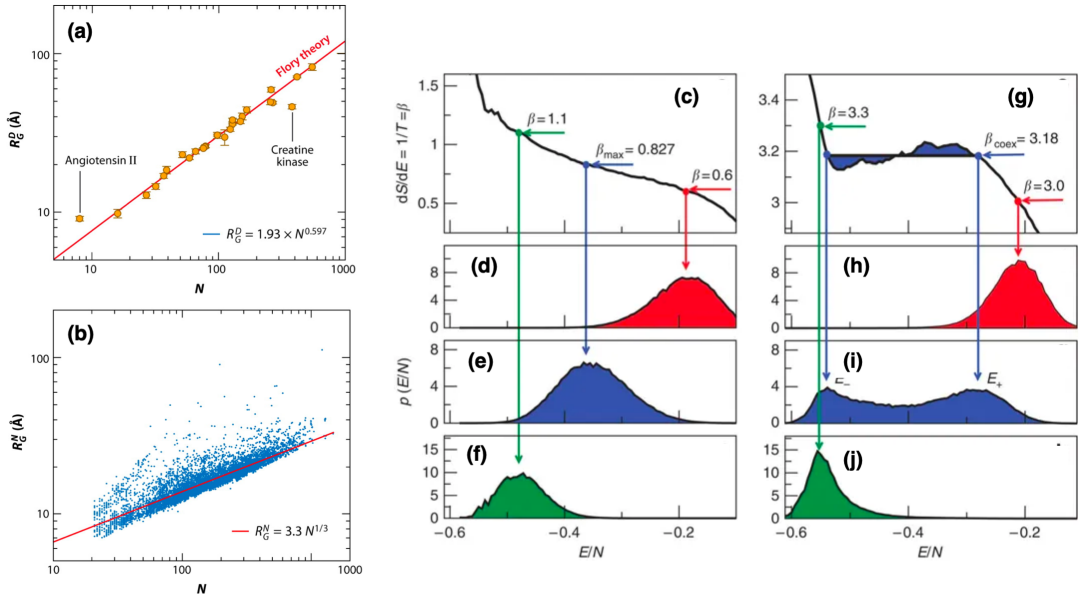

Figure 1.3 Radial distribution function for (a) unfolded, R_g^D, and (b) folded proteins, R_g^N, following Flory polymer theory. Maxwell transition curves and corresponding energy distributions for (c-f) continuous collapse transition and for (g-j) abrupt collapse transition. (a-b) From ref. [28], and (c-j) from ref. [29].

random heteropoylmers that collapse in a continuous manner [26, 27]. Systems that undergo a continuous transition have a monotonically decreasing change in entropy (S) with respect to a change in energy (E). In Fig. 1.3(c), $\frac{\mathrm{d}S}{\mathrm{d}E}$ monotonically decreases as energy increases. Additionally, the distribution of E continuous moves at specific values of $\frac{\mathrm{d}S}{\mathrm{d}E}$ [Fig. 1.3(d-f)]. In contrast, systems that undergo a first-order transition will have an inflection in the $\frac{\mathrm{d}S}{\mathrm{d}E}$ as is shown in Fig. 1.3(g). At a transition temperature, a bimodal distribution of E occurs [Fig. 1.3(i)] indicating that both phase coexist. These polymer physics and phase transition concepts play an important role in chapter 3 of this thesis.

Through polymer theory, we understand how proteins become a compact structure from a random coil; however, collapsed polymers are still disordered. Thus, folding into a unique (approximately) configuration may be described as an order-disorder transition [30]. And If

Levinthal's paradox is ture, then the energy landscape of proteins must be relatively random. To understand this transition process, we borrow ideas from spin glasses [23, 24, 25]. A randomly chosen sequences will have random interactions between residues. The prototypical system of random interactions is the Sherrington-Kirkpatrick spin-glass model[23], which has a Hamiltonian of the form

$$\mathcal{H} = -\sum_{ij} J_{ij}\sigma_i\sigma_i \tag{1.2}$$

where J_{ij} is a random energy and Ising spin $\sigma_i = \pm 1$ at site i.

The simplest example system is a three-spin triangle model as illustrated in Fig. 1.4(a). If two interactions are positive and one negative as portrayed in the figure, the system will have *frustration*. Meaning, the top spin in the current configuration can only optimally satisfy an interaction with one of its neighbors and not both.

By extending the concept of frustration to a random heteropolymer with random interaction energies between residues, $U(\boldsymbol{r}_{ij})$, we expect many near-ground states that are relatively for apart on the landscape and often with large energy barriers between them. Such rugged landscape for our random polymer model is depicted in Fig. 1.4(b). The dynamics on this surface would be glassy and a configuration may become stuck in a local minimum for spans of time that exceed experimental capabilities.

In contrast, real proteins must fold on biologically relevant time scales to be useful for living cells. Therefore, through understanding spin glasses, we arrive at the conclusion that the energy landscape of proteins must be funnel with a "flow" towards a global minimum (being the folded state) and must be minimally frustrated [31, 32, 33, 34] as shown in Fig. 1.5(a). This principle of minimal frustration means that the ratio between the folding temperature, T_f, and glass transition temperature, T_g, is maximized, allowing for rapid folding. This is directly related to the energy gap, ΔE, between order (folded) and disorder

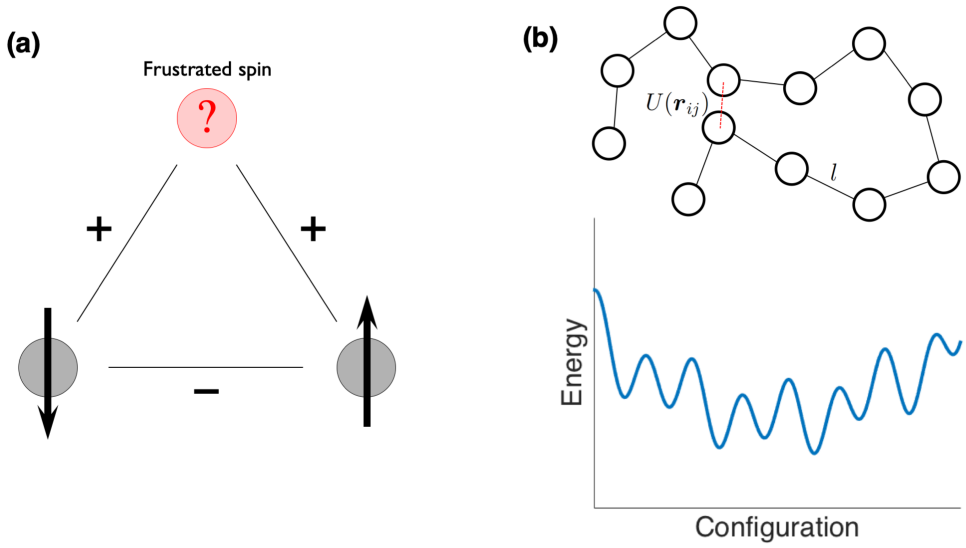

Figure 1.4 (a) Spin glass example and (b) energy landscape of frustrated polymer.

(unfolded) phases and its ratio with the energetic roughness, δE, of the landscape. The solution to Levinthal's paradox did not need "kinetic control" pathways as was mentioned in the previous section. The folding of a protein does not follow a single, specific pathway; a protein explore an ensemble of pathways on a minimally-frustrated, funnel landscape to reach its folding phase. The solution to Levinthal's paradox is that Nature selected sequences produce and energy landscape with maximized $\Delta E/\delta E$.

At the same time though, not every detail of the sequence is important. Two sequences with over 80% the sequence alignment my not have the same structure, and on the other hand, two sequences with less than 20% the sequence alignment my have homologous structures. Therefore the energy landscape emerges from the total underlying sequence and is not finely tuned [31, 35].

This energy landscape at T_f gives rise to a simple free energy plot in Fig. 1.5(b) that we are accustom to in the study of phase transitions of statistical physics. This double-well free

energy, G, with respect to an order parameter, is that of a first-order phase transition when the free energy difference between both phases $\Delta G = 0$. As such, to find the phase diagram of a protein on the pressure (P) and temperature (T) axis, we solve for $\Delta G(T, P) = 0$. A small change in Gibbs free energy difference with respect to T and P is defined as,

$$\mathrm{d}\Delta G = -\Delta S \mathrm{d}T + \Delta V \mathrm{d}P, \tag{1.3}$$

where ΔV is the difference in volume and ΔS is the difference in entropy. Upon integration of this equation from an arbitrarily chosen reference point T_0 and P_0 to T and P, one obtains

$$\begin{aligned}\Delta G(T,P) = \frac{\Delta\kappa}{2}(P-P_0)^2 + \Delta\alpha\,(T-T_0)(P-P_0) - \Delta C_P\left[T\left(\ln\frac{T}{T_0}-1\right)+T_0\right] \\ + \Delta V_0\,(P-P_0) - \Delta S_0\,(T-T_0) + \Delta G_0.\end{aligned} \tag{1.4}$$

By using a second order Taylor expansion around T_0 and P_0:

$$\Delta C_P\left[T\left(\ln\frac{T}{T_0}-1\right)+T_0\right] \approx \frac{\Delta C_P}{2T_0}(T-T_0)^2 \tag{1.5}$$

where κ is the compressiblity, α is the thermal expansion coefficient, C_P is the heat capcity, and $\Delta G_0 = \Delta G(T_0, P_0)$. This forms an ellipse on the the P-T phase diagram of a protein, where the inside is the folded phase and the outside is the unfolded phase as shown in Fig. 1.5(C).

1.3 Cellular Environment: Crowding Effect and Quinary Interaction

The theory of protein folding in a dilute solution (test-tube) is an idealized view of protein folding in the cell[36]; namely, we expect the general principles to be the same, but new

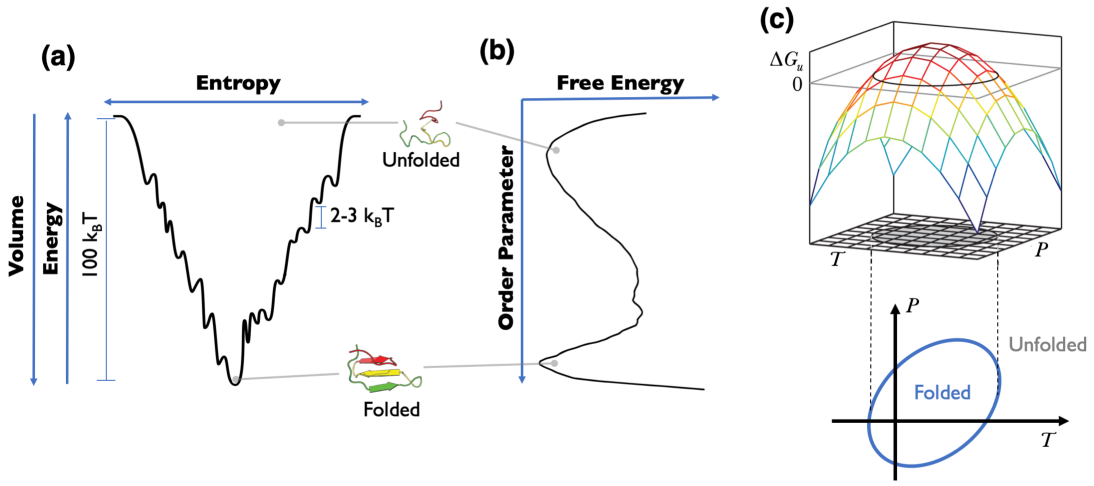

Figure 1.5 (a)Funnelled energy landscape and (b) corresponding free energy with respect to folding phase order parameter. (c)Two-state pressure-temperature phase diagram

behavior will arise with added complexity of the cellular environment [37]. The cellular environment differs from the test tube by two main forms (see Fig.1.6): (i) entropic effects from crowding, packing shapes, and solvent dynamics, and (ii) enthalpic effects from electrostatics, quinary van der Waals interactions, and chemical perturbants. Since cells are highly crowded by macromolecules, non-trivial collective effects arise. Additionally, protein sequence information can contain quinary interactions [4] that guide proteins towards self-assembly of transiently stable complexes to perform various functions. In order to bridge the gap from single protein dynamics to subcellular network dynamics, we first need to understand the behavior of single (or few) molecule(s) placed in the cellular environment.

Several in vivo experiments have begun to scratch the surface of understanding protein folding and dynamics in vivo. By studying proteins in an actual cellular environment with all of its complexity, we gain key phenomenology. Motivated by the possibility of living

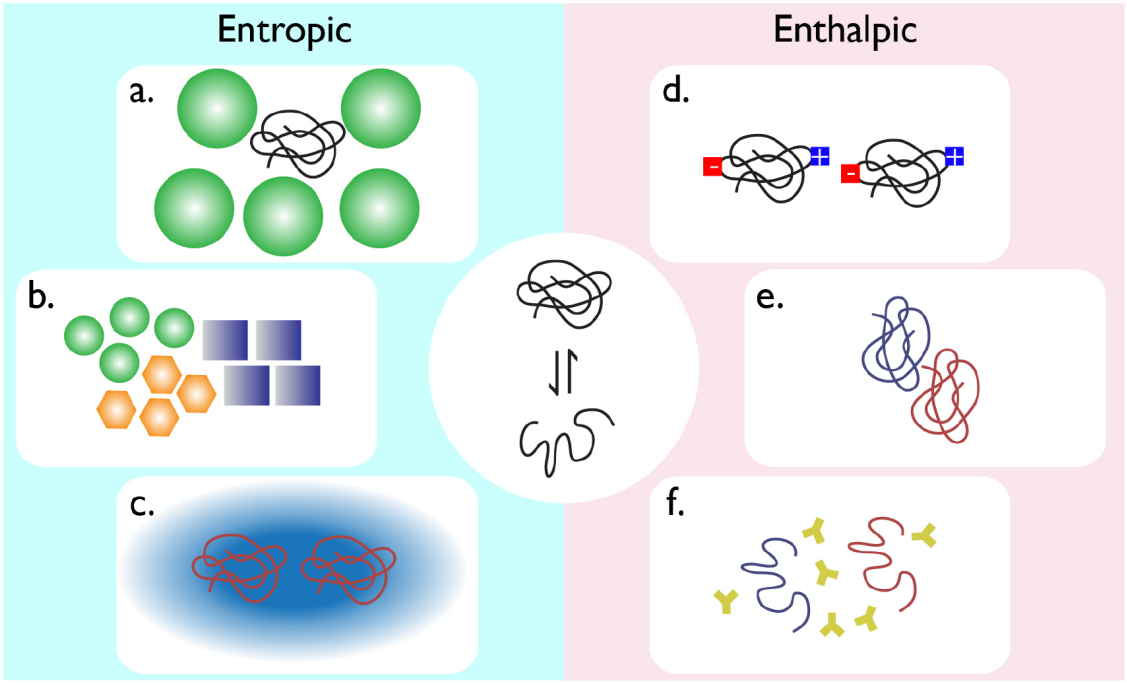

Figure 1.6 Protein folding and dynamics under the influence of the entropic and enthalpic contributions of the cellular environment. (a) Macromolecular crowding effect, (b) packing shape, (c) hydration or solvent fluctuations, (d) electrostatics, (e) quinary van der Waals interactions, and (f) chemical perturbants.

cells controlling the function and dynamics of their proteome through modulating their proteins' landscapes, these studies investigate protein stability inside living cells. Studies in [38, 39, 40, 41, 42] focus on the effects of the environmental perturbations, on the order of a $k_{\mathrm{B}}T$, to proteins. These weak interactions produce mixed results, shifting the free energy, positively, negatively or have little effect compared to in vitro folding. Other in vivo experiments [43, 44, 45] focus on the quinary interactions encoded in the protein sequence in order to understand the protein folding stability changes. These in vivo results are also corroborated by all-atom simulations of the cytoplasm of a bacterium [46].

The development of a theory is needed for the in-living-cells protein folding experiments; however, it is difficult to develop principles of in vivo folding and dynamics without a firm grounding of the interactions and entropic effects. Macromolecular crowding is arguably the most universal effect, since all molecules occupy volume. During the three decades since the 'macromolecular crowding effect [47]' was coined in 1981, the native state of proteins has been modeled as incompressible hard cores in theories that evaluate the impact of volume exclusion on protein stability and dynamics. This view has been critically challenged by collaborative work between the experimentalist group of Pernilla Wittung-Stafshede and the Cheung group. Using computer simulations and theories of statistical physics, we predicted that a structurally complex protein resembles a 'pom–pom' instead of a hard core in solutions. When the macromolecular crowding effects are considered, the stability of the native state is enhanced by the mechanistic interactions between surrounding macromolecules and the native state of a protein. This prediction has been validated experimentally in studies of a globular apoflavodoxin protein in the presence of the synthetic macromolecule Ficoll 70, which acts as a crowding agent [48]. Our group further predicted the change in folding routes of apoflavodoxin caused by different shapes and sizes of crowders [49]. The folding routes in the bulk solutions experience a higher percentage of unproductive folding intermediates

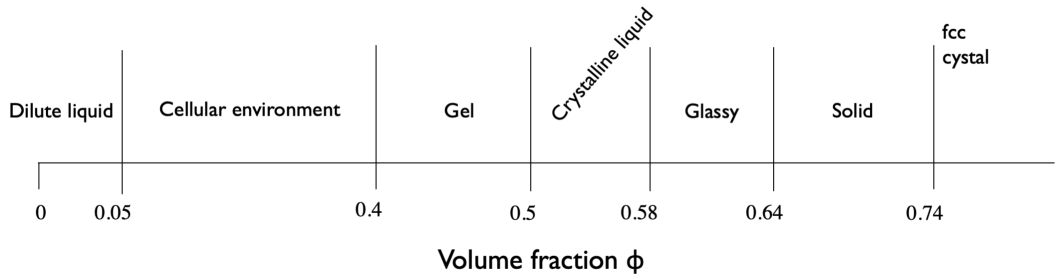

Figure 1.7 Excluded volume fraction scale. Volume fractions at which transitions arise are: freezing, $\phi_{\rm F} = 0.494$; melting, $\phi_{\rm M} = 0.545$; glassy, $\phi_{\rm G} = 0.58$. Also random close packing, $\phi_{\rm RCP} = 0.637$ and face-centered cubic array, $\phi_{\rm FCC} = 0.74$.

than in the crowded environment [49].

An E. coli cell is about 1 cubic micrometer in volume. This cubic micrometer contains on the order of 10^4 ribosomes and mRNA molecules, 10^6 proteins (thousands of types), and 10^7 DNA base pairs. The exclude volume fraction from these macromolecules can be up to 40% of the total cell volume. For additional perspective on the amount of crowding, the scale in Fig. 1.7 gives the volume fraction of various material phases. Thus, the cell is more like a heterogeneous gel instead of a dilute liquid. Even though the volume fraction at which crystallization occcurs ($\phi_{\rm F} = 0.494$) is a difference of 0.1 from the cell environment, the effects may still be felt in the cell due to local fluctuation and finite size effects.

The macromolecular crowding effect is a result of volume exclusion by surrounding macromolecules. It has been known since the 1950s that density fluctuations of macromolecules create a void where a compact protein resides and compact conformations are statistically

favored over extended conformations [50]. However, the determination of the 'native' state becomes non-trivial when the conformation of a compact protein is malleable and can be easily changed by interactions with other objects. In collaboration with Professor Stafshede-Wittung's group, we showed that the shape of aspherical VlsE proteins could be changed under cell-like conditions, such as those generated by chemical and thermal denaturation [51]. In crowded milieus, distinct conformational changes from an olive shape to a sphere in VlsE are accompanied by secondary structure alterations that lead to exposure of a hidden antigenic region. This work demonstrates the unprecedented malleability of 'native' proteins and implies that crowding-induced shape changes may be important for protein function and malfunction in vivo.

The idea of a protein's 'native' state in a crowded cell has established an important milestone towards protein folding inside the cell. In collaboration with Professor Martin Gruebele, we identified the enzymatic mechanism of a protein, phosphoglycerate kinase (PGK), with the shape of a 'pac-man' in cell-like conditions [52]. Our coarse-grained models showed that PGK adopts a closed 'pac-man' conformation in a crowded cell, bringing the two lobes together to react. Indeed, experiments have shown that the enzymatic activity of PGK is a remarkable 15-fold higher in a crowded environment.

Further theoretical developments of the macromolecular crowding effect by scaling relationships [53] allows us to understand the relationship between the ratio of crowder size to a generic biopolymer radius of gyration and the collapse of that generic biopolymer. This was also validated experimentally in a general colloid polymer solution [54] and with intrinsically disordered proteins (IDPs) in a crowded environment [55]. In addition to the size of crowders, the shape also has a significant entropic effect [56, 57]. Even though much progress has been made in understanding the macromolecular crowding effect, there is still much we do not fully understand such as the effects of heterogeneous sizes and shapes as is found in

the cytoplasm.

The last entropic effect that is necessary for a complete understanding of in vivo folding deals with the behavior of water in the cell. As shown for protein folding in the test tubes, the dynamical fluctuations of water at the protein's surface [58, 59], desolvation [60] between residues, and hydrodynamic interactions [61, 62] will also play an important role in the cell. When departing from the dilute solution limit to high volume fraction of macromolecules, the behavior of water will change entropically [63, 64], kinetically [65, 66], and possibly even quantum mechanically [67].

A great challenge is to understand how weak, seemingly random, quinary interactions [4] can lead to non-random arrangements of protein organization. These interactions can alter the thermodynamic folding barriers leaving the protein marginally stable. Both theoretical and experimental developments have been made to understand macromolecular crowding with enthalpic contributions [68, 69, 70, 71, 72]. Being able to separate out the entropic from enthalpic effects is a vital step towards the development of general principles of in vivo protein folding. This computational approach advanced the next step of research in understanding protein folding in cytoplasmic media [73].

Somewhere along this axis (Fig. 1.8) the property which we call living emerges. So by exploring the various phases across these scales, we may be able to understand what it means to be alive.

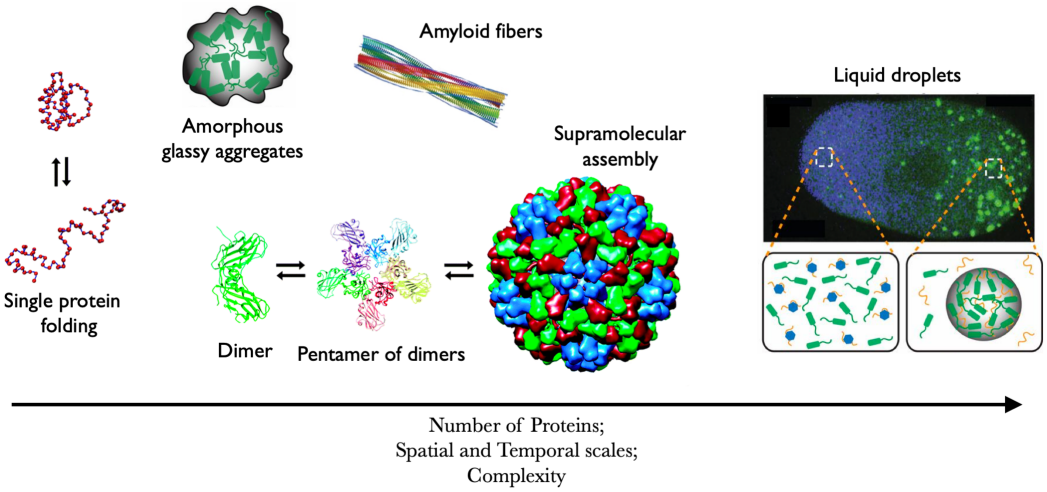

Figure 1.8 Various protein phases across scales of size, time, number, and complexity. From left to right, at the single protein scale is protein folding transitions; then at the multi-protein scale is aggregation, dimerization, fiber formation, and supramolecular assembly; at the cellular scale is liquid-liquid phase separation. Adapted from refs. [3, 74, 75].

Exploring the Phase Space of a Single Isolated Protein

αεὶ ο θεὸς ο μέγας γεωμετρεῖ τὸ σύμπαν.
(Always the great God applies geometry to the universe.)

—Plutarch

2

Understanding the Temperature-Pressure Phase Diagram of a Protein*

Proteins unfold not only under high heat but also in the presence of high hydrostatic pressure. This effect has been known since the early 1900s [76, 77], and the equation of state on the pressure-temperature plane was formalized by Hawley in 1971 [78]. While protein unfolding by heat is more intuitive, pressure denaturation can be explained from Le Chatelier's principle, in which pressure unfolds proteins due to a negative volume change. The molecular origin of this negative volume change was recently discovered to be the penetration of water into the hydrophobic core, causing loss of the protein's cavities [79, 80].

Computational simulations are essential to gain further insight into specific pressure-perturbed folding mechanisms. Pressure denaturation of proteins has been studied by all-atom molecular dynamics simulation [81, 82, 83, 84]. However, it is computationally costly

*Contents of this chapter has been published in *J. Phys. Chem. B* (2020) **124**, 1619-1627. AG Gasic is first author.

since pressure unfolds proteins on a longer timescale than heat denaturation [85] and often requires sampling tricks [86]. An alternative to all-atom models are structure-based coarse-grained models. Structure-based models render an energy landscape with minimal frustration and contain a funneled landscape with a dominant basin of attraction corresponding to an experimentally determined configuration [87, 88]. The mechanism that drives protein folding from unfolded conformations to few unique conformations where the hydrophobic residues coalesce to a "hydrophobic core" is similar to what drives oil separating from water. As such, these models are computationally inexpensive, allowing long-timescale simulations even for large proteins and complex systems. Similar to how the "Ising model" is used to develop the general theories of phase transitions, or how the "ideal gas" illustrates the basic notions of fluid behavior, structure-based models of proteins are used to understand fundamental aspects of protein folding and dynamics [88, 30]. Furthermore, to understand large systems, such as protein folding in vivo [89, 90], structure-based minimalist models are essential for developing new theories.

Pressure-dependent hydrophobic interactions in coarse-grained protein models are approximated by using the potential of mean force (PMF) between two methane molecules at various hydrostatic pressures [91, 92, 93, 94], since methane molecules are a simple model for the interaction of hydrophobic residues in a protein. Understanding the behavior of two contacting methane molecules in aqueous solution provides insight into the mechanism of hydrophobic collapse during protein folding. In general, this PMF contains a contact well, a solvent-mediated well, and a desolvation barrier between the two wells, as shown in Fig. 2.2. The presence of the desolvation barrier is due to the free-energy cost for two methane molecules to penetrate the first hydration shell between them and move from the water-mediated contact well to reach the direct contact well. Pressure (P) and temperature (T) can cause changes in the depths of wells and the height of the barrier. However, the exact

description of the PMF is still debated between two methane molecules [92, 95, 96, 97, 98] and in proteins[99, 100, 101, 102, 103]. Hummer et al.[92] find a pressure-dependent PMF with a contact well energy that weakens favoring the water-mediated well as pressure increases. Opposing Hummer et al. [92], Dias and Chan show that the contact well deepens as pressure increases [93]. Because of the different pressure-dependencies of these two potentials, they produce different folding-phase behaviors of proteins on the P-T plane. Here, we investigate these two different pressure-dependent desolvation potentials and compare it to the known behavior of a well-studied protein under high hydrostatic pressure, phosphoglycerate kinase (PGK), based on experimental evidence [104, 105, 106]. Before presenting the results, we give an overview of pressure denaturation relevant to the current investigation.

2.1 Overview of Pressure-Denaturation

Hawley's theory dictates that the contributions of the changes in entropy ($\Delta S \equiv S_{\mathrm{U}} - S_{\mathrm{F}}$) and volume ($\Delta V \equiv V_{\mathrm{U}} - V_{\mathrm{F}}$) between unfolded (U) and folded (F) protein phases result in an elliptical-shaped coexistence ($\Delta G = 0$) curve in the P-T space (Fig. 2.1) [78]. From the Clausius-Clapeyron relation, the slope of the coexistence curve between the two phases is

$$\left.\frac{\mathrm{d}P}{\mathrm{d}T}\right|_{\Delta G=0} = \frac{S_{\mathrm{U}} - S_{\mathrm{F}}}{V_{\mathrm{U}} - V_{\mathrm{F}}} = \frac{\Delta S}{\Delta V} \tag{2.1}$$

Since the coexistence curve is elliptical, both ΔV and ΔS can be positive, negative, or zero depending on the P and T. The quadrants of the ellipse are broken up by the dashed $\Delta V = 0$ and dotted $\Delta S = 0$ lines in Fig. 2.1, where $\Delta V < 0$ or $\Delta V > 0$ if above or below the $\Delta V = 0$ line and $\Delta S < 0$or $\Delta S > 0$ if to the left or right of the $\Delta S = 0$ line, respectively. Fig. 2.1 contains two different examples of elliptical curves to help breakdown

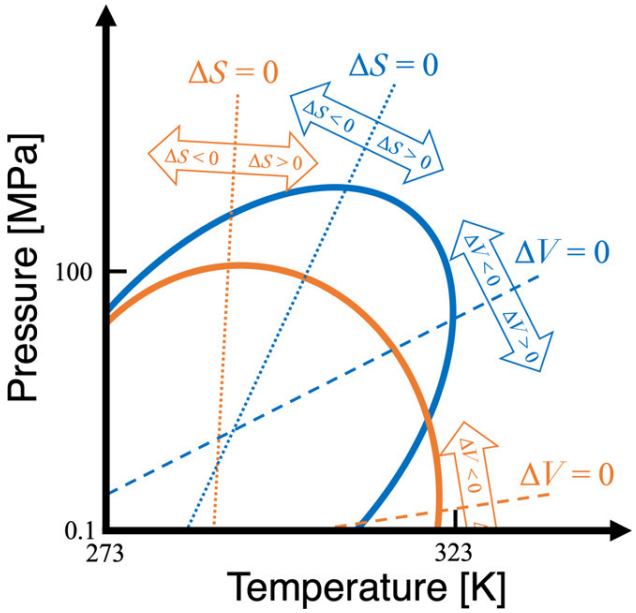

Figure 2.1 Schematic P–T phase diagrams for two proteins with different elliptical coexistence ($\Delta G = 0$) curves. The dotted lines represent $\Delta S = 0$ (left of the line, ΔS is negative; right of the line, ΔS is positive), and the dashed lines represent $\Delta V = 0$ (above the line, ΔV is negative; below the line, ΔV is positive). The two ellipses differ by a shift in the pressure direction and a shear of the ΔV and ΔS axes. The origin of the plot is at 0.1 MPa (1 atm) and 273 K. For the orange curve, ΔV is mainly negative for experimentally relevant P and T; whereas, the blue curve has a large region where ΔV is positive. Examples of proteins with a similar elliptical coexistence curves are ribonuclease and phosphoglycerate kinase (PGK) in orange and chymotrypsinogen in blue.

the contributions of entropy and volume in the discussions to follow in this section and other parts of this chapter.

From Le Chatelier's principle, high pressure shifts the equilibrium toward the phase with the smallest volume to minimize the free energy. Therefore, for hydrostatic pressure to unfold a protein, the partial molar volume of the unfolded phase (V_{U}) must be smaller than that of the folded phase (V_{U}); that is, the change in volume must be negative ($\Delta V < 0$) [107, 108].

In general, ΔV is negative at room temperature [109]; however, positive ΔV can be observed specifically at high temperature [78] or for α-helix peptides [110, 111]. For example, chymotrypsinogen [78] has a coexistence curve similar to the blue curve in Fig. 2.1, and applying a medium pressure (below 50 MPa) at high T (320 K) will fold the protein due to the positive change in volume. However, chymotrypsinogen will unfold again at a higher pressure (above 100 MPa) after crossing the $\Delta V = 0$ line.

One of the key factors in producing the elliptical phase diagram for protein folding stability in the P-T plane is the temperature dependence of ΔV. Formally, ΔV is given as

$$\begin{aligned} \Delta V &= \int_{T_0}^{T} \Delta\alpha \mathrm{d}T' + \int_{P_0}^{P} \Delta\kappa \mathrm{d}P' \\ &= \Delta V_0 + \Delta\alpha\,(T - T_0) + \Delta\kappa\,(P - P_0) \end{aligned} \tag{2.2}$$

where the change in thermal expansivity is $\Delta\alpha$, the change in compressibility is $\Delta\kappa$, and $\Delta V_0 \equiv \Delta V(T_0, P_0)$. Usually, the magnitude of $\Delta\kappa$ is small, giving an almost negligible pressure dependence to ΔV. At room temperature, generally $\Delta\kappa < 0$. Additionally, $\Delta\alpha$ tends to be greater than zero, which may be due to the greater degree of hydration of unfolded proteins [112].

The P and T dependencies of ΔS is the other factor that contributes to the elliptical phase diagram. Since a folded protein has lower entropy than an unfolded protein, heat shifts

the population to the phase with the highest entropy. At low temperatures, the low-entropy state of the solvent is favored, which is reconciled by unfolding the protein [113, 114]. Note that in this thesis, we will not consider cold denaturation.

The change in partial molar volume stems from changes in void (or solvent inaccessible cavity), van der Waals, and hydration volumes [115]

$$\Delta V = \Delta V_{\text{vdW}} + \Delta V_{\text{void}} + \Delta V_{\text{hyd}} \tag{2.3}$$

However, $\Delta V_{\text{vdW}} \approx 0$ for most cases, which means ΔV_{vdW} can be ignored.

For partial molar volume changes in void (ΔV_{void}), since folded proteins are not perfectly packed, dry voids (or cavities) of varying sizes and shapes are distributed heterogeneously throughout a protein [116, 117]. High pressure induces unfolding by introducing solvent into the protein structures, eliminating the cavities, and decreasing the overall solvent-accessible volume [79, 80]. Water penetrates the hydrophobic core of proteins because of the reduced solvent-solute interfacial free energy when pressure increases [118].

For partial molecular volume changes in hydration (ΔV_{hyd}), hydration of newly solvent-exposed side chains upon unfolding also contributes to a change in partial molar volume by changing the density of the hydration layer. This effect depends on the chemical properties of the side chains and the surface area topography. Usually, $\Delta V_{\text{hyd}} > 0$ and is small in magnitude, which may be due to polar and apolar hydration influences canceling each other. Thus, the decrease in the void volume (ΔV_{void}) overcomes the increased hydration volume (ΔV_{hyd}) in order for pressure denaturation to occur [119]. Regardless of the hydration effect, the key mechanism that unfolds a protein under high pressure is the elimination of solvent-excluded cavities because of water penetrating the hydrophobic core; therefore, $\Delta V < 0$ to unfold a protein under high pressure.

2.2 Pressure-dependent desolvation potentials

In this study, we use two models to investigate pressure-dependencies. The first model, termed "the CHOGG model" (for Cheung-Hummer-Onuchic-Garcia-Gasic), is based on a work from Hummer and co-workers [91, 92], which is derived from the information theory, and later expressed for off-lattice simulations [106, 60] of protein folding under high pressure.

The information theory model accounts for the association, solvation, and conformational equilibria of hydrophobic solutes, such as residues in the core of a protein. These calculations are described by the probability of hydrophobic solutes, forming a cavity in an aqueous solvent. As a result of increasing pressure, the desolation barrier rises because of the increased free-energy cost of forming a small cavity between two solutes. An information-theoretic model will not encounter the usual systematic errors due to overfitting of incorrect simulation data. Based on the Hummer et al.'s information theory calculation[91, 92] and Hillson et al.'s simulation [120], Cheung and co-workers [106, 60] created a desolvation barrier for structure-based models.

The second model for the investigation of pressure dependencies, termed "the DC model" (for Dias-Chan), is motivated by a work from Dias and Chan [93] that calculates the potential of mean force between to methane molecules directly from simulation data.

The values of ϵ, ϵ', and ϵ'' are related to the magnitude of pressure by the following

$$\epsilon(P) = \epsilon_0 + \upsilon P + \xi P^2 \tag{2.4a}$$

$$\epsilon'(P) = \epsilon'_0 + \upsilon' P \tag{2.4b}$$

$$\epsilon''(P) = \epsilon''_0 + \upsilon'' P + \xi'' P^2, \tag{2.4c}$$

where ϵ'_0 and ϵ''_0 are the water-mediated well energy and the barrier height at ambient P,

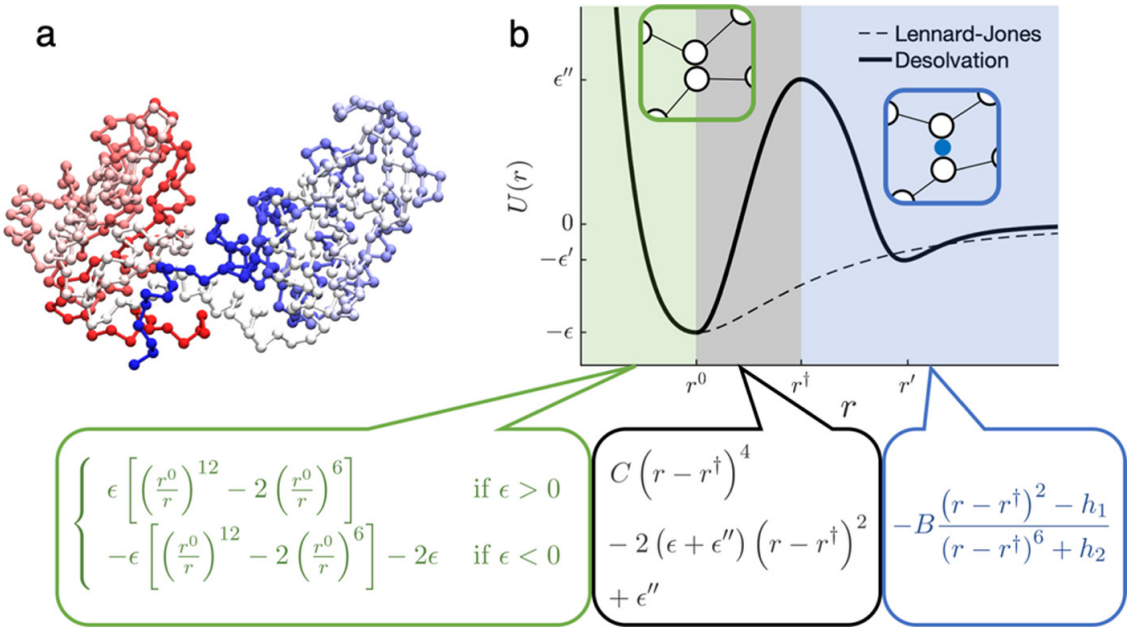

Figure 2.2 Coarse-grained representation of PGK (PDB ID: 1QPG). Each amino acid is coarse-grained to a single bead. The N- and C-terminus domains are in red and blue, respectively. (b) Desolvation potential (solid line) compared to the Lennard-Jones potential (dashed line). Here r^0, $r^\dagger$, and r' are the positions of the minimum of the first well, maximum of the desolvation barrier, and the minimum of the second well, respectively. The separation between r^0 and r' is the size of a single water molecule, 0.8σ, and $r^\dagger = (r^0 + r')/2$. The piecewise terms of the desolvation potential are shown below pointing to their corresponding colored sections, where constants $C = \frac{(\epsilon+\epsilon'')}{\left(r^\dagger - r^0\right)^2}$, $B = 3\epsilon'\left(r' - r^\dagger\right)^4$, $h_1 = \frac{2}{3}\frac{\left(r'-r^\dagger\right)^2}{\epsilon'/\epsilon''+1}$, and $h_2 = 2\frac{\left(r'-r^\dagger\right)^6}{\epsilon''/\epsilon'+1}$. The green section ($r < r^0$) is the Lennard-Jones potential for positive or negative ϵ. The attractive part of the contact well is highlighted in gray ($r^0 \leq r < r^\dagger$), and the water-mediated minimum is highlighted in blue ($r^\dagger \leq r$).

Table 2.1 Values of Constants in Pressure-Dependent Contact Well, Water-Mediated Well, and Barrier Height Energies

Constants [units]	CHOGG model	DC model
ϵ_0' $[\epsilon_0]$	0.33	0.19
ϵ_0'' $[\epsilon_0]$	1.33	0.3
υ $[\sigma^3]$	-0.127	0.039
ξ $[\sigma^3/\epsilon_0]$	0	0.007
υ' $[\sigma^3]$	0	0.042
υ'' $[\sigma^3]$	0.211	0.03
ξ'' $[\sigma^3/\epsilon_0]$	0	0.007

respectively. The constants for both models are given in Table 2.1. The constants for the CHOGG model are taken from ref. [120], and the constants for the DC model are from fitting to the PMF's (from ref. [93]) minima and barrier height versus P. Noting the values in Table 2.1, the CHOGG model only uses the linear P dependencies unlike the DC model that has second-order P terms.

Since these P-dependent energies result in two very different P-dependent desolvation potentials, as shown in Fig. 2.3, the contributions of the barrier height at $r^\dagger$ and well depth at r^0 will have different effects on the protein's behavior. With the CHOGG model, the desolvation barrier increases, and the free-energy gap between the two minima tilts to favor the water-mediated contact as pressure increases, leading to water penetrating the hydrophobic core and unfolding of a protein; whereas, with the DC model, the free-energy gap between the two minima tilts to favor the contact well as pressure increases, leading to a stabilization of the protein. Additionally, the slight softening of the contact well in the CHOGG model as pressure increases will have little effect on the global thermodynamics of the protein. This is because the water-mediated contact well (at r') is thermodynamically favored at high

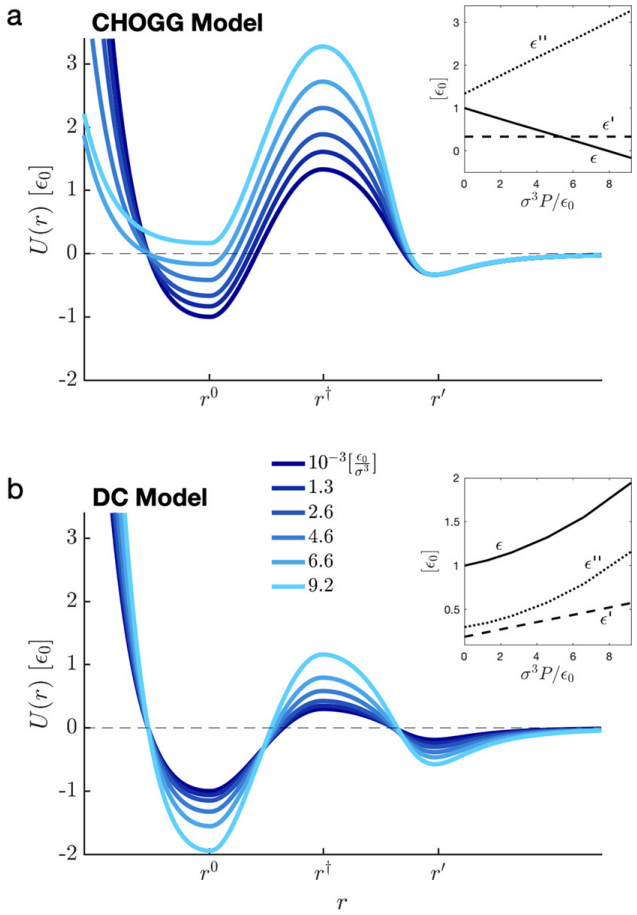

Figure 2.3 Pressure-dependent desolvation PMF between two residues while varying P, including Lennard-Jones and solvent contributions. Results are shown for $\sigma^3 P/\epsilon_0 = 10^{-3}$ to 9.2 for the (a) CHOGG model and (b) DC model. Insets show the values of ϵ, ϵ', and ϵ'' as a function of P (in reduced units, where $\epsilon_0/\sigma^3 \approx 76$ MPa).

pressure, and the residues are less likely to feel the softening of the contact well (at $r < r^0$) than they are to feel the repulsive part of the desolvation barrier (at $r^\dagger \leq r < r'$).

2.3 The phase behavior of the two desolvation models

In order to compare the behavior of PGK modeled by the two pressure-dependent desolvation potentials, we calculated the average radius of gyration (R_g) of the protein ensembles at a

wide range of P and T. We first analyze the coexistence curves ($\Delta G = 0$) of the two models on their P-T phase diagrams to show their general thermodynamic properties. We construct a P-T phase diagram for both models, shown in Fig. 2.4, using the average radius of gyration, R_g (normalized by the R_g of the crystal structure, R_g^0), as the order parameter to describe the folding phase. The black line is the coexistence curve separating the folded and unfolded phases, and on the curve, both phases are equally populated due to having a zero change in free energy ($\Delta G = 0$). The stability dependence on P and T is shown by the slope of the coexistence curve on the P- T phase diagram. The negative slope, in Figure Fig. 2.4(a), shows that both increasing P and T destabilize the protein with the CHOGG model; whereas, the positive slope, in Fig. 2.4(b), shows that only increasing T destabilizes the protein with the DC model. From Eq. 2.1, the slope of the coexistence curve is $\frac{\Delta S}{\Delta V}$. Because both models have a positive ΔS (entropy increases upon unfolding), therefore, ΔV must be negative for the CHOGG model and positive for the DC model. Thus, the CHOGG model captures the protein denaturation of PGK under pressure.

From Fig. 2.4, the slopes of the black line indicate that $\Delta V < 0$ for the CHOGG model [Fig. 2.4(a)] and $\Delta V > 0$ for the DC model [Fig. 2.4(b)]. The behavior on the phase diagram in Fig. 2.4(a) for the CHOGG model may occur in real proteins if the $\Delta V = 0$ line is shifted to a low or negative pressure and the slope is flattened similar to the orange curve in Fig. 2.1 (to the right of the $\Delta S = 0$ line); whereas, the phase diagram of the DC model in Fig. 2.4(b) is similar to the blue curve in Fig. 2.1 at a high temperature below the $\Delta V = 0$ line (i.e., $\Delta V > 0$).

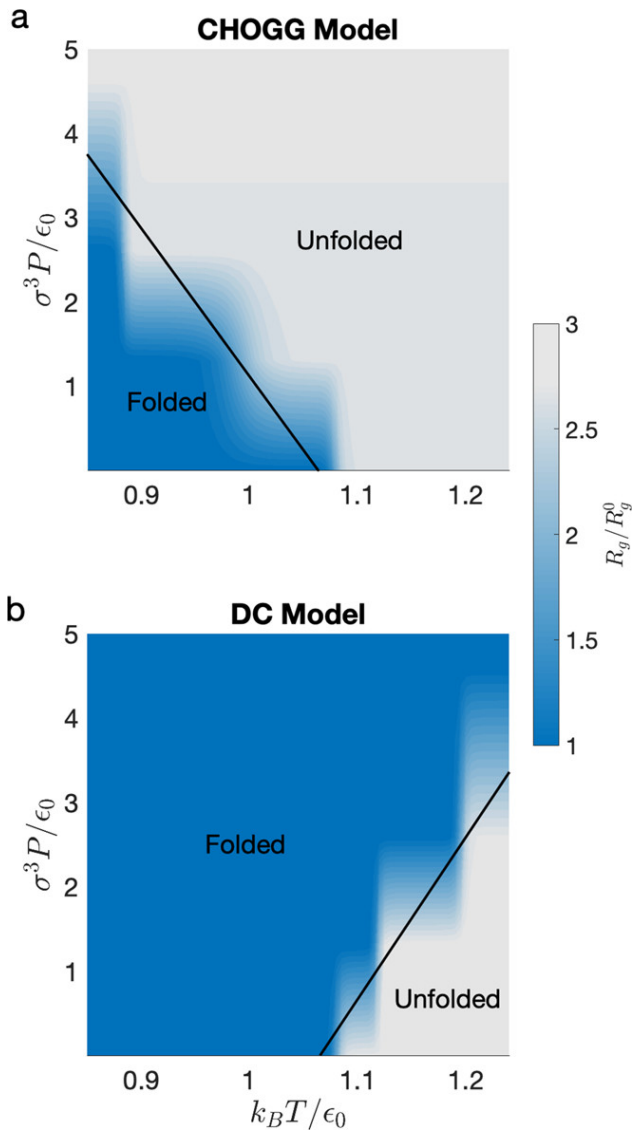

Figure 2.4 P-T phase diagram for the (a) CHOGG model and (b) DC model. The blue region corresponds to compact values of R_g/R_g^0, and the gray region corresponds to extended values of R_g/R_g^0, which signify folded and unfolded phases, respectively. The normalization (R_g^0) is the radius of gyration of the crystal structure. The black line is a linear approximation of the phase boundary between folded and unfolded protein phases. For reduced units, $\epsilon_0/\sigma^3 \approx 76$ MPa and $\epsilon_0/k_B \approx 303$ K.

2.4 CHOGG Model Captures Protein Denaturation under Pressure unlike the DC Model

Both CHOGG and the DC models do not capture the full range of the P-T plane from Hawley's theory; however, our simulation results conclusively show that the CHOGG model is best suited for studying hydrostatic pressure unfolding when in the $\Delta V < 0$ regime, which the DC model does not accurately reflect the phenomenological behavior.32 Indeed, the CHOGG model accurately describes the unfolding behavior of PGK under high pressure, including the nontrivial existence of a stable intermediate at low T and high P. The phase diagrams created by the CHOGG model were validated by experiments and by an analytical theory.32 The DC model is by no means incorrect though. The DC model may be more suited for studying the effects of pressure in the $\Delta V > 0$ regime such as an α-helix peptide, which has a smaller volume of folded than unfolded.

By combining aspects of both models, an elliptical coexistence curve can be achieved with the desired center and ΔV, ΔS axes rotation or shearing, as seen in Fig. 2.1. A protein such as chymotrypsinogen3 (blue curve in Fig. 2.1) is unfolded at high T, but it folds when a medium pressure is applied (below 50 MPa), and then it unfolds again at a high pressure (approximately 100 MPa). This is due to the fact that the $\Delta V = 0$ line of the elliptical coexistence curve is centered higher and has a steeper slope than the orange curve. For the medium pressure behavior of chymotrypsinogen, the DC model would reproduce the correct results because $\Delta V > 0$ until the pressure reaches the $\Delta V = 0$ line. Above the $\Delta V = 0$ line, the CHOGG model would be needed to ensure unfolding at higher pressures. Since PGK (orange curve in Fig. 2.1) has a $\Delta V = 0$ line with a low-pressure center and flat slope, $\Delta V < 0$ for almost all P and T, which is why the CHOGG model captures PGK's pressure

denaturation correctly.

2.5 PMFs of Methane Molecules May Not Be Fully Represent the PMFs for Proteins

These two PMFs (Fig. 2.3) may not conflict with each other because the temperature will also affect the pressure-dependence. The change in the contact well depth, whether it increases or decreases, depends on the temperature, which is shown by Ashbaugh et al [94]. through calculating the second virial coefficient. They show that at high T, the second virial coefficient increases (less attraction) as pressure increases; whereas, at low T, the second virial coefficient decreases (more attraction) as pressure increases. The pressure-dependence of the contact well of the PMF between the methanes can be interpreted as becoming shallower or deeper as pressure increases at high T or low T, respectively, which is individually captured by the CHOGG or DC model.

Furthermore, using perfectly detailed PMFs from two methane molecules as the pairwise hydrophobic interaction of a protein will not be the "true" PMF of the protein for two reasons: (i) simply, protein residues are not methane molecules and (ii) many-body and emergent effects of having many hydrophobic molecules together will change the PMF. For example, the PMF between two hydrophobic plates has a different pressure-dependence compared to that between two methane molecules [121]. For the PMF between the graphene plates immersed in TIP4P/2005 water at $T = 300$ K, as pressure increases, the contact well initially deepens until 800 MPa and becomes shallower at 1200 MPa.

Another example showing where the PMF of methane molecules contradicts the PMF of proteins: Dias also finds a different pressure-dependence in the PMF of a simple protein-

water model [101] compared to that of the two methane molecules that he and Chan propose in ref [93] (the DC model in this study). The PMF of a simple protein-water model from Dias' work [101] is comparable to Fig. 2.3(a) (the CHOGG model), which opposes the trend of Figure Fig. 2.3(b) (the DC model). The protein-water model from Dias27 and the CHOGG model both have a contact well that weakens as pressure increases; whereas, in the DC model, the contact well strengthens with increased pressure. Dias and Chan even state that proteins will have a different pressure-dependent hydrophobic interaction compared to methanes:

> Conceptually, however, it is important to recognize that two- and three-methane PMFs do not, by themselves, necessarily provide an adequate physical picture of pressure denaturation because the two- and three-body contact minima retain significant water exposure. Hence, the adequacy of these configurations as models for the sequestered folded protein core can be limited. [93]

In their conclusion of their paper [93], they also describe the difficulty of correlating the combined pressure and temperature dependencies of the methane PMF to those of real proteins. Therefore, one should understand the ΔV regimes of the protein under investigation to know whether to use the CHOGG or DC model as described in the previous subsection.

2.6 Merit of Coarse-Grained Modeling over All-Atom Simulations

Structure-based coarse-grained models have a funneled energy landscape with minimal frustration[87, 88]. As we learned from the energy landscape theory of protein folding,58 structure-based models provide fruitful insights. Why is that?—because the funneled energy landscape itself is emergent and microscopic details cease to matter [35]. This is evident with mutation experiments; few perturbations of the residue sequence retain the same topological structure.

Thus, the exact molecular scale details and chemistry are not as important as the essential physics provided by structure-based models.

Coming back to the problem at hand, regardless of the exact details, the essential physics of the CHOGG model is that the desolvation barrier increases and the free-energy gap between the two minima tilts to favor the water-mediated contact as pressure increases, leading to water penetrating the hydrophobic core and unfolding of a protein. This physics is captured by the PMF calculated by Hummer and coworkers [91, 92, 122] and has been used by others to understand pressure denaturation [114, 120, 123, 124, 125, 126]. In our recent study in collaboration with experimentalists (which inspired this work)[106], we capture important thermodynamic trends in the pressure denaturation of PGK. These trends include the evidence for critical behavior of a protein (note that criticality is another well-known emergent phenomenon [127]; not discussed here), which is not explicit in our model.

Through our investigation, we have rigorously compared two different pressure-dependent desolvation potentials and settled the debate between the two models. We have also shown how the two models may be used in structure-based minimalist models to further understand the full P-T phase diagrams of real proteins, as intended by Hawley's original work.

Exploring the Phase Space of Proteins in Crowded, Cell-like Environments

It is not the strongest of the species that survives, nor the most intelligent, but rather the one most responsive to change.

—Charles Darwin

3

Critical Phenomena in the Phase Diagram of a Protein*

Complex processes in nature often arise at an order-disorder transition [128, 30, 129, 130]. In proteins, this complexity arises from an almost perfect compensation of entropy by enthalpy: molecular interactions that create structural integrity are on the same scale as thermal fluctuations from the environment. The resulting marginal stability of proteins suggests that they could behave like fluids near a critical point [131]—their structures fluctuate considerably subject to small perturbations without overcoming a large activation barrier.

The concept of first-order and critical phase transitions does not rigorously apply to nano-objects such as proteins; nevertheless, it is a useful one to classify folding transitions. For example, folding of small model proteins has been described as an abrupt, cooperative transition between the folded and unfolded phase for some proteins (the below-critical point

*Contents of this chapter has been published in *Phys. Rev. X* (2019) **9**, 041035. AG Gasic is co-first author.

scenario), or as a gradual barrier-less 'downhill' transition for other proteins (the above-critical point scenario) [9]. Even though critical behavior of proteins has been previously hinted [132, 133, 134], there has not been a direct observation of a critical point where one of these abrupt transitions simply disappears at T_c and P_c. In larger proteins, such as phosphoglycerate kinase (PGK) ([135] in section S1), the situation can get even more complex: different parts or 'domains' of a large protein are more likely to be able to undergo separate order-disorder events [136], delicately poised between folded and partially unfolded structures to carry out their functions [52].

Proteins must fold and function while crowded by surrounding macromolecules [137], which perturb the structure of the proteins at physiological conditions in the cell. The volume exclusion from macromolecules [50], which places shape and size (or co-volume) [47, 138] constraints on the conformational space [Fig. 3.1(a)], complicates protein folding and dynamics in living cells [36]. How the competing properties of a protein arise—being both stable yet dynamically sensitive to its environment—is mostly unknown; however, we show that the crowded environment provides a unique solution by placing PGK near a critical regime.

We use pressure P, temperature T, and crowder-excluded volume fraction ϕ, to map PGK's folding energy landscape [52, 139] and its critical regime on the T–P–ϕ phase diagram. Temperature can induce heat unfolding by favoring states of high conformational entropy, or cold denaturing by favoring reduced solvent entropy when hydrating core amino acids in the protein [113, 114]. Since folded proteins contain heterogeneously distributed small, dry cavities due to imperfect packing of their quasi-fractal topology [140, 116, 117], high pressure also induces unfolding by introducing water molecules (as small granular particles) into the cavities in protein structures, leading to a reduced overall solvent-accessible volume of the unfolded protein [80]. Finally, in the presence of high crowding (large ex-

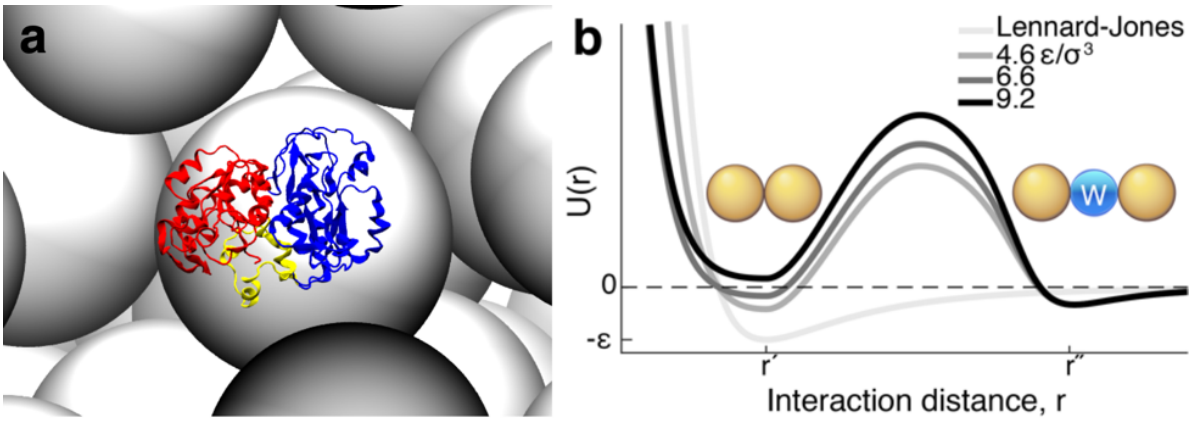

Figure 3.1 PGK surrounded by crowders and the desolvation potential between residues. (a) A snapshot from the coarse-grained molecular simulation of PGK's spherical compact state (Sph) surrounded by crowders (gray) at the volume fraction of 40%. N-, C-domain, and hinge are in red, blue, and yellow, respectively. (b) The pressure-dependent desolvation potential at $\sigma^3 P/\epsilon_0 = 4.6$, 6.6, and 9.2, contains a *desolvation barrier* with a width ($|r' - r''|$) the size of a water molecule (blue). This incorporates the entropic cost of expelling a solvent molecule between two residues (gold). The Lennard-Jones potential is plotted in light grey for comparison.

cluded volume fraction ϕ), compact desolvated (crystal) states are favored over less compact solvated (unfolded) states [52].

To investigate the opposing impact of macromolecular volume exclusion and solvation water on protein conformation, we utilized a minimalist protein model (see Appendix A and [135] in section S2.2) that incorporates the free energy cost of expelling a water molecule between a pair of residues in a contact termed the desolvation potential [Fig. 3.1(b)] [60]. This potential has a barrier that separates two minima that account for a native contact and a water-mediated contact. As pressure increases, the desolvation barrier increases and the free energy gap between the two minima tilts to favor the water-mediated contact, leading to an unfolding of a protein, capturing the main feature of pressure denaturation. Despite the model's simplicity without all the detailed chemistry in a residue [93], this desolvation model predicts a folding mechanism involving water expulsion from the hydrophobic core, which

has been observed by all-atomistic molecular dynamics [141] and validated by experiments in which the volume or polarity of amino acids is changed by mutation [142]. We previously employed a similar model without desolvation potential to investigate compact conformations of PGK induced by macromolecular crowding [52]. Now, by studying the competition of temperature, pressure and crowding on the energy landscape, we observe a costly barrier between two specific phases disappears, along a critical line on top of the isochore surface. As such, the current investigation demonstrations a richer ensemble of PGK states (Fig. 3.2) than our previous study [52].

To test our computational model, we observe structural transitions of PGK by fluorescence to construct the experimental T–P–ϕ phase diagram (Fig. 3.3). Experiment verifies the predicted existence of a critical point where T_c moves to a lower temperature T as the crowding volume fraction ϕ increases. Furthermore, we derive a critical line $T_c(\phi)$ using scaling arguments from polymer physics and present a unified phase diagram (Fig. 3.4) to investigate the underlying physical origin of such transition. As a consequence of being near the critical regime, PGK exhibits large structural fluctuations at physiologic conditions, which may be advantageous for enzymatic function. The current investigation is transforming the typical "structure-function" problem in proteins to a novel paradigm of a "structure-function-environment" relationship and is a step toward developing universal thermodynamic principles of protein folding in living cells.

3.1 Computational T–P–ϕ Phase Diagram of PGK

We investigated the conformations of PGK, a large, 415 amino acid, two-domain protein ([135] in section S1 for more information on PGK), in an environment containing Ficoll 70, which acts as a crowding agent mimicking cell-like excluded volume. Ficoll 70 is computa-

tionally modeled as a hard sphere, as it is known to be inert to proteins and behaves as a semirigid sphere [143, 144]. From prior FRET (Förster Resonance Energy Transfer) experiments and molecular simulations, we gained knowledge of several PGK conformations that denote a phase diagram in the ϕ–T plane [52]. It includes four states: C (crystal structure), CC (collapsed crystal), Sph (spherically compact state), and U (unfolded structures). In the C state, there is a linker that separates the N-terminal and C-terminal domains, resembling an open "Pacman". The CC state is a closed "Pacman". The Sph state involves a twisting of one of the domains with respect to the other and becomes more spherical than the CC state. A complete description of the structures of these states is in Supplemental Material [135] section S3.

By changing hydrostatic pressure P and volume fraction of crowders ϕ at several temperatures T, we have identified two new states on the ϕ–P isothermal phase plane (Fig. 3.2): I (folding intermediate), and SU (swollen compact unfolded structure). The criteria to define the six distinctive conformations are in Table S3.1 [135]. The I state is an ensemble of structures containing one folded domain (C-terminus) and one unfolded domain (N-terminus), making a specific prediction as to which domain is least stable on its own (N-terminal). SU is completely denatured but exhibits many water-mediated contacts [Fig. 3.2(b)]. Thus, the SU state is structurally more compact than the U state.

The microscopic mechanism of the pressure-induced unfolding of PGK depends on the T and ϕ. Fig. 3.2 shows the P–ϕ phase plane at low T in Fig. 3.2(a) and high T in Fig. 3.2(c). At sufficiently low T and $\phi = 0$ (no crowders) the unfolding of PGK is a multi-state transition between crystal state C [Fig. 3.2(a) & 3.2(b)] and unfolded state U via an intermediate state

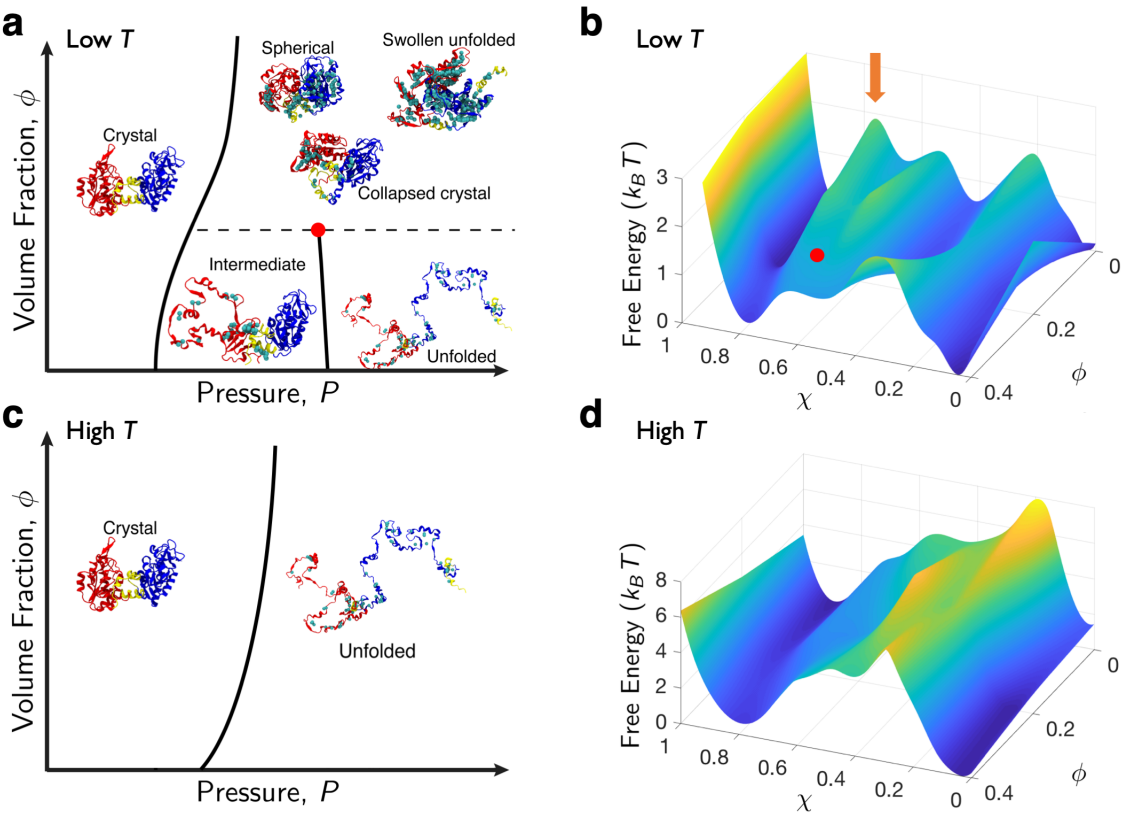

Figure 3.2 Solvation and crowding give rise to an intricate phase diagram of PGK. (a & c) Schematics of PGK's behavior in the crowding volume fraction-pressure (ϕ–P) phase plane and (b & d) corresponding free energy with respect to the overlap χ and crowding volume fraction ϕ at the folding pressure at low (a & b) and high (c & d) temperatures. Solid lines represent the division between distinct configurational phases that are separated by a free energy barrier from simulations at $\phi = 0$, 0.2, and 0.4, and pressures from $\sigma^3 P/\epsilon_0 = 10^3$ to 23. The dashed line (a) represents a continuous transition along ϕ and red dots (a & b) represent an approximate position of the critical points. The orange arrow (b) marks the peak of the barrier that diminishes until it disappears after the critical point. Collapsed crystal, spherical, and swollen unfolded states are indistinguishable in terms of free energy. These configurations were reconstructed from coarse-grained models to all-atomistic protein models for illustration purposes. N-, C-domain, and hinge are in red, blue, and yellow, respectively. A cyan sphere was inserted in between residues to show water-mediated contacts.

I. We capture the folding process using the overlap parameter χ,

$$\chi \equiv 1 - \frac{1}{N^2 - 5N + 6} \sum_{i=1}^{N-3} \sum_{j=i+3}^{N} \Theta\left(1.2 r_{ij}^0 - r_{ij}\right), \tag{3.1}$$

where N is the number of residues ($= 415$), Θ is the Heaviside step-function, r_{ij} is the distance between the residues i and j for a given conformation, and r_{ij}^0 is that corresponding distance in the crystal structure. It characterizes similarity to the crystal structure, C state. χ ranges from 0 to 1 where 0 represents the C state. In Fig. 3.2(b), the I state has $\langle\chi\rangle = \chi_I \approx 0.35$, and the U state has $\langle\chi\rangle = \chi_U \approx 0.9$. The state I is a consequence of the heterogeneous distribution of cavities, causing uneven pressure-denaturation where N-domain unfolds, and C-domain remains intact. Since the total cavity volume of the N-terminal domain ($\approx 171\text{Å}^3$) is about a third larger than that of the C-terminal domain ($\approx 132\text{Å}^3$), the former is more vulnerable to high pressure. Moreover, two antiparallel β-strands m and n of the N-terminal domain are totally exposed to the solvent ([135] in section S5 and Fig. S1.1). Under high pressure, they act as a channel for water to fill the N-terminal domain's cavities.

At sufficiently high ϕ and low T [Fig. 3.2(a), above the red critical point], there is only a single transition due to pressure between a crystal state and several compact states (Sph, CC, and SU) without the I state. The transition from C to Sph or CC states involves domain rearrangement when the linker "cracks" [136] and forms a disordered hinge. Whereas, high pressure competing with crowding gives rise to another compact unfolded conformations where up to half of the contacts becomes swollen with water that forms a "wet interface" (swollen unfolded states, SU). As the limited void formed by the density fluctuations of crowders inhibits extended conformations [145], the U state is unfavorable due to macromolecular crowding [51]. The protein only needs to subtly reduce its volume as it expels water molecules out of this wet core to return to the Sph or CC state from the SU state.

There are effectively no barriers between the Sph, CC, and SU states, which are thus located in the same region of the phase diagram (see Fig. 3.2(a) and 3.2(b) at $\chi = 0.4$ to 0.8, and S5.2 [135]). This data supports the hypothesis that protein dynamics is governed by the solvent motion [146], and water inside the protein "lubricates" the transitions between conformations without significant free energy costs [60].

Similarly, at high T [in Fig. 3.2(c) & 3.2(d)] ranging from low to high ϕ, PGK also undergoes a single pressure-denaturation transition, but it is between the C and U states. Due to the increase in T, the U state is entropically more favorable than all other states. As such, the U state's entropy considerably compensates the C state's energy, causing an increase in the free energy barrier between $\chi = 0$ (C state) and 0.8–0.9 (U state) in Fig. 3.2(d).

Our model predicts from these P–ϕ slices at various T that crowding makes the folding of PGK two-state, whereas lack of crowding produces a multi-state transition below a critical temperature T_c. Therefore, PGK undergoes a critical transition through by either of two directions on the T–P–ϕ phase space. One direction is by increasing ϕ at low T and sufficiently high P surpassing a critical volume fraction ϕ_c as shown in Fig. 3.2(a) at the red critical point. This transition is clearly seen by the diminishing of the free energy barrier in Fig. 3.2(b) pointed by an orange arrow, and

$$\lim_{\phi \to \phi_c^-} \chi_U(\phi) - \chi_I(\phi) = 0, \tag{3.2}$$

where ϕ_c is between 0.2 and 0.4. The second way is by increasing T at low ϕ and sufficiently high P surpassing a critical temperature T_c. Take $\phi = 0$ as an example; the free energy barriers in Fig. 3.2(b) pointed by an orange arrow must diminish, in order for the multi-state free energy to become two-state resembling the high T free energy shown in Fig. 3.2(d).

This also means,

$$\lim_{T \to T_c^-} \chi_U(T) - \chi_I(T) = 0. \tag{3.3}$$

Thus, these two directions that cause a critical transition suggest that T_c is a line on T–P–ϕ phase space and the value decreases as ϕ increases.

3.2 Experimental T–P–ϕ Phase Diagram of PGK

To validate the computed phase diagram, we measured the P–T phase diagram of PGK at various Ficoll 70 crowder concentrations to obtain the full P–T–ϕ information experimentally (Fig. 3.3). While one cannot expect the exact temperatures and pressures to agree, identical topologies of the experimental phase diagrams validate the general conclusions from simulations. Changes of the states of PGK were detected by tryptophan fluorescence because tryptophan mean fluorescence wavelength is sensitive to water exposure as the protein unfolds. We scanned T from 283 to 318 K at constant P, and P from 0 to 250 MPa at constant T with 0, 25, 50, 100, 150, and 200 mg/mL of Ficoll 70 concentration ($\phi = 0$ to ≈ 0.56) to cover the complete phase diagram. Each transition produces a sigmoidal step in the plot of mean tryptophan fluorescence wavelength λ_m vs. P (Fig. S2.1 [135]).

In the absence of crowder at sufficiently low T [Fig. 3.3(a), blue trace], there are two steps in λ_m as a function of P, signaling two separate transitions among three states. These steps are straightforwardly revealed by plotting $\partial\lambda_m/\partial P$ and identifying peaks (see Fig. 3.3(a), and Fig. S2.1 [135]). We assign the first peak to the C to I transition, the second to the I to U transition. At sufficiently high T, at ≥ 303 K and $P \approx 170$ MPa, one of the peaks disappears [Fig. 3.3(c), blue trace], leaving only one transition between two states. We assign this to a direct transition from C to U, as shown in Fig. 3.2(c), corresponding to a critical point at

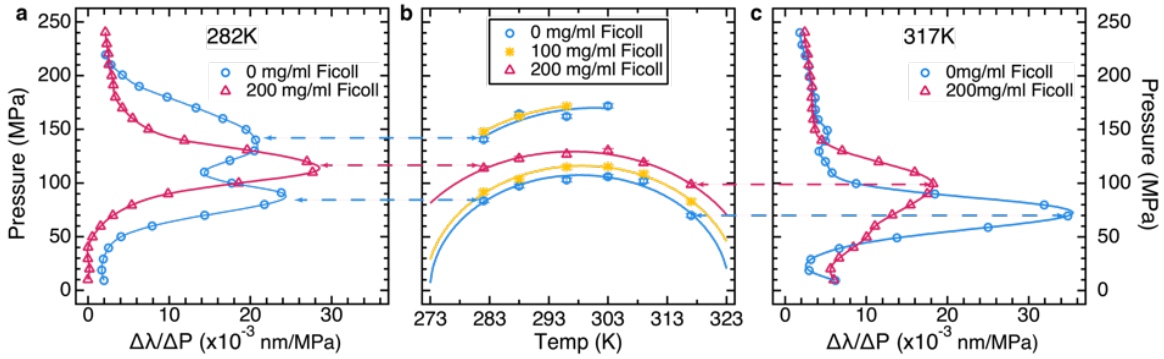

Figure 3.3 Experimental T–P–ϕ phase diagram of PGK. (a) The derivative of the mean tryptophan fluorescence wavelength vs. pressure of PGK at 282 K calculated from fluorescence spectra. Two of six Ficoll 70 concentrations are shown. The markers show the data points and the solid line shows a cubic spline interpolation. The blue curve (0 mg/ml Ficoll 70) has two peaks as pressure increases, signifying two transitions; the magenta curve (200 mg/ml Ficoll 70) has only one peak point, signifying only one transition when pressure is applied. The dashed lines point from transition midpoints to the corresponding point in the phase diagram. (b) P–T phase diagrams at several ϕ obtained by fitting the fluorescence data to obtain the inflection points of $\lambda_m(P)$ (peaks in the derivative $\partial\lambda_m/\partial P$). Three of six Ficoll 70 concentrations are shown. Circles represent midpoint pressures measured at 282, 288, 296, 303, 309 and 317K in absence of Ficoll 70 (0 mg/ml), asterisks represent transitions for the middle Ficoll 70 concentration (100 mg/ml) and triangles represent transitions for the highest Ficoll 70 concentration (200 mg/ml). At high T, or upon increasing Ficoll 70 concentration, the second (higher P) transition disappears, mapping out a critical point that moves to lower T_c at higher Ficoll 70 concentration. Solid elliptical curves going through the circles are fits to Eq. (3.4) representing the $\Delta G = 0$ curves. (c) Equivalent data as in (a) at 317 K. Note that the second (higher P) transition is never present at high T.

$T_c = 306 \pm 3$ K. Finally, when crowder is added, T_c moves to lower temperature, until the apparent three-state transition is no longer observed at all at 200 mg/ml Ficoll 70 [Fig. 3.3(a) & 3.3(c), red traces]. We assign this to the transition between C and SU/Sph/CC as shown in Fig. 3.2(a). Accurate transition midpoints (T_m, P_m, ϕ_m) were obtained from each trace by fitting to sigmoidal two- or three-state models (solid curves in Fig. 3.3(a) & 3.3(c); see Appendix B; all data traces are shown [135] in section S4). Singular value decomposition analysis ([135] in section S4) also strongly supports the conclusions obtained from analyzing λ_m.

We constructed P–T planes of the phase diagram at all crowder concentrations as follows: First, the transition midpoints were plotted on P–T slices at constant ϕ as shown in Fig. 3.3(b). These points correspond to zero free energy difference, $\Delta G = 0$, for the first-order transition, where concentrations of C and I, I and U, or C and U (depending on the location on the phase diagram) are equal. Then the transitions were fitted to Hawley's elliptical P–T phase curve for proteins [78],

$$\Delta G(T,P) = \frac{1}{2}\Delta\kappa\,(P-P_0)^2 + \Delta\alpha\,(T-T_0)\,(P-P_0) - \Delta C_P\left[T\left(\ln\frac{T}{T_0}-1\right)+T_0\right] + \Delta V_0\,(P-P_0) - \Delta S_0\,(T-T_0) + \Delta G_0, \quad (3.4)$$

at each value of ϕ (fits for all ϕ and parameter definitions in [135] section S4). Here $\Delta\kappa$, $\Delta\alpha$, ΔC_P, ΔV_0, and ΔS_0, signify changes in compressibility, thermal expansion coefficient, heat capacity, volume, and entropy, respectively. The resulting experimental phase diagram in Fig. 3.3(b) agrees with the computational data as both exhibit two pressure transitions at low T, one at high T, and a shift from two to three transitions at a value of T_c that decreases with increased crowding.

The simulation predicts that in the state I, the N-terminus would be unfolded, and the C-terminus folded. We truncated the protein to the N-terminal domain and indeed found it to be unfolded with long tryptophan fluorescence wavelength and no cooperative transition (Fig. S4.4 [135]). It is known from the literature [105] that the C-terminal domain of PGK is stable by itself. These two observations combined strongly support the computational assignment of the I state with the N-terminal domain primarily unfolded, and the C-terminal domain mostly folded. Thus, experiment and simulations are in agreement both on the disappearance of the difference between two phases at high T or high ϕ, as well as the general structural features of the I state formed at low crowding.

3.3 Unified T–P–ϕ Phase Diagram of PGK

The three-dimensional (3D) T–P–ϕ phase diagram in Fig. 3.4(c) presents a unified picture of the computational and experimental results. This unified phase diagram includes two surfaces: the blue surface represents C–I (or C–U, depending on T and ϕ) and the red surface represents I–U coexistence surfaces, respectively (the calculations of the surfaces can be found in Appendix 3.4 and [135] section S6). The projection of this 3D coexistence surface onto a 2D ϕ–P plane in Fig. 3.4(a) shows a low and high T slice as found computationally in Fig. 3.2. When projected on P–T plane in Fig. 3.4(b), it shows a low ϕ and high ϕ slice as found experimentally in Fig. 3.3. As temperature increases, the second transition surface terminates at a critical line [bold red line on the red I–U coexistence surface in Fig. 3.4(c)]. As the crowding volume fraction increases, the critical point on each P–T slice shifts towards lower temperatures. Thus, from the experiment, above $\phi = \phi_c \approx 0.5$ or $T = T_c^0 \approx 306$ K the pressure-induced folding transition contains only two apparent phases. Whereas at low ϕ and T, PGK exhibits apparent three-state folding.

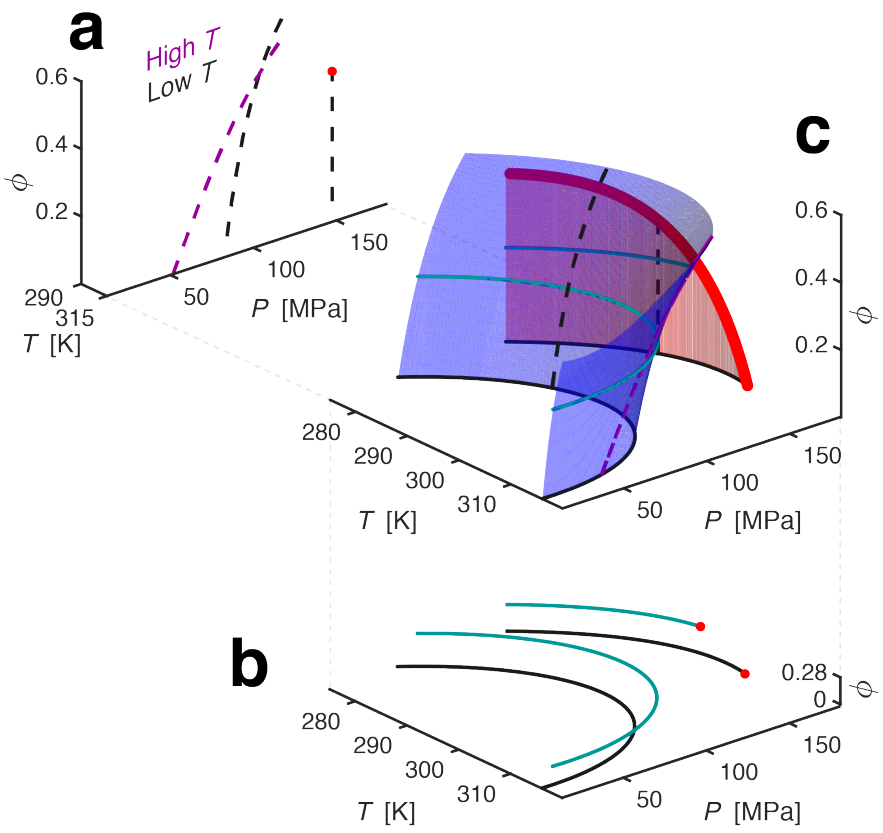

Figure 3.4 T–P–ϕ phase diagram of PGK from theory mapped onto the experimental data. (a) PGK's ϕ–P phase plane at high (magenta) and low (black) T. Dotted lines represent the division between distinct configurational phases. The red dot signifies a critical point. (b) Slices of P–T phase diagram observed experimentally at no Ficoll 70 (black) and 100 mg/mL Ficoll 70 (cyan). Note that I-U coexistence curve terminates at the critical point (in red dots) and shifts T_c to a lower temperature in the presence of Ficoll 70. Solid elliptical curves going through the circles are the fits representing the $\Delta G = 0$ curves. (c) A T–P–ϕ phase diagram of PGK. The blue and red surfaces are the C–I (or C–U, depending on T and ϕ) and I–U coexistence surfaces, respectively. The dashed magenta and black line are the ϕ–P cross-section from Fig. 4(a), and the solid black and cyan are the P–T cross-section from Fig. 4(b). The bold, red line bordering the red surface is the critical line.

To quantitate the unified phase diagram, we modified Hawley's theory by incorporating the free energy change due to crowding, as is similarly treated in Minton's theory [138], to construct the first transition surface [blue surface in Fig. 3.4(c)]. As for the critical line on the second transition surface [in red in Fig. 3.4(c)], we used scaling arguments to derive the equation for the critical line,

$$P - P_c = a_1 \left(T - T_c(\phi)\right) + a_2 \left(T - T_c(\phi)\right)^2 + \mathcal{O}\left(\Delta T^3\right), \tag{3.5}$$

by treating the protein's U to I transition similar to in the coil-globule transition of theory [147, 21]. Here,

$$T_c(\phi) = T_c^0 \left(1 - \frac{\phi}{\phi_c}\right)^{\gamma}, \tag{3.6}$$

where T_c^0 is the critical temperature without crowding, ϕ_c is the critical crowding volume fraction, P_c ($\approx$ 170 MPa) is the critical pressure taken from our experiment at T_c^0, and $a_1 = \frac{dP}{dT}\Big|_{T=T_c^0}$ and $a_2 = \frac{d^2P}{dT^2}\Big|_{T=T_c^0}$. From the fitting to experimental critical points at all slices of ϕ, we found $\gamma = 0.40 \pm 0.01$, which is the predicted scaling exponent of a polymer collapse due to crowders, $\gamma = 2/5$ [148, 53] (see Appendix 3.4 and [135] section S6). From this phase diagram, we can see the protein moves through a diverse phase space, suggesting different folding mechanisms that depend on how the phase diagram is traced out [149, 150].

3.4 Construction of the Phase Diagram

We derived the critical line [Eq. (3.5) & (3.6); red line in Fig. 3.4(c)] on the T-P-ϕ phase diagram using arguments based on the coil-globule transition [147, 21] of a polymer. Beginning with a Landau-Ginsberg free energy [151], $F = -r(T, \phi)\Psi^2 + u\Psi^4 + F_0$, to describe the critical transition, where Ψ is the order parameter, which is a scaled and shifted R_g (radius

of gyration) so that $\Psi = -\Psi_0$ for the I state and $\Psi = +\Psi_0$for the U state. Since pressure is only involved with the first-order transitions, it can be ignored for now. At the critical temperature, the barrier between the I and U states vanishes, meaning $r = 0$; therefore, a reasonable function is $r(T, \phi) = -r_0[T - T_c(\phi)]$, where the critical temperature T_c is a function of ϕ, and r_0 is positive constant. To find the ϕ-dependence of T_c, we used the scaling relationship,

$$\frac{R_g(\phi)^2}{R_g(0)^2} \sim (1 - c_0\phi)^\gamma \,, \tag{3.7}$$

which relates R_g at a given ϕ to R_g without crowders for the collapse of a coil to globule transition [148]. The scaling exponent γ, is shown to be 2/5 in Refs. [148] and [53]. Since the collapse of the polymer, or in the current case the protein, is dependent on ϕ, and since $\Psi^2 \sim R_g(\phi)^2/R_g(0)^2$, the critical temperature $T_c(\phi)$ causing the free energy barrier between I and U to disappear must also scale as Eq. (3.7), giving Eq. (3.6) (see [135] section S6 for more details). We fit Eq. (3.6) to the experimental critical point values at all Ficoll 70 concentrations to find γ and ϕ_c (or $1/c_0$). We fit Eq. (3.5) to experimental values of the I to U transition surface to find the Taylor expansion coefficients.

Lastly, we modified Hawley's equation [78] to fit the C to I (or U, depending on T and ϕ) transition surface (in blue in Fig. 3.4(c)) by adding a ϕ-dependent $\Delta G_{crowd}(\phi)$ term to Eq. (3.4),

$$\Delta G_{crowd}(\phi) = g\left(\frac{\phi}{1-\phi}\right) + \mathcal{O}\left(\phi^2\right), \tag{3.8}$$

making the 3D free energy change $\Delta G(T, P, \phi) = \Delta G(T, P) + \Delta G_{crowd}(\phi)$. This term is similar to Minton's theory [138], which treats the folded and unfolded proteins as effective hard spheres and employs scaled particle theory (SPT) to estimate the change in folding free energy as the difference between the insertion free energy for the folded and the unfolded states. Eq. (3.8) adds one more fitting parameter, g, to the total free energy change compared

to Eq. (3.4).

This appendix will explain the derivation of the critical line (red curve in Fig. 3.4). First, we will briefly go through the polymer model used in Ref. [148], and then calculate the crowding-dependent mean-square end-to-end distance $\langle R_{ee}^2 \rangle$. Lastly, we derive the crowding-dependent critical temperature $T_c(\phi)$ using the calculations from the previous sections.

Following Ref. [148], the Hamiltonian for the isolated polymer in a crowded solution is formulated to be,

$$\begin{aligned}\mathcal{H}[r(s)] = \frac{3}{2l}\int_0^L \left(\frac{dr}{ds}\right)^2 ds + \omega \int_0^L ds \int_0^L ds' \delta\left(r(s) - r\left(s'\right)\right) \\ + \sum_{i=1}^{N} \int_0^L v[r(s) - R_i] ds.\end{aligned} \tag{3.9}$$

The model uses a continuous curve $r(s)$, parametrized by the variable s, to describe the conformation of the polymer of length L. The strength of the excluded-volume interaction is controlled by the parameter ω. The last term is the crowder potential and is given by,

$$v\left(r - R_i\right) = \beta v_0 \delta\left(r - R_i\right)/l \tag{3.10}$$

and is divided by the Kuhn length l.

For this model without crowders, which is the Edwards polymer model [152], the mean-squared end-to-end distance $\langle R_{ee}^2 \rangle$ scales by [147],

$$\left\langle R_{ee}^2 \right\rangle \sim L^{2\nu}, \tag{3.11}$$

where the exponent $\nu = 3/5$ for $\omega \neq 0$. Whereas, $\langle R_{ee}^2 \rangle$ with crowders is given by,

$$\left\langle R_{ee}^2 \right\rangle = \frac{1}{Z(\omega, \phi, L)} \int \mathcal{D}r(s) |r(L) - r(0)|^2 e^{-\mathcal{S}[r(s)]} \tag{3.12}$$

$Z(\omega, \phi, L)$ being the partition function and an effective action $\mathcal{S}[r(s)]$ for the polymer Hamiltonian in a crowded solution Eq. (1), with volume fraction ϕ.

In order to evaluate the above path integral in Eq. (4), Ref. [148] employs the self-consistent variational approach introduced by Edwards and Singh [153]. The key reasoning behind this approach comes from choosing an effective reference action with an appropriately renormalized step length l_1, such that $\langle R_{ee}^2 \rangle \equiv L l_1$. This ensures that all correction terms to the relation be zero, by definition. The evaluation of the path integral in Eq. (4) results in a self-consistent equation for l_1 as a function of l, ω, ϕ, and L (Appendix of Ref. [148]):

$$L l_1^2 \left(\frac{1}{l} - \frac{1}{l_1} \right) = 2 c_1 (1 - c_0 \phi) \frac{L^{3/2}}{l_1^{1/2}}, \tag{3.13}$$

where $c_1 = 2\omega\sqrt{6/\pi^3}$ and $c_0 = \frac{1}{\omega} \left(\frac{\beta v_0}{l} \right)^2$. Therefore,

$$l_1^{5/2} \left(\frac{1}{l} - \frac{1}{l_1} \right) = 2 c_1 (1 - c_0 \phi) L^{1/2}, \tag{3.14}$$

resulting in the scaling relation,

$$l_1 \sim (1 - c_0 \phi)^{2/5} L^{1/5}. \tag{3.15}$$

When substituting Eq. (7) into $\langle R_{ee}^2 \rangle = L l_1$, it becomes a ϕ-dependent version of the

well-known Flory scaling relation,

$$\langle R_{ee}^2 \rangle \sim (1 - c_0\phi)^{2/5} L^{6/5} \tag{3.16}$$

In the case without crowders, $\langle R_{ee}^2 \rangle \sim L^{6/5}$; therefore, the ratio between with and without crowders becomes,

$$\frac{\langle R_{ee}^2 \rangle(\phi)}{\langle R_{ee}^2 \rangle(0)} \sim (1 - c_0\phi)^{2/5}. \tag{3.17}$$

The critical volume fraction is then [148],

$$\phi_c = \omega \left(\frac{l}{\beta v_0} \right)^2 = \frac{1}{c_0}. \tag{3.18}$$

Since $\langle R_{ee}^2 \rangle \sim \langle R_g^2 \rangle$, the same relation in Eq. (9) also holds for the ratio $\langle R_g^2 \rangle$.

We derived the ϕ-dependent critical temperature, $T_c(\phi)$ on the T-P-ϕ phase diagram using a simple statistical mechanical model. To begin, the Landau-Ginsberg free energy,

$$F(\Psi, T, \phi) = -r(T, \phi)\Psi^2 + u\Psi^4 + F_0 \tag{3.19}$$

is used to describe the critical transition, where Ψ is the order parameter, which is scaled and shifted R_g so that $\Psi = -\Psi_0$ for the I phase and $\Psi = +\Psi_0$ for the U phase. Since we are only interested in the critical transition, we can ignore the odd powered terms. To find the free energy minima, we take the derivative with respect to Ψ, and solve for the zeros of the equation:

$$\frac{\partial F}{\partial \Psi} = -2r\Psi + 4u\Psi^3 = 0 \tag{3.20}$$

$$\Psi = \begin{cases} \pm\sqrt{\frac{r}{2u}} \\ 0 \end{cases} \tag{3.21}$$

At the critical temperature T_c or at critical crowding volume fraction ϕ_c, the two phases merge together (I and U) at the free energy minimum $\Psi = 0$, meaning $r = 0$. Furthermore, since $\Psi^2 \sim \langle R_g^2 \rangle$, then

$$\Psi^2 = \frac{r}{2u} \sim \left(1 - \frac{\phi}{\phi_c}\right)^{\gamma} \tag{3.22}$$

To find a reasonable function for $r(T, \phi)$, it must satisfy Eq. (14) and the following limits:

$$\lim_{T \to T_c^-} \Psi = 0 \tag{3.23}$$

$$\lim_{\phi \to \phi_c^-} \Psi = 0 \tag{3.24}$$

$$\lim_{\phi \to 0} \Psi = \sqrt{\frac{r}{2u}} \tag{3.25}$$

Therefore, a reasonable function is

$$r(T, \phi) = -r_0 \left[T - T_c^0 \left(1 - \frac{\phi}{\phi_c}\right)^{\gamma} \right], \tag{3.26}$$

where T_c^0 is the critical temperature at $\phi = 0$, and r_0 is a positive constant. Examples of the F, from Eq. (11), using Eq. (19) are plotted in Fig. 2. When, $r = 0$, $T = T_c^0 \left(1 - \frac{\phi}{\phi_c}\right)^{\gamma}$; thus, we can define a ϕ-dependent T_c as,

$$T_c(\phi) = T_c^0 \left(1 - \frac{\phi}{\phi_c}\right)^{\gamma}, \tag{3.27}$$

recovering a critical temperature that decreases in value as ϕ increases.

To find γ, we linearly fit Eq. (20) on a log-log scale of the experimental $T_c(\phi)$ values. The

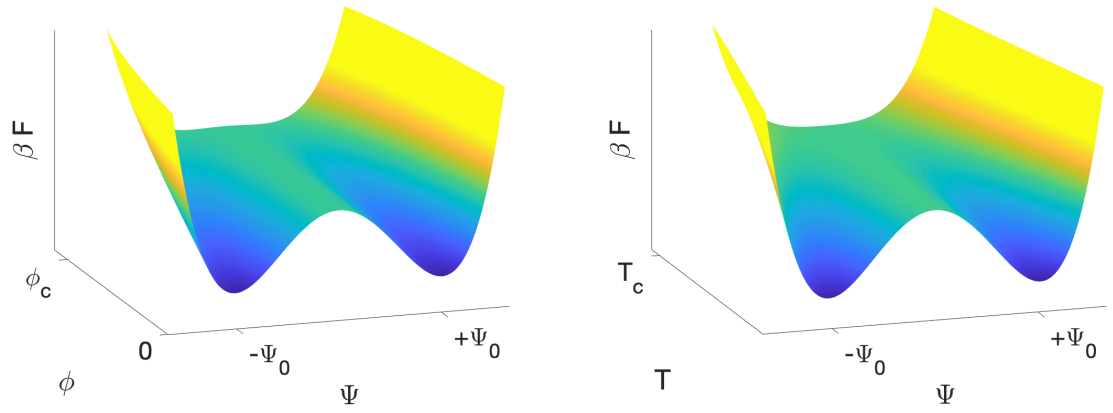

Figure 3.5 Example Landau-Ginsberg free energies with respect to order parameter Ψ, and (a)T with constant $\phi < \phi_c$ or (b) ϕ with constant $T < T_c^0$. The form of $r(T, \phi)$ from Eq. (11) is given by Eq. (19) with $\gamma = 2/5$. Critical transitions occur when surpassing their respected critical points, signifying the disappearance of the barrier between the two phases (I and U), and two phases become thermodynamically indistinguishable.

linear fit is best when $\phi_c = 0.5$, resulting in $\gamma = 0.40 \pm 0.01$, which is a perfect match with the exponent from Eq. (9).

To find the critical line on the T-P-ϕ phase diagram, we expand $P(T)$ around T_c to attain the second-order line (red curve in Fig. 1),

$$P - P_c = a_1 \left(T - T_c(\phi)\right) + a_2 \left(T - T_c(\phi)\right)^2 + \mathcal{O}\left(\Delta T^3\right), \tag{3.28}$$

where P_c is the critical pressure taken from experiments at T_c^0, and $a_1 = \frac{dP}{dT}\Big|_{T=T_c^0}$ and $a_2 = \frac{d^2P}{dT^2}\Big|_{T=T_c^0}$. Both a_1 and a_2 are found by fitting experimental values.

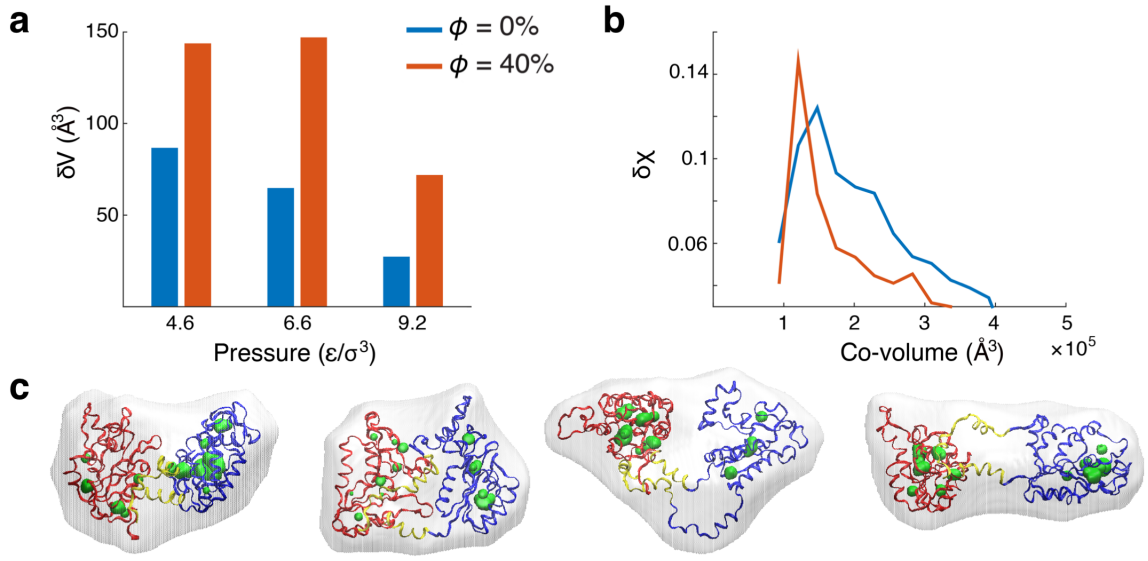

Figure 3.6 Cavity volume and structural fluctuations near the critical regime. (a) Cavity volume fluctuations, $\delta V^2 = \langle V^2 \rangle - \langle V \rangle^2$, (or proportionally compressibility) of PGK at $k_B T/\epsilon_0 = 0.97$ and $\sigma^3 P/\epsilon_0 = 4.6$, 6.6, and 9.2 with (orange) and without (blue) crowding. (b) Overlap fluctuations, $\delta\chi^2 = \langle \chi^2 \rangle - \langle \chi \rangle^2$, as a function of co-volume at pressure $\sigma^3 P/\epsilon_0 = 6.6$. (c) Conformations from the ensemble in presence of crowding at $k_B T/\epsilon_0 = 0.97$ and $\sigma^3 P/\epsilon_0 = 6.6$ with co-volumes $\approx 1.1 \times 10^5 \text{Å}^3$ and cavities $\approx 200 \text{Å}^3$. Most left conformation is a crystal state. The following structures from left to right have a $\chi = 0.31$, 0.38 and 0.48. N-, C-domain, and hinge are in red, blue, and yellow, respectively. Co-volumes are shown as translucent surfaces surrounding the protein and cavity surfaces are shown in green.

3.5 Consequences of Criticality

In Fig. 3.6, we explore the impact of P and ϕ on the folding of PGK. The consequences of the critical regime are revealed by the ensemble distributions of the cavity volume (conjugate variable of P) and co-volume (conjugate variable of ϕ) [47, 138] from our simulations (also see Fig. S5.1 [135]). In the critical regime, small perturbations in crowding ϕ, P, or T will significantly affect the system.

We investigate the response of the conformational distribution of structures close to the critical region by comparing the cavity volume fluctuations ($\delta V^2 = \langle V^2 \rangle - \langle V \rangle^2$) (or proportionally, the compressibility) and structural fluctuations ($\delta \chi^2 = \langle \chi^2 \rangle - \langle \chi \rangle^2$) in the presence and absence of crowding agent. PGK has larger δV [Fig. 3.6(a)] at $\phi = 0.4$ with a peak at 6.6 ϵ_0/σ^3 than that of $\phi = 0$. We suspected that the critical regime is between $\phi = 0.2$ and 0.4 and between pressures 4.6 ϵ_0/σ^3 and 6.6 ϵ_0/σ^3 at a temperature of 0.97 ϵ_0/k_BT in the computational model, which qualitatively agrees with the experiment. Even though $\delta\chi$ is large in the presence of crowders, structures lie in a narrow range of co-volumes, making them indistinguishable to macromolecular crowding effects if shape can be neglected to the 0^{th} order [Fig. 3.6(b)]. A sample of the diverse structures with similar cavity volumes and co-volumes are shown in Fig. 3.6(c).

Not only does crowding shift the population of structures to more compact states such as CC or Sph (Fig. 3.2 and [52]), where the two ligand binding sites (for ADP and 1,3-DPG) come into close proximity of each other, but it also increases the structural fluctuations of the compact states by bringing PGK closer to the critical regime, as shown in Fig. 3.6. Both of these properties would most likely facilitate enzymatic activity. This is corroborated by previous FRET experiments that show an increase in PGK's enzymatic activity as Ficoll

70 concentration increases [52]. These results suggest that criticality assists the enzymatic function of a protein.

3.6 Conclusion

In summary, we have shown direct evidence of equilibrium critical-like behavior on the T–P–ϕ phase diagram of a protein by computational simulations, by fluorescence spectroscopy, and by a theoretical argument based on polymer physics. Despite the simplicity of the computational and theoretical model, all three different approaches agree with one another, validating the trends on the T–P–ϕ phase diagram and presence of the critical regime. Above the critical line in Fig. 3.4(c) (by increasing T, ϕ, or both at $P_c \approx 170$ MPa), the difference between the I and U phases disappears. This is due to the loss of the free energy barrier between the two phases [orange arrow in Fig. 3.2(b)] and is reaffirmed by the high-pressure fluorescence measurements (Fig. 3.3).

What is the origin of the critical behavior in proteins? To answer this question, two concepts need to be rationalized together. Firstly, proteins are biopolymers that often undergo an abrupt or first-order-like transition to a compacted folded state from an expanded unfolded state or coil at a folding temperature, T_F. Secondly, the coil-globule transition seen in other polymers is a continuous transition at a specific temperature called θ-temperature, T_θ [147, 21]. Therefore, the first order transition in protein folding must be occurring near the collapse transition ($T_F \approx T_\theta$), meaning it normally is *tricritical* [132]. In the current system, the pressure perturbation may cause $T_F \neq T_\theta$, separating the continuous and first-order transitions. When going from a continuous to a first-order transition, there are signatures of passing through a critical point [154, 29]. Finally, when ϕ is high ($\phi > \phi_c$), the protein is already collapsed even when it is unfolded. Our theoretical model in Eq. (3.6) and

Appendix 3.4 (also [135] section S6) captures this postulation of the basis of criticality in proteins.

Furthermore, our computational and experimental results are in accord with the capillarity picture of folding [155], which posits a wetting interface between folded and unfolded parts of a protein, giving rise to a diverse phase space. Strong macromolecular crowding, which drives conformational changes to favor compact states, roughens that wetting interface, allowing cavities to spread throughout the conformation of the protein, with two major consequences. A roughened interface reduces activation barriers for folding, driving multi-state transitions towards apparent two-state transitions. It also creates a critical state where heterogeneous conformations coexist, as the front of wetting interface moves across the protein.

We conclude that large structural fluctuations (Fig. 3.6) and merging of protein phases are consequences of being close to a critical point [Fig. 3.3(c)]. At such a point, the barrier separating states vanishes (here: between I and U). Critical behavior has been proposed for protein folding at the onset of downhill folding [133], but its manifestation has been challenging to demonstrate computationally and experimentally [156]. Macromolecular crowding shifts the critical point to a lower temperature [Eq. (3.6)], indicating that such criticality could be physiologically important [129, 130]: a protein near a critical regime could access a wide range of conformations without significant activation barriers for functional purposes inside the cell.

Further work will be needed to provide stronger evidence for the universality of critical behavior in proteins. Due to their complexity, proteins are not like other conventional condensed matter systems, and conventional tools, such as finite size scaling [157] or renormalization group theory [158], are not clearly applicable. The current investigation is a starting point toward developing universal principles of protein folding relevant to the en-

vironmental perturbations inside living cells and is an inspiration to create new tools to understand critical phenomena in these complex systems.

It turns out that an eerie type of chaos can lurk just behind a facade of order—and yet, deep inside the chaos lurks an even eerier type of order.

—Douglas R. Hofstadter

4

Competition of individual Protein folding with Inter-protein Interactions*

An important question in protein dynamics is how proteins manage to fold in the presence of many other biomolecules with which they could interact instead inside the cell [159]. For example, it has been discussed extensively how proteins can transiently aggregate during their folding process, thus mimicking the existence of monomeric folding intermediates [160].

Repeat proteins are particularly interesting subjects for studying the interplay between folding and aggregation [161]. The proximity of tethered domains with identical or near-identical folds enhances protein–protein interactions [162, 163]. It was shown by Borgia et al. for immunoglobulin-like oligomeric repeats that identical neighbors transiently misfold more readily than neighbors of lower sequence identity [164, 165]. Thus evolutionary pressure reduces sequence similarity between adjacent repeat domains, and many natural repeat

*Contents of this chapter has been published in *Phys. Chem. Chem. Phys.* (2019) **21**, 24393-24405. AG Gasic is co-first author.

proteins contain folds that do not interact too strongly. Such sequences can also be engineered: the energy landscape of some ankyrin repeat proteins, especially of consensus [166] sequences, enables parallel folding of the domains [167, 168].

Here we study the competition between folding and transient (or permanent) aggregation of engineered WW domain oligomers, with $n = 1$ to 5 domains tethered by short glycine-serine sequences. We chose "GS linkers" because they have been well characterized, are highly soluble, and allow monomers to interact [169]. We use coarse-grained simulated annealing simulations to obtain structural information about the misfolded oligomers. A variety of interesting structures emerges, from individual misfolded domains, to chimeric misfolds (where two proteins intermingle), to entirely new beta-sheet structures, and finally even to alpha-helical structures that bear no resemblance to the original domains making up the oligomer.

4.1 Decrease of thermal stability from monomer to tetramer

we used AWSEM to perform simulated annealing starting with unfolded structures at 650 K and gradually reduced the temperature to 300 K to sample increasingly folded structures. We first looked at the WW monomer. Fig.4.1 compares the simulated annealing trajectories for the three Models with $\lambda_{\mathrm{FM}} = 0.4\,\mathrm{kJ\,mol^{-1}}$(Model I), $0.8\,\mathrm{kJ\,mol^{-1}}$ (Model II) and $1.2\,\mathrm{kJ\,mol^{-1}}$ (Model III). As discussed in Methods, λ_{FM} defines the strength of the fragment memory Hamiltonian, with large values favoring folding over domain interactions. As λ_{FM} increases, the WW-domain collapses and folds earlier and at higher temperature. At $Q \approx 0.35$, only a single β-hairpin is formed. At $Q \approx 0.65$, all three β-strands form, but sidechains are not quite natively packed yet. At $Q \approx 0.95$, the protein is folded. Model I does not fold sufficiently well to be consistent with experiment, whereas Model II and III completely fold,

consistent with fully native structure of the monomer. We favor Model II because it has less weighting on the fragment memory interactions (smaller value of λ_{FM}) than Model III, thus achieving complete folding of the monomer without overweighting native interactions.

Simulations were next performed on tethered repeat proteins to get a higher resolution picture of the type of misfolded structures that may form. In agreement with experiment, simulation revealed that as the number of domains increases, the probability of misfolding increases. A gallery of monomers and oligomers with varying degrees of folding/misfolding from simulated annealing is shown in Fig. 4.2, and annealing results are shown in ESI Section S8. As the number of domains increases (MFiP35 to QFiP35), the folding of the individual domains competed less effectively with inter-domain interactions. A common feature of the misfolded structures is that one β-strand unfolds, and the remaining two β-strands from separate domains come together to make a larger β-sheet [e.g. structures Fig. 4.2(E & H)]. This misfolding mechanism is clearly seen in the annealing trajectory of a trimer (Fig. S11–S13, ESI). The tetramer exhibited an additional type of misfolding [structure Fig. 4.2(K)] by forming chimeric β-sheets with domain-swapped structures. Thermodynamically, the non-fragment memory terms of the Hamiltonian in eqn (4), primarily the Ramachandran term, $\mathcal{H}_{\mathrm{RAMA}}$ (see eqn (S5), ESI), the β-strand hydrogen bonding term, $\mathcal{H}_{\beta}$ (see eqn (S6), ESI), and parallel-antiparallel cooperative hydrogen bonding term, $\mathcal{H}_{\mathrm{P-AP}}$ (see eqn (S7), ESI), are responsible the formation of the β-sheets across multiple domains.

4.2 Misfolding propensity increases with oligomer size

The experimental expression yield trends are supported by coarse-grained simulations of the tethered systems. Fig. 4.3 shows the probabilities for m or more domains in the n-mer to misfold, or $P_{\mathrm{misfold}}(m \geq \mu|n)$, for models I, II, and III. The scaling factor λ_{FM} controls the

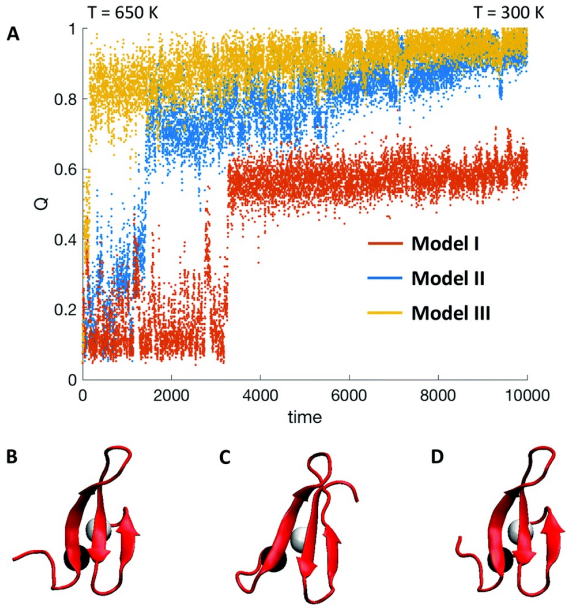

Figure 4.1 (A) Simulated annealing trajectories with respect to fraction of native contacts Q for WW-domain monomer for model I, II, and III. Trajectories start at 650 K and are gradually cooled to 300 K. Time is represented in units of 103 timesteps. Below the trajectory are the annealed monomer structures, from left to right: Q = 0.72 for model I (B), Q = 0.95 for model II (C), and Q = 1.0 for model III (D). The black and white beads are the C_β atoms of Trp 8 and Tyr 20, respectively.

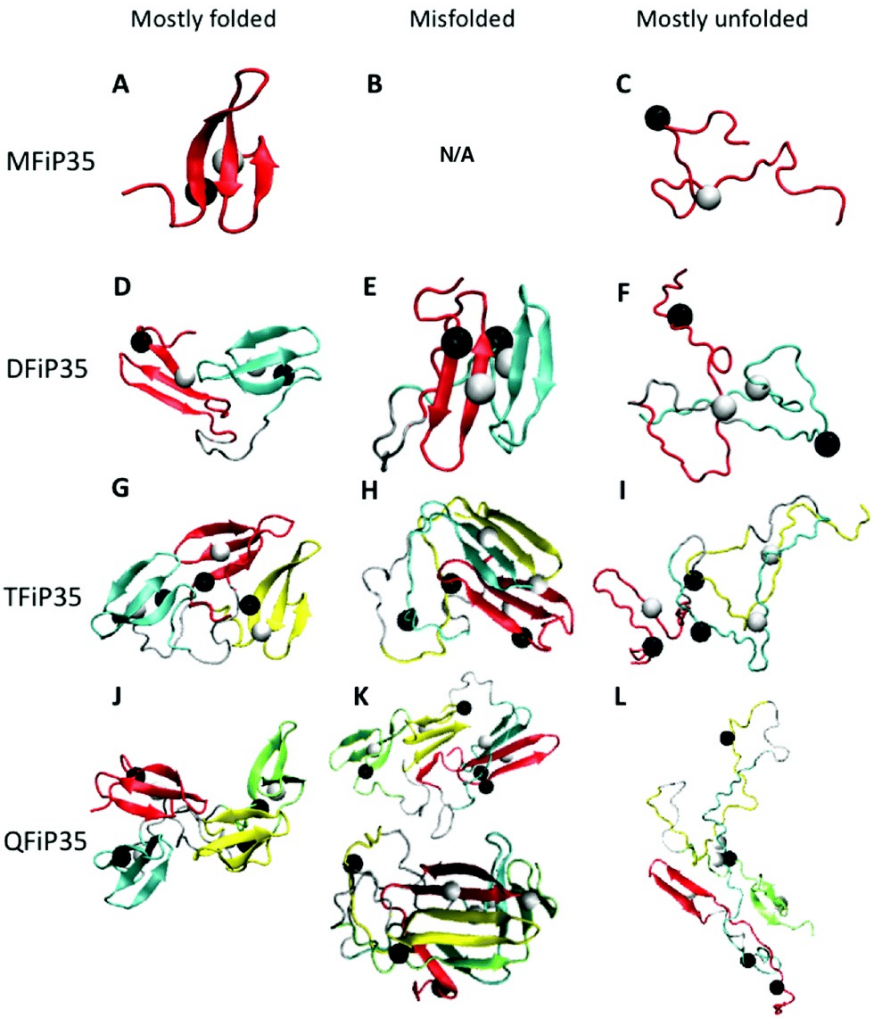

Figure 4.2 Gallery of oligomers. Examples of predicted WW-domain monomer,dimer, trimer, and tetramer structures (from top to bottom) with varying amounts of folding/misfolding of domains. The domains are colored red-cyan-yellow-green from the N-terminus, and linkers are gray. The black and white beads are the C_β atoms of Trp 8 (or 51, 94, 137) and Tyr 20 (or 63, 109,152), respectively. Oligomers states corresponding roughly to the discrete global fitting model for the experimental data: (A) N, (B) no long-lived misfolded state observed, (C) U, (D) NN, (E) MM, (F) UU, (G) NNN,(H) NMM, (I) UUU, (J) NNNN, (K) NNMM and MMMM, (L) UUUU.

bias towards the monomer native structure, with smaller values leading to more interaction among domains. For smaller λ_{FM} (model I), $P_{\mathrm{misfold}}(m \geq \mu|n)$ is driven towards 1 for smaller repeat proteins. The probability of misfolding of at least one domain is > 0.5 for the tetramer in model II, which fits well with the observed intracellular environmental sensitivity of the tetramer as seen by decrease in yield and sensitivity to the type of purification tag being attached. Model II shows no significant effect on monomer and dimer, consistent with the onset of lower melting temperature for the trimer in the thermal melts performed on the tethered proteins (Table 1). Even in model III, which has the strongest domain folding propensity, the tetramer has at least one domain misfolded with a probability of 0.3.

With increasing number of domains, the probability of misfolding increases due to increased competition of interdomain interactions with folding, shifting equilibrium towards misfolded states. Another possible reason is that because not all the β-strands form at the same time, misfolding can occur when β-sheets of neighboring domains interact and become kinetically trapped beyond the time scale of the experiments. The gallery of n-mers in Fig. 4.2 is consistent with both scenarios, although we favor the equilibrium scenario for two reasons: in the model, extensive simulated annealing was applied; and in the global fitting model (Fig. 6), equilibrium is achieved while accurately fitting the experimental data. Furthermore, the strong coupling between domains reflected by P_{misfold} increasing with n (Fig. 4.3) and the mis/partly folded structures in Fig. 4.2 validate the global fitting model assumptions (Section 2.5) and fitting results (Section 3.5). It is reasonable for state M to represent non-native structure of a domain due to domain interactions, rather than an isolated misfolded state of WW domain.

Fig. 4.4 presents the averages of three order parameters with respect to number of domains for models I, II, III. The number of contacts made by tryptophan [Fig. 4.4(C)] can be correlated with the fluorescence spectroscopy, since fewer contacts imply more solvent expo-

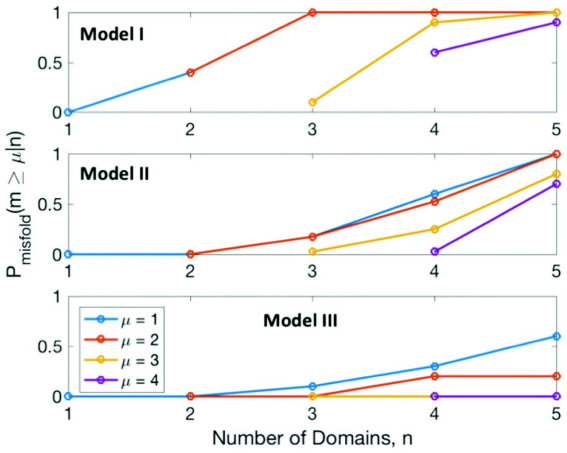

Figure 4.3 Probability $P_{\text{misfold}}(m \geq \mu|n)$ of misfolding m or more domains,given the size of the n-mer from $n = 1$ to 5. Models I, II, and III are shown from top to bottom. Probabilities are calculated from structure predictions of simulated annealing runs.

sure and red-shifted fluorescence. The number of Trp contacts is highest for the monomer signifying a stable native structure. This is also verified with the change in energy per domain [Fig. 4.4(A)], which shows the monomer as the most stable compared to the other oligomers. Fig. 4.4(B) shows that more C_β contacts form as the number of domains increases, signifying an increase in hydrogen bonding between b stands of different domains. This increase in Cb contacts can be visualized as an increase in chimeric β-sheets seen in Fig. 4.2(E, H & K). This analysis is consistent with the experimental results obtained by CD and fluorescence spectroscopy.

The CD spectra in Fig. 2 vary in intensity, but generally have the same shape, indicating similar local backbone configurations. Fig. 4.5 plots Φ and Ψ Ramachandran angle probabilities for the different n-mers. Even though there is a clear change in the structures globally as more domains are added (Fig. 4.2), the local secondary structure landscape is preserved in Fig. 4.5. As expected from the high amount of β-sheet formation (either within a single domain or across multiple), the most probably angles are those that are prone to form β-

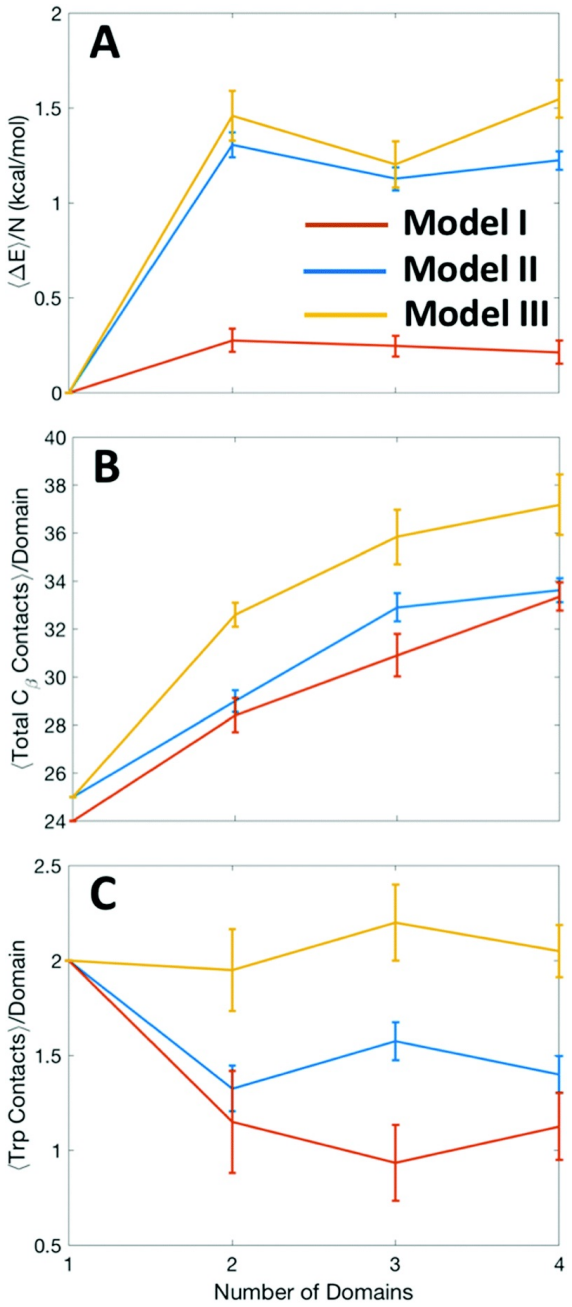

Figure 4.4 (A) Energy change (from monomer) per residue vs. number of tethered domains. (B) Total number of C_β to C_β contacts per domain vs. number of domains. (C) Number of Trp contacts per domain vs. number of domains, for model I in orange, model II in blue, and model III in yellow.

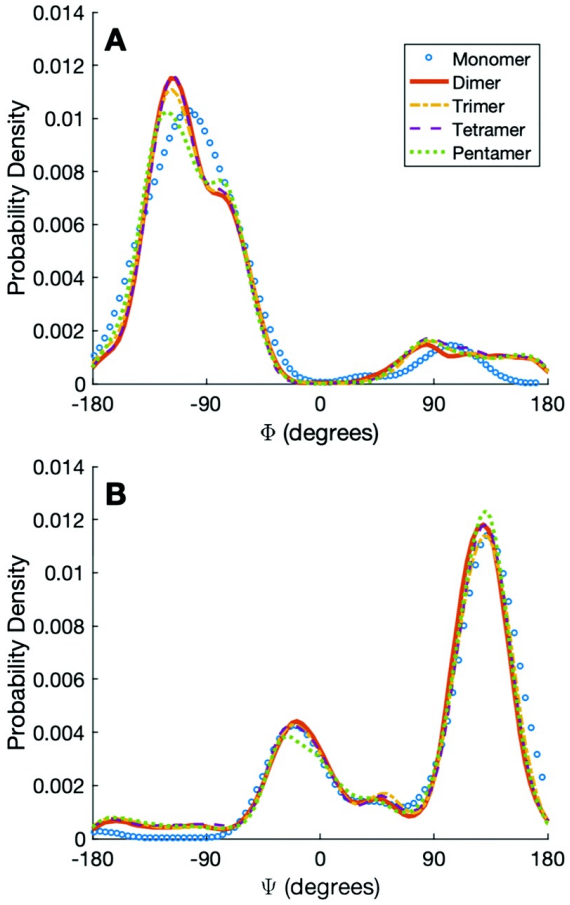

Figure 4.5 Probability density of (A) Φ and (B) Ψ angles for monomer, dimer,trimer, tetramer, and pentamer for model II.

sheets. The Ramachandran histogram also populates angles that have a high propensity of forming α-helices ($-70° > \Psi > 20°$) even though none of the proposed structures ($n = 1$ to 4), in Fig. 4.1 or Fig. 4.2, contain an α-helix.

However, the angles with helical tendencies lead to an actual α-helix only in the coarse-grained simulated annealing of the pentamer ($n = 5$) in Fig. 4.6, which could not be expressed in experiments. The pentamer forms a new type of misfolded structure compared to the ones seen in Fig. 4.2 for $n = 2$ to 4: an α-helix containing a Trp residue in the second domain, which is surrounded by β-sheets, emerges in $\approx 20\%$ of predicted pentamer structures. This

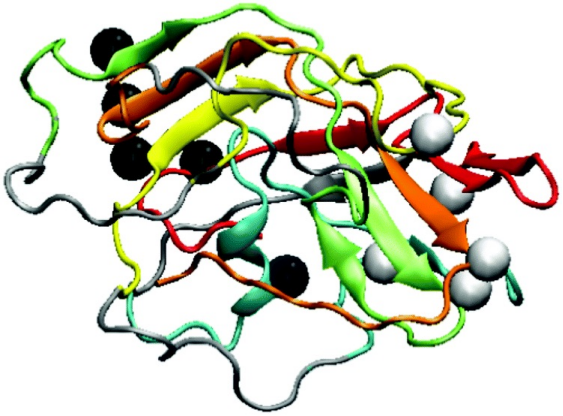

Figure 4.6 Example of a predicted WW-domain pentamer structure with all domains misfolded. Domains are colored, and linkers are gray. The order of the colors starting from N-terminus is red, cyan, yellow, green, orange. The black and white beads are the C_β atoms of Trp (8, 51, 94, 137, 180) and Tyr (20, 63, 109, 152, 195), respectively. An α-helix containing a Trp residue appears in the cyan domain.

suggests that the extra domains stabilize an α-helix formed by residues with Ψ angles that are in the range $(-70° > \Psi > 20°)$. The extra domains provide more tertiary contacts allowing for side-chains to align correctly into an α-helix from an already twisted β-strand.

Additionally, the pentamer has a probability very close to 1 of at least one domain being misfolded, and 0.6 even for the conservative model III (Fig. 4.3). The lack of pentamer expression again suggests that Model II represents a fairly accurate balance between fragment memory and inter-domain interactions.

4.3 Conclusion

Natural and engineered repeat proteins have provided many insights to relate folding, misfolding and function. Evolution for folding, which does not favor repeats with similar sequences adjacent in a multi-domain protein [164, 165, 170], goes hand-in-hand with evolution for function, which sometimes favors multiple domains of similar structure.21,41 Ankyrin

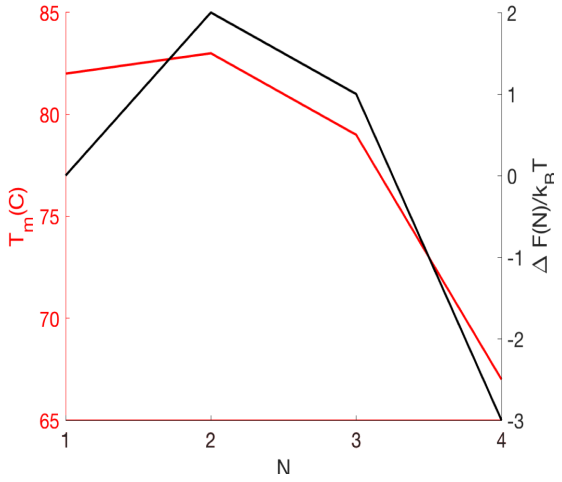

Figure 4.7 Melting temperature and free energy change vs. number of tethered domains from experiments

domains42,43 and TPR motifs44 in particular have shown how nearly identical folds can co-exist with the right balance of sequence similarity. These results have been complemented by studies on consensus repeat sequences,8,45 which have shown evidence of a highly parallel, but not completely homogeneous, folding process capable of generating stable native states.3,9,46

Our results for repeats of sequence-identical WW domains show that above $n = 3$, a critical number of repeats is reached: individual domains are destabilized and likely to form non-native states. While the stability of dimer and trimer is at most slightly smaller than that of the monomer, the tetramer is noticeably less stable thermodynamically and sensitive to the purification tag used, whereas the pentamer cannot be expressed in significant quantities, presumably due to domain interactions that lead to misfolding. Thus, 2 to 3 identical repeat domains lead to a stable native WW repeat protein, but more identical domains in series foster misfolding. The observation that consensus ankyrin sequences (α-helical secondary structure) can form longer repeat folds than WW domain (β-sheet secondary

structure) indicates that certain folds and sequences have substantially lower propensity for misfolding than others when tethered together. This may be the reason why WW-domains are mainly observed as tandem repeats in nature,21 whereas natural ankyrins can contain many additional repeats.47

Transient aggregates have been proposed as a step during the folding of many non-repeat proteins, masquerading as folding intermediates. For example, the RNA-binding protein U1A forms such transient aggregates.2 We have shown that when U1A is tethered into a repeat protein, transient aggregation is enhanced and leads to irreversible (on the time scale of the experiments) aggregation when too many repeats are tethered together.47 U1A is a very aggregation-prone protein, and we found that the size of its irreversible aggregation nucleus is only $n = 2.10$ WW domain is not prone to aggregation (as evidenced by facile NMR structures obtained at mM concentration).33 Here we find that the size of the FiP35 irreversible aggregation nucleus lies at $n = 4$. Thus, if a range of $n \approx 2$ to 4 is likely for the aggregation nuclei of most proteins; oligomeric aggregates may be formed rather easily. The 'intramolecular amyloids' we observe when repeats interact [e.g. Fig. 4.2(H)] may be examples of what oligomeric aggregates in non-tethered proteins look like. Indeed, it has been shown for protein U1A that addition of an Alzheimer sequence increases transient aggregation and allows stable dimers to form.48

The WW tetramer highlights how protein folding can be sensitive to the environment, in tandem with current in-cell folding experiments.49,50 The type of purification tag used for WW tetramer (histidine vs. GST) determines whether a nativelike or a non-native secondary structure is recovered. Thus, the local environment is critical for the folding of the tetramer. In-cell experiments have shown that proteins can be stabilized or destabilized in the cellular environment, depending on protein surface electrostatics,51 or the organelle environment.52,53 While these effects are small, they can be critical in regulating signaling and

other protein–protein interactions, which are often weak (on the order of a few kJ mol^{-1}).1 Such sticking or 'quinary structure' of proteins,54–56 of which only the tip of the iceberg has been characterized,57,58 may well account for the large majority of in-cell protein–protein or protein–nucleic acid interactions.

Repeat proteins may assist in the evolution of new folds.59–62 Our structural simulations of identical repeats highlight one possible path towards the evolution of more complex protein folds. For example the tetramer [Fig. 4.2(K)] forms larger-stranded beta sheets by combining strands from different domains. Since the WW-domain monomer contains three beta-strands with strong curvature (seen in Fig. 4.1), the beta-strands that form the disordered loops in the oligomers are prone to form helices. Such loops could evolve to form helical structure (Fig. 4.5), yielding a protein whose beta sheets have large contact order63 because they are separated by other secondary structure elements (loops, helices). The latter is a very common structural motif. Indeed, longer repeats can form entirely novel structures, such as the one shown in Fig. 4.6. Although the pentamer has many disordered regions, the combination of beta sheets and an alpha helix showed up in $\approx 20\%$ of simulated structures. If the loops were optimized by shortening, or mutated to favor additional alpha helices, Fig. 4.6 would represent a compact alpha/beta fold. Although not the subject of this paper, it would be interesting to take a sequence that forms simulated compact misfolded structure such as in Fig. 4.6, truncate the loops or increase their helix propensity, and see if improved expression and a well-defined tertiary structure could be obtained.

Eccentric, intervolved, yet regular
Then most, when most irregular they seem;
And in their motion harmony divine.
—John Milton

5

Other Cytoplasmic Effects: Crowding Shape* and Hydrodynamics†

5.1 Shape Packing Entropy

We investigated the geometry of voids, in terms of shape and size, formed by crowders of the various models modulated by the protein-crowder interactions. The voids are the depletion zones due to nonoverlapping volume exclusion between a protein and crowders. The distribution of the asphericity of the void Δ^{void} is virtually unchanged between spherical and rodlike crowders without attractive interactions [Fig. 5.2(a)]. Additionally, as ϕ_c increases, the peaks of the distributions are unperturbed. These results indicate that, without attractive interactions, the variation in the shape effect from rodlike crowders has been averaged

*The rod-like crowder part of this chapter has been published in *J. Phys. Chem. B* (2019) **123**, 3607-3617. AG Gasic is third author.

†The hydrodynamic interactions part of this chapter has been published in *Phys. Rev. E* (2018) **97**, 032402. AG Gasic is fourth author.

in the ensemble. However, when attractive interactions are added to the crowders, there is a clear shift rightward (toward higher asphericity) of the peak. The spread of the distributions increases with AR when the shape of a crowder becomes elongated [Fig. 5.2(b)]. These effects are greatest for $\phi_c = 5\%$.

We show that the distribution of the size of a void ((R_g^{void}) depends on the volume fraction of crowders, the shape of a crowder, and interaction between the protein and crowder. To make a fair comparison between spherical and rodlike crowders, we plotted the distribution of $(R_g^{\text{void}} - R)/\sigma$ in Fig. 5.2(c & d) where R is the radius of a sphere in a crowder model. The overall size of $(R_g^{\text{void}} - R)/\sigma$ is smaller for rodlike crowders (DX20-AR8) than spherical crowders (DX20-S) by roughly a third at each ϕ_c when a protein and crowders interact through hard-core interactions [Fig. 5.2(c)]. However, with a weak attraction between a protein and crowders ($\lambda = 0.83$), $(R_g^{\text{void}} - R)/\sigma$ is reduced by half across all crowder types [Fig. 5.2(d)] compared to those in [Fig. 5.2(d)]. The peaks in the distribution of $(R_g^{\text{void}} - R)/\sigma$ for rodlike crowders (DX20A-AR8) at all ϕ_c nearly overlap at the same position around 10.

These two properties, both crowder shapes and solvent-mediated interactions in manipulating the geometry of a protein, are pivotal to understand the full complexity of proteins in the cytoplasm.

5.2 Hydrodynamic Interactions

We investigated the impact of hydrodynamic interactions (HI) on protein folding using a coarse-grained model. The extent of the impact of hydrodynamic interactions, whether it accelerates, retards, or has no effect on protein folding, has been controversial. Together with a theoretical framework of the energy landscape theory (ELT) for protein folding that describes the dynamics of the collective motion with a single reaction coordinate across a fold-

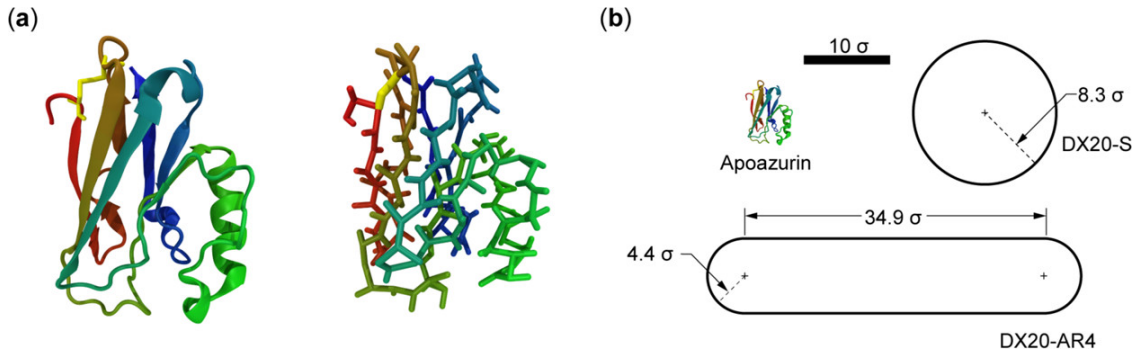

Figure 5.1 (a) Representations of the folded structure of apoazurin (PDB ID: 1E65). The folded structure is illustrated in a cartoon representation and a side-chain C_α representation. These protein illustrations were produced using VMD. The algorithm DSSP was used to assign the secondary structures. The residues that form the disulfide bond are shown in yellow. (b) Illustrations of two models for dextran 20 with the same volume: a spherical model (DX20-S) and a spherocylinder model with an aspect ratio (AR) of 4 (DX20-AR4). As reference, the folded structure of apoazurin is shown. The reduced unit of length is σ.(Created by Fabio Zagarra))

ing barrier, we compared the kinetic effects of HI on the folding rates of two protein models that use a chain of single beads with distinctive topologies: a 64-residue α/β chymotrypsin inhibitor 2 (CI2) protein, and a 57-residue β-barrel α-spectrin Src-homology 3 domain (SH3) protein. When comparing the protein folding kinetics simulated with Brownian dynamics in the presence of HI to that in the absence of HI, we find that the effect of HI on protein folding appears to have a "crossover" behavior about the folding temperature. This means that at a temperature greater than the folding temperature, the enhanced friction from the hydrodynamic solvents between the beads in an unfolded configuration results in lowered folding rate; conversely, at a temperature lower than the folding temperature, HI accelerates folding by the backflow of solvent toward the folded configuration of a protein. Additionally, the extent of acceleration depends on the topology of a protein: for a protein like CI2, where its folding nucleus is rather diffuse in a transition state, HI channels the formation of contacts by favoring a major folding pathway in a complex free energy landscape, thus accelerating

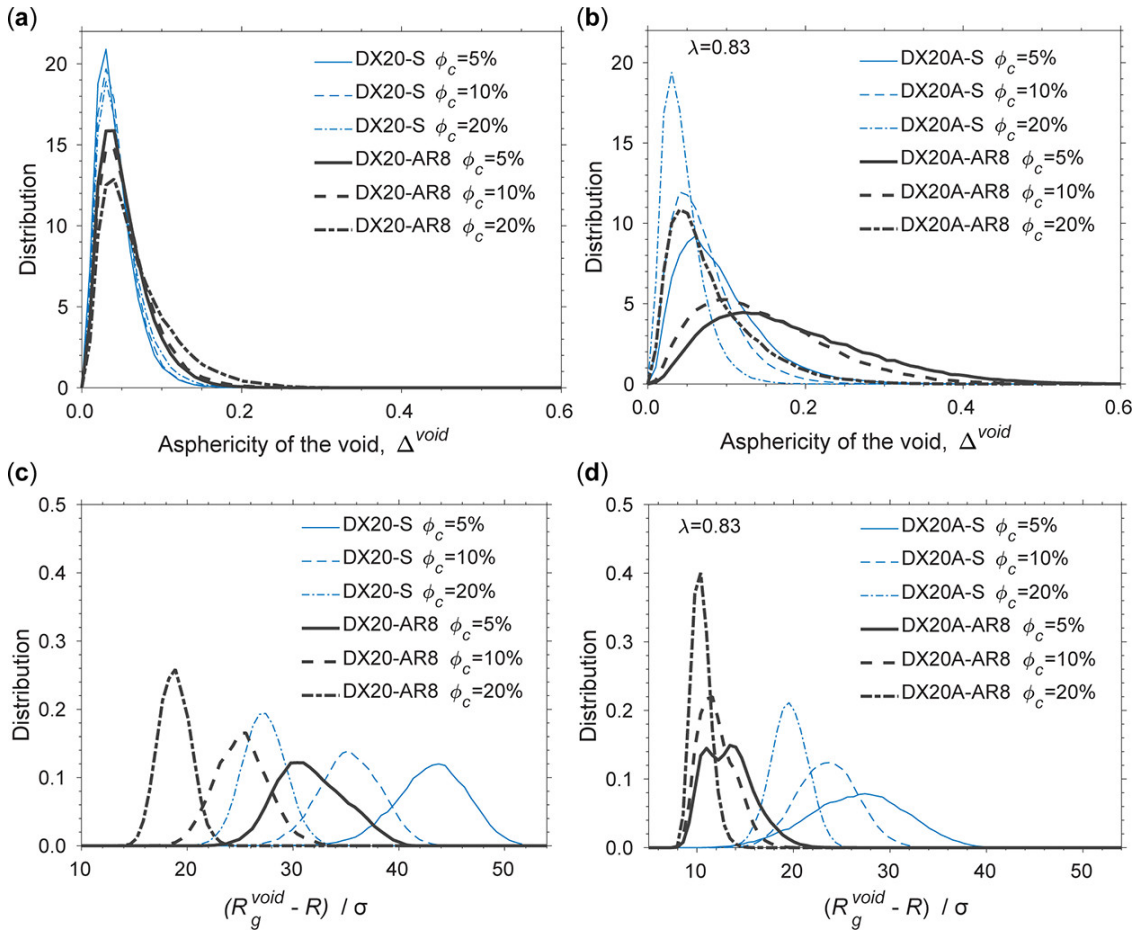

Figure 5.2 Distribution of the asphericity of the void, Δ^{void}, between the protein and crowders with (a) steric repulsive interactions and (b) nonspecific attractive interactions between the protein and crowders. The distribution of the radius of gyration of the void, $(R_g^{\text{void}} - R)/\sigma$, between the protein and crowders with (c) steric repulsive interactions and (d) nonspecific attractive interactions between the protein and crowders.

folding. For a protein like SH3, where its folding nucleus is already specific and less diffuse, HI matters less at a temperature lower than the folding temperature. Our findings provide further theoretical insight to protein folding kinetic experiments and simulations.

Our motivation is to reconcile the differences in reported influences of HI on protein folding over a wide range of temperature from the viewpoint of the folding energy landscape theory[17,18], particularly with a funnel-shaped energy landscape[19]. We used a computer protein model that is guaranteed to fold into the native state from any unfolded conformation [20]. We tracked its collective motion on a single reaction coordinate, the fraction of the native contact formation Q either on a thermodynamic free energy barrier or by kinetic trajectories. We studied the effects of HI on folding of two well-studied model proteins with distinctive topologies: one is the 64-residue α/β protein chymotrypsin inhibitor 2 (CI2) [21], and the other is the 57-residue β-barrel α-spectrin Src-homology 3 (SH3) domain [22]. The two proteins fold and unfold in a two-state manner and have been used for studying folding mechanisms from other computational studies [23–27]. We simulated the Brownian dynamics of particles including HI by implementing the algorithm developed by Ermak and McCammon [28]. The effects of HI are approximated through a configuration-dependent diffusion tensor $\mathbf{D}$ used in the Brownian equation of motion.

Our study shows that the effect of HI on folding rates can both accelerate protein folding at a temperature lower than the folding temperature and retard protein folding speed at a temperature higher than the folding temperature, in comparison with the folding dynamics without HI. Since HI affects the kinetic ordering of contact formation, for a protein with multiple viable folding pathways like CI2, HI will favor a particular folding route in a complex folding energy landscape. In that sense, energy landscape theory (ELT) is short of fully predicting folding rates. From Secs. IIIB to III E, we investigate the cause of this temperature dependence of the effect of HI on folding rates and the implications for energy

landscape theory.

We explored the folding kinetics of CI2 and SH3 by comparing t_{fold} with BD over a broad range of temperatures. Both proteins exhibit non-Arrhenius [46] behavior against temperature as shown in Figs. 2(a) and 2(b). At high temperatures, t_{fold} increases because the thermal fluctuations are higher than the stability of the protein, and at low temperatures, t_{fold} increases due to the fact that the protein is trapped in a local energy minimum [18,46,47]. The temperature that renders the fastest t_{fold} is at 0.95 T_f^{CI2} and 0.91 T_f^{SH3} for CI2 and SH3, respectively. We computed t_{fold} for the proteins with BDHI over a narrow range of temperatures around T_f of the proteins in Figs. 2(c) and 2(d) in dashed lines. Our study shows that the impact of HI on t_{fold} is small within an order of magnitude, but statistically significant. What is most interesting is that HI either increases or decreases the folding time depending whether the temperature is higher or lower than T_f . This distinctive "crossover" behavior occurs in the proximity of the folding temperature of CI2 ($\approx 1.03\ T_f^{\text{CI2}}$) and SH3 ($\approx 0.98\ T_f^{\text{SH3}}$). Thus, the impact of HI on protein folding kinetics is temperature dependent. However, the acceleration of the folding is more prominent for CI2 than for SH3 at $T < T_f$. Therefore, HI effects also depend on the topology of a protein.

We will further investigate the role of topology in the extent of impact from HI on protein folding in the following subsection at two temperatures for each protein: below T_f (0.95 T_f^{CI2} for CI2 and 0.91 T_f^{SH3} for SH3) and above T_f (1.06 T_f^{CI2} for CI2 and 1.03 T_f^{SH3} for SH3).

Can we capture this folding behavior using a global order parameter? A theoretical estimation of the folding kinetic rate k (the rate is the inverse of folding time t_{fold}) depends on the shape of free energy surface and the effective diffusion coefficient D^{eff} of an order

parameter on the free energy surface [46–48] as such,

$$k = \frac{1}{t_{\text{fold}}} = \left(\frac{\beta}{2\pi}\right)^{1/2} D^{\text{eff}} \omega\omega^{\dagger} \exp\left(-\beta\Delta F^{\dagger}\right) \tag{5.1}$$

where ω and $\omega^{\dagger}$ are the curvatures of the unfolded state free energy well and barrier, respectively, β is the inverse temperature, and $\Delta F^{\dagger}$ is the free energy barrier height with respect to the unfolded state free energy. However, since the Hamiltonian for BD and BDHI are identical rendering the same free energy profiles, the change in the folding kinetic rates should be explained by the change in the diffusion of the order parameter.Here, the order parameter is the fraction of native contact formation Q. The mean-squared displacement (MSD) of Q is obtained as a function of time. D^{eff} of Q is estimated from the linear region of the MSD of Q as a function of lag time

If Q is a perfectly good reaction coordinate that captures the collective dynamics of a complex system, the folding rates computed directly from the folding kinetic simulations should be the same as the rates predicted by the energy landscape theory. When the simulated rate deviates from the prediction, it infers that the dynamics of a complex system of many degrees of freedom might not be adequately described by using only a single reaction coordinate. We obtained $k_{\text{BDHI}}/k_{\text{BD}}$ (or $t_{\text{fold}}^{\text{BD}}/t_{\text{fold}}^{\text{BDHI}}$) from the kinetic simulations for CI2 and SH3 in Table 5.1. Additionally, because HI only influences the pre-exponential factor but not the barrier height in Eq. (14), the ratio of the predicted rates from the energy landscape theory is equivalent to $D_{\text{BDHI}}^{\text{eff}}/D_{\text{BD}}^{\text{eff}}$. To test this, we compare the ratio of $D_{\text{BDHI}}^{\text{eff}}/D_{\text{BD}}^{\text{eff}}$ from MSD calculation to the ratio of $t_{\text{fold}}^{\text{BD}}/t_{\text{fold}}^{\text{BDHI}}$ from kinetic simulations. We found that indeed the ratio of $D_{\text{BDHI}}^{\text{eff}}/D_{\text{BD}}^{\text{eff}}$ is not equal to the ratio of $t_{\text{fold}}^{\text{BD}}/t_{\text{fold}}^{\text{BDHI}}$ in Table 5.1, although it shows the right trend of the crossover behavior. We speculate that a mean-field description of overall folding with the collective order parameter of Q along an energy landscape may not

fully grasp the kinetic principle of HI on folding.

Table 5.1 Folding time from kinetic simulations ($t_{\rm fold}$) and the effective diffusion coefficient ($D^{\rm eff}$) of Q for CI2 or SH3 using BD or BDHI.

Protein	T	From kinetic simulations			From MSD analysis		
		$t_{\rm fold}^{\rm BD}$ $[10^6\tau]$	$t_{\rm fold}^{\rm BDHI}$ $[10^6\tau]$	$\frac{k_{\rm BDHI}}{k_{\rm BD}} = \frac{t_{\rm fold}^{\rm BD}}{t_{\rm fold}^{\rm BDHI}}$	$D_{\rm BD}^{\rm eff}$ $[10^{-9}1/\tau]$	$D_{\rm BDHI}^{\rm eff}[10^{-9}1/\tau]$	$\frac{k_{\rm BDHI}}{k_{\rm BD}} = \frac{D_{\rm BDHI}^{\rm eff}}{D_{\rm BD}^{\rm eff}}$
	$[T_f^{\rm CI2}]$						
CI2	0.95	0.52 ± 0.02	0.38 ± 0.01	1.37 ± 0.06	89.16 ± 0.08	101.22 ± 0.09	1.14 ± 0.00
	1.06	3.64 ± 0.13	4.39 ± 0.14	0.83 ± 0.04	2.29 ± 0.01	0.35 ± 0.00	0.15 ± 0.00
	$[T_f^{\rm SH3}]$						
SH3	0.91	0.19 ± 0.01	0.16 ± 0.01	1.19 ± 0.10	217.72 ± 0.30	236.71 ± 0.24	1.09 ± 0.00
	1.03	1.08 ± 0.05	1.65 ± 0.07	0.65 ± 0.04	6.39 ± 0.01	2.56 ± 0.01	0.40 ± 0.00

The most exciting phrase to hear in science, the one that heralds the most discoveries, is not 'Eureka!' (I found it!) but 'That's funny...'

—Isaac Asimov

6

Perspectives and Outlook

Life is complex. We all know a living thing when we see it, but do we really understand it? Currently, no scientist or engineer can come close to creating a chemical reaction or machine that is as complex as one of the simplest living cells.

The collective interactions of these nano-machines, i.e. proteins, give this living matter the ability to repair itself, digest food, and sense its surroundings in order to make decisions. Proteins are not alive, but somewhere between one protein and many, "life" emerges. Therefore, through studying the cytoplasm and how a single protein operates in this complex environment, we can connect the physics of one protein to many, with the ultimate aspiration to understand the emergent properties that makes living matter "alive".

This thesis has investigated how these machines work in such a crowded and chaotic environment of the cytoplasm.

6.1 Main Conclusions

The main overarching conclusion from the previous chapters are that the environment of the cell gives rise to new non-trivial phase behavior of proteins that would not be possible as an isolated system. In short: *More proteins are different.*[*] The collective effects of the protein and surrounding many-body system (i.e., the cytoplasm) are emergent and essential for biological function. In summary:

- **Crowding effects the stability of the individual folding phases and the barriers between them.** In chapter 3, a critical transition is observed in PGK's folding phase diagram. This is due to the loss of a barrier between to phases, which is controlled by the amount of crowding volume fraction.

- **Neighboring proteins can destabilize folding phases and new ordered structures can emerge.** In chapter 4, as the number of tethered WW-domains increased, the probability of misfolding increased as well. This was mainly due to the competition of inter- and intra-domain interactions. Additionally, a totally new structure emerges that was not expected based on the underlying Hamiltonian.

- **Crowder shape can break depletion force symmetry.** From chapter 5, the rod-like crowders produced an elongated conformation of apoazurin, which is due to the directional entropic forces of the crowders.

- **Protein kinetics are affected by hydrodynmaic interactions.** Also in chapter 5, a proteins folding kinetics increases at low temperatures and decreases at high temperatures. This can also be extrapolated to protein binding, protein-complex assembly,

[*]A play on Anderson's famous essay "More is different" (cite).

and beyond.

6.2 Outcome and Future Directions

Universality of critical phenomena in proteins. The main goal is to capture the P-T-ϕ phase diagram of PGK using the same methodology used in Thirumalai's paper [171], and extend this to other proteins in hopes of finding the origin of criticality in proteins and possibly universality.

Protein abundance dictates protein-complex assembly and state. We can think of the cytoplasm as a collection of proteins which have interactions J_{ij} and chemical potentials μ_i and the change in the state of a cell (from A to B) depends on these two properties:

$$\text{State}_A(\{\{J_{ij}^A\},\{\mu_i^A\}\}) \rightarrow \text{State}_B(\{\{J_{ij}^B\},\{\mu_i^B\}\}) \tag{6.1}$$

The change in $\{J_{ij}^A\} \rightarrow \{J_{ij}^B\}$ may be due to change in protein conformation or mutations, and $\{\mu_i^A\} \rightarrow \{\mu_i^B\}$ may be due to change in gene expression level. Protein interactions may also affect gene expression and vise versa.

Living cells can contain on the order of 10^4 distinct types of proteins and other macromolecules at a given time. In this many-component mixture of the subcellular environment, macromolecules assemble into complexes and organize hierarchically into spatial networks. In fact, these unfathomably complex networks give rise to the emergence of all biological functions and ultimately the properties of life [172, 173, 174, 175].

The specific arrangements of macromolecules are thought to emerge from the vast amount of weak "quinar" and entropic interactions[159, 176, 3, 89]. The most intuitive conception of protein biophysics in this crowded environment is that of volume exclusion[177] exerted

on a given protein by surrounding macromolecules. Volume exclusion creates entropic forces that depend on the shapes and sizes in the crowd of macromolecules. Additionally, the presence of "quinary structures"[4, 44] undercuts the assumption that volume exclusion from surrounding crowders is the dominant physics principle in predicting protein dynamics in a cell. Proteins interact weakly and form unconventional quinary complexes through counteracting forces between favorable electrostatic interactions and unfavorable solvation energies provided by their metabolites[46].

However, the physical mechanism of these complex assemblies is still unclear. In addition to protein-protein interactions and macromolecular crowding, another entropic effect must be considered: fluctuations in the number of particles. Changing the chemical potentials, μ, of the different types of proteins in the multi-component mixture can fundamentally alter its properties and display complex transitions such as liquid-liquid demixing or granular body formation[178, 179]. Without this understanding of effects of particle number fluctuations, the physics of the assembly and hierarchical organization of macromolecules will remain elusive.

Here we propose a cellular cytoplasm model to understand hierarchical protein-complex assembly using a Grand canonical Monte Carlo simulation (GCMC) and the principle of maximum entropy to solve the inverse statistical mechanics problem [180] of finding the correct chemical potentials for each protein type. Our approach is built upon the coarse-grained modeling of macromolecules that exert volume exclusion in a crowded environment [47], and infer pair-wise interactions between macromolecules using proteomic experimental data gathered by chemical cross-linking mass spectroscopy (XL-MS)[181]. Since the number of macromolecules in the cell is not static, the Grand Canonical (GC) ensemble is uniquely suited to allow for fluctuation in particle number by keeping the set of chemical potentials $\{\mu_\alpha\}$ fixed.

Furthermore, the cytoplasm inside a cell is comprised of thousands of different types of biomolecular components causing complicated behaviors of the system. Previous work by Sear and Cuesta[182] developed a new statistical approach to estimate the condition whereby the mixtures of a large number of components unintentionally phase separate (or demix) through randomly interacting biomolecules. In conjunction with GCMC simulations, Jacobs and Frenkel[183] further showed that the changes in the mixing entropy of distinguishable multi-components direct the transition between a condensation phase and a demixing phase. This phenomena is one of the driving forces in liquid-liquid phase separation[184] in cell biology. Another investigation from the Leibler group[185, 179] also used GCMC to show that by tuning the chemical potential, μ, diverse structures of proteins self-organize into distinct groups of assemblies. One urgent challenge from these GCMC studies is to justify that $\{\mu_\alpha\}$ determines the distribution of particle numbers for each type of biomolecule in an open system. We infer the chemical potentials of specific proteins using the principle of maximum entropy to match experimentally derived mean particle number or protein abundance from the protein abundance database (PaxDB) [186]. Similar approaches have been used various areas of biological physics to infer the interaction energies of a Canonical ensemble (constant particle number) such as in understanding chromosome architecture [187], protein-protein interactions [188], or correlations in neural networks [189]. In contrast, here we are estimating the particle distribution of a GC ensemble where $\{\mu_\alpha\}$ is inferred.

The ultimate goal of this investigation will lead to understanding the of role of protein-protein interactions at a proteomic level that establishes the hierarchical assembly of macromolecular complexes, and its role in controlling cellular function.

A

Coarse-grained Computational Models

A.1 Structure-based Models

Our simulations use a structure-based model, which is minimalist protein model ("beads on a chain") that incorporates experimentally derived structural information [190], to investigate the mechanism of protein folding dynamics optimally. The emergence of structure and function from a protein sequence makes the modeling of proteins from first principle (*ab initio* models) computationally and theoretically prohibitive. Therefore, experimentally derived structural information is needed (even in models termed "all-atom", which refine *ab initio* force field parameters to fit experimentally known structures) to capture key features in protein folding and dynamics [86]. A structure-based model is often utilized as the "ideal gas" of protein folding for the investigation of a wide range of folding mechanisms [30, 191].* This model renders an energy landscape [87] with minimal frustration and

*The importance of ideal models is best stated by Alexander Y. Grosberg at the 2019 APS March meeting, "perfectly detailed simulations can only reproduce Nature, but not explain it."

contains a dominant basin of attraction, corresponding to an experimentally determined configuration [88]. As such, the model carries the bonus of being computationally inexpensive, enabling long-timescale simulations to be obtained for a large protein and macromolecular crowding system. Long-timescale simulations are also crucial for high-pressure unfolding since pressure unfolds proteins at an order of magnitude (or more) slower than heat unfolding; therefore, structure-based, minimalist-model simulations provide statistically significant results. Lastly, structure-based models tend to capture unfolded protein scaling laws better than all-atom models [192], which is necessary to characterize the various non-crystal states of PGK correctly.

A.2 Desolvation Potential and Crowder Hamiltonian

Similar to adding specific complexity to the ideal gas model to study specific phenomena, we add the desolvation barrier [60] to the native interactions that accounts for the free energy cost to expel a water molecule in the first hydration shell between two hydrophobic residues [120] to study pressure unfolding, leading to the appearance of a partially folded intermediate. The use of this model has been validated in other systems [142]. The total system is described by the Hamiltonian $\mathcal{H}_{\text{tot}} \equiv \mathcal{H}_p + \mathcal{H}_{pc} + \mathcal{H}_{cc}$, which accounts for the interaction within the protein ($\mathcal{H}_p$), between the protein and crowders ($\mathcal{H}_{pc}$), and between

crowders ($\mathcal{H}_{cc}$). The Hamiltonian of this structure-based protein model, $\mathcal{H}_p$, is as follows:

$$\begin{aligned}\mathcal{H}_p\left(\Gamma,\Gamma^0\right) =& \sum_{i<j} K_r\left(r_{ij}-r_{ij}^0\right)^2\delta_{j,i+1} + \sum_{i\in \text{ angles}} K_\theta\left(\theta_i-\theta_i^0\right)^2 \\ &+ \sum_{i\in \text{ dihedrals}} K_\phi\left(\left\{1-\cos\left[\phi_i-\phi_i^0\right]\right\}+\frac{1}{2}\left\{1-\cos\left[3\left(\phi_i-\phi_i^0\right)\right]\right\}\right) \\ &+ \sum_{i,j\in \text{ native}} U\left(r_{ij},r_{ij}^0,\epsilon,\epsilon''\right) + \sum_{i,j\notin \text{ native}} \epsilon_0\left(\frac{\sigma}{r_{ij}}\right)^{12}, \end{aligned} \tag{A.1}$$

where Γ is a configuration of the set r, θ, ϕ. The r_{ij} term is the distance between i^{th} and j^{th} residues, θ is the angle between three consecutive beads, and ϕ is the dihedral angle defined over four sequential residues. δ is the Kronecker delta function. $\Gamma^0 = \{\{r^0\},\{\theta^0\},\{\phi^0\}\}$ is obtained from the crystal structure configuration. Lastly, $U(r_{ij},r_{ij}^0,\epsilon,\epsilon'')$ is the desolvation potential in Fig. 3.1(b) (or Fig. S2.2 [135]), which contains a P-dependent contact well energy (ϵ) and barrier height energy (ϵ'') as,

$$\epsilon(P) = \epsilon_0 - \upsilon_1 P, \tag{A.2a}$$

$$\epsilon''(P) = \epsilon_0'' + \upsilon_2 P, \tag{A.2b}$$

where ϵ_0 is the solvent averaged energy and ϵ_0'' is the barrier height at ambient P. The constants υ_1 and υ_2 are taken from Ref. [120]. The potential is described as piecewise function,

$$U\left(r,r^0,\epsilon,\epsilon''\right) = \begin{cases} \epsilon\left[\left(\frac{r^0}{r}\right)^{12}-2\left(\frac{r^0}{r}\right)^6\right] & \text{if } r<r^0 \text{ and } \epsilon>0 \\ -\epsilon\left[\left(\frac{r^0}{r}\right)^{12}-2\left(\frac{r^0}{r}\right)^6\right]-2\epsilon & \text{if } r<r^0 \text{ and } \epsilon<0 \\ C\left(r-r^\dagger\right)^4-2\left(\epsilon+\epsilon''\right)\left(r-r^\dagger\right)^2+\epsilon'' & \text{if } r^0\le r<r^\dagger \\ -B\frac{\left(r-r^\dagger\right)^2-h_1}{\left(r-r^\dagger\right)^6+h_2} & \text{if } r^\dagger\le r, \end{cases} \tag{A.3}$$

where constants C, B h_1 and h_2 are,

$$C = \frac{(\epsilon + \epsilon'')}{(r^\dagger - r^0)^2}, \tag{A.4}$$

$$B = 3\epsilon' \left(r'' - r^\dagger\right)^4, \tag{A.5}$$

$$h_1 = \frac{2}{3}\frac{\left(r'' - r^\dagger\right)^2}{\epsilon'/\epsilon'' + 1}, \tag{A.6}$$

$$h_2 = 2\frac{\left(r'' - r^\dagger\right)^6}{\epsilon''/\epsilon' + 1}. \tag{A.7}$$

Here, $r'' = r^0 + 0.8\sigma$ is the position of the water-mediated potential well minimum, and $r^\dagger = 0.5(r^0 + r'')$ is the position of the desolvation barrier maximum. Crowders are modeled as hard spheres with Hamiltonians $\mathcal{H}_{pc}$ and $\mathcal{H}_{cc}$ with the following form:

$$\mathcal{H}_{pc}(r_{ij}) = \sum_i^N \sum_j^n \epsilon_0 \left(\frac{\sigma_{ij}}{r_{ij}}\right), \tag{A.8a}$$

$$\mathcal{H}_{cc}(r_{ij}) = \sum_{i<j}^n \epsilon_0 \left(\frac{\sigma_{ij}}{r_{ij}}\right), \tag{A.8b}$$

where N and n are the number of residues ($= 415$) and crowders, respectively; $\sigma_{ij} = 0.5(\sigma_i + \sigma_j)$, is distance between any two particles in direct contact.

The complete descriptions of a structure-based protein model, desolvation potential, and simulations of PGK in a periodic cubic box of Ficoll 70 are provided in the Supplemental Material [135]. All simulations were performed using GROMACS 2016.3 molecular dynamics software [193].

A.3 Associative Memory, Water Mediated, Structure and Energy Model

The tethered WW-domains were computationally simulated using the Associative memory, Water mediated, Structure and Energy Model (AWSEM).29 The model predicts structures and helps understand the competition between folding and interdomain interactions by providing polymeric insights into the formation of contacts according to physico-chemical features of protein residues (sample structures in Fig. 1).

AWSEM is a coarse-grained protein model with transferable energy functions that have been optimized to predict tertiary structures of globular proteins. AWSEM has been used in globular protein structure prediction, binding predictions of protein dimers, and amyloid fibril formation, through simulated annealing.13,30–32 The AWSEM potential is a combination of both knowledge-based and physics-based terms. It uses a three-bead per amino-acid coarse-graining (C_α, C_β, and O atoms) that generates the coordinates of other heavy atoms along a backbone. It includes independent and cooperative hydrogen bonding, water-mediated tertiary interactions, and biasing local structural preferences based on short fragment memories. Although AWSEM lacks atomistic resolution, this model is sufficient in sampling a wide conformational space involving folding and binding contributing to the folded, the unfolded, or the misfolded states. Our model uses the native state of an individual WW domain as a reference state, and thus conservatively assesses the population of misfolded states.

The total Hamiltonian consists of a backbone term $\mathcal{H}_{\mathrm{BB}}$, a potential of mean force $\mathcal{H}_{\mathrm{PMF}}$,

and a fragment memory term $\mathcal{H}_{\mathrm{FM}}$:

$$\mathcal{H}_{\mathrm{AWSEM}} = \mathcal{H}_{\mathrm{BB}} + \mathcal{H}_{\mathrm{PMF}} + \mathcal{H}_{\mathrm{FM}} \tag{A.9}$$

$\mathcal{H}_{\mathrm{BB}}$ constrains the backbone chain to physically realistic heteropolymer conformations (see ESI Section S3 for details). The potential of mean force $\mathcal{H}_{\mathrm{PMF}}$ depends on the identities of the interacting residues and contains direct contacts, water mediated contacts, burial, and hydrogen bonding terms (see ESI Section S3 for details). The $\mathcal{H}_{\mathrm{BB}}$ and $\mathcal{H}_{\mathrm{PMF}}$ terms do not depend on the knowledge of the native structure and only depend on the sequence of residues; thus, the two terms allow for non-native and long-sequence distance interactions and are responsible for the formation of non-native structure across multiple domains. Model parameters are chosen to minimize misfolding of the WW-domain by itself, in accord with experimental observation (see Section 3.6).

The fragment memory term $\mathcal{H}_{\mathrm{FM}}$ is particularly important in the context of single domain folding, as it contains local sequence interactions using the knowledge of the native structure. Memories are sequences with known structures (typically obtained from the protein data bank). The fragment memory potential sums over all memories m from short sequences, and all pairs of atoms (not residues) i and j such that the atoms have a sequence separation $3 \leq |I - J| \leq 9$, having the form

$$\mathcal{H}_{\mathrm{FM}} = -\lambda_{\mathrm{FM}} \sum_{m} \sum_{i,j \ni 3 \leq |I-J| \leq 9} \exp\left[-\frac{\left(r_{ij} - r_{ij}^{m}\right)^2}{2\sigma_{IJ}^2} \right] \tag{A.10}$$

where r_{ij} and r_{ij}^{m} are the distances between atoms i and j of the simulated structure and of the memory structure, respectively. In this study, we use a single memory, which is the folded experimental structure of the isolated FiP35 WW domain (PDB ID: 2F21).33 The

well width $\sigma_{IJ} = \lambda_\sigma |I - J|^{0.15}$, and we fixed $\lambda_\sigma = 0.2$ Å in all simulations. Unlike Gō or structure-based models,34 HFM only acts less than 10 residues apart within the monomer, affecting mainly secondary structure. The other non-backbone terms act both locally and in long-sequence distances, affecting tertiary and interdomain structure also. A more detailed description of the AWSEM Hamiltonian terms can be found in ref. 29, 35 and 36.

The fragment memory terms contain a scaling parameter λ_{FM} in eqn (5) adjusting the interaction strength. λ_{FM} allows us to tune the aggregation propensity relative to the folding propensity. In this study, we use three different values of λ_{FM} to compare folding vs. aggregation of the tethered domains as bias towards the folded crystal structure is decreased ($\lambda_{\mathrm{FM}} = 0.4\,\mathrm{kJ\,mol^{-1}}$, Model I), or increased ($\lambda_{\mathrm{FM}} = 1.2\,\mathrm{kJ\,mol^{-1}}$, Model III) compared to the standard value ($\lambda_{\mathrm{FM}} = 0.8\,\mathrm{kJ\,mol^{-1}}$, Model II). All other parameters were kept at default settings.29 The full set of AWSEM parameters used is shown in ESI (Section S4).

B

Simulation Methods

B.1 Langevin Dynamics

Langevin dynamics (LD) adds a friction and a noise term to Newton's equations of motion, as

$$m\dot{\boldsymbol{v}}(t) = -\nabla\mathcal{H}(\boldsymbol{r}) - \eta\boldsymbol{v}(t) + \boldsymbol{\xi}(t), \tag{B.1}$$

where

$$\langle\boldsymbol{\xi}(t)\cdot\boldsymbol{\xi}(t')\rangle = 2\eta k_B T\delta(t-t'), \tag{B.2a}$$

$$\langle\boldsymbol{\xi}(t)\rangle = \mathbf{0}. \tag{B.2b}$$

In GROMACS there is one simple and efficient implementation. Its accuracy is equivalent to the normal MD leap-frog and Velocity Verlet integrator. It is nearly identical to the common way of discretizing the Langevin equation, but the friction and velocity term are

applied in an impulse fashion [194]. It is described in Algo. B.1, where $\alpha = 1 - e^{-\gamma \Delta t}$.

Algorithm B.1 Langevin Dynamics Integrator Algorithm

procedure LD($\{\boldsymbol{r}_i(t), \boldsymbol{v}_i(t - \frac{1}{2}\Delta t), \boldsymbol{F}_i(t)\}, \Delta t, \alpha$) ▷ run LD to update configuration
  for all i **do** ▷ repeat for all particles
    $\boldsymbol{v}_i^\star \leftarrow \boldsymbol{v}_i\left(t - \frac{1}{2}\Delta t\right) + \frac{1}{m}\boldsymbol{F}_i(t)\Delta t$ ▷ compute 1/2-step velocity
    $\boldsymbol{\xi}_i \leftarrow \mathcal{N}(\boldsymbol{0}, \boldsymbol{1})$ ▷ attain from 3D Gaussian distribution
    $\Delta \boldsymbol{v}_i \leftarrow \alpha \boldsymbol{v}_i^\star + \sqrt{\frac{k_B T}{m}(1 - \alpha^2)}\boldsymbol{\xi}_i$ ▷ compute impulse term
    $\boldsymbol{r}_i(t + \Delta t) \leftarrow \boldsymbol{r}_i(t) + \left(\boldsymbol{v}_i^\star + \frac{1}{2}\Delta \boldsymbol{v}_i\right)\Delta t$ ▷ update positions
    $\boldsymbol{v}_i\left(t + \frac{1}{2}\Delta t\right) \leftarrow \boldsymbol{v}_i^\star + \Delta \boldsymbol{v}_i$ ▷ update velocities
  end for
  return $\{\boldsymbol{r}_i(t + \Delta t), \boldsymbol{v}_i(t + \frac{1}{2}\Delta t)\}$
end procedure

The global scheme for MD is given in Algo. B.2

Algorithm B.2 Global Molecular Dynamics Algorithm

procedure MD($\mathcal{H}, \{\boldsymbol{r}_i(0), \boldsymbol{v}_i(0)\}, \tau, \Delta t$) ▷ run MD given $\mathcal{H}$ and the initial conditions
  while $t \leq \tau$ **do** ▷ repeat for $\tau/\Delta t$ steps
    for all i **do**
      $\boldsymbol{F}_i(t) \leftarrow -\sum_j \frac{\partial \mathcal{H}}{\partial \boldsymbol{r}_{ij}}$ ▷ compute forces
    end for
    $\{\boldsymbol{r}_i(t + \Delta t), \boldsymbol{v}_i(t + \frac{1}{2}\Delta t)\} \leftarrow$ LD from Algo. B.1 ▷ update configuration
    write outputs
    $t \leftarrow t + \Delta t$
  end while
end procedure

Initial velocities are randomly generated from the Maxwell-Boltzmann velocity distribution:

$$\wp\left(\boldsymbol{v}_i\right) = \sqrt{\frac{\beta m}{2\pi}} \exp\left(-\frac{\beta m v_i^2}{2}\right) \tag{B.3}$$

B.2 Replica Exchange Method

Replica exchange method (REM) can be used to speed up the sampling of any type of simulation, especially if conformations are separated by relatively high energy barriers. It involves simulating multiple replicas of the same system at different temperatures and randomly exchanging the complete state of two replicas at regular intervals with the probability:

$$\mathcal{P}(1 \leftrightarrow 2) = \min\left(1, e^{(\beta_1-\beta_2)(E_1-E_2)}\right) \tag{B.4}$$

where β_1 and β_2 are the reference inverse temperatures and E_1 and E_2 are the instantaneous potential energies of replicas 1 and 2 respectively. After exchange the velocities are scaled by $(T_1/T_2)^{\pm 0.5}$ and a neighbor search is performed the next step. This combines the fast sampling and frequent barrier-crossing of the highest temperature with correct Boltzmann sampling at all the different temperatures [195, 196].

B.3 Simulated Annealing

We built the single memory configuration using atomic coordinates provided in the WW-domain of the human FiP mutant crystal structure with PDB ID: 2F21. We matched the sequences of the WW-domain oligomers used in the experiments (Table S1 in ESI). Each individual domain in the oligomers used the single memory of the monomer, and linkers joining domains were not influenced by the fragment memory term.

We performed all simulations in the canonical ensemble (NVT) using the Nosé-Hoover thermostat implemented using the LAMMPS molecular dynamics software.39 To predict the structures, we performed annealing simulations starting from a linear extended peptide

structure at a temperature of 650 K, and slow cooled over 10 million time-steps to 300 K (where a time-step is approximately 5 fs). Initial velocities were chosen randomly from a Boltzmann distribution with the average squared velocity equal to $3k_{\mathrm{B}}T/m$, where k_{B} is the Boltzmann constant, m is the mass, and the temperature T is set equal to 650 K. The simulated annealing was repeated 40 times for each oligomer and the three λ_{FM} values (Models I, II, and III). The temperature range was chosen to be approximately 150 K above and below the folding temperature of the monomer.

Bibliography

[1] H. Garcia-Seisdedos, C. Empereur-Mot, N. Elad, and E. D. Levy, "Proteins evolve on the edge of supramolecular self-assembly," *Nature*, vol. 548, no. 7666, p. 244, 2017.

[2] M. Chen and P. G. Wolynes, "Aggregation landscapes of huntingtin exon 1 protein fragments and the critical repeat length for the onset of huntington's disease," *Proceedings of the National Academy of Sciences*, vol. 114, no. 17, p. 4406, 2017.

[3] Y. Shin and C. P. Brangwynne, "Liquid phase condensation in cell physiology and disease," *Science*, vol. 357, no. 6357, 2017.

[4] E. H. McConkey, "Molecular evolution, intracellular organization, and the quinary structure of proteins," *Proceedings of the National Academy of Sciences*, vol. 79, no. 10, p. 3236, 1982.

[5] F. Huber, J. Schnauß, S. Rönicke, P. Rauch, K. Müller, C. Fütterer, and J. Käs, "Emergent complexity of the cytoskeleton: from single filaments to tissue," *Advances in physics*, vol. 62, no. 1, p. 1, 2013.

[6] K. G. Wilson, "The renormalization group: Critical phenomena and the kondo problem," *Reviews of modern physics*, vol. 47, no. 4, p. 773, 1975.

[7] R. Phillips, J. Kondev, J. Theriot, and H. Garcia, *Physical biology of the cell*. Garland Science, 2012.

[8] R. Milo and R. Phillips, *Cell biology by the numbers*. Garland Science, 2015.

[9] J. N. Onuchic, P. G. Wolynes, Z. Luthey-Schulten, and N. D. Socci, "Toward an outline of the topography of a realistic protein-folding funnel," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 92, no. 8, p. 3626, 1995.

[10] V. S. Pande, A. Y. Grosberg, and T. Tanaka, "Heteropolymer freezing and design: towards physical models of protein folding," *Reviews of Modern Physics*, vol. 72, no. 1, p. 259, 2000.

[11] L. Pauling, R. B. Corey, and H. R. Branson, "The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain," *Proceedings of the National Academy of Sciences*, vol. 37, no. 4, p. 205, 1951.

[12] L. Pauling and R. B. Corey, "Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 37, no. 11, p. 729, 1951.

[13] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, no. 4096, p. 223, 1973.

[14] D. Guin and M. Gruebele, "Weak chemical interactions that drive protein evolution: crowding, sticking, and quinary structure in folding and function," *Chemical reviews*, vol. 119, no. 18, p. 10691, 2019.

[15] C. Levinthal, "Are there pathways for protein folding?," *Journal de chimie physique*, vol. 65, pp. 44–45, 1968.

[16] C. Levinthal, "How to fold graciously," in *Mossbauer Spectroscopy in Biological Systems* (. M. Debrunner, Tsibris, ed.), (Urbana), p. 22, University of Illinois Press, 1969. Proceedings of a Meeting held at Allerton House, Monticello, IL.

[17] K. A. Dill and J. L. MacCallum, "The protein-folding problem, 50 years on," *science*, vol. 338, no. 6110, p. 1042, 2012.

[18] F. H. Stillinger and P. G. Debenedetti, "Glass transition thermodynamics and kinetics," *Annu. Rev. Condens. Matter Phys.*, vol. 4, no. 1, p. 263, 2013.

[19] T. Araki and H. Tanaka, "Nematohydrodynamic effects on the phase separation of a symmetric mixture of an isotropic liquid and a liquid crystal," *Physical review letters*, vol. 93, no. 1, p. 015702, 2004.

[20] P. Hohenberg and J. Swift, "Metastability in fluctuation-driven first-order transitions: Nucleation of lamellar phases," *Physical Review E*, vol. 52, no. 2, p. 1828, 1995.

[21] P.-G. De Gennes and P.-G. Gennes, *Scaling concepts in polymer physics*. Cornell university press, 1979.

[22] P. J. Flory, *Principles of polymer chemistry*. Cornell University Press, 1953.

[23] D. Sherrington and S. Kirkpatrick, "Solvable model of a spin-glass," *Physical review letters*, vol. 35, no. 26, p. 1792, 1975.

[24] S. F. Edwards and P. W. Anderson, "Theory of spin glasses," *Journal of Physics F: Metal Physics*, vol. 5, no. 5, p. 965, 1975.

[25] K. Binder and A. P. Young, "Spin glasses: Experimental facts, theoretical concepts, and open questions," *Reviews of Modern physics*, vol. 58, no. 4, p. 801, 1986.

[26] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, "Funnels, pathways, and the energy landscape of protein folding: a synthesis," *Proteins: Structure, Function, and Bioinformatics*, vol. 21, no. 3, p. 167, 1995.

[27] N. D. Socci and J. N. Onuchic, "Kinetic and thermodynamic analysis of proteinlike heteropolymers: Monte carlo histogram technique," *The Journal of chemical physics*, vol. 103, no. 11, p. 4732, 1995.

[28] D. Thirumalai, E. P. O'Brien, G. Morrison, and C. Hyeon, "Theoretical perspectives on protein folding," *Annual review of biophysics*, vol. 39, p. 159, 2010.

[29] C. Maffi, M. Baiesi, L. Casetti, F. Piazza, and P. De Los Rios, "First-order coil-globule transition driven by vibrational entropy," *Nat. Commun.*, vol. 3, p. 1065, 2012.

[30] P. C. Whitford, K. Y. Sanbonmatsu, and J. N. Onuchic, "Biomolecular dynamics: order–disorder transitions and energy landscapes," *Rep. Prog. Phys.*, vol. 75, no. 7, p. 076601, 2012.

[31] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, "Theory of protein folding: the energy landscape perspective," *Annual review of physical chemistry*, vol. 48, no. 1, p. 545, 1997.

[32] J.-E. Shea, J. N. Onuchic, and C. L. Brooks, "Exploring the origins of topological frustration: design of a minimally frustrated model of fragment b of protein a," *Proceedings of the National Academy of Sciences*, vol. 96, no. 22, p. 12512, 1999.

[33] S. S. Cho, Y. Levy, J. N. Onuchic, and P. G. Wolynes, "Overcoming residual frustration in domain-swapping: the roles of disulfide bonds in dimerization and aggregation," *Physical Biology*, vol. 2, no. 2, p. S44, 2005.

[34] D. U. Ferreiro, J. A. Hegler, E. A. Komives, and P. G. Wolynes, "Localizing frustration in native proteins and protein assemblies," *Proceedings of the National Academy of Sciences*, vol. 104, no. 50, p. 19819, 2007.

[35] R. B. Laughlin, D. Pines, J. Schmalian, B. P. Stojković, and P. Wolynes, "The middle way," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 97, no. 1, p. 32, 2000.

[36] M. Gruebele, K. Dave, and S. Sukenik, "Globular protein folding in vitro and in vivo," *Annu. Rev. Biophys.*, vol. 45, p. 233, 2016.

[37] K. Luby-Phelps, "Cytoarchitecture and physical properties of cytoplasm: volume, viscosity, diffusion, intracellular surface area," in *International review of cytology*, vol. 192, p. 189, Elsevier, 1999.

[38] I. Guzman, H. Gelman, J. Tai, and M. Gruebele, "The extracellular protein vlse is destabilized inside cells," *Journal of molecular biology*, vol. 426, no. 1, p. 11, 2014.

[39] M. Gao, D. Gnutt, A. Orban, B. Appel, F. Righetti, R. Winter, F. Narberhaus, S. Müller, and S. Ebbinghaus, "Rna hairpin folding in the crowded cell," *Angewandte Chemie International Edition*, vol. 55, no. 9, p. 3224, 2016.

[40] I. König, A. Zarrine-Afsar, M. Aznauryan, A. Soranno, B. Wunderlich, F. Dingfelder, J. C. Stüber, A. Plückthun, D. Nettels, and B. Schuler, "Single-molecule spectroscopy of protein conformational dynamics in live eukaryotic cells," *nAture methods*, vol. 12, no. 8, pp. 773–779, 2015.

[41] S. Sukenik, P. Ren, and M. Gruebele, "Weak protein–protein interactions in live cells are quantified by cell-volume modulation," *Proceedings of the National Academy of Sciences*, vol. 114, no. 26, p. 6776, 2017.

[42] J. Danielsson, X. Mu, L. Lang, H. Wang, A. Binolfi, F.-X. Theillet, B. Bekei, D. T. Logan, P. Selenko, H. Wennerström, *et al.*, "Thermodynamics of protein destabilization in live cells," *Proceedings of the National Academy of Sciences*, vol. 112, no. 40, p. 12402, 2015.

[43] X. Mu, S. Choi, L. Lang, D. Mowray, N. V. Dokholyan, J. Danielsson, and M. Oliveberg, "Physicochemical code for quinary protein interactions in escherichia coli," *Proceedings of the National Academy of Sciences*, vol. 114, no. 23, p. E4556, 2017.

[44] W. B. Monteith, R. D. Cohen, A. E. Smith, E. Guzman-Cisneros, and G. J. Pielak, "Quinary structure modulates protein stability in cells," *Proceedings of the National Academy of Sciences*, vol. 112, no. 6, p. 1739, 2015.

[45] A. E. Smith, L. Z. Zhou, A. H. Gorensek, M. Senske, and G. J. Pielak, "In-cell thermodynamics and a new role for protein surfaces," *Proceedings of the National Academy of Sciences*, vol. 113, no. 7, p. 1725, 2016.

[46] I. Yu, T. Mori, T. Ando, R. Harada, J. Jung, Y. Sugita, and M. Feig, "Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm," *Elife*, vol. 5, p. e19274, 2016.

[47] A. P. Minton, "Excluded volume as a determinant of macromolecular structure and reactivity," *Biopolym. Orig. Res. Biomol.*, vol. 20, no. 10, p. 2093, 1981.

[48] L. Stagg, S.-Q. Zhang, M. S. Cheung, and P. Wittung-Stafshede, "Molecular crowding enhances native structure and stability of α/β protein flavodoxin," *Proceedings of the National Academy of Sciences*, vol. 104, no. 48, p. 18976, 2007.

[49] D. Homouz, L. Stagg, P. Wittung-Stafshede, and M. S. Cheung, "Macromolecular crowding modulates folding mechanism of α/β protein apoflavodoxin," *Biophysical Journal*, vol. 96, no. 2, p. 671, 2009.

[50] S. Asakura and F. Oosawa, "On interaction between two bodies immersed in a solution of macromolecules," *J. Chem. Phys.*, vol. 22, no. 7, p. 1255, 1954.

[51] D. Homouz, M. Perham, A. Samiotakis, M. S. Cheung, and P. Wittung-Stafshede, "Crowded, cell-like environment induces shape changes in aspherical protein," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 105, no. 33, p. 11754, 2008.

[52] A. Dhar, A. Samiotakis, S. Ebbinghaus, L. Nienhaus, D. Homouz, M. Gruebele, and M. S. Cheung, "Structure, function, and folding of phosphoglycerate kinase are strongly perturbed by macromolecular crowding," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 107, no. 41, p. 17586, 2010.

[53] H. Kang, P. A. Pincus, C. Hyeon, and D. Thirumalai, "Effects of macromolecular crowding on the collapse of biopolymers," *Phys. Rev. Lett.*, vol. 114, no. 6, p. 068303, 2015.

[54] S. Palit, L. He, W. A. Hamilton, A. Yethiraj, and A. Yethiraj, "Combining diffusion nmr and small-angle neutron scattering enables precise measurements of polymer chain compression in a crowded environment," *Physical Review Letters*, vol. 118, no. 9, p. 097801, 2017.

[55] A. Soranno, I. Koenig, M. B. Borgia, H. Hofmann, F. Zosel, D. Nettels, and B. Schuler, "Single-molecule spectroscopy reveals polymer effects of disordered proteins in crowded environments," *Proceedings of the National Academy of Sciences*, vol. 111, no. 13, p. 4874, 2014.

[56] G. van Anders, D. Klotsa, N. K. Ahmed, M. Engel, and S. C. Glotzer, "Understanding shape entropy through local dense packing," *Proceedings of the National Academy of Sciences*, vol. 111, no. 45, p. E4812, 2014.

[57] M. Spellings, M. Engel, D. Klotsa, S. Sabrina, A. M. Drews, N. H. Nguyen, K. J. Bishop, and S. C. Glotzer, "Shape control and compartmentalization in active colloidal cells," *Proceedings of the National Academy of Sciences*, vol. 112, no. 34, p. E4642, 2015.

[58] H. Frauenfelder, G. Chen, J. Berendzen, P. W. Fenimore, H. Jansson, B. H. McMahon, I. R. Stroe, J. Swenson, and R. D. Young, "A unified model of protein dynamics," *Proceedings of the National Academy of Sciences*, vol. 106, no. 13, p. 5129, 2009.

[59] D. Russo, A. Laloni, A. Filabozzi, and M. Heyden, "Pressure effects on collective density fluctuations in water and protein solutions," *Proceedings of the National Academy of Sciences*, vol. 114, no. 43, p. 11410, 2017.

[60] M. S. Cheung, A. E. García, and J. N. Onuchic, "Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 99, no. 2, p. 685, 2002.

[61] F. C. Zegarra, D. Homouz, Y. Eliaz, A. G. Gasic, and M. S. Cheung, "Impact of hydrodynamic interactions on protein folding rates depends on temperature," *Physical Review E*, vol. 97, no. 3, p. 032402, 2018.

[62] H. Tanaka, "Roles of hydrodynamic interactions in structure formation of soft matter: protein folding as an example," *Journal of Physics: Condensed Matter*, vol. 17, no. 31, p. S2795, 2005.

[63] P.-h. Wang, I. Yu, M. Feig, and Y. Sugita, "Influence of protein crowder size on hydration structure and dynamics in macromolecular crowding," *Chemical Physics Letters*, vol. 671, p. 63, 2017.

[64] U. Raviv, P. Laurat, and J. Klein, "Fluidity of water confined to subnanometre films," *Nature*, vol. 413, no. 6851, p. 51, 2001.

[65] T. Ando and J. Skolnick, "Crowding and hydrodynamic interactions likely dominate in vivo macromolecular motion," *Proceedings of the National Academy of Sciences*, vol. 107, no. 43, p. 18457, 2010.

[66] C. Echeverria and R. Kapral, "Macromolecular dynamics in crowded environments," *The Journal of chemical physics*, vol. 132, no. 10, p. 104902, 2010.

[67] G. F. Reiter, A. Deb, Y. Sakurai, M. Itou, and A. I. Kolesnikov, "Quantum coherence and temperature dependence of the anomalous state of nanoconfined water in carbon nanotubes," *The Journal of Physical Chemistry Letters*, vol. 7, no. 22, p. 4433, 2016.

[68] A. P. Minton, "Explicit incorporation of hard and soft protein–protein interactions into models for crowding effects in protein mixtures. 2. effects of varying hard and soft interactions upon prototypical chemical equilibria," *The Journal of Physical Chemistry B*, vol. 121, no. 22, p. 5515, 2017.

[69] T. Hoppe and A. P. Minton, "Incorporation of hard and soft protein–protein interactions into models for crowding effects in binary and ternary protein mixtures. comparison of approximate analytical solutions with numerical simulation," *The Journal of Physical Chemistry B*, vol. 120, no. 46, p. 11866, 2016.

[70] C. Ota and K. Takano, "Behavior of bovine serum albumin molecules in molecular crowding environments investigated by raman spectroscopy," *Langmuir*, vol. 32, no. 29, p. 7372, 2016.

[71] D. Nilsson, S. Mohanty, and A. Irbäck, "Markov modeling of peptide folding in the presence of protein crowders," *The Journal of Chemical Physics*, vol. 148, no. 5, p. 055101, 2018.

[72] I. A. Shkel, D. Knowles, and M. T. Record Jr, "Separating chemical and excluded volume interactions of polyethylene glycols with native proteins: Comparison with peg effects on dna helix formation," *Biopolymers*, vol. 103, no. 9, p. 517, 2015.

[73] Q. Wang and M. S. Cheung, "A physics-based approach of coarse-graining the cytoplasm of escherichia coli (cgcyto)," *Biophysical journal*, vol. 102, no. 10, p. 2353, 2012.

[74] T. C. Michaels, A. Šarić, J. Habchi, S. Chia, G. Meisl, M. Vendruscolo, C. M. Dobson, and T. P. Knowles, "Chemical kinetics for bridging molecular mechanisms and macroscopic measurements of amyloid fibril formation," *Annual review of physical chemistry*, vol. 69, p. 273, 2018.

[75] B. J. Pieters, M. B. Van Eldijk, R. J. Nolte, and J. Mecinović, "Natural supramolecular protein assemblies," *Chemical Society Reviews*, vol. 45, no. 1, p. 24, 2016.

[76] P. Bridgman, "The coagulation of albumen by pressure," *Journal of Biological Chemistry*, vol. 19, no. 4, p. 511, 1914.

[77] W. Kauzmann, "Advances in protein chemistry (anfinson, jr., anson, ml, bailey, k., and edsall, jt, eds) vol. 14," 1959.

[78] S. A. Hawley, "Reversible pressure-temperature denaturation of chymotrypsinogen," *Biochemistry*, vol. 10, no. 13, p. 2436, 1971.

[79] K. J. Frye and C. A. Royer, "Probing the contribution of internal cavities to the volume change of protein unfolding under pressure," *Protein Science*, vol. 7, no. 10, p. 2217, 1998.

[80] J. Roche, J. A. Caro, D. R. Norberto, P. Barthe, C. Roumestand, J. L. Schlessman, A. E. García, and C. A. Royer, "Cavities determine the pressure unfolding of proteins," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 109, no. 18, p. 6945, 2012.

[81] D. B. Kitchen, L. H. Reed, and R. M. Levy, "Molecular dynamics simulation of solvated protein at high pressure," *Biochemistry*, vol. 31, no. 41, p. 10083, 1992.

[82] L. Meinhold, J. C. Smith, A. Kitao, and A. H. Zewail, "Picosecond fluctuating protein energy landscape mapped by pressure–temperature molecular dynamics simulation," *Proceedings of the National Academy of Sciences*, vol. 104, no. 44, p. 17261, 2007.

[83] D. Paschek and A. E. García, "Reversible temperature and pressure denaturation of a protein fragment: a replica exchange molecular dynamics simulation study," *Physical review letters*, vol. 93, no. 23, p. 238105, 2004.

[84] Y. Liu, M. B. Prigozhin, K. Schulten, and M. Gruebele, "Observation of complete pressure-jump protein refolding in molecular dynamics simulation and experiment," *Journal of the American Chemical Society*, vol. 136, no. 11, p. 4265, 2014.

[85] A. J. Wirth, Y. Liu, M. B. Prigozhin, K. Schulten, and M. Gruebele, "Comparing fast pressure jump and temperature jump protein folding experiments and simulations," *Journal of the American Chemical Society*, vol. 137, no. 22, p. 7152, 2015.

[86] P. L. Freddolino, C. B. Harrison, Y. Liu, and K. Schulten, "Challenges in protein-folding simulations," *Nat. Phys.*, vol. 6, no. 10, p. 751, 2010.

[87] J. D. Bryngelson and P. G. Wolynes, "Spin glasses and the statistical mechanics of protein folding," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 84, no. 21, p. 7524, 1987.

[88] P. E. Leopold, M. Montal, and J. N. Onuchic, "Protein folding funnels: a kinetic approach to the sequence-structure relationship," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 89, no. 18, p. 8721, 1992.

[89] M. S. Cheung and A. G. Gasic, "Towards developing principles of protein folding and dynamics in the cell," *Physical biology*, vol. 15, no. 6, p. 063001, 2018.

[90] M. Boob, Y. Wang, and M. Gruebele, "Proteins:"boil'em, mash'em, stick'em in a stew"," *The Journal of Physical Chemistry B*, vol. 123, no. 40, p. 8341, 2019.

[91] G. Hummer, S. Garde, A. E. García, A. Pohorille, and L. R. Pratt, "An information theory model of hydrophobic interactions," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 93, no. 17, p. 8951, 1996.

[92] G. Hummer, S. Garde, A. E. García, M. E. Paulaitis, and L. R. Pratt, "The pressure dependence of hydrophobic interactions is consistent with the observed pressure denaturation of proteins," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 95, no. 4, p. 1552, 1998.

[93] C. L. Dias and H. S. Chan, "Pressure-dependent properties of elementary hydrophobic interactions: ramifications for activation properties of protein folding," *J. Phys. Chem. B*, vol. 118, no. 27, p. 7488, 2014.

[94] H. S. Ashbaugh, K. Weiss, S. M. Williams, B. Meng, and L. N. Surampudi, "Temperature and pressure dependence of methane correlations and osmotic second virial coefficients in water," *J. Phys. Chem. B*, vol. 119, no. 20, p. 6280, 2015.

[95] T. Ghosh, A. E. García, and S. Garde, "Molecular dynamics simulations of pressure effects on hydrophobic interactions," *J. Am. Chem. Soc.*, vol. 123, no. 44, p. 10997, 2001.

[96] G. Graziano, "Hydrostatic pressure effect on hydrophobic hydration and pairwise hydrophobic interaction of methane," *J. Chem. Phys.*, vol. 140, no. 9, p. 094503, 2014.

[97] M. S. Moghaddam and H. S. Chan, "Pressure and temperature dependence of hydrophobic hydration: Volumetric, compressibility, and thermodynamic signatures," *The Journal of chemical physics*, vol. 126, no. 11, p. 03B613, 2007.

[98] K. Koga and N. Yamamoto, "Hydrophobicity varying with temperature, pressure, and salt concentration," *The Journal of Physical Chemistry B*, vol. 122, no. 13, p. 3655, 2018.

[99] V. Bianco and G. Franzese, "Contribution of water to pressure and cold denaturation of proteins," *Phys. Rev. Lett.*, vol. 115, no. 10, p. 108101, 2015.

[100] V. Bianco, G. Franzese, C. Dellago, and I. Coluzza, "Role of water in the selection of stable proteins at ambient and extreme thermodynamic conditions," *Phys. Rev. X*, vol. 7, no. 2, p. 021047, 2017.

[101] C. L. Dias, "Unifying microscopic mechanism for pressure and cold denaturations of proteins," *Phys. Rev. Lett.*, vol. 109, no. 4, p. 048104, 2012.

[102] E. Van Dijk, P. Varilly, T. P. Knowles, D. Frenkel, and S. Abeln, "Consistent treatment of hydrophobicity in protein lattice models accounts for cold denaturation," *Phys. Rev. Lett.*, vol. 116, no. 7, p. 078101, 2016.

[103] G. Graziano, "On the effect of hydrostatic pressure on the conformational stability of globular proteins," *Biopolymers*, vol. 103, no. 12, p. 711, 2015.

[104] S. Osváth, L. M. Quynh, and L. Smeller, "Thermodynamics and kinetics of the pressure unfolding of phosphoglycerate kinase," *Biochemistry*, vol. 48, no. 42, p. 10146, 2009.

[105] S. Osváth, J. J. Sabelko, and M. Gruebele, "Tuning the heterogeneous early folding dynamics of phosphoglycerate kinase," *J. Mol. Biol.*, vol. 333, no. 1, p. 187, 2003.

[106] A. G. Gasic, M. M. Boob, M. B. Prigozhin, D. Homouz, C. M. Daugherty, M. Gruebele, and M. S. Cheung, "Critical phenomena in the temperature-pressure-crowding phase diagram of a protein," *Physical Review X*, vol. 9, no. 4, p. 041035, 2019.

[107] L. Smeller, "Pressure–temperature phase diagrams of biomolecules," *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, vol. 1595, no. 1-2, p. 11, 2002.

[108] J. Roche and C. A. Royer, "Lessons from pressure denaturation of proteins," *Journal of The Royal Society Interface*, vol. 15, no. 147, p. 20180244, 2018.

[109] C. A. Royer, "Revisiting volume changes in pressure-induced protein unfolding," *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, vol. 1595, no. 1-2, p. 201, 2002.

[110] T. Takekiyo, A. Shimizu, M. Kato, and Y. Taniguchi, "Pressure-tuning ft-ir spectroscopic study on the helix–coil transition of ala-rich oligopeptide in aqueous solution," *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, vol. 1750, no. 1, p. 1, 2005.

[111] R. B. Best, C. Miller, and J. Mittal, "Role of solvation in pressure-induced helix stabilization," *The Journal of chemical physics*, vol. 141, no. 22, p. 12B621_1, 2014.

[112] K. L. Schweiker, V. W. Fitz, and G. I. Makhatadze, "Universal convergence of the specific volume changes of globular proteins upon unfolding," *Biochemistry*, vol. 48, no. 46, p. 10846, 2009.

[113] C. L. Dias, T. Ala-Nissila, M. Karttunen, I. Vattulainen, and M. Grant, "Microscopic mechanism for cold denaturation," *Phys. Rev. Lett.*, vol. 100, no. 11, p. 118101, 2008.

[114] B. J. Sirovetz, N. P. Schafer, and P. G. Wolynes, "Water mediated interactions and the protein folding phase diagram in the temperature–pressure plane," *J. Phys. Chem. B*, vol. 119, no. 34, p. 11416, 2015.

[115] K. Heremans and L. Smeller, "Protein structure and dynamics at high pressure," *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, vol. 1386, no. 2, p. 353, 1998.

[116] S. Reuveni, R. Granek, and J. Klafter, "Proteins: coexistence of stability and flexibility," *Phys. Rev. Lett.*, vol. 100, no. 20, p. 208101, 2008.

[117] P. D. Chowdary and M. Gruebele, "Molecules: what kind of a bag of atoms?," *J. Phys. Chem. A*, vol. 113, no. 47, p. 13139, 2009.

[118] J. K. Cheung, P. Shah, and T. M. Truskett, "Heteropolymer collapse theory for protein folding in the pressure-temperature plane," *Biophysical journal*, vol. 91, no. 7, p. 2427, 2006.

[119] C. R. Chen and G. I. Makhatadze, "Molecular determinant of the effects of hydrostatic pressure on protein folding stability," *Nature communications*, vol. 8, no. 1, p. 1, 2017.

[120] N. Hillson, J. N. Onuchic, and A. E. García, "Pressure-induced protein-folding/unfolding kinetics," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 96, no. 26, p. 14848, 1999.

[121] J. Engstler and N. Giovambattista, "Comparative study of the effects of temperature and pressure on the water-mediated interactions between apolar nanoscale solutes," *J. Phys. Chem. B*, vol. 123, no. 5, p. 1116, 2018.

[122] S. Garde, G. Hummer, A. E. Garcia, M. E. Paulaitis, and L. R. Pratt, "Origin of entropy convergence in hydrophobic hydration and protein folding," *Phys. Rev. Lett.*, vol. 77, no. 24, p. 4966, 1996.

[123] R. Perezzan and A. Rey, "Simulating protein unfolding under pressure with a coarse-grained model," *J. Chem. Phys.*, vol. 137, no. 18, p. 185102, 2012.

[124] S. Sarupria, T. Ghosh, A. E. García, and S. Garde, "Studying pressure denaturation of a protein by molecular dynamics simulations," *Proteins*, vol. 78, no. 7, p. 1641, 2010.

[125] M. Wojciechowski and M. Cieplak, "Coarse-grained modelling of pressure-related effects in staphylococcal nuclease and ubiquitin," *Journal of Physics: Condensed Matter*, vol. 19, no. 28, p. 285218, 2007.

[126] G. Zuo, J. Wang, M. Qin, B. Xue, and W. Wang, "Effect of solvation-related interaction on the low-temperature dynamics of proteins," *Physical Review E*, vol. 81, no. 3, p. 031917, 2010.

[127] K. G. Wilson, "Problems in physics with many scales of length," *Sci. Am.*, vol. 241, no. 2, p. 158, 1979.

[128] J. P. Crutchfield, "Between order and chaos," *Nat. Phys.*, vol. 8, no. 1, p. 17, 2012.

[129] T. Mora and W. Bialek, "Are biological systems poised at criticality?," *J. Stat. Phys.*, vol. 144, no. 2, p. 268, 2011.

[130] M. A. Munoz, "Colloquium: Criticality and dynamical scaling in living systems," *Rev. Mod. Phys.*, vol. 90, no. 3, p. 031001, 2018.

[131] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, "Navigating the folding routes," *Science*, vol. 267, no. 5204, p. 1619, 1995.

[132] M. S. Li, D. K. Klimov, and D. Thirumalai, "Finite size effects on thermal denaturation of globular proteins," *Phys. Rev. Lett.*, vol. 93, no. 26, p. 268107, 2004.

[133] S. S. Cho, P. Weinkam, and P. G. Wolynes, "Origins of barriers and barrierless folding in bbl," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 105, no. 1, p. 118, 2008.

[134] Q.-Y. Tang, Y.-Y. Zhang, J. Wang, W. Wang, and D. R. Chialvo, "Critical fluctuations in the native state of proteins," *Phys. Rev. Lett.*, vol. 118, no. 8, p. 088102, 2017.

[135] See Supplemental Material at [URL will be inserted by publisher] for simulation and experiment details, and further analysis.

[136] O. Miyashita, J. N. Onuchic, and P. G. Wolynes, "Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 100, no. 22, p. 12570, 2003.

[137] S. B. Zimmerman and S. O. Trach, "Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of escherichia coli," *J. Mol. Biol.*, vol. 222, no. 3, p. 599, 1991.

[138] A. P. Minton, "Models for excluded volume interaction between an unfolded protein and rigid macromolecular cosolutes: macromolecular crowding and protein stability revisited," *Biophys. J.*, vol. 88, no. 2, p. 971, 2005.

[139] T. Q. Luong, S. Kapoor, and R. Winter, "Pressure—a gateway to fundamental insights into protein solvation, dynamics, and function," *ChemPhysChem*, vol. 16, no. 17, p. 3555, 2015.

[140] T. G. Dewey, *Fractals in molecular biophysics*. Oxford University Press, 1998.

[141] J.-E. Shea, J. N. Onuchic, and C. L. Brooks, "Probing the folding free energy landscape of the src-sh3 protein domain," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 99, no. 25, p. 16064, 2002.

[142] A. Fernandez-Escamilla, M. Cheung, M. Vega, M. Wilmanns, J. Onuchic, and L. Serrano, "Solvation in protein folding analysis: combination of theoretical and experimental approaches," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 101, no. 9, p. 2834, 2004.

[143] K. Luby-Phelps, P. E. Castle, D. L. Taylor, and F. Lanni, "Hindered diffusion of inert tracer particles in the cytoplasm of mouse 3t3 cells," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 84, no. 14, p. 4910, 1987.

[144] D. Venturoli and B. Rippe, "Ficoll and dextran vs. globular proteins as probes for testing glomerular permselectivity: effects of molecular size, shape, charge, and deformability," *Am. J. Physiol. Physiol.*, vol. 288, no. 4, p. F605, 2005.

[145] M. R. Shaw and D. Thirumalai, "Free polymer in a colloidal solution," *Phys. Rev. A*, vol. 44, no. 8, p. R4797, 1991.

[146] P. W. Fenimore, H. Frauenfelder, B. H. McMahon, and F. G. Parak, "Slaving: solvent fluctuations dominate protein dynamics and functions," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 99, no. 25, p. 16047, 2002.

[147] I. Lifshitz, A. Y. Grosberg, and A. Khokhlov, "Some problems of the statistical physics of polymer chains with volume interaction," *Rev. Mod. Phys.*, vol. 50, no. 3, p. 683, 1978.

[148] D. Thirumalai, "Isolated polymer molecule in a random environment," *Phys. Rev. A*, vol. 37, no. 1, p. 269, 1988.

[149] P. I. Zhuravlev, M. Hinczewski, S. Chakrabarti, S. Marqusee, and D. Thirumalai, "Force-dependent switch in protein unfolding pathways and transition-state movements," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 113, no. 6, p. E715, 2016.

[150] C. A. Pierse and O. K. Dudko, "Distinguishing signatures of multipathway conformational transitions," *Phys. Rev. Lett.*, vol. 118, no. 8, p. 088101, 2017.

[151] L. D. Landau, E. M. Lifshitz, and L. P. Pitaevskii, *Course of Theoretical Physics: Statistical Physics Pt. 1.* Pergamon P., 1980.

[152] S. F. Edwards, "The statistical mechanics of polymers with excluded volume," *Proc. Phys. Soc.*, vol. 85, no. 4, p. 613, 1965.

[153] S. F. Edwards and P. Singh, "Size of a polymer molecule in solution. part 1.—excluded volume problem," *J. Chem. Soc., Faraday Trans 2: Mol. Chem. Phys.*, vol. 75, p. 1001, 1979.

[154] M. P. Taylor, W. Paul, and K. Binder, "All-or-none proteinlike folding transition of a flexible homopolymer chain," *Phys. Rev. E*, vol. 79, no. 5, p. 050801, 2009.

[155] P. G. Wolynes, "Folding funnels and energy landscapes of larger proteins within the capillarity approximation," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 94, no. 12, p. 6170, 1997.

[156] J. Sabelko, J. Ervin, and M. Gruebele, "Observation of strange kinetics in protein folding," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 96, no. 11, p. 6031, 1999.

[157] K. Binder, "Finite size scaling analysis of ising model block distribution functions," *Zeitschrift für Phys. B Condens. Matter*, vol. 43, no. 2, p. 119, 1981.

[158] K. G. Wilson, "Renormalization group and critical phenomena. i. renormalization group and the kadanoff scaling picture," *Phys. Rev. B*, vol. 4, no. 9, p. 3174, 1971.

[159] A. J. Wirth and M. Gruebele, "Quinary protein structure and the consequences of crowding in living cells: Leaving the test-tube behind," *BioEssays*, vol. 35, no. 11, p. 984, 2013.

[160] M. Silow and M. Oliveberg, "Transient aggregates in protein folding are easily mistaken for folding intermediates," *Proceedings of the National Academy of Sciences*, vol. 94, no. 12, p. 6084, 1997.

[161] D. U. Ferreiro, C. F. Cervantes, S. M. Truhlar, S. S. Cho, P. G. Wolynes, and E. A. Komives, "Stabilizing iκbα by "consensus" design," *Journal of molecular biology*, vol. 365, no. 4, p. 1201, 2007.

[162] M. E. Zweifel and D. Barrick, "Studies of the ankyrin repeats of the drosophila melanogaster notch receptor. 2. solution stability and cooperativity of unfolding," *Biochemistry*, vol. 40, no. 48, p. 14357, 2001.

[163] C. H. Croy, S. Bergqvist, T. Huxford, G. Ghosh, and E. A. Komives, "Biophysical characterization of the free iκbα ankyrin repeat domain in solution," *Protein Science*, vol. 13, no. 7, p. 1767, 2004.

[164] M. B. Borgia, A. Borgia, R. B. Best, A. Steward, D. Nettels, B. Wunderlich, B. Schuler, and J. Clarke, "Single-molecule fluorescence reveals sequence-specific misfolding in multidomain proteins," *Nature*, vol. 474, no. 7353, p. 662, 2011.

[165] C. F. Wright, S. A. Teichmann, J. Clarke, and C. M. Dobson, "The importance of sequence diversity in the aggregation and evolution of proteins," *Nature*, vol. 438, no. 7069, p. 878, 2005.

[166] E. R. Main, S. E. Jackson, and L. Regan, "The folding and design of repeat proteins: reaching a consensus," *Current opinion in structural biology*, vol. 13, no. 4, p. 482, 2003.

[167] T. Aksel and D. Barrick, "Direct observation of parallel folding pathways revealed using a symmetric repeat protein system," *Biophysical journal*, vol. 107, no. 1, p. 220, 2014.

[168] F. Liu and M. Gruebele, "Mapping an aggregation nucleus one protein at a time," *The Journal of Physical Chemistry Letters*, vol. 1, no. 1, p. 16, 2010.

[169] X. Chen, J. L. Zaro, and W.-C. Shen, "Fusion protein linkers: property, design and functionality," *Advanced drug delivery reviews*, vol. 65, no. 10, p. 1357, 2013.

[170] A. Borgia, K. R. Kemplen, M. B. Borgia, A. Soranno, S. Shammas, B. Wunderlich, D. Nettels, R. B. Best, J. Clarke, and B. Schuler, "Transient misfolding dominates multidomain protein folding," *Nature communications*, vol. 6, no. 1, p. 1, 2015.

[171] H. S. Samanta, P. I. Zhuravlev, M. Hinczewski, N. Hori, S. Chakrabarti, and D. Thirumalai, "Protein collapse is encoded in the folded state architecture," *Soft Matter*, vol. 13, no. 19, p. 3622, 2017.

[172] M. E. Cusick, N. Klitgord, M. Vidal, and D. E. Hill, "Interactome: gateway into systems biology," *Human molecular genetics*, vol. 14, no. suppl_2, pp. R171–R181, 2005.

[173] H. Wu and M. Fuxreiter, "The structure and dynamics of higher-order assemblies: amyloids, signalosomes, and granules," *Cell*, vol. 165, no. 5, p. 1055, 2016.

[174] S. Kühner, V. van Noort, M. J. Betts, A. Leo-Macias, C. Batisse, M. Rode, T. Yamada, T. Maier, S. Bader, P. Beltran-Alvarez, *et al.*, "Proteome organization in a genome-reduced bacterium," *Science*, vol. 326, no. 5957, p. 1235, 2009.

[175] J. D. O'Connell, A. Zhao, A. D. Ellington, and E. M. Marcotte, "Dynamic reorganization of metabolic enzymes into intracellular bodies," *Annual review of cell and developmental biology*, vol. 28, p. 89, 2012.

[176] P. Chien and L. M. Gierasch, "Challenges and dreams: physics of weak interactions essential to life," *Molecular biology of the cell*, vol. 25, no. 22, p. 3474, 2014.

[177] H.-X. Zhou, G. Rivas, and A. P. Minton, "Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences," *Annu. Rev. Biophys.*, vol. 37, p. 375, 2008.

[178] W. M. Jacobs and D. Frenkel, "Predicting phase behavior in multicomponent mixtures," *The Journal of chemical physics*, vol. 139, no. 2, p. 024108, 2013.

[179] P. Sartori and S. Leibler, "Lessons from equilibrium statistical physics regarding the assembly of protein complexes," *Proceedings of the National Academy of Sciences*, vol. 117, no. 1, p. 114, 2020.

[180] H. C. Nguyen, R. Zecchina, and J. Berg, "Inverse statistical problems: from the inverse ising problem to data science," *Advances in Physics*, vol. 66, no. 3, p. 197, 2017.

[181] F. J. O'Reilly and J. Rappsilber, "Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology," *Nature structural & molecular biology*, vol. 25, no. 11, p. 1000, 2018.

[182] R. P. Sear and J. A. Cuesta, "Instabilities in complex mixtures with a large number of components," *Physical review letters*, vol. 91, no. 24, p. 245701, 2003.

[183] W. M. Jacobs and D. Frenkel, "Phase transitions in biological systems with many components," *Biophysical journal*, vol. 112, no. 4, p. 683, 2017.

[184] S. F. Banani, H. O. Lee, A. A. Hyman, and M. K. Rosen, "Biomolecular condensates: organizers of cellular biochemistry," *Nature reviews Molecular cell biology*, vol. 18, no. 5, p. 285, 2017.

[185] A. Murugan, Z. Zeravcic, M. P. Brenner, and S. Leibler, "Multifarious assembly mixtures: Systems allowing retrieval of diverse stored structures," *Proceedings of the National Academy of Sciences*, vol. 112, no. 1, p. 54, 2015.

[186] M. Wang, C. J. Herrmann, M. Simonovic, D. Szklarczyk, and C. von Mering, "Version 4.0 of paxdb: protein abundance data, integrated across model organisms, tissues, and cell-lines," *Proteomics*, vol. 15, no. 18, p. 3163, 2015.

[187] B. Zhang and P. G. Wolynes, "Topology, structures, and energy landscapes of human chromosomes," *Proceedings of the National Academy of Sciences*, vol. 112, no. 19, p. 6062, 2015.

[188] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, "Identification of direct residue contacts in protein–protein interaction by message passing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 1, p. 67, 2009.

[189] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, "Weak pairwise correlations imply strongly correlated network states in a neural population," *Nature*, vol. 440, no. 7087, p. 1007, 2006.

[190] C. Clementi, H. Nymeyer, and J. N. Onuchic, "Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? an investigation for small globular proteins," *J. Mol. Biol.*, vol. 298, no. 5, p. 937, 2000.

[191] G. A. Papoian, *Coarse-Grained Modeling of Biomolecules*. CRC Press, 2017.

[192] J. Hu, T. Chen, M. Wang, H. S. Chan, and Z. Zhang, "A critical comparison of coarse-grained structure-based approaches and atomic models of protein folding," *Phys. Chem. Chem. Phys.*, vol. 19, no. 21, p. 13629, 2017.

[193] H. J. Berendsen, D. van der Spoel, and R. van Drunen, "Gromacs: a message-passing parallel molecular dynamics implementation," *Comput. Phys. Commun*, vol. 91, no. 1-3, p. 43, 1995.

[194] N. Goga, A. Rzepiela, A. De Vries, S. Marrink, and H. Berendsen, "Efficient algorithms for langevin and dpd dynamics," *J. Chem. Theory Comput.*, vol. 8, no. 10, p. 3637, 2012.

[195] K. Hukushima and K. Nemoto, "Exchange monte carlo method and application to spin glass simulations," *J. Phys. Soc. Jpn.*, vol. 65, no. 6, p. 1604, 1996.

[196] Y. Sugita and Y. Okamoto, "Replica-exchange molecular dynamics method for protein folding," *Chem. Phys. Lett.*, vol. 314, no. 1-2, p. 141, 1999.