Understanding the Structure and Formation of Protein Crystals Using Computer Simulation and Theory

by

Irem Altan

Department of Chemistry
Duke University

Date: __________________________
Approved:

Patrick Charbonneau, Advisor

David Beratan

Dewey G. McCafferty

Scott C. Schmidler

Kevin Welsher

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Chemistry in the Graduate School of Duke University 2019
Abstract

Understanding the Structure and Formation of Protein Crystals Using Computer Simulation and Theory

by

Irem Altan

Department of Chemistry
Duke University

Date: ______________________

Approved:

__________________________
Patrick Charbonneau, Advisor

__________________________
David Beratan

__________________________
Dewey G. McCafferty

__________________________
Scott C. Schmidler

__________________________
Kevin Welsher

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Chemistry in the Graduate School of Duke University 2019
Abstract

The complexity of protein-protein interactions enables proteins to self-assemble into a rich array of structures, such as virus capsids, amyloid fibers, amorphous aggregates, and protein crystals. While some of these assemblies form under biological conditions, protein crystals, which are crucial for obtaining protein structures from diffraction methods, do not typically form readily. Crystallizing proteins thus requires significant trial and error, limiting the number of structures that can be obtained and studied. Understanding how proteins interact with one another and with their environment would allow us to elucidate the physicochemical processes that lead to crystal formation and provide insight into other self-assembly phenomena. This thesis explores this problem from a soft matter theory and simulation perspective.

We first attempt to reconstruct the water structure inside a protein crystal using all-atom molecular dynamics simulations with the dual goal of benchmarking empirical water models and increasing the information extracted from X-ray diffraction data. We find that although water models recapitulate the radial distribution of water around protein atoms, they fall short of reproducing its orientational distribution. Nevertheless, high-intensity peaks in water density are sufficiently well captured to detect the protonation states of certain solvent-exposed residues.

We next study a human \( \gamma \)D-crystallin mutant, the crystals of which have inverted solubility. We parameterize a patchy particle and show that the temperature-dependence of the patch that contains the solubility inverting mutation reproduces
the experimental phase diagram. We also consider the hypothesis that the solubility is inverted because of increased surface hydrophobicity, and show that even though this scenario is thermodynamically plausible, microscopic evidence for it is lacking, partly because our understanding of water as a biomolecular solvent is limited.

Finally, we develop computational methods to understand the self-assembly of a two-dimensional protein crystal and show that specialized Monte Carlo moves are necessary for proper sampling.
# Contents

Abstract iv

List of Tables x

List of Figures xi

Acknowledgements xiii

## 1 Introduction 1

1.1 Protein Self-Assembly and Crystallization 1

1.2 Typical Protein Phase Diagrams 2

1.3 Biomolecular Solvation 4

1.4 Outline 5

## 2 Methods 6

2.1 Introduction 6

2.2 Materials 7

2.3 Methods 9

2.3.1 Effective Protein-Protein Interactions Through Umbrella Sampling 9

2.3.2 Phase Diagram 14

2.4 Notes 26

## 3 Water in Protein Crystals 30

3.1 Introduction 30
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2</td>
<td>Methods</td>
<td>34</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Protein and Setup</td>
<td>34</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Water Models</td>
<td>35</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Molecular Dynamics Simulations</td>
<td>36</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Analyzing the Solvent Structure</td>
<td>39</td>
</tr>
<tr>
<td>3.2.5</td>
<td>Inferring Protonation States</td>
<td>43</td>
</tr>
<tr>
<td>3.3</td>
<td>Results and Discussion</td>
<td>45</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Real-Space Comparison of Water Densities</td>
<td>45</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Reciprocal Space</td>
<td>49</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Inferring Protonation States</td>
<td>51</td>
</tr>
<tr>
<td>3.4</td>
<td>Conclusion</td>
<td>54</td>
</tr>
<tr>
<td>4</td>
<td>Parameterizing a Patchy Model to Understand Inverted Solubility in γD-Crystallin</td>
<td>57</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>58</td>
</tr>
<tr>
<td>4.2</td>
<td>Materials and Methods</td>
<td>61</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Preparation and Characterization of Double Mutant</td>
<td>61</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Crystallization and Data Collection</td>
<td>61</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Solubility Measurements</td>
<td>62</td>
</tr>
<tr>
<td>4.2.4</td>
<td>Data Processing and Structure Solution</td>
<td>62</td>
</tr>
<tr>
<td>4.2.5</td>
<td>Description of the Model</td>
<td>63</td>
</tr>
<tr>
<td>4.2.6</td>
<td>Phase Diagram Determination</td>
<td>65</td>
</tr>
<tr>
<td>4.3</td>
<td>Results and Discussion</td>
<td>65</td>
</tr>
<tr>
<td>4.4</td>
<td>Conclusions</td>
<td>71</td>
</tr>
<tr>
<td>4.5</td>
<td>Simulation Methods</td>
<td>72</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Calculation of Model Parameters</td>
<td>72</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Model Geometry</td>
<td>75</td>
</tr>
</tbody>
</table>
5 Testing the Hydrophobicity Scenario as a Microscopic Basis for Inverted Solubility in \( \gamma \)-D-Crystallin

5.1 Introduction

5.2 Experimental Context for Hydrophobicity

5.3 Theory and Computational Details

5.3.1 Patchy Model

5.3.2 Inverted Solubility Models

5.3.3 Crystal Solubility Determination

5.4 Results and Discussion

5.4.1 Inverted Solubility from Hydrophobicity Models

5.4.2 Solubility Lines for Temperature Deactivated Patches

5.5 Conclusion

6 Coarse-Grained Models to Understand Self-Assembly and Stability of 2D-Crystals

6.1 Introduction

6.2 Methods

6.2.1 Patchy Model and Self-Assembly

6.2.2 Specialized Monte Carlo Moves

6.2.3 The Simplest Model for a Crystalline Membrane

6.3 Results

6.3.1 Performance Increase due to Wave Moves

6.3.2 Consistency Check for Area Moves

6.3.3 Crumpling Transition in the Triangular Mesh Model

6.4 Outlook

7 Conclusions
List of Tables

2.1 Predictor-corrector algorithms ........................................... 25
3.1 Default vs inverted protonation states ................................. 44
4.1 Data Collection and Refinement Statistics ............................. 62
4.2 Patch parameters .......................................................... 78
4.3 Patch geometry ............................................................. 78
4.4 Geometry of crystal unit cells ........................................... 79
# List of Figures

1.1 A typical protein phase diagram .............................. 3
2.1 The vicious cycle of protein crystallization .................. 8
2.2 Umbrella sampling windows .................................... 10
2.3 Histograms generated with WHAM and the resulting PMF ...... 14
2.4 Schematic of the patchy model ................................. 16
2.5 The fluid equation of state and integration from an ideal gas .... 20
2.6 Integration from an ideal Einstein crystal and finding the fluid-crystal coexistence ........................................ 24
2.7 Evolution of densities of the liquid and gas phases in Gibbs Ensemble simulations ............................................ 27
2.8 The final rubredoxin phase diagram ............................ 28
3.1 Typical water models that are used in biomolecular simulations .... 31
3.2 Averaged RDFs for surface N, O, and C atoms from different water models ...................................................... 46
3.3 Radial, averaged radial, angular, and real-space correlation coefficients 48
3.4 Recall and prediction scores for peaks in water density ........... 49
3.5 Comparison of water densities in reciprocal space ............... 51
3.6 Comparison of water density distribution for different protonation states 56
4.1 Experimental phase diagram for the P23T + R36S mutant of HGD .... 66
4.2 Interactions of Thr23 and Ser36 in the crystal contacts ............ 67
4.3 Interactions and flexibility of the C-terminal domain ............. 69
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4</td>
<td>Theoretical phase diagram obtained from patchy particles</td>
<td>71</td>
</tr>
<tr>
<td>5.1</td>
<td>Hydrophobicity estimates for different crystal structures of single and double mutants of human γD-crystallin</td>
<td>87</td>
</tr>
<tr>
<td>5.2</td>
<td>Schematic of the patchy model with torsion</td>
<td>89</td>
</tr>
<tr>
<td>5.3</td>
<td>Arrangement of patches for the HGD double mutant</td>
<td>90</td>
</tr>
<tr>
<td>5.4</td>
<td>The effect of assuming all bonds are active in the crystal on the resulting chemical potentials</td>
<td>96</td>
</tr>
<tr>
<td>5.5</td>
<td>Plot that illustrates the $n_w$ parameter in the MLG model required for inverted solubility</td>
<td>100</td>
</tr>
<tr>
<td>5.6</td>
<td>Theoretical phase diagram obtained with the MLG model</td>
<td>101</td>
</tr>
<tr>
<td>5.7</td>
<td>Theoretical phase diagrams obtained with the Wentzel-Gunton model</td>
<td>102</td>
</tr>
<tr>
<td>5.8</td>
<td>Theoretical phase diagram using the temperature-deactivated patchy model, with various sets of perturbed patch energies</td>
<td>104</td>
</tr>
<tr>
<td>5.9</td>
<td>The effect of changing $\tau$ on the DBI solubility line and the possibility of a temperature-independent solubility line</td>
<td>105</td>
</tr>
<tr>
<td>5.10</td>
<td>The liquid-liquid binodal regions for the temperature-deactivated patch model for various Patch 4 energies, $\varepsilon_4$</td>
<td>106</td>
</tr>
<tr>
<td>6.1</td>
<td>The model of the protein that assembles into sheets</td>
<td>111</td>
</tr>
<tr>
<td>6.2</td>
<td>Autocorrelation function for a small $q$ out-of-plane fluctuation mode, with and without wave moves</td>
<td>115</td>
</tr>
<tr>
<td>6.3</td>
<td>Imposed frame tension vs. measured surface tension</td>
<td>115</td>
</tr>
<tr>
<td>6.4</td>
<td>Crumpling transition for the triangular mesh model</td>
<td>117</td>
</tr>
<tr>
<td>6.5</td>
<td>System size dependence of radius of gyration at $\kappa = 0$ and $\kappa = 2$</td>
<td>117</td>
</tr>
</tbody>
</table>
Acknowledgements

I would not be able to write this dissertation had it not been for the support and guidance of several people. Firstly, I would like to thank my advisor Patrick Charbonneau. It is because of his guidance and support that I have become a better researcher and will get the chance to improve even further. I am also thankful that he has included me in many teaching/outreach activities, allowing me to learn about the science of food.

I also thank Diana Fusco, not only for her friendship but also for being incredibly patient in answering my questions even back when I did not know how to ask questions properly.

I was lucky to have terrific friends in my cohort who always supported me: Anastasia Gant, Anita Donlic, and Steven Conklin. I could not have asked for better (and crazier) comrades.

I could not have done it without Aslı Cansunar, who was always there to guide me through major, as well as minor and silly, life decisions and to prevent me from getting anxious and nervous about things I have absolutely no control over, such as the Sun becoming a red giant in about 5 billion years. She has been doing so since the first day we stepped on the basketball court together almost ten years ago.

Together with Ash, Necdet Emir Demirtaş, Serkant Adigüzel, and Arlo Cansunar have been my family away from home. Thank you for your continuous strife of getting me to live more spontaneously.
I also thank Utku Cansu for his support and being able to philosophize even the most pedestrian things, such as a game of table tennis.

I thank Agostino Migliore and Athraa Zaki; Deniz and Bülent Ender; as well as Kerem and Merve Yıldırım for their friendship and support over the years.

Without the support of Burcu “BB” Beykal, Cemre Özmenci, and Miray Yılmaz, I would have failed in a multitude of unimaginable ways. Their friendship and support have spanned numerous continents and timezones.

I have to thank the old Mesoamerican civilizations for coming up with the bizarre idea of cutting open cacao fruits and fermenting their seeds.

I thank all of my teachers and professors over the years, particularly “Sir” Ersin Yurtsever, my undergraduate advisor. Without his vision and guidance, I may not even have made it into a Ph.D. program.

Last but not least, I thank my family for always and unconditionally supporting me. My parents Neslihan and Sarper Altan have worked selflessly to help me get where I am, and perhaps more importantly, given me my sense of humor. My sister, Gizem Altan, has been helping me become a better person since 1994. My aunt, Selen Altan, has been my number one supporter since day one (i.e., May 29, 1991). My grandmother, Sevinç Altan, was the original scientist in our family, and it is thanks to her that I grew up with interest in science, and never for a moment questioned whether I could become a scientist when I grew up. I am forever indebted to them.
1

Introduction

1.1 Protein Self-Assembly and Crystallization

Proteins self-assemble into a diverse set of structures, such as virus capsids [78], amyloid fibers [161, 45], amorphous aggregates [58], and protein crystals. This richness results from the complexity of protein-protein interactions. While some of these assemblies carry out a biological function, such as virus capsids and the S-crystallin gel that control the refractive index of the squid eye lens [38], others result from changing the chemical environment, or from mutations, and can lead to diseases. For instance, amyloid fibers have been linked with neurodegenerative troubles like Alzheimer’s disease, and the aggregation or crystallization of eye lens proteins results in cataracts disease [90]. Thus, elucidating the types of protein-protein interactions, as well as solution conditions that lead to self-assembly generally informs disease-causing assemblies.

Protein crystals, by contrast, typically emerge in extreme chemical conditions [62]. Proteins have indeed evolved to carry out their biological function in a non-crystalline form, and crystallization is typically deleterious to their biological function. Protein
crystals are nonetheless crucial for protein structure determination, as X-ray diffraction is still the primary tool in this context. Two other methods that do not require crystals are also used, NMR and cryoEM [13], although the former is limited to smaller proteins [76, 179], and the latter has yet to reach resolutions comparable to X-ray diffraction for globular proteins.

Finding the solution conditions that are suitable for protein crystal formation is still mainly a brute-force effort. Typically, on the order of a thousand different chemical conditions are first tested, and these conditions are then further refined to improve the quality of the obtained crystals [7]. The origin of this difficulty is twofold: (i) interactions that lead to protein crystals are typically weak, and (ii) there is a large number of chemical conditions that can be adjusted [120]. Elucidating physicochemical processes that lead to protein crystallization would help us understand how the microscopic details of protein-protein interactions translate to protein phase behavior, and thus would lead to the design of more specific and rational screens. It would also pave the way to design self-assembling biomaterials by engineering protein-protein interactions [168, 173]. The study of protein crystallization thus has broad implications in understanding protein self-assembly.

1.2 Typical Protein Phase Diagrams

A protein solution behavior depends on many factors, including temperature, protein concentration, ionic strength, pH, and the concentration of various additives. It is nonetheless possible to represent the phase diagram of a protein on a temperature-density phase diagram, accounting for the effect of additives and solvent implicitly. The resulting typical protein phase diagram is surprisingly similar to that of a simple fluid (Fig. 1.1). The main difference is that in proteins, the critical temperature is metastable with respect to the solubility line. This is a consequence of the narrowness of the protein-protein interaction range relative to the protein size. Inside the gas-
liquid binodal proteins often form a metastable, yet long-lived gel that prevents crystal formation. This phase separation is a direct consequence of the short range of attraction because as the temperature is decreased, particles rearrange to reduce the energy, which has a high entropic cost due to the narrow interaction range. The system thus minimizes the free energy by forming a low-density phase that maximizes the entropy, alongside the high-density phase. The region under the solubility line but above the critical temperature is called the nucleation zone, as supersaturating the protein solution in this regime is conducive to crystal nucleation without the risk of phase separation. For a generic system, knowing the protein phase diagram would allow for determining the optimum conditions for crystal formations. Unfortunately, experimental phase diagrams have been obtained for only a handful of proteins [23, 121, 89, 67, 75].

![Figure 1.1](image)

**Figure 1.1**: (a) Phase diagram of a simple fluid [79], and (b) a typical protein phase diagram.

Coarse-grained models of proteins provide a computationally tractable way of sketching such phase diagrams. In addition, they are an invaluable tool for pinpointing which microscopic details of protein-protein interactions are relevant to the protein phase behavior. Simple globular proteins can be modeled by hard spheres with short-ranged, anisotropic interactions. The interaction range causes the critical point to become metastable with respect to the solubility line, as expected from 3
experiments, while the anisotropy is necessary to explain the low packing fraction of protein crystals and to account for the asymmetry of the gas-liquid binodal \cite{113}. Modeling atypical proteins require other modifications to this simple model, such as shape anisotropy for the case of virus capsids \cite{78} and amyloids \cite{161}, or temperature-dependence for the case of proteins with inverted solubility \cite{165,194} (see Chapters 4 and 5). It has also been shown that shape anisotropy can give rise to striking phase behaviors, such as a closed-loop binodal \cite{69}.

Patchy models can be parameterized starting from a crystal structure, from which the protein phase diagram can be obtained. This method has been shown to produce phase diagrams that are consistent with the conditions at which the proteins were crystallized \cite{65}. While it might seem counterproductive for protein structures to be needed to generate phase diagrams that guide protein crystallization, this exercise, along with a survey of protein-protein interactions in the crystal contacts, allows for making sense of the observed phase behavior in terms of these interactions.

1.3 Biomolecular Solvation

Water is the medium in which all of these protein self-assembly phenomena take place. Because water molecules often directly interact with proteins and affect protein-protein interactions \cite{100} and function \cite{190}, accounting for water-protein interactions is imperative for fully grasping the microscopic causes of protein phase behavior. In order for protein crystals to form, most of the water molecules solvating attractive patches need to be displaced to the bulk, typically creating an entropic cost for crystal formation. After the crystal has been formed, some of the solvating water molecules remain in crystal contacts and mediate the interaction between polar and charged residues of two protein copies.

It has also been argued that the presence of hydrophobic patches on the surface of the protein can cause a net entropy gain and stabilize the crystal upon moving
the solvating water molecules to the bulk [165, 194, 138]. The thermodynamics of this argument is founded in experimental observations and theoretical predictions regarding hydrocarbon solvation [125]. It is unclear, however, whether microscopic arguments made about water solvating fully hydrophobic solutes [128, 108] are also valid for only partially hydrophobic protein surface residues [147].

Understanding water-protein interactions is also significant for protein structure determination through X-ray and neutron crystallography, as water typically occupies around 50% of the unit cell by volume, and thus contributes significantly to the diffraction signal [192, 4]. Because protein structure solution involves iterative refinement of a model of the protein unit cell, a poor representation of the water structure limits the accuracy of the final protein structure [84]. This in turn puts limits on our investigation of how the protein structure dictates its phase behavior.

1.4 Outline

In Chapter 2, the methodology for obtaining protein phase diagrams starting from the all-atom representation is briefly described. Chapter 3 explores how protein-water interactions are vital in obtaining structural information from X-ray diffraction data of protein crystals. This is done by comparing the structure of water within the unit cell of a protein derived from X-ray diffraction to that generated from molecular dynamics (MD) simulations. In Chapter 4, we describe a patchy model for a protein with inverted solubility and obtain the solubility lines for this protein with simulations. Chapter 5 further analyzes modifications to this model in attempts to elucidate the microscopic mechanism responsible for the inversion of solubility. Finally, Chapter 6 presents the methodological aspects and tests for the simulation of a two-dimensional protein crystal.
2

Methods

Reprinted (adapted) with permission from


2.1 Introduction

While all-atom simulations of a single solvated protein are now fairly run-of-the-mill, simulating protein crystallization in a similar way is far beyond computational reach. The operation would require simulating hundreds of copies of the macromolecule, over very long timescales. In addition, such simulations would contain so much information that teasing out the relevant physico-chemical features that drive crystal assembly would itself be challenging. To circumvent these obstacles, coarse-grained models are used to capture protein-protein interactions in an effective manner, and
thus to hide from view (and from computations) most of the obfuscating details. The key operations for such coarse-graining are: (i) determining and characterizing the relevant features of protein-protein interactions, and (ii) solving the properties of the resulting effective model. The first is done using all-atom simulations, and the second with the coarse-grained model alone. From a conceptual viewpoint, the key difficulties consist of identifying the relevant features and of choosing an appropriate coarse-graining scheme. Here, we describe a procedure developed over the last few years that focuses on crystal contacts between proteins for the former, and uses patchy models for the latter.

An attentive reader will notice that this approach gives information about the crystal assembly of a protein of known structure, and thus that has presumably already been crystallized. This is the vicious cycle of protein crystallization (see Fig. 2.1). Despite this obvious drawback for the study of a specific protein, such a scheme can be employed to understand the generic features that control protein crystallization, and thus the different classes of macromolecular assembly.

In this chapter, we detail the methodology for obtaining effective representations of protein-protein interactions, patchy models, and the steps involved in determining their phase diagram. For each of these tasks, we also briefly summarize the underlying theory. We conclude by discussing common complications and workarounds.

2.2 Materials

It may sound paradoxical but the most essential information needed to study protein crystal assembly is the protein crystal structure itself (Fig. 2.1). While high quality crystal structures are preferable, if the positions of some of the side chains cannot be resolved, it is still possible to add them on using tools such as KiNG [43], and then minimize the energy of the resulting configuration in order to avoid steric clashes. Once that has been resolved, the expensive computational work can begin.
Figure 2.1: The vicious cycle of protein crystallization. The phase diagram of a given patchy model can be straightforwardly determined. The resulting phase information can in turn be used to optimize crystallization screens to reliably obtain protein crystals. However, constructing patchy models requires knowing protein-protein interactions, for which the protein structure itself is needed. Going through this process nevertheless results in a better understanding of how the microscopic properties of the protein control its phase behavior.

In order to run all-atom simulations, molecular dynamics (MD) packages, such as Gromacs [21] or Amber [39], are essential. These packages include a variety of protein force fields and water models and are designed for sharing the computational load with graphical processing units (GPUs). This is especially useful for simulating systems that contain a large number of non-bonded interactions, e.g., interactions that involve solvent molecules.

Once model parameters have been determined from all-atom MD simulations, the remainder of the work uses Monte Carlo (MC) simulations. Note that no generic MC code distribution is widely available, but the relevant methods for patchy models can be straightforwardly implemented based on Ref. [61]. Rovigatti et al. have also recently published a detailed review of the specific MC methods used for simulating patchy particles, along with an educational package for performing various such
2.3 Methods

In this section we first describe how effective protein-protein interactions are obtained (Sec. 2.3.1), and how the nature of these interactions leads to a minimal patchy model. We then detail the process of obtaining the phase diagram of this model (Sec. 2.3.2). For the sake of concreteness, we use Gromacs for the first step and illustrate the overall process with a specific rubredoxin mutant [65] (Protein data bank [24] ID: 1YK4 [32]).

2.3.1 Effective Protein-Protein Interactions Through Umbrella Sampling

The change in free energy upon forming or destroying a protein-protein contact is determined from simulations that mimic experimental conditions and thus include solvent molecules and ions. In general, the free energy difference between two states along a reaction coordinate, \( \xi \), is called the potential of mean force (PMF). For protein-protein interactions in particular, one needs to determine the PMF as a function of distance between two proteins, given a specific crystal contact. The natural choice for the reaction coordinate is then the protein-protein distance.

At equilibrium, the probability that the system is found at a given \( \xi \) is

\[
Q(\xi)d\xi = \frac{\int \delta(\xi(r^N) - \xi)e^{-\beta U(r^N)}dr^N}{\int e^{-\beta U(r^N)}dr^N}d\xi,
\]

where \( \delta \) is the Dirac delta function, \( r^N \) denotes the coordinates of the \( N \) particles of the system, and \( U(r^N) \) is the potential energy of a given configuration. This configuration is observed with a probability proportional to its Boltzmann weight, i.e., \( e^{-\beta U(r^N)} \) at inverse temperature \( \beta \equiv 1/k_B T \) where \( k_B \) is the Boltzmann constant. The constrained (Helmholtz) free energy, \( A(\xi) \), as a function of the reaction coordinate,
−βA(ξ) = ln Q(ξ), corresponds to the PMF. While it is theoretically possible to sample Q(ξ) in a single molecular dynamics (MD) simulation, in practice the small Boltzmann weight of the dissociated configurations makes sampling exceedingly difficult. It is thus advantageous to introduce a series of biasing potential, w_i(ξ), and to simulate the system with modified energy functions,

\[ U'_i(r^N) = U(r^N) + w_i(ξ). \]  

(2.2)

Sampling all ξ is then possible because the original energy barriers are lowered, or equivalently, the Boltzmann weight of the associated configurations is increased. A convenient choice of bias is a harmonic potential, i.e., a spring,

\[ w_i(ξ) = k(ξ_i - ξ)^2, \]  

(2.3)

where ξ_0 is the imposed protein-protein distance and k is the spring constant. Using this potential, we separate the protein-protein center of mass distance into M umbrella sampling windows, with different ξ_i (Fig. 2.2).

**Figure 2.2**: Two proteins are pulled away from each other to generate umbrella sampling windows centered at ξ_i. Each window is sampled by MD simulations using a biased interaction potential \( U'_i(r^N) \). The results for the different windows are then joined to reconstruct the overall PMF as a function of ξ.

In what follows, we detail the computational steps involved in this procedure for a given protein. Note that a detailed tutorial for the process is available for Gromacs (see note [1]). We here provide the details that are not mentioned in that tutorial or that differ for crystal contacts.
1. **Determine contacts.** Crystal contacts can be determined using PISA [106], an online tool that identifies protein-protein interfaces for a given .pdb (protein data bank) file. PISA takes into account protein symmetry and crystal periodicity, in addition to using a distance cutoff for determining contacts. For each contact, it also lists the residues involved in hydrogen bonding and salt bridges, and provides an estimate for the contact free energy. This estimate, however, is fairly rough, because many contributions, such as interactions with the crystallization cocktail or side-chain and backbone conformational changes, are not explicitly considered.

2. **Add any missing or incomplete residues.** When starting from a crystal structure, it is possible that entire residues or some of their side chains might be missing because they could not be crystallographically resolved. These should be added manually to the protein structure. Note that the accuracy of the orientation and the conformation of these residues is not critical at this stage (within chemical reasonableness), because the protein structure is minimized in subsequent steps, eliminating any steric clash that might arise. The interaction strength between patches also depends on the protonation states of the contained residues. A rough estimate for the protonation states can be obtained with propKa [154], keeping in mind that in most cases the solution conditions for crystallization experiments is such that the protein carries no net charge.

3. **Generate input files for each contact.** For each contact, separate .pdb files with two copies of the protein forming the contact should be generated. These .pdb files should then be converted to .gro files (the default structure file format for Gromacs) using the Gromacs command pdb2gmx. This command also prompts the user to choose a force field and a water model. Once the .gro file is created, it is convenient to rotate the protein assembly so that the
z-axis corresponds to the pull direction and to center them in the simulation box. The assembly should be at least 1 nm away from the box sides, in order to prevent one protein from interacting with its copy across the (periodic) box boundary, given that the cutoff for neighbor lists, electrostatic and van der Waals interactions is less than 2 nm. This centering and resizing can be achieved with the following command

```
gmx editconf -f box.gro -o newbox.gro -c -d 1.0
```

The next step is to elongate the box along the pull direction, making sure that the box is at least twice the pull direction plus the original box size. Suppose that the box size in `newbox.gro` is \(10 \times 10 \times 10\) nm. In order to pull one of the protein copies 5 nm away one should run

```
gmx editconf -f newbox.gro -o newbox2.gro -center 5 5 5 -box 10 10 20,
```

for the box to be resized to \(10 \times 10 \times 20\) nm.

4. **Add solvent and ions.** Once the box dimensions are selected and the proteins are positioned, solvent and ions are inserted with the commands `gmx solvate` and `gmx genion`, respectively. Ideally, one should use the same ion type and concentration as in the crystallization cocktail. If the force field parameters for these specific ions are not readily available, however, one might consider replacing them with simple generic ions such as sodium and chloride. This replacement at least matches the ionic strength of the solvent and thus the extent of charge screening in the experimental setup.

5. **Minimize energy and equilibrate.** The energy of the resulting system should be minimized before running any simulation, because the positions of the waters and ions placed in the box, as well as the conformation and orientation of the inserted residues and side chains, need to be relaxed. To further relax
the system, a short additional simulation should be performed in which the number of particles, pressure, and temperature are kept constant (constant $NPT$), before generating input configurations for umbrella sampling. Note that keeping the center of mass of the protein-protein complex fixed for this step facilitates the remainder of the procedure.

6. **Generate configurations for umbrella sampling.** In order to generate initial configurations for umbrella sampling, a simulation is run in which one protein is pulled away from the other at a constant rate, using a harmonic potential with force constant $k$, while the other protein is restrained. (A convenient approach is to restrain three or four backbone atoms.) Typically, $k$ ranging from 1,000 to 10,000 kJ mol$^{-1}$nm$^{-2}$ is appropriate for pulling the proteins apart. Since the resulting configurations are not necessarily equilibrated they need to be further relaxed after the pull, before being used as inputs to umbrella sampling simulations.

7. **Choose umbrella sampling windows and run simulations.** The $M$ chosen configurations should cover the whole distance interval of interest, and their pair separation, $\Delta \zeta$, (Fig. 2.2) should be such that the resulting umbrella sampling windows overlap sufficiently. In particular, a significant portion of the tails of distributions of $\xi$ for each neighboring pair of windows should overlap (see Fig. 2.3b). The constant $k$ for the umbrella simulations should be strong enough to keep the proteins at roughly the desired separation, but not so strong that the resulting distributions are overly narrow. Additional windows, as permitted by the box size, can nonetheless be added after the PMF is generated, if any pair of windows does not sufficiently overlap. Note that if the overlap is poor, one often observes discontinuities in the resulting PMF. This serves as a diagnostic.
8. **Generate PMF.** The PMF is constructed from the output of each umbrella sampling simulation using the command `gmx wham` (Fig. 2.3). The input files necessary for this step are the `.tpr` (The Gromacs executables for individual umbrella sampling simulations) and `pullf*.xvg`, which contain the force information for each window.

![Figure 2.3](image)

**Figure 2.3:** (a) Histograms generated by the weighted histogram analysis method (`gmx wham`) [88]. The distribution for each window overlaps well with those of their neighbors. (b) The PMF for a contact of rubredoxin as a function of pull distance, \( \xi \), at \( T = 300K \) [65]. Infinite separation sets the zero of the energy. The square-well potential generated from this PMF by fitting the second virial coefficient is shown in red (see Section 2.3.2).

### 2.3.2 Phase Diagram

Our ultimate goal is to capture the solution and the assembly behavior of a protein from the simplest possible physical model. This is not only useful in making the simulations computationally tractable, but also serves as a consistency check for the microscopic features we previously identified as relevant. In that context, we consider globular proteins to be roughly spherical objects with anisotropic interactions that are dictated by their surface amino acids. These key features, together with the
assumption that the relevant surface amino acids are involved in crystal contacts, suggest a minimal model comprising a hard sphere with attractive surface patches, i.e. a patchy model. Patchy models based on the Kern-Frenkel potential \cite{96} and others have indeed been shown to recapitulate the phase behavior of various globular proteins \cite{65, 64, 164, 194, 50}. The location, interaction range and strength of the patches, as well as their angular width, can be determined from all-atom simulations of the crystal structure. Note that this model is suitable for short-range interactions. That is, the protein should either be uncharged, or the ionic strength of the crystallization cocktail should be sufficiently high for charge-charge interactions to be screened. These conditions are precisely those used in crystallization experiments. Once this minimal model is parameterized, its phase diagram can be obtained using MC simulations. This section first describes a model that captures the properties of the effective protein-protein interactions computed in Sec. 2.3.1 (Sec. 2.3.2). The general idea of thermodynamic integration, which is used for calculating the free energy of a system from a reference state is then introduced (Sec. 2.3.2). The calculation of fluid (Sec. 2.3.2) and crystal (Sec. 2.3.2) free energies are subsequently described, as well as the procedure for obtaining a coexistence point between these two phases, and the Gibbs-Duhem integration scheme (Sec. 2.3.2) for obtaining the complete crystal solubility curve. Finally, we discuss how Gibbs Ensemble simulations can be used to obtain the (metastable) gas-liquid binodal (Sec. 2.3.2).

Patchy Models

In a patchy model, the interaction potential between two patchy particles $i$ and $j$ is

$$U(r_{ij}, \Omega_i, \Omega_j) = U_{HS} + \sum_{\alpha,\beta=1}^{n} U_{\alpha\beta}(r_{ij}, \Omega_i, \Omega_j), \quad (2.4)$$

where $U_{HS}$ is the hard sphere repulsion, which prohibits overlaps between particles
of diameter $\sigma$, $\alpha$ and $\beta$ label one of the $n$ surface patches, $r_{ij}$ is the distance between particles, and $\Omega_1$ and $\Omega_2$ describe the particle orientations (either in terms of Euler angles or quaternions). The attractive part of the potential, $U_{\alpha\beta}$, is then

$$U_{\alpha\beta} = v_{\alpha\beta}(r_{ij})f_{\alpha\beta}(\Omega_i, \Omega_j),$$

(2.5)

with

$$v_{\alpha\beta}(r_{ij}) = \begin{cases} -\epsilon_{\alpha\beta}, & \sigma < r_{ij} \leq \lambda_{\alpha\beta}\sigma \\ 0, & \text{otherwise} \end{cases},$$

(2.6)

$$f_{\alpha\beta}(\Omega_i, \Omega_j) = \begin{cases} 1, & \cos \theta_\alpha \geq \cos \delta_\alpha \text{ and } \cos \theta_\beta \geq \cos \delta_\beta \\ 0, & \text{otherwise} \end{cases}. $$

(2.7)

In other words, a square-well potential of range $\lambda_{\alpha\beta}\sigma$ controls the radial part of the attraction, and the widths of patches $\alpha$ and $\beta$, $\delta_\alpha$ and $\delta_\beta$, respectively (Fig. 2.4) set their angular range. That is, patches only attract when the vector joining their centers of mass passes through the patch of both particles.

**Figure 2.4:** Two particles do not interact unless their patches overlap in distance and orientation. Here, $\hat{e}_\alpha$ and $\hat{e}_\beta$ point to the centers of patches $\alpha$ and $\beta$, respectively, hence they interact if both $\hat{e}_\alpha \cdot \hat{r}_{ij} = \cos \theta_\alpha \geq \cos \delta_\alpha$ and $\hat{e}_\beta \cdot \hat{r}_{ij} = -\cos \theta_\beta \leq \cos \delta_\beta$, with $r_{ij} \leq \lambda_{\alpha\beta}\sigma$. 

16
The parameters for the radial part of the attraction are obtained from the PMF computed in Sec. 2.3.1 (Fig. 2.3a). The depth of the square-well potential, $\epsilon_{\alpha\beta}$ is that of the PMF for contact between patches $\alpha$ and $\beta$. The range of the square-well attraction is obtained by matching its contribution to $B_2$, the second virial coefficient, to that of the PMF, where

$$B_2 = -\frac{1}{2} \int (e^{-\beta U(r)} - 1) dr,$$

This integral is evaluated numerically for the PMF (its value denoted $I$) and analytically for the $\alpha-\beta$ contact. For a given contact, the interaction range, $\lambda_{\alpha\beta}$, is found by equating the two results,

$$\lambda_{\alpha\beta} = \left(\frac{3I}{e^{\beta\epsilon_{\alpha\beta}} - 1} + 1\right)^{1/3}.$$

Finally, the angular breadth of the interaction is set by running simulations in which the distance between the two proteins is fixed, but not their relative orientation. This is achieved by constraining the center of mass distance with a harmonic spring, at the equilibrium bonding distance. The deviation of the patch vectors from the center of mass axis is tracked in terms of the angle $\delta$ between them. The angular breadth, $\cos \delta_{\alpha}$, for patch $\alpha$ is taken to be the mean of the computed distribution for that angle, and the same for $\cos \delta_{\beta}$ of patch $\beta$.

**Thermodynamic Integration**

Once the patchy model is parameterized, various types of MC simulations are employed to trace out its phase diagram. For two or more phases to be in coexistence, their temperature, pressure, and chemical potential, $\mu$, must all be equal. While $P$ and $T$ can be straightforwardly enforced, $\mu$ is more challenging. Simulations, like
experiments, can only determine the change in free energy along a transformation, not its absolute value. One thus needs a reference state of known free energy and a transformation from that reference to the system of interest to calculate its free energy [61, 182]. Reference states that are of particular interest for us are the ideal gas and the Einstein crystal. Integrating from any of these states along an isotherm yields for the Helmholtz free energy

$$A(\rho, T) = A(\rho_0, T) + N \int_{\rho_0}^{\rho} \frac{P(\rho')}{\rho'^2} d\rho', \quad (2.10)$$

where $\rho_0$ is the density of the reference system and $P(\rho)$ is the equilibrium pressure of the system at a density, $\rho$. Here, we consider the number density, such that $\rho \equiv N/V$, where $V$ is the volume of the system. For numerical convenience, if the reference system is an ideal gas, this expression is rewritten as [153]

$$A_{\text{fluid}}(T, \rho) = A_{\text{ideal}}(\rho) + N \int_{0}^{\rho} \left[ \frac{P}{\rho'^2} - \frac{1}{\beta \rho'} \right] d\rho', \quad (2.11)$$

where $A_{\text{ideal}}$ is the free energy of the ideal gas (see note 2)

$$\frac{A_{\text{ideal}}(\rho)}{N} = \frac{1}{\beta} \log(\rho \Lambda^3) - 1 + \frac{1}{N} \log(2\pi N). \quad (2.12)$$

where the de Broglie wavelength, $\Lambda^3$, is set to unity without loss of generality. Another option is to integrate at constant pressure, i.e., along an isobar, varying the temperature of the system

$$\beta \mu(T, P) = \beta \mu(T_0, P) - \int_{T_0}^{T} \frac{H(T')}{Nk_B T'} dT', \quad (2.13)$$

18
where $H$ is the enthalpy of the systems, or along an isochore,

$$
\beta \frac{A(T,V)}{N} = \beta \frac{A(T_0,V)}{N} - \int_{T_0}^{T} \frac{U(T')}{Nk_BT'} dT',
$$

(2.14)

where $U$ is the internal energy of the system.

**Free Energy of the Fluid Phase**

Using the principles of thermodynamic integration introduced above, the following steps summarize how we obtain the free energy of the fluid phase. To integrate along the isotherm from the ideal gas, we run $NPT$ simulations at a set of pressures $\{P_1, \cdots, P_m\}$, where $P_1$ is a very low pressure and $\rho(P_m)$ is the density of interest. Figure 2.5(a) shows the numerical equation of state of the fluid phase for the patchy model of rubredoxin [65]. The integrand of Eq. (2.11) is calculated from these data points (Fig. 2.5b). The integrand gets noisy as pressure decreases, because both $1/\rho^2$ and $1/\rho$ diverge as $\rho \to 0$, hence the numerical error is then amplified. In that regime, one can use the fact that the integrand converges to $B_2$, Eq. (2.8), as $\rho \to 0$ to increase numerical accuracy. There are three options for the rest of the thermodynamic integration: (i) continue integrating along the same isotherm to obtain the free energy as a function of pressure, (ii) integrate along an isobar using Eq. (2.13), or (iii) integrate along an isochore using Eq. (2.14).

**Free Energy of the Crystal Phase**

The Frenkel-Ladd method [60] is a thermodynamic integration scheme to obtain the free energy of a crystal using an Einstein crystal with a fixed center of mass as a reference. Particles are then restrained to their equilibrium lattice positions and orientations, and do not otherwise interact. The interaction energy of this system is
Figure 2.5: (a) Equation of state (density as a function of pressure), and (b) the integrand of Eq. (2.11) for the fluid phase of the patchy model of rubredoxin at $\beta = 0.2$ (in units of $1/k_B T$) [65]. The star denotes $B_2$, the $y$-intercept of the integrand Eq. (2.8). Curves are polynomial fits to the data.

$$U_E = \sum_{i=1}^{N} \kappa (\mathbf{r}_i - \mathbf{r}_{i,0})^2 + \sum_{i=1}^{N} \kappa g(\Omega_i). \quad (2.15)$$

The first term restrains the positions $\mathbf{r}_i$ of the $N$ particles to their positions $\mathbf{r}_{i,0}$ by a harmonic potential with spring constant $\kappa \in [0, \infty)$. The second term penalizes particles that deviate from their equilibrium orientations, by the potential $g(\Omega) = 1 - \cos(\psi_{i\alpha}) + 1 - \cos(\psi_{i\beta})$, where $\psi_{i\alpha} (\psi_{i\beta})$ is the angle between the vector that defines patch $\alpha (\beta)$ in its equilibrium and instantaneous orientations, and patches $\alpha$ and $\beta$ are chosen arbitrarily among the surface patches (see note 3). The free energy of the ideal Einstein crystal with fixed center of mass, $A_{E,t}$, has both translational and orientational contributions $A_E = A_{E,t} + A_{E,o}$ where [153][182]

$$\beta \frac{A_{E,t}}{N} = -\frac{3}{2} \frac{N - 1}{N} \log \left( \frac{\pi}{\beta \kappa} \right) - \frac{3}{2N} \log(N) \quad (2.16)$$

$$\beta \frac{A_{E,o}}{N} = -\log \left( \frac{1}{8\pi^2} \int d\Omega \, e^{-\kappa g(\Omega)/k_B T} \right). \quad (2.17)$$

This reference system is converted to the interacting protein crystal in three steps.
1. **Switch on interactions.** The free energy change in this step is

\[ \Delta A_1 = -\log \left(\frac{\langle e^{-\beta(\tilde{U}-U_0)} \rangle_E}{U_0} \right), \]

where \( \langle \cdot \rangle_E \) denotes an averaging over ideal Einstein crystal configurations, \( \tilde{U} \) is the energy of the interacting Einstein crystal without the harmonic spring contributions, and \( U_0 \) is the interacting crystal ground state energy. For large enough \( \kappa \) (denoted \( \kappa_{\text{max}} \)) the contribution from the thermal average vanishes, i.e. \( \Delta A_1 \approx U_0 \). This condition sets \( \kappa_{\text{max}} \).

2. **Turn off position and orientation restraints.** In this step, the springs are turned off, i.e. \( \kappa \to 0 \). The change in free energy of this process is

\[ \Delta A_2 = -\int_{0}^{\kappa_{\text{max}}} d\kappa' \left( \frac{\partial U_E}{\partial \kappa'} \right)_{NVT\kappa'} = -\int_{0}^{\kappa_{\text{max}}} d\kappa' \left( \sum_{i=0}^{N} (r_i - r_{i,0})^2 + \eta \sum_{i=0}^{N} g(\Omega_i) \right)_{NVT\kappa'}. \]

Note that because \( \kappa_{\text{max}} \) can be very large, it is often more convenient to use \( \log \kappa \) as the integration variable

\[ \Delta A_2 = -\int_{-\infty}^{\log \kappa_{\text{max}}} d(\log \kappa') \left( \sum_{i=0}^{N} (r_i - r_{i,0})^2 + \eta \sum_{i=0}^{N} g(\Omega_i) \right)_{NVT\kappa'}. \]
Quadrature with 20 to 40 logarithmically-spaced points. Because the integrand vanishes when $\log \kappa \to -\infty$ the integration can start from a small (non-zero) $\kappa$. Evaluating the translational contribution to the Einstein crystal free energy is straightforward Eq. (2.16), and although the orientational part Eq. (2.17) cannot be calculated analytically, for large $\kappa_{\text{max}}\eta$, one can use the saddle point approximation

$$\int d\Omega e^{-\beta\kappa_{\text{max}}\eta g(\Omega)} \approx e^{-\beta\kappa_{\text{max}}\eta g(\Omega_0)} \sqrt{\frac{2\pi}{\beta \kappa_{\text{max}}\eta g''(\Omega_0)}}.$$  (2.21)

Here we have approximated $g(\Omega)$ with its second order Taylor expansion, and $\Omega_0$ is the orientation that minimizes $g(\Omega)$. The orientational contribution to the Einstein crystal free energy then becomes

$$\beta A_{E,o} \approx \frac{3}{2} \log(\beta \kappa_{\text{max}}\eta) + \frac{1}{2} \left\{8\pi \det\left(H[g(\Omega_0)]\right)\right\},$$  (2.22)

where $\det(H[g(\Omega_0)])$ is the determinant of the Hessian computed at the minimum of $g(\Omega)$ [65]. Note that one should check whether $\kappa_{\text{max}}$ is indeed large enough by verifying that higher order terms in the Taylor expansion are negligible.

An estimate for the plateau value for the integrand of Eq. (2.20) is can be analytically estimated for sufficiently large $\kappa$, and thus serves as a consistency check. The orientational contribution to this quantity is calculated using the saddle point approximation to $A_{E,o}$, Eq. (2.22),

$$\kappa\eta\left\langle \sum_{i=0}^{N} g(\Omega_i) \right\rangle_{\kappa} = \kappa\eta \frac{\partial A_{E,o}}{\partial (\kappa\eta)} = \kappa\eta \frac{3}{2\beta} \frac{\beta}{2\beta} \frac{3N}{2\beta},$$  (2.23)
and the translational contribution can be estimated using the expression derived for the hard sphere mean square displacement [60]. In the limit of very large $\kappa$, this result should be exact because translational and orientational displacements are then too small to break any bond.

3. **Release the crystal center of mass.** Removing the constraint over the center of mass finally gives

$$\Delta A_3 = \frac{1}{\beta} \log(\rho). \quad (2.24)$$

Cumulating these results gives the absolute free energy of the crystal of patchy particles

$$A = A_E + \Delta A_1 + \Delta A_2 + \Delta A_3 \quad (2.25)$$

at a given density and temperature. Integration along an isobar, isotherm, or an isochore within the crystal phase can then be performed to explore a broad range of conditions within that phase.

*Gibbs-Duhem Integration*

Given a coexistence point between the crystal and the liquid, the Gibbs-Duhem relation,

$$t(P(\beta), \beta) \equiv \left( \frac{dP}{d\beta} \right)_{\text{coex}} = -\frac{H_{\text{cryst}}/N - H_{\text{liq}}/N}{\beta(1/p_{\text{cryst}} - 1/p_{\text{liq}})} = -\frac{\Delta H/N}{\Delta(1/\rho)}, \quad (2.26)$$

can be integrated to obtain coexistence points at different temperatures [103]. This can be done using a numerical method, such as predictor-corrector algorithms, and evaluating the thermodynamic quantities via MC simulations. The general idea is as follows.
1. Start from a known coexistence point \((P_0, \beta_0)\), consider a system at \(\beta_1 = \beta_0 + \Delta \beta\), where \(\Delta \beta\) is small (see below).

2. Guess the coexistence pressure at this temperature according to the appropriate predictor formula.

3. Run \(NPT\) simulations of the two phases simultaneously to equilibrate \(\Delta H/N\) and \(\Delta (1/\rho)\). Correct the pressure prediction using these quantities according to the appropriate corrector formula.

4. Repeat 2 and 3 until convergence.

The chosen integration scheme depends on how many coexistence points are known (see Table 2.1). At the start of the process, only one such point, \((P_0, \beta_0)\), is known. A short simulation is run for both phases to obtain \(t(P_0, \beta_0)\). The guess for the pressure, \(P_1\), for the next coexistence temperature, \(\beta_1 = \beta_0 + \Delta \beta\), and its correction are then given by the trapezoid rule, after calculating \(t(P_1, \beta_1)\) with the initial guess. The third and fourth coexistence points can be calculated using the
Table 2.1: Predictor-corrector algorithms, where \( t_i \equiv t(P_i, \beta_i) \) for \( \beta_i = \beta_0 + i\Delta \beta \), is defined in Eq. (2.26).

<table>
<thead>
<tr>
<th>Method</th>
<th>Predictor</th>
<th>Corrector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trapezoid</td>
<td>( P_{i+1} = P_i + \Delta \beta t_i )</td>
<td>( P_{i+1} = P_i + \frac{\Delta \beta}{2} (t_i + t_{i+1}) )</td>
</tr>
<tr>
<td>Midpoint</td>
<td>( P_{i+2} = P_i + 2\Delta \beta t_{i+1} )</td>
<td>( P_{i+2} = P_i + \frac{\Delta \beta}{3} (t_{i+2} + 4t_{i+1} + t_i) )</td>
</tr>
<tr>
<td>Adams</td>
<td>( P_{i+4} = P_{i+3} + \frac{\Delta \beta}{24} (55t_{i+3} - 59t_{i+2} + 37t_{i+1} - 9t_i) )</td>
<td>( P_{i+4} = P_{i+3} + \frac{\Delta \beta}{24} (9t_{i+4} + 19t_{i+3} - 5t_{i+2} + t_{i+1}) )</td>
</tr>
</tbody>
</table>

midpoint rule. Once four points are obtained, additional ones can be found iteratively using Adams rule (see note 4). While \( \Delta \beta \) should be large enough to allow for an efficient tracing using a too large a value causes numerical instability [104]. One way to validate the resulting coexistence line is to repeat this procedure starting from different, well separated coexistence points.

Obtaining the Gas-Liquid Binodal

Coexistence points on the gas-liquid binodal can certainly be obtained by slightly modifying the above procedure, but a more straightforward approach is to use Gibbs Ensemble simulations [136, 61], which is specifically designed for identifying coexistence between homogeneous phases of intermediate density. In this method, two boxes of fluid are simulated simultaneously. Their total volume and number of particles are kept constant but boxes can exchange volume as well as particles between each other. The density of each box then converges to the gas or the liquid density, \( \rho_g \) and \( \rho_l \) respectively (see note 5). The binodal is then obtained as follows.

1. **Obtain a few coexistence points from Gibbs Ensemble simulations.**

   Gibbs Ensemble simulations are run for a set of temperatures below the estimated critical temperature, \( T_c \), from generalized law of corresponding states [134], starting from an intermediate fluid density, \( \rho \approx 0.3 \).

2. **Fit coexistence data to obtain the full binodal.** The physical universality of the gas-liquid transition allows for tracing the binodal using only a few
coexistence points. The full binodal, including the critical point, can then be calculated by fitting two universal equations: a scaling law and the law of rectilinear diameters [61]. The former gives an estimate of $T_c$,

$$\log(\rho_l - \rho_g) = \log B + \beta \log(T - T_c),$$

(2.27)

where $\beta = 0.32$ (not to be confused with the inverse temperature) is the magnetization exponent for the Ising universality class, and $B$ is a proportionality constant [31]. Once $T_c$ is found, the critical density, $\rho_c$, can be obtained from the law of rectilinear diameters, which describes the asymmetry of the gas-liquid binodal away from $T_c$,

$$\frac{\rho_l + \rho_g}{2} = \rho_c + A(T - T_c),$$

(2.28)

where $A$ and $\rho_c$ are determined by fitting. The coexistence binodal (see Fig. 2.8) is then given by

$$\rho = \rho_c + A(T - T_c) \pm \frac{B(T - T_c)^\beta}{2}.$$  \hspace{1cm} (2.29)

2.4 Notes

The above procedure results in the phase diagram for a simple globular protein as seen in Fig. 2.8. In what follows we discuss a number of geometric issues and how they can be avoided, as well as briefly mention possible ways of increasing the model complexity.
Figure 2.7: Evolution of $\rho_l$ and $\rho_g$ throughout the Gibbs Ensemble simulations for various temperatures. Note that the densities converge to their coexistence values after roughly $2 \times 10^5$ MC sweeps. Average densities are calculated (dashed lines) after the equilibration period. These three pairs of data points are those that appear in the final phase diagram (Fig. 2.8).

1. The tutorial prepared by Justin Lemkul can be accessed at http://www.bevanlab.biochem.vt.edu/Pages/Personal/justin/gmx-tutorials/umbrella/index.html.

2. While the simplest method for obtaining the free energy of a fluid is Widom insertion [61], at high density it is more accurate to use Eq. (2.11).

3. The parameter $\eta$ is a proportionality constant chosen such that the strengths of both restraints can be tuned by $\kappa$ alone.

4. It is often not advantageous to use higher order predictor-corrector formulas. These not only require a higher number of coexistence points but also exhibit stability issues.

5. If the temperature is close to the critical temperature the small density dif-
Figure 2.8: The $T - \rho$ phase diagram of a patchy model of rubredoxin [65]. Blue points denote the solubility line and are obtained from Gibbs-Duhem integration. Black points are gas-liquid coexistence points obtained from the Gibbs Ensemble simulations. The fit to the gas-liquid binodal (gray dashed line) terminates at the resulting critical point (black star). Below this line, the system exhibits a metastable liquid-liquid phase coexistence regime in which protein solutions often gel in experiments. This long-lived state often precludes crystallization. The region between the solubility line and $T_c$ is called the nucleation zone because supersaturated solutions in this region are more likely to produce crystals by avoiding gelation.

Reference can cause the boxes to switch identity (liquid↔gas) (Fig. 2.7), which limits the efficacy of the approach in this regime.

6. Geometric Constraints: Because globular proteins are not actually spherical, the onset of harsh repulsion for each contact PMF can occur at slightly different distances. In the scheme above, the chosen particle diameter should be the same for all contacts. The smallest center of mass distance should then be taken as the hard sphere diameter and the other PMFs should be translated such that the onset of attraction coincides with that diameter.

Another problem that can arise due to deviations from sphericity is that a simple projection of the patch position on the sphere may not result in all
patches interacting within the relevant crystal symmetry. In this case, patch vectors and unit cell parameters can be perturbed slightly to ensure that all bonds are satisfied in the crystal phase. This modification is known to have but a very limited effect on the phase diagram [63].

7. **Increasing Model Complexity** The simple patchy model described here does not capture the phase behavior of certain proteins. Numerous modifications patchy models have been proposed to capture these effects. The impact of shape anisotropy [176, 201, 70], patch mobility [28], and the interaction potential form [48, 198] have been investigated in the context of general self-assembly. Such features can be considered if the microscopic properties of the protein of interest suggest that more complex models are required. The application of these features to specific protein systems is still an open area of research.

8. Data and scripts relevant to this work have been archived and can be accessed at [http://dx.doi.org/doi:XX.XXXX/xxxx].
3

Water in Protein Crystals

Reprinted (adapted) with permission from


Copyright (2018) American Chemical Society.

3.1 Introduction

Water is not only a medium for biological processes, but an active participant[16]. It mediates interactions between proteins and small-molecule inhibitors [100] [14], and enables the enzymatic transfer of a proton to a protein residue [188]. Moreover ice-binding proteins alter the ordering of water around them, affecting ice nucleation[186]. A reliable physico-chemical description of water in the vicinity of biomolecules is thus needed both to properly solvate these complex objects and
to comprehend their function. Yet, despite continued advances to our microscopic understanding of the properties of bulk water [169, 135, 180], including its many phases [99, 160, 57] and hydrophobicity [40], our grasp of biomolecular solvation still markedly lags behind [15, 116]. The intricate interplay between the mosaic of hydrophobic and hydrophilic surface residues, steric hindrance, and side-chain dynamics requires a careful balance of the various intermolecular interactions in order for a structurally accurate description of solvation to emerge. Standard water models, which are rigid, non-polarizable and parameterized to reproduce a standard set of bulk properties, attempt to do just that [180, 77] (Figure 3.1), but it is unclear how they fare at solvating biomolecules [170]. The lack of reliable experimental information about solvation has thus far rendered this problem intractable.

Figure 3.1: Typical water models used in biomolecular simulations vary mostly in the number of point charges they use to capture intermolecular interaction. All include charges at the hydrogen positions and a Lennard-Jones potential on the oxygen atom, but (a) three-site models contain an additional point charge on the oxygen atom (e.g., SPC [22] and SPC/E [20]), while (b) four-site models use a virtual site (V) (e.g., TIP4P [93], with [85] and without Ewald summation, and TIP4P/2005 [2]), and (c) five-site models split the charge between two virtual sites (e.g., TIP5P [118]). Although six-site models also exist, they are not commonly used.

A possible experimental headway into this problem comes from protein crystallography. Protein crystal unit cells contain a significant fraction of water (between 26% and 90% by volume with an average around 50% [192, 193]), hence the inhomogeneous distribution of the solvent impacts the diffracting radiation – be it X-ray [92, 41, 52], neutron [130], or electron [131, 132]. The phase problem of crystallogra-
phy actually makes the accurate reconstruction of water density profile an essential
cOMPONENT of most protein structure determinations [52]. Full structure factors –
amplitude and phase – are needed to determine atomic densities within a unit cell, yet
only amplitudes can typically be measured directly. Even when some of the phases
can be gleaned from multiple intensity measurements or molecular replacement [177],
many phase values can still go missing. Phases must thus be obtained by iteratively refining the unit cell description and the phase estimates. Because all atoms,
i.e. both the macromolecule and the solvent, contribute to all structure factors, an
accurate model of the solvent structure is required for this iterative refinement. Obtaining the structure of the protein chain therefore requires a careful treatment of
water density fluctuations.

However, the description of the unit cell structure from refinement is far from
perfect. The extent of the mismatch is commonly quantified by R-factors,

$$R = \frac{\sum_{k \in S} |F_{\text{obs}}(k) - F_{\text{model}}(k)|}{\sum_{k \in S} F_{\text{obs}}(k)},$$

where $F_{\text{obs}}(k)$ and $F_{\text{model}}(k)$ are the experimentally observed and model structure factor amplitudes, respectively, from the set $S$ of observed reflections $k$, where $k = 2\pi n$
is a vector containing the Miller indices of the reflection. Even for the highest-quality protein structures, $R$-factors average 15%, which is an order of magnitude larger than for small molecules [200]. Although part of the difference is attributable to experimental noise, the weaker agreement between model and experiment is more generally ascribed to the limited sophistication of the structural model of the unit cell.

---

1 If the set $S$ contains all the measured structure factors, the resulting $R$-factor is $R_{\text{work}}$. As a measure of overfitting, crystallographers also calculate $R_{\text{free}}$ [35] by choosing $S$ to be a small set of structure factors that are not included in any stage of structure determination, including refinement. Note that in the crystallography literature the vector $k$ is conventionally written without the factor of $2\pi$, and an extra $2\pi$ then appears in the Fourier transform (see Eq. 3.2).
content\textsuperscript{[37]}, and especially of the solvent\textsuperscript{[34, 37]}. The water model used for structural refinement is assembled from the sum of (i) localized crystal water molecules and (ii) delocalized bulk water regions. A solvent model that improves the description of the water structure should increase agreement between model and data, and ultimately improve the quality of the protein structures obtained crystallographically. Whether molecular dynamics (MD) simulations, which allow a continuum of description between (i) and (ii), can complement the diffraction data\textsuperscript{[91]} and thus improve the description of biomolecular solvation is largely unexplored.

In this work, we make an attempt in this direction by comparing the MD and refinement-derived hydration structure of a single protein, a Yb\textsuperscript{3+}-substituted mannose binding protein (PDB ID: 1YTT)\textsuperscript{[37]}. Like\textsuperscript{[author?]}, who previously studied this protein to probe the surrounding water structure, we choose this system because its X-ray structure was determined from multi-wavelength anomalous diffraction (MAD) and a full set of experimental phases was experimentally extracted. This rare occurrence enables us to determine the experimental solvent density profile unbiased by the refinement process. This comparison also allows for benchmarking the water models used in MD simulations. From a methodological standpoint, our comparison relies on an ergodic-like hypothesis that the signal from diffraction techniques is spatially averaged over the configurations of water in the various unit cells, and thus can be recovered by averaging over water configurations obtained from long MD trajectories of a single unit cell. In the following, we first describe the test protein (Sec. 3.2.1), the water models used in the study (Sec. 3.2.2), as well as the MD simulation (Sec. 3.2.3), and comparison (Sec. 3.2.4), and protonation schemes (Sec. 3.2.5) before detailing the results of our analysis in Section 3.3.
3.2 Methods

This section presents the technical aspects of the experimental system and of the MD simulations as well as the solvent density analysis scheme.

3.2.1 Protein and Setup

We study the Yb$^{+3}$-substituted mannose binding protein (PDB ID: 1YTT) solved by MAD phasing up to a resolution of 1.8Å\cite{37} from a crystal with space group symmetry P2$_1$2$_1$2$_1$ \cite{34}. The unit cell contains four protein dimers related by symmetry operations, and thus totals eight protein copies (see SI). The model deposited in the Protein Data Bank (PDB) nearly two decades ago had $R_{\text{work}} = 0.185$ and $R_{\text{free}} = 0.206$\cite{24}, but methodological advances achieved since by Phenix \cite{3} (version phenix-dev-2405) have allowed us to make substantial improvements to the structural refinement and to update the assigned crystal waters. The biochemical reasonableness of the resulting structure was nevertheless verified by MolProbity \cite{44}. No crystal waters were found to clash with protein atoms and all were at a reasonable hydrogen bonding distance from other crystal waters. Careful examination of the local difference density maps, however, led us to manually remove six water molecules that resulted in an excess electron density compared to the experimental data. Keeping the remaining 254 crystal waters per protein in place, an additional iteration of structural refinement gave $R_{\text{work}} = 0.159$ and $R_{\text{free}} = 0.183$. The robustness of this result to experimental and refinement noise indicates that the final structure is slightly overfitted but nevertheless a reasonable starting point for this study (see SI).

From the set of optimal structure factors obtained from the refinement process, the electron density at each point $\mathbf{r}$ within the unit cell can formally be computed
as,

\[
\rho(\mathbf{r}) = \frac{1}{v} \sum_{\mathbf{k} \in S} F_{\text{obs}}(\mathbf{k}) \, e^{i\varphi(\mathbf{k}) - k \cdot \mathbf{r}},
\]

(3.2)

where \(v\) is the volume of the unit cell. However, because \(F(0)\) cannot be extracted experimentally – it is coincident with the transmitted beam – the density profile can only be determined up to an unknown constant, \(\bar{\rho}\), and the sum is truncated at high \(k\) once experimental peaks become unresolvable (see SI\[59\]).

3.2.2 Water Models

The water models considered in molecular simulations are: (i) SPC\[22\], (ii) SPC/E\[20\], (iii) TIP3P\[93\], (iv) TIP4P\[93\] with Ewald summation\[85\], (v) TIP4P/2005\[2\], and (vi) TIP5P\[118\] (see Fig. 3.1). The first five have three planar charges (TIP4P and TIP4P/2005 have a negative charge off the oxygen atom), while the sixth has four tetrahedrally-distributed charges. All overestimate the gas phase dipole moment of water, in order to treat some of the many-body contributions in condensed phases in an effective way\[180\]. SPC and SPC/E, unlike TIP3P, have an O-H bond length and a H-O-H bond angle that differs from the gas phase water geometry for a similar reason. The charge distribution in SPC/E also effectively takes into account the polarization correction to the energy\[20\]. Note that the only difference between TIP4P and TIP4P/2005 is that their parameters were optimized to match different sets of thermodynamic properties.

These models describe bulk water with varying degrees of success. For instance, TIP4P is better than SPC and TIP3P at reproducing the structure of the gas phase dimer as well as the water density, enthalpy of vaporization, and peak structure of the oxygen-oxygen radial distribution function\[93\]. TIP5P reproduces the oxygen-oxygen radial distribution function even better than TIP4P\[181\], while TIP4P/2005

\[\text{Throughout this paper, we refer to TIP4P with Ewald summation as TIP4P.}\]
reproduces better the phase diagram of water than any other models of this type\textsuperscript{2}. Although (author?), judged TIP4P/2005 to be generally superior, their analysis mainly highlighted that all of such models result from compromises. Whether similar distinctions between these models exist for the structure of water near a protein surface, however, has not yet been tested.

3.2.3 Molecular Dynamics Simulations

The numerical solvent density profile was extracted from molecular dynamics simulations. Systems are initialized by first placing copies of the crystal structure (obtained in Sec. 3.2.1) of the protein following the crystal symmetry, within a simulation box that has the same dimensions as the crystal unit cell (see SI). Preserving the protein within its unit cell rather than solvating it within a larger simulation box more closely captures the confinement conditions within the crystal as well as the impact of protein-protein interfaces on water ordering. This choice, however, also introduces computational difficulties. In particular, sampling configurations near protein-protein interfaces can be sluggish, and tuning the water density in confinement is nontrivial. Errors in the latter may result in a water activity that is quite different from that of a crystal grown in an experimental cocktail\textsuperscript{3}. In order to minimize the impact of both of these problems on the water density profile we run four simulations, each containing a different three-protein dimer copy subset of the unit cell. The absence of a protein dimer copy both accelerates sampling and endogenously introduces a reservoir of solvent that brings its activity near that of the bulk. Note that because only seven protein surface atoms (out of 1769) per chain lie at the interface of four protein dimers, the impact of this removal on the analysis of the solvent structure should be negligible. Water initialization is done by the solvate

\textsuperscript{3} We assume that only water and small ions are present in the crystallization cocktail. In practice, other additives are often included\textsuperscript{7, 142}. 

36
module in Gromacs, which results in a water density within the bulk region of the simulation box that deviates at most by 1% from its standard value, 1.00 g/mL, at temperature $T=298\text{K}$. \textbf{[author?] have found that the ionic strength does not noticeably affect the structure of water within the unit cell }\textbf{[83], therefore our simulations use 0.05M NaCl, which is within the typical range of ionic strengths encountered in crystallization experiments.}

The protonation states of side chains were at first automatically assigned by Gromacs \textbf{[21]} (version 5.1.2), based on the hydrogen-bonding network analysis of the software package\textbf{[109].} In order to assess the impact of protonation on the surrounding water structure, we also generated variants with opposite protonation states for histidines, glutamates, lysines, and aspartates (see Sec. 3.2.5).

The protein chain was modeled using the Amber99sb biomolecular force field\textbf{[86].} Parameters for Yb$^{3+}$ ions, which are not defined in this force field, were constructed from the Lennard-Jones parameters for sodium ions, which has a similar ionic radius, but a charge of +3. Although this crude treatment cannot fully capture the rich coordination chemistry of a transition metal ion, only a small subset of nearby surface atoms are affected by this choice.

MD simulations were then run with various restraints. To minimize possible deviations from the experimentally-refined protein model, carbon and nitrogen atoms on the backbone were kept immobile (restrained), while oxygens were allowed to move, as their position does not affect the overall protein backbone shape. Yb$^{+3}$ ions were also restrained, in order not to bias the simulation results with the approximate parameters described above. In order to facilitate the sampling of water configurations near the protein surface, heavy atoms (i.e. all protein atoms except hydrogens) in the side chains as well as backbone oxygen atoms were constrained harmonically with a force constant of 1000 kJ nm$^{-2}$ mol$^{-1}$, which is found to be weakest restraint that prevents side chains from changing conformation over the course of the simulations.
Although these constraints slightly bias the final water density, they are required for a reliable comparison of the resulting MD water density with the experimentally observed density. Hydrogen atoms, water molecules and ions were allowed to move freely.

The simulations were thermostatted at 298 K. Although the crystallographic data was obtained at 110K upon flash freezing the crystal sample, we assume that the unit cell configuration at the crystallization temperature was preserved by this quench and energy minimization has but a marginal impact on the structure. Amorphous water at that temperature in the protein crystal is indeed glassy. The strong confinement experienced by water in the crystal is expected to leave water in a low-density amorphous (LDA) ice with a structure similar to that of liquid water from which it was quenched. (Neutron diffraction results suggest that the local spatial distribution around a given water molecule in LDA is closely related to that of the liquid phase.) We thus here assume that at distances comparable to the size of the solvent cavities in the protein unit cell, the amorphous structure of water closely resembles that of room-temperature liquid water. This temperature is also optimal for the protein force fields and water models that were used, and facilitates the sampling of solvent configurations (see SI).

We optimize the sampling frequency and computational time by first equilibrating the systems for 30 ns, and then saving configuration snapshots every 3 ns. This provides a total of 40 fairly well decorrelated solvent configurations (see SI). As a consistency check, we compare the water distribution surrounding a given protein atom with that of its symmetric counterparts by computing the real-space correlation coefficients around these atoms (as detailed in section 3.2.4). Less than 6% of the surface atoms were found to have a sampling error larger than 10%, which suggests

---

4 The crystallization temperature for this protein was not reported but most structures deposited in the PDB are crystallized at room temperature, and a lack of experimental details suggests that an atypical experimental procedure is unlikely.
that a thorough sampling was achieved.

### 3.2.4 Analyzing the Solvent Structure

Electron density maps are extracted from the MD simulations by averaging over the atomic densities obtained from the individual snapshots, and also from the diffraction dataset, using the Computational Crystallography Toolbox (CCTBX) library [74], upon which Phenix is based. This algorithm uses a three-dimensional grid that spans the unit cell, with a grid spacing that is roughly one fourth of the maximum resolution of the dataset (see SI[5]). The contribution of water to the overall MD electron density, $\rho_{\text{solvent}}(r)$, is then estimated (using a standard Phenix routine), by centering an isotropic Gaussian on the positions of oxygen atoms, with a standard deviation determined by the given atomic $B$-factor (see SI).

In order to reconstruct the solvent density from the set of simulation boxes that contain only parts of the unit cell (see Sec. 3.2.3), we use the density information about protein-protein interfaces from the simulation box that contains the relevant copies of the protein dimer. In other words, we select the protein dimer copy that contains the given atom, and the two neighboring protein dimer copies that are closest to that atom. The densities are joined by first partitioning the unit cell, such that each grid point is assigned to the protein atom that is closest to it (considering the refined protein coordinates), and then by copying the density within the partition associated with each atom.

We compare the spatial distribution of water around protein atoms in both experimental and simulated systems using the grid described above. Because the nitrogens and carbons in the backbone are kept immobile, the protein structure in the various MD snapshots only differ from that of the refined structure in its side-chain positions. The radial distribution functions (RDF), which capture the average solvent density profile as a function of distance from a protein heavy atom, offers the lowest-order
correction to the bulk solvent description near an interface \[79\]. For a subset of atoms \(A\) and the grid described above, it is computed following a scheme similar to that of (author?). For an atom \(i \in A\), let \(\chi_i\) be the set of grid points assigned to that atom, and \(X = \bigcup_{i \in A} \chi_i\) be the set of grid points assigned to atoms in \(A\). Then,

\[
g_A(r) = \frac{1}{\langle \rho_{\text{solvent}} \rangle} \frac{\sum_{i \in A} \sum_{p \in \chi'_i} \rho(p) \Theta[\rho(p)]}{\sum_{i \in A} \sum_{p \in \chi'_i} \Theta[\rho(p)]},
\]

where \(p\) is a grid point, \(\rho(p)\) is the electron density at that grid point, \(\chi'_i\) is the subset of \(\chi_i\) that contains grid points \(r \pm \Delta r\) away from atom \(i\), \(\Theta\) is the Heaviside function, and \(\langle \rho_{\text{solvent}} \rangle\) is the average electron density in the solvent region. The chosen shell thickness, \(\Delta r = 0.3\,\text{Å}\), is only slightly smaller than the grid spacing derived from the maximal resolution of the protein data, \(d_{\text{min}} = 1.8\,\text{Å}\), which ensures that a statistically sufficiently number of grid points is captured within each shell, without overly coarsening the data.

RDFs are obtained both for separate sets of surface N, O, and C atoms and for individual surface atoms, in both cases considering only surface atoms that are well localized, i.e., with \(|\chi_i| \geq 500\) and \(B_i \leq 24\,\text{Å}^2\) (the \(B\)-factor of atom \(i\)), which roughly corresponds to a mean-squared displacement of 0.3 Å, and follows the definition used in Ref. \[37\]. We discard surface atoms that are within 6 Å of Yb\(^{3+}\) ions due to the strong Fourier ripples that surround them (see SI). For the sets of surface N, O, and C atoms, an average radial correlation coefficient of the RDFs are computed for \(2.4\,\text{Å} < r < 6\,\text{Å}\) away from the protein atoms. For \(r < 2.4\,\text{Å}\) it is not possible to deconvolute protein and solvent contributions to the observed electron density, whereas for \(r > 6\,\text{Å}\) statistical noise and diffraction artifacts dominate because less than 2% of the grid points fall beyond that distance. The correspondence between the RDFs from experimental and MD-generated densities is assessed by the Pearson
correlation coefficient \([1.51]\). For individual surface atoms, we construct the set of RDFs, \(\{(g_{i,MD}(r), g_{i,obs}(r))\}\) for all \(i \in A\), and all radial bins. The Pearson correlation coefficient of this set of ordered pairs is also computed. In order to compare the radial position of a given peak in the RDFs, its 95% confidence interval is estimated by drawing 1000 perturbed RDFs according to the error margin in each radial bin.

Because RDFs average out information about the orientation of water molecules, we also consider angular distribution functions (ADFs), which depend on the hydrogen bond network in each configuration, and thus encode three-body and higher-order correlations. Only grid points within the first solvation shell, i.e., for \(2.4\text{Å} < r < 4.8\text{Å}\), are considered for this computation. The heavy atom is placed at the origin, and then the orientation of each point around this atom is determined using spherical coordinates, \((\theta, \phi)\), with the axis orientations following the PDB conventions\([1]\),

\[
\gamma_i(\phi, \theta) = \frac{\sum_{p \in \chi_i(I_\phi, I_\theta)} \rho(p) \Theta[\tilde{r}_p - r_1] \Theta[r_2 - \tilde{r}_p]}{\sum_{p \in \chi_i(I_\phi, I_\theta)} \Theta[\tilde{r}_p - r_1] \Theta[r_2 - \tilde{r}_p]},
\]

where \(\tilde{r}_p\) gives the distance from the grid point to the heavy atom, \(\chi_i(I_\phi, I_\theta)\) is the set of grid points assigned to \(i\) and are oriented such that \(\phi \in [\phi - \Delta \phi / 2, \phi + \Delta \phi / 2]\) and \(\theta \in [\theta - \Delta \theta / 2, \theta + \Delta \theta / 2]\). We set \(\Delta \phi = \Delta \theta = \pi / 30\), which corresponds to an arc-length of 0.25Å at 2.4Å, and 0.5Å at 4.8Å, comparable to the radial binning of the RDF. The comparison between the angular distribution functions in experiments and simulations is also done using the Pearson correlation coefficient of \(\Gamma_{obs}(i, \phi, \theta) = \gamma_i,obs(\phi, \theta)\) and \(\Gamma_{MD}(i, \phi, \theta) = \gamma_i,MD(\phi, \theta)\), considering only cases in which both quantities are nonzero.

The real-space distribution of the water density combines information about both the radial and angular components. It thus provides an overall comparison of the solvent structure. Using the three-dimensional grid on which the electron density
is calculated, we consider correlations between each grid point within $2.4\text{Å} < r < 6\text{Å}$ of a surface atom. Because grid points are roughly $0.4\text{Å}$ apart, the resulting coarsening is similar to that of both the RDF and the ADF, allowing for a meaningful comparison between their correlation coefficients. The discrepancy between the real-space simulation and experimental maps is also measured separately for the first solvation shell and for protein-protein contacts. Note that the latter are defined as the grid points at least $2.4\text{Å}$ and at most $3.0\text{Å}$ away from a pair of N or O atoms situated on different protein dimer copies.

Because the real-space distribution of water is calculated by averaging over exact electron densities calculated from MD snapshots, the peak shapes and locations are affected by the precise motion of the water molecules. In order to compare the water density peak locations, we eliminate the role of water density widths and shapes by selecting only peaks that appear above a given threshold density, $\rho_{th}$. We additionally deconstruct the solvent density by focusing exclusively on crystal waters, which by definition are associated with an observed local electron density well above experimental noise. This comparison thus deconvolutes the role of peak shape from that of peak location in assessing the density profile. Following\[
\text{(author?)},
\]
we define a prediction $A_{\text{pred}}(\rho_{th})$ and a recall $A_{\text{rec}}(\rho_{th})$ score. The former yields the fraction of crystal waters that are within a distance smaller than the water radius, i.e., $\sim 1.4\text{Å}$, of an MD peak above the threshold, while the latter gives the fraction of MD peaks above the threshold that are within $\sim 1.4\text{Å}$ of a crystal water,

$$A_{\text{pred}}(\rho_{th}) = \frac{\sum_{\alpha \in P_{\text{MD}}} \Theta[\rho(r_\alpha) - \rho_{th}] \{1 - \prod_{\beta \in P_{\text{CW}}} [1 - w(|r_\alpha - r_\beta|)]\}}{\sum_{\beta \in P_{\text{MD}}} \Theta[\rho(r_\alpha) - \rho_{th}]},$$

$$A_{\text{rec}}(\rho_{th}) = \frac{\sum_{\beta \in P_{\text{CW}}} \{1 - \prod_{\beta \in P_{\text{CW}}} [1 - w(|r_\alpha - r_\beta|)]\Theta[\rho(r_\alpha) - \rho_{th}]\}}{|P_{\text{CW}}|},$$

(3.5)
where $P_{\text{MD}}$ is the set of MD peaks, $P_{\text{CW}}$ is the set of crystal waters, $\rho(r_\alpha)$ is the density that corresponds to peak $\alpha$, with $w(r) \equiv \Theta(1.4 - r)$ the overlap function defined in terms of the Heaviside $\Theta$ function, $|r_\alpha - r_\beta|$ is the distance between the MD peak $\alpha$ and the crystal water $\beta$, and $|P_{\text{CW}}|$ is the total number of crystal waters in the refined protein structure. In other words, $A_{\text{pred}}$ is the true positive rate, while $A_{\text{recall}}$ is the true negative rate. Note that to assess the structural significance of the measured signal, we further compute these scores with a random distribution of crystal waters with the same number density in the solvent region. This null model results in a constant $A_{\text{pred}}(\rho_{\text{th}}) = 0.1$, and a $A_{\text{rec}}(\rho_{\text{th}})$ that steadily decays from 0.2 as $\rho_{\text{th}}$ increases, both values being well below the level of the measured signal.

Finally, we compare the experimental and MD densities in reciprocal space by generating a model of the protein unit cell that combines the simulated density with the protein model (see SI[4]). Comparing the resulting $R_{\text{work}}$ of this model with that of the original protein model determines whether or not the simulated densities improve the agreement with the experimental data. For this analysis, we estimate the error in the $R_{\text{work}}$ values due to measurement errors to be one part in ten thousand (95% confidence interval, see SI). This analysis can also be performed by partitioning the set of reflections into different resolution bins and calculating $R_{\text{work}}$ for each. Because higher resolution bins correspond to more structured parts of the unit cell, such as the protein atoms and ordered water molecules around the protein surface, while lower resolution bins correspond to regions with flatter electron density, such as the bulk solvent[192], this analysis provides insight into the regions of MD-generated solvent density that better agree with experimental data.

### 3.2.5 Inferring Protonation States

The solvent distribution is a reflection of its environment. Given sufficiently accurate solvation information, it should thus be possible to determine the protonation state
Table 3.1: Default vs inverted protonation states

<table>
<thead>
<tr>
<th>residue</th>
<th>default</th>
<th>inverted</th>
</tr>
</thead>
<tbody>
<tr>
<td>histidine</td>
<td>$N_{\delta 1}$ protonated or $N_{\epsilon 2}$ protonated</td>
<td>$N_{\epsilon 2}$ protonated or $N_{\delta 1}$ protonated</td>
</tr>
<tr>
<td>lysine</td>
<td>$N_{\zeta}$ has 3 protons</td>
<td>$N_{\zeta}$ has 2 protons</td>
</tr>
<tr>
<td>aspartate</td>
<td>both $O_{\delta 1}$ and $O_{\delta 2}$ deprotonated</td>
<td>either $O_{\delta 1}$ or $O_{\delta 2}$ has 1 proton</td>
</tr>
<tr>
<td>glutamate</td>
<td>both $O_{\epsilon 1}$ and $O_{\epsilon 2}$ deprotonated</td>
<td>either $O_{\epsilon 1}$ or $O_{\epsilon 2}$ has 1 proton</td>
</tr>
</tbody>
</table>

of a residue. To test this hypothesis, different MD simulations were run for alternative side-chain protonation states, and the resulting water density was compared with the experimental density. The default Gromacs protonation states for a subset of histidine, glutamate, aspartate and lysine residues were inverted in different simulations. The default and inverted protonation states for the residue types we study are summarized in Table 3.1. For glutamates, aspartates and lysines, the residues to be (de)protonated were chosen, such that: (i) they have one surface side chain oxygen or nitrogen; (ii) they are at least 6Å away from another residue chosen for protonation analysis in the same simulation to avoid interference between the solvent distribution of one residue with the other; and (iii) do not neighbor a Yb$^{3+}$ ion and thus are not affected by the approximations to its force field. We further verify that the water density in the vicinity of these examples is well sampled by making sure that all the surrounding water molecules decorrelate in at most $\sim 1$ ns, and that observations are consistent for all four protein dimer copies. This whole set of simulations was run with the TIP4P water model.
3.3 Results and Discussion

In this section, the experimental and MD solvent information is used to assess the quality of the MD description first by comparing density profiles, and second by using standard crystallographic observables. The potential to infer the protonation state of residues from MD solvent density is also examined.

3.3.1 Real-Space Comparison of Water Densities

The RDF, which is a quintessential quantity in liquid state theory [79], has been utilized as main observable by most prior studies of macromolecular solvation [37, 115, 111, 185]. Some of these have even attempted to reconstruct protein hydration from RDFs alone [111, 115, 185]. It is therefore a natural starting point for our evaluation.

Comparing the RDF for different atom types and water models reveals that the various descriptions qualitatively agree with one another (Fig. 3.2). In particular, a clear first solvation shell is noted, and hints of a second shell can be gleaned, although the number of available grid points beyond 6Å is too small to obtain a reliable profile of that shell. Because of experimental noise and artifacts, such as Fourier ripples (see SI), it is difficult to determine whether the first peak position of the various simulation models match that of the experimental RDF. The first peak positions of all water models, however, agree with each other within the error margin, with the exception of TIP5P for surface oxygens, for which the peak is pushed further out. For nitrogens and oxygens, the peak amplitude is significantly higher in simulations than in experiment. One might be tempted to ascribe the sensitivity of this feature to the choice of $B$-factor for water. Some water molecules are indeed less localized than others, especially near fairly mobile surface protein atoms. Hence, no single $B$ can reliably describe all water molecules. The fact that neither nitrogens ($B = 20.2\text{Å}^2$)
nor oxygens \((B = 18.7 \AA^2)\) have significantly higher average \(B\)-factors than carbons \((B = 19.3 \AA^2)\) does not rule out this possibility, as \(B\)-factors are unreliable estimates of thermal motion in protein crystals \[\text{[157]}\]. It is also possible that the peak intensity could be weakened by experimental noise and artifacts (see SI).

![Figure 3.2: Averaged RDFs for surface (a) N, (b) O, and (c) C atoms, from different water models. Results obtained from different water models agree well with each other, as well as with the experimental RDFs.](image)

The overall shape of the RDF should nevertheless be insensitive to these effects. The Pearson correlation coefficients between the averaged RDFs of surface N, O, and C atoms reveal that the water density in the vicinity of surface oxygens and carbons is more accurately reproduced than around nitrogens (dashed lines in Fig. 3.3a). However, the radial correlation coefficients of RDFs for individual atoms (solid lines in Fig. 3.3a) suggest that the distributions around oxygens are significantly worse. The radial distribution of water around individual oxygen atoms appears to depend more sensitively on the chemical environment than around nitrogens and carbons. We also conclude that the distribution of water around each atom is far from universal. Efforts to reconstruct water density using averaged radial distribution functions – as was previously attempted \[\text{[115, 185]}\] – therefore have serious shortcomings. Interestingly, all water models perform identically within the estimated error, for both average and regular radial correlation coefficients. We get back to this point below.
Angular correlation coefficients are generally slightly lower than their radial counterparts. This effect is consistent with the latter being a higher-order structural feature. One might nonetheless expect that a model parameterized to more accurately reproduce the subtle orientational order of the various bulk water crystal phases\cite{2}, such as TIP4P/2005, or a model like TIP5P, which explicitly treats tetrahedral point charges, to improve the description of ADFs. Neither TIP5P nor TIP4P/2005, however, perform significantly better than the other water models, including TIP4P.

Angular correlation coefficients are generally larger for nitrogens and oxygens than for carbons, which is particularly interesting. The orientation of water molecules around surface nitrogens and oxygens indeed mostly results from direct hydrogen bonding, while that of water molecules around carbons are affected by their interplay with the broader hydrogen-bond network and are thus less constrained by the protein force field. The resulting hydrophobicity is structurally more subtle to capture, which likely explains why water models struggle to capture this effect (Fig. \ref{fig:3.3}b). Water models that account more accurately for many-body correlations in water, such as E3B \cite{174} and E3B2 \cite{175}, might improve the orientational description of water in these systems. Direct tests, however, are not immediately possible because these models have not yet been parameterized for macromolecular solvation.

Because it contains higher-order structural information, the real-space distribution generally gives rise to significantly lower correlations than either the radial or the angular correlation coefficients (Fig. \ref{fig:3.3}). While water models capture the radial distribution of water around carbons equally well as around nitrogens, they rank last in spatial correlation coefficients. This is consistent with their poor performance describing angular correlations. Similarly, models reproduce the angular distribution around oxygens as well as around nitrogens, but perform worse for real-space correlations.

To gain further insight into the aspects of water models that increase their propen-
Figure 3.3: (a) Radial (solid) and averaged radial (dashed), (b) angular, and (c) real-space correlation coefficients for surface N (blue), O (red) and C (yellow) atoms. Real-space correlation coefficients for first-layer (green) and contact waters (black) are also given in (c). Error bars denote 95% confidence intervals. The lines connecting the data points are solely a guide for the eye and have no physical meaning.

Density to capture water structure around biomolecules, we compare the spatial distribution of water in different regions of space. We first calculate real-space correlations separately for contact and first-layer waters. Correlations for the first shell are consistent with the overall real-space correlations for surface N, O, and C. Beyond the first layer, errors get amplified by structural imprecisions in the first layer, a situation further worsened by the reduced number of grid points in that region of space. Protein-protein contacts, by contrast, show fairly good structural agreement. This likely results from the surface atoms in these regions being much less mobile than elsewhere, and from steric constraints there playing a larger role in dictating the solvent structure. The position and orientation of water molecules in protein contacts are thus likely less sensitive to water model and protein force field parametrizations.

We next consider the recall and prediction scores (Eq. (3.5)) of the MD peak locations with the assigned crystal waters. At low threshold densities many MD peaks are identified and a high fraction of crystal waters are recovered, although only a few of these MD peaks are near crystal waters. As \( \rho_{th} \) increases, the number
Figure 3.4: Prediction (solid) and recall scores (dashed), as defined in Eq. (3.5). At low threshold densities, too many MD peaks are identified, resulting in a high recall score but a low prediction score. As the threshold increases, MD peaks with stronger signals persist, which at high densities predict roughly 70% of the crystal waters. However, the recall scores fall as the density increases, suggesting that a significant fraction of crystal waters do not overlap with an MD peak.

of MD peaks decreases, but a greater fraction of the remaining ones overlap with crystal waters, decreasing the rate of false negatives. This encouragingly suggests that the strongest predictions (and interactions) of the MD model correlate with crystal waters with reasonably high accuracy (70%). The recall scores, however, fall steadily with increasing $\rho_{th}$, and thus many crystal waters are not predicted by MD simulations. In other words, a low true positive rate is obtained. For all water models, the highest $A_{\text{pred}}$ and $A_{\text{rec}}$ is $\sim 0.7$. The discrepancy between MD and experiments is thus not purely due to imprecisions in the MD description of the shape and width of the density peaks, but also in their location. Some of this error is likely attributable to the protein force fields, as the location of high density peaks in the MD density are affected by the average positions of the nearby protein atoms throughout the simulation.
3.3.2 Reciprocal Space

The agreement between MD and experiments is assessed in reciprocal space by first combining MD densities with the refined PDB coordinates of the protein without the crystal waters. If MD simulations were to reproduce water densities reasonably well, the resulting $R_{\text{work}}$ would be less than that of the refined PDB model. Yet for the best water model (SPC) we obtain $R_{\text{work}} = 0.208$, which is significantly higher than $R_{\text{work}} = 0.159$ obtained for the refined protein model (Fig. 3.5). The difference in $R_{\text{work}}$ is also greater at higher resolution, suggesting that highly ordered solvent regions are not adequately captured. If we instead consider the entire solvent region to have a flat electron density, $R_{\text{work}} = 0.219$, which is about 1% worse than the best water model. Note that this increase is orders of magnitude larger than the estimated error in $R_{\text{work}}$ (see SI). Hence, although the MD models contain some information about the water density within the unit cell, a significant fraction of it is inaccurate.

To check whether MD simulations capture solvent structure that is complementary to that of the crystal waters assigned from the experimental data, we combine MD densities with the refined PDB coordinates, including crystal waters. (The MD electron density of crystal waters is thus removed.) This strategy reduces $R_{\text{work}}$ for water models to $R_{\text{work}} = 0.162$, which is an improvement over the previous scheme yet still appreciably higher than the refined model $R_{\text{work}} = 0.159$. The gap between $R_{\text{work}}$ values at lower resolutions at least is then closed.

It is important to note that $R_{\text{work}}$ at high resolution is affected not only by the water density around each protein atom, but also by the average protein atom positions in the MD simulation being slightly different from that of the refined protein structure. Although refined coordinates are used for this analysis, the water density is affected by the slightly perturbed protein atom locations throughout the MD simulations, resulting in possible overlaps between the solvent density and refined
atom positions.

Figure 3.5: $R_{\text{work}}$ in different resolution bins for the original model (EXP, black), and for models constructed by combining MD densities with the protein model. The overall $R_{\text{work}}$ is as given in the legend. Dashed lines correspond to $R_{\text{work}}$ for models with the refined crystal waters combined with the MD density.

Retaining crystal waters in the refined protein structure results in substantially lower $R_{\text{work}}$ at high resolution than the MD density results alone. This strategy thus yields results comparable to the original protein model, but as resolution decreases, $R_{\text{work}}$ becomes significantly worse than for the original protein model, which once again confirms that the refined model describes the electron density in the unit cell more accurately than the MD solvent density.

3.3.3 Inferring Protonation States

A complication that hinders the improvement of the solvent description in the analysis of X-ray diffraction experiments is that hydrogens, which are surrounded by relatively small electron clouds, cannot be detected unless a remarkably high diffraction resolution, i.e., better than roughly 0.7-1.0Å, is achieved. Although the position of many of the hydrogens on the protein chain can be inferred based on an elementary
description of bonding (partly explaining the success of structure validation tools, such as MolProbity [44]), side chain protonation states can remain somewhat ambiguous, as do the positions of side chain hydrogens with a rotational degree of freedom. This problem is especially acute for side chains that contribute to an enzymatic pathway [188] or to protein-protein interactions [187], such as salt-bridges [9, 49, 65]. Prediction servers have thus been developed to infer pKₐ values and titration curves of individual side chains, based on the electrostatic properties of neighboring residues [73, 154]. Other software packages rely on less involved algorithms to assign protonation states. For instance, MolProbity picks the most suitable protonation state and hydrogen atom position that minimizes clashes, while Gromacs [21] analyses the hydrogen bonding network around it [109]. Yet because the presence or absence of protons affects the solvent distribution around these sites, probing the solvent distribution around such residues should allow one to determine their protonation state more systematically.

The preceding analysis suggests that reconstructing the solvent density, and hence predicting every density peak, is not possible using existing water models. We are nevertheless encouraged by the fact that MD simulations reproduce a significant fraction of the high intensity peaks associated with crystal waters. It may thus be possible to infer protonation states by comparing the overlap between MD peaks and crystal water, if changing the protonation state of a residue gives rise to or eliminates such high intensity peaks in the MD solvent density.

In most cases considered here, either residues have insufficient solvent exposure to conduct the analysis, no significant difference in solvation is observed, or both sets of density patterns are similarly mismatched with the refined structure. Despite this, a few examples for which inverting the protonation state of a residue significantly affects the surrounding water distribution can be found.

Removing one of the three protons of the default +1 charged LYS 145 in chain B
results in a slightly better overlap with two crystal waters (CW), labeled CW1 and CW2 in Figure 3.6a. There is, however, a third crystal water, CW3, within hydrogen bonding distance to the nitrogen that is unexplained by either protonation state. Although this lysine residue is relatively well localized and its average position does not deviate from that in the refined structure, MD models completely miss CW3. In addition, both protonation states result in an MD peak with the same orientation as the crystal waters, because removing a proton does not drastically change the geometry of the remaining two hydrogens. The MD peaks in the deprotonated case are, however, pushed away from the protein, likely due to the altered charge distribution in the residue. We conclude that a neutral lysine with two protons at this position leads to a water density that is more consistent with the experimental density, although with caveats.

The case of ASP 200 in chain A is slightly more complicated. The MD peak resulting from the protonated case agrees better with CW1, compared to the peak resulting from the simulation in which the residue is unprotonated (Fig. 3.6b). However, two crystal waters (CW2 and CW3) overlap only with high intensity peaks in the density obtained with the standard protonation state. It is therefore likely that this residue is not protonated, but it is unclear why the protonated case better explains the CW1 peak.

For both GLU 130 in chain B and GLU 218 in chain A, the unprotonated case gives better agreement between MD peaks and crystal waters. Protonating the former results in a loss of an MD peak that overlaps the crystal water (Fig. 3.6c). Similarly, protonating GLU 218 in chain A results in the loss of an MD peak that overlaps CW1, but retains those on CW2 and CW3 (Fig. 3.6d). This is likely because CW3 is still within hydrogen bonding distance to the residue, and CW2 is within hydrogen bonding distance to CW3. However, the disappearance of the density peak on CW1 is unexpected because the protonated oxygen could still form a hydrogen
bond to a water at that location.

While these results are encouraging, their robustness with respect to protein atom positions remains untested. The location of high density peaks in the MD density is likely affected by both the equilibrium position of protein atoms and the degree to which they are localized. In the case of CW1 near GLU 218 in chain A, for instance, the MD-density peak on CW1 might be missing in the protonated case (even though a water molecule at this location would be within hydrogen bonding distance), because other water molecules might be forming a more stable hydrogen bond network nearby. In addition, the success of these inferences ultimately depends on our ability to reliably reconstruct the water density around proteins by MD simulations. Using this method to detect protonation states thus relies on being well above the noise inherent to the structural analysis.

3.4 Conclusion

Using a protein with a high-quality dataset from X-ray crystallography, we have attempted to extract complementary information about water structure in protein crystals from diffraction data and MD simulations. This work improved upon earlier efforts in a few different ways. (i) We used a simulation box equivalent to the protein unit cell, containing multiple protein copies in order to capture water structure in the protein-protein interfaces. (ii) We ran significantly longer simulations, which enabled the solvent configurations to be ergodically sampled. (iii) We used reference diffraction data, for which experimental phases is available. Thanks to these advances, it was possible for us to compare more detailed aspects of the water structure and to explore the role solvent structure plays around titratable residues.

Comparison of experimental and MD densities in real space revealed that although water models are relatively good at capturing the radial distribution of water near the protein surface, they struggle to predict angular distributions and
are somewhat deficient at reconstructing the overall water density. The relatively poor distribution of water around carbons, in particular, suggests that the structural consequences of hydrophobic effect are inadequately captured by these models. Remarkably, all water models we considered were found to behave rather similarly at the structural level.

Although MD water models are insufficient for reconstructing biomolecular hydration with a precision sufficient to conduct structural refinement, they nonetheless capture the position of a fraction of the crystal waters. In optimal hydration circumstances, these models may thus help assign protonation states to some of the protein side chains. The robustness of these predictions with respect to the choice of parameters, including the protein force field and the protonation state of the nearby residues, is untested, but these findings nevertheless suggest MD simulations can provide information complementary to what is available from X-ray crystallography. It may further be possible to devise refinement schemes that utilize this information to improve biomolecule structure quality, similar to some already existing schemes[36, 115].

Our results suggest that it may be necessary to add more features to the common water models in order to reconstruct accurately the solvent structure around biomolecules. A re-parametrization of the existing models taking into account properties pertaining to protein-water interactions might improve the description of these interactions. Such a re-parametrization may not, however, adequately capture both bulk and interfacial water properties at once. A single, fixed dipole moment might indeed not be able to capture the behavior of water in both of these regions [202]. Considering more complex models that include polarizability[180] or include three-body interactions[46], might provide a more robust starting point. To model a process that depends sensitively on the position of water molecules, it might be preferable to consider even higher-accuracy models of water that include \textit{ab initio} descriptions.
The use of models such as those based on quantum mechanics \cite{204} may eventually become computationally tractable, allowing for a refinement process in which simulations are run at each refinement step.
Figure 3.6: Comparison of water density distribution for simulations that contain different protonation states for (a) LYS 145 in chain A (the blue blob is behind the water and does not overlap it), (b) ASP 200 in chain A, (c) GLU 130 in chain B, and (d) GLU 218 in chain A. The water density from the default protonation state simulations are shown in blue wireframe, and the alternate protonation state simulations are shown in red wireframe. For all snapshots the isosurfaces are contoured at 0.88 e\textsuperscript{−}/Å\textsuperscript{3}. Crystal waters (CW) from the refined protein structure are denoted with red spheres.
Parameterizing a Patchy Model to Understand Inverted Solubility in \( \gamma \)D-Crystallin

Reprinted (adapted) with permission from


Amir Khan, Susan James, Michelle Quinn, and Jennifer McManus obtained the experimental phase diagram and solved the crystal structures of the double mutant. Irem Altan and Patrick Charbonneau developed and parameterized the coarse-grained model and obtained the theoretical phase diagram.

4.1 Introduction

The rationalization of protein crystallization remains a major obstacle to efficient structure determination — a requirement to understand the molecular basis for many diseases and to pinpoint targets for new drug development [122]. Sampling hundreds (or sometimes thousands) of solution conditions (i.e., mixtures of different buffers, salts, and precipitants) is often the most productive strategy to identify lead conditions for protein crystallization. Even when coupled with rational design strategies such as surface entropy reduction [72], this approach can be time-consuming and costly because screening methods often fail to produce crystalline material or diffraction-quality crystals. Protein phase diagrams that map how a given protein behaves across sets of solution conditions dramatically improve the success of the process and narrow the screening required for producing diffraction-quality crystals but have only been measured for a small number of proteins ([120] and references therein). These reference studies have identified key challenges in guiding and improving protein crystallization.

An excellent such reference is human γD-crystallin (HGD), a major structural protein found in the eye lens. HGD is unusually stable in the eye lens in mixtures with α- and β-crystallins, often over a whole lifetime [205]. Its phase behavior is otherwise generally similar to that of a large group of important globular proteins that includes hemoglobin [66], immunoglobulins [190], lysozyme [129], and thaumatin [10]. These phase diagrams are defined by net attractive short-range interactions that result in liquid-liquid phase separation and crystallization. Although native HGD itself does not readily form crystals, several of its genetic- cataract-related single-amino-acid substitutions do so easily, without any major structural changes [80, 171, 140, 139, 19, 17]. The P23T substitution—which is a naturally occurring mutation associated with congenital cataracts—however, has unusual phase behavior in that its aggregates
have inverted solubility, i.e., they melt as temperature is decreased [137, 121]. As a result, the protein is insoluble at physiological concentration and temperature, leading to eye-lens opacity. In the related P23V mutant, both aggregates and crystals are observed, both also with inverted solubility [137, 121]. Yet, crystallization of the P23T mutant under physiological conditions has remained elusive. Numerous structural and biophysical studies, including x-ray structures at pH 4.5 and NMR solution studies, have failed to unambiguously identify major structural changes in the P23T mutant, and hence a full explanation for its anomalous inverted solubility is still unknown [18, 94, 90].

Physicochemical insights into protein phase behavior – both normal and anomalous – are often gleaned from colloidal science. Simple colloidal models do capture key features of protein phase diagrams, such as their metastable critical point [11]. However, protein phase diagrams cannot be completely rationalized without including some level of anisotropy, in terms of the directional contacts between proteins in solution or within a crystal lattice [113, 51, 164] or of shape anisotropy [172]. This anisotropy gives rise to rich protein phase diagrams and is more widely exploited for the controlled assembly of biological and biomimetic materials [162]. It has even been proposed that these types of interactions are important in controlling liquid-liquid phase separation in cells [133], with important implications in understanding stress responses, RNA processing, and gene expression. However, understanding and predicting anisotropic protein-protein interactions ab initio is not yet possible because of the extreme heterogeneity of amino acid side chains on the protein surface. Although measurements indicative of net protein-protein interactions such as the osmotic second virial coefficient, $B_{22}$, or the diffusivity constant, $k_D$, can provide some insight, they reflect the averaged pair interactions between proteins. These parameters are typically insufficient to trace back the specific, directional protein-protein interactions that control the dramatic (and often unpre-
dictable) changes in protein assembly upon mutagenesis [17, 121]. Enhanced numerical models that capture the details of anisotropic protein-protein interactions may allow for the prediction of protein phase diagrams and hence optimal crystallization conditions [17, 11, 113, 51, 164, 29, 47, 148, 172, 149, 65, 64, 63]. To identify the microscopic origin of inverted solubility, however, we need high-resolution structural information detailing the underlying anisotropic interactions, using, for instance, crystal structures of the protein of interest.

To design a P23T mutant that crystallizes at pH 7, we focused our interest on HGD mutant structures that do not form specific protein-protein contacts near proline 23. One such mutant, R36S, readily crystallizes by forming a crystal lattice contact at position 36. By combining the R36S and P23T substitutions, we reasoned that crystals of the double mutant would display inverted solubility based on a comparison of the phase diagrams for the single-mutant proteins, thus providing insights into the mechanism for the P23T mutant retrograde solubility. Remarkably, the double mutant P23T + R36S formed two distinct crystal forms—one with normal solubility and one with inverted solubility [89]. Although inverted solubility in proteins has been previously observed, a protein that forms two distinct crystal lattices, each with opposite temperature dependence of the solubility line, had not, and therefore, this double mutant offers a rare opportunity to access the microscopic origin of solubility inversion, which we now probe further.

Here, we report the x-ray structures of the two crystal forms of the P23T + R36S mutant of HGD. We find that the two are polymorphs with different unit cells and crystal contacts and that it is possible to interchange between them solely by varying the solution temperature. In the inverted solubility crystal, a lattice contact involving the cataract-associated Thr23 residue is formed. This is a new contact with the same binding energy determined from a statistical mechanics analysis of the chemical potentials of the solubility lines in earlier work [121]. We have used both the phase
diagram for P23T + R36S and crystal structures to design a custom patchy particle model that incorporates specific contacts formed in the crystal lattice. We find that when temperature-dependent patchy interactions are included, the temperature dependence of the solubility lines for both crystal lattices can be reproduced by simulations performed using the custom model. Specifically, we show that a change to the contact that contains the 23rd residue in the inverted solubility crystal is sufficient to cause inverted solubility. This contact becomes engaged as temperature increases, stabilizing the inverted solubility crystal phase and thus revealing the molecular origin of the inverted solubility for P23T.

4.2 Materials and Methods

4.2.1 Preparation and Characterization of Double Mutant

The double mutant was created, expressed, and purified as described previously [89]. Sodium dodecyl sulfate-polyacrylamide gel electrophoresis and size-exclusion high performance liquid chromatography were used to confirm protein purity at >98%. The intact molecular weight for the mutant protein was analyzed by electrospray ionization mass spectrometry (Finger Prints Proteomics Facility, College of Life Sciences, University of Dundee, Dundee, UK), which confirmed a molecular mass of 20,541 ± 1 Da for the P23T + R36S mutant.

4.2.2 Crystallization and Data Collection

The crystals from the P23T + R36S double mutant of HGD protein were obtained and grown in capillaries in 100 mM sodium phosphate buffer (pH 7) in the presence of 20 mM dithiothreitol. The solution concentration of protein was in the range of 1–2 mg/mL, and there was no additional precipitant in the solution. Crystals of the double mutant with inverted solubility (DBI) were grown at 310 K, whereas crystals with normal solubility (DBN) grew at 277 K. Both crystal types formed within a
few hours of incubation at the relevant temperature. Crystals were harvested from capillaries and mixed with 25% glycerol, flash-cooled in liquid nitrogen, and subjected to x-ray diffraction. Data sets from two crystals, one grown at 310 K and the second grown at 277 K, were collected at the PX2 beamline at Le Soleil Synchrotron (Saint-Aubin, France) on an ADSC Q315 detector (Area Detector Systems Corporation, Poway, CA).

4.2.3 Solubility Measurements

Protein solutions were prepared initially by diafiltration against 100 mM sodium phosphate buffer (pH 7.0) using Ultrace1 10 KDa ultrafiltration disks (Merck Millipore, Tullagreen, Ireland). Protein concentrations for the double mutant was measured by ultraviolet absorbance using the extinction coefficient value of 2.09 mg\(^{-1}\)mL cm\(^{-1}\) after filtration through 0.22 mm Millex-GV Millipore (Merck Millipore) syringe-driven filters. When required, protein solutions were further concentrated by ultrafiltration using Amicon Ultra-4 centrifugal filter units (Merck Millipore) and the protein concentration reestablished by ultraviolet absorbance.

4.2.4 Data Processing and Structure Solution

The structure of DBN was solved using the model that contains the R36S single-site mutation in HGD (Protein Data Bank (PDB): 2G98) \(^{101}\). The program Phaser \(^{119}\) provided the starting model, which was improved through cycles of manual model-building using Coot \(^{54}\) and Phenix refinement \(^3\). The structure of DBI was solved using the high-resolution 1.25Å structure of wild-type HGD (PDB: 1HK0 \(^{19}\)). The obtained structure was further refined using the same refinement procedure as for DBN. Statistics from the data collection and refinement strategies are detailed in Table 1. Crystal contacts determined from the structural analysis were used to determine the patch-patch interactions for the phase diagram, as described
Table 4.1: Data Collection and Refinement Statistics

<table>
<thead>
<tr>
<th></th>
<th>DBI</th>
<th>DBN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution range</td>
<td>44.01 - 1.197</td>
<td>48.16 - 2.20</td>
</tr>
<tr>
<td></td>
<td>(1.24 - 1.197)</td>
<td>(2.277 - 2.20)</td>
</tr>
<tr>
<td>Space group</td>
<td>P 1 2 1 1</td>
<td>P 2 1 2 1</td>
</tr>
<tr>
<td>Unit cell (Å)</td>
<td>44.02 31.70 52.50</td>
<td>54.04 82.10 106.25</td>
</tr>
<tr>
<td></td>
<td>90 91.29 90</td>
<td>90 90 90</td>
</tr>
<tr>
<td>Total reflections</td>
<td>89029 (8455)</td>
<td>164792</td>
</tr>
<tr>
<td>Unique reflections</td>
<td>45768 (4504)</td>
<td>24656 (2391)</td>
</tr>
<tr>
<td>Multiplicity</td>
<td>1.9 (1.9)</td>
<td>6.7</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>99.46 (99.03)</td>
<td>99.63 (98.03)</td>
</tr>
<tr>
<td>Mean I/sigma(I)</td>
<td>8.93 (1.22)</td>
<td>9.2 (1.4)</td>
</tr>
<tr>
<td>Wilson B-factor</td>
<td>10.73</td>
<td>38.69</td>
</tr>
<tr>
<td>R-merge</td>
<td>0.03956 (0.5534)</td>
<td>0.129</td>
</tr>
<tr>
<td>R-meas</td>
<td>0.05595 (0.7827)</td>
<td>0.139</td>
</tr>
<tr>
<td>R-pim</td>
<td>0.03956 (0.5534)</td>
<td>0.053</td>
</tr>
<tr>
<td>Reflections used for R-free</td>
<td>1999 (190)</td>
<td>1233 (119)</td>
</tr>
<tr>
<td>R-work</td>
<td>0.1493 (0.2662)</td>
<td>0.2304 (0.2993)</td>
</tr>
<tr>
<td>R-free</td>
<td>0.1794 (0.2835)</td>
<td>0.2658 (0.3614)</td>
</tr>
<tr>
<td>Number of heavy atoms</td>
<td>1687</td>
<td>2888</td>
</tr>
<tr>
<td>macromolecules</td>
<td>1479</td>
<td>2776</td>
</tr>
<tr>
<td>solvent</td>
<td>208</td>
<td>112</td>
</tr>
<tr>
<td>Protein residues</td>
<td>173</td>
<td>341</td>
</tr>
<tr>
<td>RMS(bonds)</td>
<td>0.013</td>
<td>0.025</td>
</tr>
<tr>
<td>RMS(angles)</td>
<td>1.21</td>
<td>1.49</td>
</tr>
<tr>
<td>Ramachandran favored (%)</td>
<td>98.83</td>
<td>94.93</td>
</tr>
<tr>
<td>Ramachandran allowed (%)</td>
<td>1.17</td>
<td>4.18</td>
</tr>
<tr>
<td>Ramachandran outliers (%)</td>
<td>0.00</td>
<td>0.9</td>
</tr>
<tr>
<td>Rotamer outliers (%)</td>
<td>1.85</td>
<td>0.69</td>
</tr>
<tr>
<td>Average B-factor</td>
<td>14.98</td>
<td>56.9</td>
</tr>
<tr>
<td>macromolecules</td>
<td>13.30</td>
<td>57.26</td>
</tr>
<tr>
<td>solvent</td>
<td>26.91</td>
<td>47.89</td>
</tr>
<tr>
<td>chain A (DBN): 1-81</td>
<td>---</td>
<td>40.16</td>
</tr>
<tr>
<td>chain A (DBN): 82-173</td>
<td>---</td>
<td>40.81</td>
</tr>
<tr>
<td>chain B (DBN): 1-81</td>
<td>---</td>
<td>51.48</td>
</tr>
<tr>
<td>chain B (DBN): 82-173</td>
<td>---</td>
<td>101.38</td>
</tr>
</tbody>
</table>
4.2.5 Description of the Model

Because transitions between the two crystal forms occur upon temperature change, we consider the phase behavior of the double mutant using a patchy particle model with temperature-dependent patches. This choice accounts for the associated change in bonding free energy [121].

Proteins are modeled as patchy particles with interactions adapted from the Kern-Frenkel model [96]. Hard spheres with a diameter $\sigma$, chosen as the largest center of mass distance between protein-protein crystal contacts, interact with directional, attractive patches of range $\lambda_{\alpha\beta}\sigma$ for each protein-protein contact $\alpha\beta$. The patch interaction potential,

$$u(r_{ij}, \Omega_i, \Omega_j) = u_{HS} + \sum_{\alpha,\beta} u_{\alpha\beta}(r_{ij}, \Omega_i, \Omega_j), \quad (4.1)$$

thus includes a factorized attractive contribution, $u_{\alpha\beta} = \nu_{\alpha\beta}(r_{ij})f_{\alpha\beta}(\Omega_i, \Omega_j)$, that depends on interparticle distance, $r_{ij}$, and particle orientations, $\Omega_i$ and $\Omega_j$. Its orientational component is

$$f_{\alpha\beta} = \begin{cases} 
1, & \theta_{\alpha,ij} \leq \delta_{\alpha} \text{ and } \theta_{\beta,ij} \leq \delta_{\beta} \\
0, & \text{otherwise}
\end{cases} \times \begin{cases} 
1, & \psi_{\alpha\beta,ij} \in [\phi_{\alpha\beta} - \Delta\phi_{\alpha\beta}, \phi_{\alpha\beta} + \Delta\phi_{\alpha\beta}] \\
0, & \text{otherwise}
\end{cases}, \quad (4.2)$$

where the first term ensures that patch vectors face each other (Fig. S1). The second restricts the torsion between the two particles (Fig. S2). Its radial component is a square-well potential

$$\nu_{\alpha\beta}(r_{ij}) = \begin{cases} 
-\varepsilon_{\alpha\beta}(T) & \sigma < r_{ij} < \lambda_{\alpha\beta}\sigma \\
0, & \text{otherwise}
\end{cases}, \quad (4.3)$$

65
where $\varepsilon_{\alpha\beta}$ is constant if the patch is not temperature-dependent and otherwise has a modulated interaction $\varepsilon_{\alpha\beta}(T) = (\tilde{\varepsilon}_{\alpha\beta}/2)(1 + \text{tanh}((T - T_a)/\tau_{\alpha\beta}))$ [48], which becomes deactivated below temperature $T_a$ over a rate set by $\tau_{\alpha\beta}$, thus capturing the change in free energy upon increasing the temperature. Model parameters were determined from all-atom molecular dynamics simulations of all patches of DBI and DBN (see Supporting Materials and Methods for methodological details and model parameters).

### 4.2.6 Phase Diagram Determination

The phase diagram of the schematic model was obtained by specialized Monte Carlo simulations: 1) the reference crystal free energies were obtained by integrating from an ideal Einstein crystal using the Frenkel-Ladd method [60]; 2) the chemical potentials of the crystal phases as a function of temperature were obtained by thermodynamic integration along isobars from the reference in 1; and 3) fluid free energies were approximated using the second virial coefficient, $B_{22}$, because of the inefficiency of traditional Monte Carlo sampling at low densities (Supporting Materials and Methods). Coexistence points between the fluid and crystal phases were determined from the intersection of chemical potential curves, and coexistence lines were then traced out using a Gibbs-Duhem integration scheme [103, 104].

### 4.3 Results and Discussion

The equilibrium phase diagram for P23T + R36S is shown in Fig. [89]. Two different crystal types are observed, distinguished by the temperature dependence of their respective solubility lines; one with normal solubility (DBN), which melts as temperature increases, and a second with inverted solubility (DBI), which forms at higher temperatures and melts as temperature is lowered. The solubility lines intersect at $\sim 303$K, where both crystal forms coexist. Remarkably, the two crystals
form under near-physiological conditions of temperature, pH, and salt, unlike the previously determined structure of P23T (pH 4.6, PEG4K) [90].

Figure 4.1: Experimental phase diagram for P23T + R36S mutant of HGD, indicating the equilibrium phase boundaries for the two crystals formed and their respective fluid phases (solid lines: data taken from [89]). The volume fraction ($\phi$) is calculated as $\phi = c \times \nu_{sp}$, where $c$ is the concentration of protein in mg/mL and $\nu_{sp}$ is the partial specific volume = $7.1 \times 10^{-4}$ mg/mL [121]. Coexistence of the two crystals is observed at the temperature at which the phase boundaries overlap (~303K).

The crystals have different morphologies: DBN crystals are rod-shaped, and DBI crystals are rhombic. The proteins remain in their fully folded globular state across the temperatures probed in this work. We further showed in previous work that this mutant protein displays no significant change to its secondary structure relative to native HGD [89]. Our determination of the structures for these two crystal forms by x-ray crystallography confirms this finding for these polymorphs.

The DBI (PDB: 6ETC) and DBN (PDB: 6ETA) structures consist of paired homologous domains that each adopt a Greek key motif. DBI crystallized as a monomer at high resolution (1.2Å), whereas DBN crystallized with two molecules in
the asymmetric unit at medium resolution (2.2 Å).

**Figure 4.2:** (A) Interactions between Thr23 and the crystal lattice. The side chain of Thr23 (gold) forms a hydrogen bond with the backbone of NH of Gly128, indicating the close contacts between the P23T locus and a symmetry-related molecule. (B) Crystal contacts near the R36S locus of DBI are shown. Ser36 does not make direct contacts in the crystal lattice. However, the side of Ser36 hydrogen bonds with Asp61, which additionally forms a salt bridge to Arg139 in a symmetry-related molecule. (C) Crystal contacts of DBN involve R36S. In this case, there is a hydrogen bond between Ser36 and Asn24.

The overall structures of DBN and DBI are otherwise highly conserved, with a root-mean-square deviation of 0.44 Å for the main chain atoms (residues 1–173; superposition of molecule A of DBN onto DBI). The side chain of Thr23 in DBI is involved in a crystal contact, within which it forms a hydrogen bond to the backbone of Gly128 in a symmetry-related molecule (Fig. 4.2A). This interaction is unique to DBI because Thr23 is not involved in lattice contacts of DBN. These polar interactions unambiguously demonstrate that the pathogenic P23T mutation enables direct interactions in the crystal lattice. By contrast, Ser36 is not involved in any direct contact in the DBI crystal (Fig. 4.2B). It only contributes a hydrogen bond within a DBN crystal contact (Fig. 4.2C).

If we are to relate our findings to the P23T single mutant, it is important to ponder whether the P23T + R36S mutant is a good model for it. The R36S contact is not activated in DBI, suggesting that it does not influence the structure of the DBI crystal to any significant extent, and the DBN crystal has the same structure
and lattice contacts as the R36S single mutant. Because P23T and R36S reside on opposite sides of the N-terminal domain, we expect the structural and energetic influence of the two to be uncoupled. The molecular interactions at the 36 locus are also distinct. In the structures of DBN and the R36S single mutant (PDB: 2G98) [17], Ser36 forms a hydrogen bond with Asn24 with a symmetry mate in the lattice. In contrast, DBI forms an intramolecular hydrogen bond with Asp62, which in turn ion-pairs with Arg140 in a symmetry-related molecule. Thus, Ser36 in DBI orients Asp62 lattice interactions, which is distinct from DBN.

There is a significant degree of flexibility in the C-terminal domain of DBN (molecule B), which likely explains why only a medium resolution structure could be obtained, as evidenced by the associated B-factors (Table 1). Other HGD mutants with medium resolution structures, namely the P23T (PDB: 4JGF) [90] and R36S (PDB: 4JGF) [139] single mutants, display comparable flexibility in the C-terminal domain. By contrast, the corresponding domain in the DBI crystal is more rigid. A stabilizing lattice contact is formed between Ser173 (Og) and a symmetry-related Gly157 (O), which is associated with the higher-resolution structure (Fig. 3A). Strikingly, the C-terminal carboxylate forms an ion pair with Arg141 from the same symmetry mate. There are also nonpolar interactions between Phe172 and Glu67 from a second symmetry-related molecule, indicative of the intimate associations between the C-termini of DBI in the crystal lattice. It is not clear, however, whether this flexibility is a result of a lack of a stabilizing crystal contact or whether, conversely, it precludes contact formation.

The formation of a hydrogen bond between Thr23 and the backbone of Gly129 in DBI suggests a molecular basis for a change in protein-protein interactions in the region of position 23 in the mutant protein. The change in the net binding energy between native HGD and the P23T single-mutant protein, calculated from the solubility data (16), corresponds to \( \sim 2.4 \kappa B T \), which is indeed the strength of
Figure 4.3: FIGURE 3 (A) Interactions at the C-terminus of DBI (gold). DBI is green, and symmetry-related molecules are gray and teal, respectively. (B) Flexibility of the C-terminal domain of DBN is shown. The two ribbon models (A and B) in the asymmetric unit are annotated by gradient colors of backbone B-factors. Regions with high values which denote flexibility are red, and ordered regions are white.

a typical hydrogen bond. Beyond this observation, there is no obvious structural basis for the inverted temperature dependence of the solubility line. Therefore, we employed a modeling strategy based on custom patchy particle colloidal models to investigate the microscopic origins of the inverted solubility of the double mutant.

The model describes proteins as having a hard, spherical core with directional, short-ranged attractive patches representing crystal contacts derived from the crystal structures (see Supporting Materials and Methods). DBI and DBN are modeled with five patches each, as determined from their crystal contacts, which we assume recapitulate the relevant physical chemistry for crystal formation (Fig. 4.4; see Supporting Materials and Methods for details, including the amino acids involved in the different contacts). Despite its very crude description of protein-protein interactions, such models can recapitulate the characteristic topology of protein phase diagrams. Because solubility inversion necessarily implies some degree of temperature dependence for the patch interactions [89], we first consider deactivating the contact that contains the 23rd residue, where the new crystal contact is formed,
around a temperature $T_a$ with rate $\tau$, set by the experimentally observed inverted solubility temperature and density ranges, respectively.

Simulations of this model with specialized Monte Carlo methods determined the equilibrium phase diagram shown in Fig. 4. Upon cooling, the model solubility line for DBI crystals reaches a minimal volume fraction, $\phi \sim 10^{-4}$, before exhibiting an inverted solubility regime, all in remarkable agreement with experimental observations. The DBN solubility line, which shows normal solubility, intersects with that of the DBI crystal around $\phi \sim 10^{-3}$, forming a triple point. Experimental results are also suggestive of a triple point for comparable densities, but the flatness of the DBI solubility line in this regime precludes its accurate determination. This model allows us to speculate about the phase behavior of other double mutants that could be designed similarly, i.e., R36S + P23S and R36S + P23V, knowing that the single mutants P23S and P23V also exhibit inverted solubility. Strengthening the patch containing the 23rd residue in the model would push the DBI solubility line to lower $\phi$, which suggests that the putative (inverted solubility) crystals of R36S + P23S and R36S + P23V may have higher solubilities than DBI. Such behavior is consistent with the binding energy estimates in [121, 89] and therefore presents a new, to our knowledge, mechanism for the inverted solubility of proteins. How common this mechanism is compared to other proposals, however, remains to be determined.

Note that although similarly deactivating a larger set of DBI patches can also reproduce the observed experimental phase behavior, no microscopic basis exists for these changes, and doing so to more than a couple of patches melts the crystal before solubility inversion can be observed. Prior experimental observations suggest that a change to surface hydrophobicity using either small molecule dyes [149, 150] or by mutagenesis at position 23 [138, 124] may give rise to entropic gain upon crystallization and could explain the lowered solubility of the mutant protein. The functional form of the temperature-dependent patch energy in our model may suggest
that additional flexibility in amino acid side chains with increasing temperature in
the solution phase may be more likely than a hydrophobic patch effect, but this
possibility cannot be excluded. However, because there is no experimental evidence
for local unfolding or structural changes, we should not exclude the possibility that
inverted solubility could have some other microscopic origin that has not yet been
considered.

4.4 Conclusions

The rational design of a double mutant based on phase diagrams of single-mutant
proteins has allowed us to produce two crystal forms of the P23T + R36S mutant
of HGD that are polymorphs with different unit cells and distinct crystal contacts.
The use of a single amino-acid substitution (R36S), previously shown to increase the
crystallization propensity of HGD and to be unrelated to the mutant under consideration (P23T), is not standard but could provide an alternative design strategy to assist large-scale crystallization screening. The crystal displaying inverted solubility (DBI) forms a hydrogen bond at position 23, which distinguishes it from other γ-crystallin structures. We employed crystallographic data for both crystals, which made further investigation of the microscopic origin of inverted solubility and greater understanding of the solution behavior of the P23T single mutant. By considering a patchy particle model parameterized for this particular system, the phase diagram for the double-mutant protein was reproduced by simulations. A single temperature-dependent contact, specifically the contact that includes the P23T mutation, is sufficient to explain the crystallization behavior for the protein. Activation of the patch that contains this mutation was found to stabilize the inverted solubility crystal. This overall analysis illustrates that although noncovalent protein-protein interactions are far from trivial and thus challenging to predict, the combination model and experimental phase diagrams could be a productive approach to rationalize and provide support for future crystallization studies.

4.5 Simulation Methods

This section details the technical aspects of model development and parameterization, as well as the final model geometry.

4.5.1 Calculation of Model Parameters

Model parameters are calculated from all-atom molecular dynamics (MD) simulations using Gromacs [1] (versions 5.1 and 2016). As input configuration, pairs of proteins were placed in their crystal contact conformation. The missing atoms and residues to the crystal structures were inserted and the resulting structure was placed in a simulation box along with water and salt. Because no force fields exist for sodium
phosphate, sodium chloride in identical ionic strength was included. The key impact of this salt on protein-protein interactions is to tune the ionic strength, hence the net result is equivalent. In addition, DTT was not included because its dominant role is to prevent disulfide bonds from forming, which is in any case impossible in our classical MD simulations. The simulation box was then relaxed by energy minimization followed by a short, 100 ps NPT simulation. Simulation parameters that are common to the various types of MD simulations are given in table S4.

In order to parameterize the patchy model, the potential of mean force (PMF) was obtained for each crystal contact. by running umbrella sampling simulations. These simulations were prepared by first pulling one of the protein copies away from the reference protein, following a direction perpendicular to the contact interface. Meanwhile, the reference protein is kept fixed by restraining the positions of three or four alpha carbons, chosen such that they are approximately equally-spaced, are not coplanar, and are not too close to the contact interface. To prevent the pulled protein from rotating, x- and y-coordinates of the same set of alpha carbons were also restrained. Pulling is done with a harmonic spring with a spring constant of 5000 kJ mol$^{-1}$ nm$^{-2}$, at a rate of 0.01 nm per ps. Input configurations for the umbrella sampling are then generated from the resulting trajectory, and then relaxed with a 100 ps NPT simulation with the same harmonic constraints as in the pull simulation, before running 20 ns-long trajectories. Force information is saved every 100 ps to generate the PMF with Weighted Histogram Analysis Method implemented within Gromacs [88].

The parameters for each patch were then determined as follows. Its square-well potential depth was chosen as the depth of the PMF from the well to the long-distance plateau. Its interaction range was calculated by matching its contribution to the second virial coefficient, $B_{2,\alpha\beta}$, of the square well potential to that of the PMF.
\[ B_{2,\alpha\beta} = \frac{1}{2} 4\pi \int \left( e^{-\beta U(r)} - 1 \right) r^2 dr, \]  
(4.4)

from which the interaction range is found to be

\[ \lambda_{\alpha\beta} = \left( \frac{3\int (e^{-\beta U_{PMF}(r)} - 1)r^2 dr}{e^{\beta \varepsilon_{\alpha\beta}} - 1} + 1 \right). \]  
(4.5)

An additional 20 ns simulation was performed to determine the patch width and the width of the torsion angle between two particles for each patch. For these simulations, one protein copy was fixed in place by restraining the same set of alpha carbons. The harmonic spring was active with the same strength, but the pull rate was set to zero and the center of mass distance was kept at the equilibrium distance. Snapshots saved every 10 ps were analyzed to determine the vector perpendicular to the protein-protein interface. The dot product of the average vector with instantaneous vectors were calculated, and their mean was chosen as \( \cos \delta_{\alpha} \), the patch width. The width of the dihedral angles, \( \Delta \phi_{\alpha\beta} \), is calculated from the same trajectory by applying the rotation of the protein copies to a reference vector and calculating the angle between the two planes defined by (i) the first rotated vector and the center of mass vector, and (ii) the second rotated vector and the center of mass vector (Fig. 5.2).

The resulting parameters that result from this procedure are as reported in table S5 for contacts labeled cY, where Y is a Roman numeral for DBN contacts, and an Arabic numeral for DBI contacts. Contact 2 (c2) of DBI and contact I (cI) of DBN were found to be very similar in terms of geometry and the nature of their interactions and were hence merged. Similarly, contact V (cV) and VII (cVII) of DBN are almost identical and were also merged. The widths, range of dihedral angles, and interaction...
ranges for some patches for DBN had to be slightly widened to accommodate the crystal geometry of the model (see Model Geometry section below for details).

The model parameters obtained by all-atom simulations of the crystal structures give rise to patch free energies slightly higher than typical for protein crystal contacts, resulting in crystals remaining stable at unusually high temperatures. These patch free energies are nevertheless consistent with contact dissociation free energy estimates from the PISA server, it is therefore possible that some of the surface regions not considered as patches might have a slightly higher repulsion than estimated by the hard-core description. Because the qualitative picture that emerges from this analysis is only affected by relative patch strengths, however, this correction is not qualitatively significant. To correct for this effect, the patch deactivation temperature, $T_a = 1.9$, is thus chosen such that the minimum packing fraction of DBI in its solubility line is around $10^{-4}$, as in experiments. (Note that the unit of temperature is such that $T = 1$ corresponds to 277K, the experimental crystallization temperature for DBN.) We thus choose $\tau = 0.05$ so that patch deactivation occurs over a temperature range of order 10K, again paralleling the experimental observation. The triple point temperature obtained from simulations is $T_{tp} = 1.89$, slightly below $T_a$, and is used in the main text to correct at once for the slight overestimate of the patch energies compared with the experimental values.

4.5.2 Model Geometry

A number of modifications were made to the crystal geometry and to the patchy model so as to accommodate the relevant crystal forms and to simplify the computations. Because the DBI structure unit cell is almost orthorhombic: $\alpha = 90^\circ, \beta = 91.29^\circ, \gamma = 90^\circ$ we approximated it as a purely orthorhombic crystal. This results in minor changes in patch locations, which were previously shown not to affect phase diagram topology [8]. The other key feature is that the protein is slightly ellipsoidal,
while a patchy particle is perfectly spherical. For DBI, all protein copies in the crystal are aligned, hence shrinking the unit cell along the $z$-direction solves the problem. For DBN, however, the structure is more complicated because the long axis of some of the eight asymmetric unit cells are aligned either in the $x$- or in the $y$-direction, thus requiring a more significant compression of the crystal. As a result, both the protein crystal and the patch positions must be perturbed to remove clashes. In addition, contacts II, cIV, and cVI must be collocated, in order to satisfy the various bonding constraints. The final geometry of both crystal models is summarized in tables S6 and S7. Interestingly, dihedral constraints are here necessary to properly separate the crystal ground states within the energy landscape.

4.5.3 Phase Diagram Determination

The phase diagram of the patchy particle model is computed by specialized MC simulations. In general, 200,000 MC sweeps each of which consists of $N$ displacement and $N$ rotational moves for constant NVT simulation and an additional two volume moves for NPT simulations. The amplitude of the displacements is chosen such that an acceptance rate around 50% is obtained. System sizes are chosen to be similar from one phase to the other, while respecting the crystal symmetry, here, $N_{\text{DBI}} = 432$ and $N_{\text{DBN}} = 512$. For convenience, these simulations report distances in units of $\sigma$, and energies in units of $k_B T$ with $T = 277 K$.

The details for obtaining crystal and fluid chemical potentials are given in Sec. 5.3.3. Coexistence points were obtained from the intersection of the crystal and fluid isobaric temperature-chemical potential curves. Solubility lines were then traced by numerical Gibbs-Duhem integration [9] [10]. Both boxes were simulated for 200,000 MC sweeps for each temperature. The slope of the solubility line was corrected every 1000 MC sweeps. The temperature was propagated in steps of $\Delta \beta = 0.003$. Note that introducing temperature-dependent patches requires modifying the standard
Clausius-Clapeyron equation \[103\]

\[
\frac{dP}{dT} = \frac{\Delta e + p\Delta v - T\Delta \langle \frac{\partial u}{\partial T} \rangle}{T\Delta v},
\]

(4.6)

where $\frac{\partial u}{\partial T}$ is the derivative of the interaction energy with respect to the temperature, and the angular brackets denote an ensemble average. This additional term here allows for the crystal solubility curve to be inverted. (Typically, $\frac{dP}{dT}$ is positive because $\Delta e$ and $\Delta v$ tend to have the same sign.)

For the patchy model studied here, the second virial coefficient, $B_2$, is given as

\[
B_2 = -\frac{1}{2V} \int d\mathbf{r} \int d\chi \int d\Omega_1 \int d\Omega_2 (e^{-\beta u(\mathbf{r} | \Omega_1, \Omega_2)} - 1),
\]

(4.7)

where the integrand is the Mayer function, and the integration is performed over particle positions, $\mathbf{r}$, orientations $\Omega$, and dihedral $\chi$. Computing this integral for our patchy model gives

\[
\frac{B_2}{B_{\text{HS}}} = 1 - \sum \left( \lambda_{ij}^3 - 1 \right) \sin^2 \left( \frac{\delta_i}{2} \right) \sin^2 \left( \frac{\delta_j}{2} \right) (e^{\beta \varepsilon_{ij}} - 1)(2\Delta \phi_{ij})
\]

\[
- (\lambda_{IV}^3 - 1)(e^{\beta (\varepsilon_{IV} + \varepsilon_{VI} + \varepsilon_{VI})} - 1) \sin^2 \left( \frac{\delta_{IV,\alpha}}{2} \right) \sin^2 \left( \frac{\delta_{IV,\beta}}{2} \right)(2\Delta \chi)
\]

\[
- (\lambda_{II}^3 - 1)(e^{\beta (\varepsilon_{II} + \varepsilon_{IV} + \varepsilon_{VI})} - 1) \sin^2 \left( \frac{\delta_{II,\alpha}}{2} \right) \sin^2 \left( \frac{\delta_{II,\beta}}{2} \right)(2\Delta \chi)
\]

\[
- (\lambda_{VI}^3 - 1)(e^{\beta (\varepsilon_{IV} + \varepsilon_{VI})} - 1)(2\Delta \chi) \left[ \sin^2 \left( \frac{\delta_{IV,\alpha}}{2} \right) \sin \left( \frac{\delta_{IV,\alpha} + \delta_{IV,\beta}}{2} \right) \sin \left( \frac{\delta_{IV,\alpha} - \delta_{IV,\beta}}{2} \right)
\]

\[
+ \sin \left( \frac{\delta_{VI,\alpha} - \delta_{IV,\alpha}}{2} \right) \sin \left( \frac{\delta_{VI,\alpha} + \delta_{IV,\alpha}}{2} \right) \sin^2 \left( \frac{\delta_{VI,\beta}}{2} \right) \right],
\]

(4.8)
Table 4.2: Patch parameters obtained from all-atom molecular dynamics simulations and slightly modified to adjust the crystal geometry

<table>
<thead>
<tr>
<th>Contact</th>
<th>$\cos \delta_\alpha$</th>
<th>$\cos \delta_\beta$</th>
<th>$\Delta \phi_{\alpha\beta}$ (rad)</th>
<th>$\varepsilon (k_B T)$</th>
<th>$\lambda (\sigma)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1</td>
<td>0.956</td>
<td>0.963</td>
<td>0.07</td>
<td>21.5</td>
<td>1.025</td>
</tr>
<tr>
<td>c2</td>
<td>0.935</td>
<td>0.935</td>
<td>0.21</td>
<td>6.3</td>
<td>1.1</td>
</tr>
<tr>
<td>c3</td>
<td>0.992</td>
<td>0.994</td>
<td>0.14</td>
<td>10.3</td>
<td>1.06</td>
</tr>
<tr>
<td>c4</td>
<td>0.997</td>
<td>0.997</td>
<td>0.06</td>
<td>15.8</td>
<td>1.025</td>
</tr>
<tr>
<td>c5</td>
<td>0.947</td>
<td>0.966</td>
<td>0.13</td>
<td>6.0</td>
<td>1.037</td>
</tr>
<tr>
<td>cI</td>
<td>0.972</td>
<td>0.981</td>
<td>0.21</td>
<td>6.3</td>
<td>1.1</td>
</tr>
<tr>
<td>cII</td>
<td>0.96</td>
<td>0.96</td>
<td>0.3</td>
<td>8.7</td>
<td>1.11</td>
</tr>
<tr>
<td>cIII</td>
<td>0.95</td>
<td>0.95</td>
<td>0.2</td>
<td>10.1</td>
<td>1.1</td>
</tr>
<tr>
<td>cIV</td>
<td>0.96</td>
<td>0.937</td>
<td>0.3</td>
<td>4.3</td>
<td>1.14</td>
</tr>
<tr>
<td>cV</td>
<td>0.95</td>
<td>0.95</td>
<td>0.21</td>
<td>18.5</td>
<td>1.153</td>
</tr>
<tr>
<td>cVI</td>
<td>0.898</td>
<td>0.927</td>
<td>0.3</td>
<td>8.6</td>
<td>1.11</td>
</tr>
<tr>
<td>cVII</td>
<td>0.95</td>
<td>0.95</td>
<td>0.21</td>
<td>18.5</td>
<td>1.53</td>
</tr>
</tbody>
</table>

where the sum is over all patch pairs excluding c_{II}, c_{IV}, and c_{VI}, which are collocated.

The terms after the sum specifically describe the contribution of these patches. Note that $2\Delta \chi = 0.6$ is the width of range of dihedral angles for these three patches, and $\Delta \phi_V = 1.0632$ because there are four valid dihedral angles for cV with overlapping ranges.
Table 4.3: Vectors defining each patch location and dihedral angle, $\phi_{\alpha\beta}$. Note that because of geometric constraints of the DBN structure, contacts cII, cIV, and cVI are collocated.

<table>
<thead>
<tr>
<th>Contact</th>
<th>$\hat{e}_1$</th>
<th>$\hat{e}_2$</th>
<th>$\hat{e}_3$</th>
<th>$\phi_{\alpha\beta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1</td>
<td>0.6614</td>
<td>0.6543</td>
<td>0.3667</td>
<td>-2.2703</td>
</tr>
<tr>
<td></td>
<td>0.6473</td>
<td>-0.2853</td>
<td>0.7068</td>
<td></td>
</tr>
<tr>
<td>c2/cI</td>
<td>-0.9854</td>
<td>0.0471</td>
<td>0.1101</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>0.9854</td>
<td>-0.0471</td>
<td>-0.1101</td>
<td></td>
</tr>
<tr>
<td>c3</td>
<td>-0.0141</td>
<td>-0.9402</td>
<td>0.3403</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>0.0141</td>
<td>0.9402</td>
<td>-0.3403</td>
<td></td>
</tr>
<tr>
<td>c4</td>
<td>0.3231</td>
<td>0.1911</td>
<td>-0.9269</td>
<td>-2.1288</td>
</tr>
<tr>
<td></td>
<td>0.3089</td>
<td>-0.7486</td>
<td>-0.5867</td>
<td></td>
</tr>
<tr>
<td>c5</td>
<td>-0.6474</td>
<td>0.2851</td>
<td>-0.7068</td>
<td>2.2698</td>
</tr>
<tr>
<td></td>
<td>-0.6613</td>
<td>-0.6543</td>
<td>-0.3669</td>
<td></td>
</tr>
<tr>
<td>cII*</td>
<td>-0.9494</td>
<td>0.3101</td>
<td>-0.0499</td>
<td>-2.4946</td>
</tr>
<tr>
<td></td>
<td>0.0011</td>
<td>0.4334</td>
<td>0.9012</td>
<td></td>
</tr>
<tr>
<td>cIII</td>
<td>-0.6455</td>
<td>-0.5123</td>
<td>-0.5665</td>
<td>-1.80</td>
</tr>
<tr>
<td></td>
<td>-0.7102</td>
<td>-0.5524</td>
<td>0.4365</td>
<td></td>
</tr>
<tr>
<td>cIV*</td>
<td>-0.9494</td>
<td>0.3101</td>
<td>-0.0499</td>
<td>-2.4946</td>
</tr>
<tr>
<td></td>
<td>0.0011</td>
<td>0.4334</td>
<td>0.9012</td>
<td></td>
</tr>
<tr>
<td>cV/VII</td>
<td>0.3162</td>
<td>-0.8026</td>
<td>-0.5058</td>
<td>-2.1529</td>
</tr>
<tr>
<td></td>
<td>0.4927</td>
<td>0.8531</td>
<td>0.1716</td>
<td></td>
</tr>
<tr>
<td>cVI*</td>
<td>-0.9494</td>
<td>0.3101</td>
<td>-0.0499</td>
<td>-2.4946</td>
</tr>
<tr>
<td></td>
<td>0.0011</td>
<td>0.4334</td>
<td>0.9012</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.4: Geometry of DBI and DBN crystals for the patchy model. Lengths are in units of $\sigma$. The rotation of each particle is calculated from Euler angles reported as $(\alpha, \beta, \gamma)$ with the $R_z(\alpha)R_x(\beta)R_z(\gamma)$ convention, where $R_z$ denotes a counterclockwise rotation through the z-axis and $R_x$ a counterclockwise rotation through the x-axis.

<table>
<thead>
<tr>
<th></th>
<th>2 asymmetric units</th>
<th>Unit cell: $(\sqrt{2}, 1, 1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBI</td>
<td>Particle Position</td>
<td>Particle rotation</td>
</tr>
<tr>
<td></td>
<td>0, 0</td>
<td>$-1.927, 1.349, -1.474$</td>
</tr>
<tr>
<td></td>
<td>$\sqrt{2}/2, 0.5, 0.5$</td>
<td>$1.927, 1.793, 1.668$</td>
</tr>
<tr>
<td>DBN</td>
<td>8 asymmetric units</td>
<td>Unit cell: $(1.02, 3.149, 3.586)$</td>
</tr>
<tr>
<td></td>
<td>Particle position</td>
<td>Particle Rotation</td>
</tr>
<tr>
<td></td>
<td>0.184, 0.470, 0.385</td>
<td>1.427, 1.681, 1.371</td>
</tr>
<tr>
<td></td>
<td>0.183, 0.073, 3.061</td>
<td>1.096, 0.186, 5.375</td>
</tr>
<tr>
<td></td>
<td>0.761, 1.024, 3.238</td>
<td>1.685, 1.547, 4.421</td>
</tr>
<tr>
<td></td>
<td>0.647, 1.438, 0.631</td>
<td>2.043, 2.916, 2.222</td>
</tr>
<tr>
<td></td>
<td>0.954, 2.030, 1.478</td>
<td>4.869, 1.420, 4.519</td>
</tr>
<tr>
<td></td>
<td>0.032, 1.704, 2.446</td>
<td>4.876, 3.029, 1.923</td>
</tr>
<tr>
<td></td>
<td>0.479, 2.623, 2.213</td>
<td>4.548, 1.581, 1.316</td>
</tr>
<tr>
<td></td>
<td>0.537, -0.203, 1.167</td>
<td>3.468, 0.065, 6.046</td>
</tr>
</tbody>
</table>
Testing the Hydrophobicity Scenario as a Microscopic Basis for Inverted Solubility in \( \gamma \)D-Crystallin

Reprinted (adapted) with permission from


Copyright (2019) American Chemical Society.

5.1 Introduction

Proteins can self-organize into a rich variety of superstructures [120], such as crystals [64], virus capsids [78], disease-forming aggregates [161], and biomaterials [173]. A key challenge is understanding how microscopic features of solvated proteins can give rise to such complex structures, and eventually to design systems that reliably assemble as such [145, 70, 87, 173, 158, 102]. In this context, coarse-grained models
are especially valuable, because they help both pinpoint and abstract the microscopic features that can reproduce the experimentally observed behavior. (Because simulating protein self-assembly typically requires hundreds to thousands of protein copies, which are themselves comprised of thousands of atoms, such models are also a computational necessity [6, 183, 153].) For example, even relatively simple models of short-ranged [178, 152], anisotropic pair interactions largely recapitulate the phase behavior of globular proteins [113, 27, 120]. Understanding the assembly of some systems, however, requires coarse-grained models with a richer set of features, such as shape anisotropy for viral capsid and amyloid fiber-forming proteins [78, 120]. Capturing certain aspects of protein crystallization, which is key to protein structure determination by diffraction methods [123], can also require enriched patchy particle models [64].

Proteins that exhibit atypical solution behaviors provide essential tests of our understanding of the physico-chemical processes that underlie their assembly. One such phenomenon is the decrease of protein solubility with increasing temperature, i.e., inverted crystal solubility, which is observed in a few proteins, such as some single mutants of γD-crystallin [141, 138], and the wild type carbomonoxy-hemoglobin C [184]. (The temperature invariant solubility of apoferritin is a limit case [143].) Thermodynamically, inverted solubility suggests that as temperature increases, the Gibbs free energy of crystallization decreases, and hence that the crystal becomes increasingly more stable than the fluid. The phenomenon is often attributed to a large and positive entropy gain upon crystallization. Crystal formation is then possible even if the enthalpy of crystallization is non-negative [184, 165, 194]. Because the solute contribution to the change in entropy is typically negative, the solvent contribution is traditionally considered to be the key microscopic determinant of the phenomenon [194, 165, 184].

The association of inverted solubility in proteins with the hydrophobic effect
also comes from our understanding of the aqueous solvation of hydrocarbons, which presents an analogous anomaly [125]. The minimal Muller-Lee-Graziano (MLG) model for this effect considers water as being in one of four states: disordered shell (ds), ordered shell (os), disordered bulk (db), and ordered bulk (ob) [108, 128]. It was used by Shiryayev et al. to estimate the phase diagram of model globular proteins with isotropic interactions assumed to be driven exclusively by hydrophobic interactions [165]. Although the resulting phase behavior does present an inverted solubility regime, it is unclear whether this feature would persist for more realistic protein models, with a complex surface mosaic of hydrophilic and hydrophobic interactions. In other words, while the hydrophobic scenario for solubility inversion in proteins is thermodynamically sound, microscopic evidence for it remains limited. The generality of the underlying physical arguments is also seemingly incompatible with the relatively rare occurrence of inverted solubility in experiments.

Here, we examine this microscopic scenario in the context of a double mutant (R36S+P23T) of the human γD-crystallin, which forms two competing crystals: a normal solubility structure (DBN, PDB ID: 6ETA) and an inverted solubility structure (DBI, PDB ID: 6ETC) [89, 98]. The solubility inversion is here most likely associated with the mutation in the 23rd residue because the single mutants P23T, P23S, and P23V also exhibit inverted solubility [121], albeit without generating diffraction quality crystals. In earlier work, we have parameterized a patchy model for the R36S+P23T mutant and obtained a solubility inversion regime by completely deactivating the patch containing the 23rd residue at low temperatures [98]. Interestingly, the DBI crystal does not present any obvious structural feature that could explain this effect, other than the formation of a hydrogen bond through the 23rd residue. Here, we critically evaluate three different temperature-dependent interaction potentials: the generic model we previously considered, and two that explicitly model the hydrophobic scenario. We use these models to test the hydrophobic sce-
nario as well as the robustness of the inverted solubility regime with respect to model parameters. We thus attempt to elucidate why inverted solubility is not more commonly observed. We further explore the relationship between the liquid-liquid critical point and the solubility curve, which has been experimentally studied for some of these systems [121]. The plan for the rest of this paper is as follows. We first survey γD-crystallin crystal structures available in literature to determine whether an increase in surface hydrophobicity can be discerned upon introducing the solubility inverting mutations (Sec. 5.2). We then introduce a patchy protein model for these proteins (Sec. 5.3.1) along with the different temperature-dependent patch models (Sec. 5.3.2), and the methods used to determine solubility lines (Sec. 5.3.3). Sections 5.4.1 and 5.4.2 provide a detailed analysis of these patchy models, and we conclude with proposals for further discerning experiments in Sec. 5.5.

5.2 Experimental Context for Hydrophobicity

As a first consideration of the reasonableness of the hydrophobicity scenario, we evaluate the surface hydrophobicity of various human γD-crystallin crystal structures. Were the P23T mutation to consistently increase surface hydrophobicity, one would infer the existence of strong evidence for the decrease in protein solubility upon mutagenesis to be driven by the hydrophobic effect. By studying the relative binding propensity of two dyes known to bind hydrophobic surfaces, Pande et al. indeed have previously inferred that P23T, P23S, and P23V mutants of human γD-crystallin do present a higher surface hydrophobicity than the wild type (WT) protein. [138] Here, in order to test the robustness and microscopic validity of this interpretation, we consider different scales that quantify hydrophobicity at the amino acid level. More specifically, we compute an average hydrophobicity of solvent-exposed residues [163] weighted by their solvent accessible surface area (SASA) [62], for five different scales: the grand average of hydropathy (GRAVY) [107], as well as the scales of Wimley
and White (ww) [199], Hessa et al. (hh) [82], Moon and Fleming (mf) [126], and Zhao and London (also known as transmembrane tendency, tt) [203]. Each of these scales assigns a hydrophobicity index to each residue type; all but hh and mf assign positive values to hydrophobic residues.

We compute hydrophobicity for three sets, $S$, of amino acids: (i) the entire protein surface, (ii) the surface of its N-terminus, i.e., the first 82 residues (including the solubility inverting 23rd residue), and (iii) the surface residues in the DBI contact that includes the 23rd residue (Patch 4 as per Sec. 5.3) [98]. The hydrophobicity, $H_\zeta$, for a given scale $\zeta$ is then obtained as

$$H_\zeta = \frac{\sum_{i \in S} f_\zeta(i) A(i)}{\sum_{i \in S} A(i)}, \quad (5.1)$$

where $f_\zeta(i)$ is the hydrophobicity index for residue $i$, and $A(i)$ is its SASA. We specifically consider: WT (PDB ID: 1HK0 [19]), the P23T single mutant (PDB ID: 4JGF [90]), the R36S single mutant (PDB ID: 2G98 [101]), the R58H single mutant (PDB ID: 1H4A [19]), DBI (PDB ID: 6ETC [98]), and DBN (PDB ID: 6ETA [98]). Of these, only WT, R36S, and R58H do not have a mutation at the 23rd residue. Note that missing residues are completed using Modeller [159] within Chimera [144], and all crystal water molecules are removed prior to running this analysis. In order to estimate the error on these measured hydrophobicities, 100 configurations per crystal structure are created by perturbing particle coordinates by a random number selected from a Gaussian distribution with a standard deviation corresponding to the coordinate error specified in the PDB file. Two assumptions are made in estimating these error bars. First, the coordinate error reported in the PDB entry is assumed to be distributed uniformly and isotropically across all protein atoms. In reality, certain domains or residues in proteins are more mobile and thus harder to resolve by X-ray
diffraction than others, but finer, residue-level information is not available. This assumption thus overestimates the error in more localized parts of the protein and underestimates the error in more mobile parts. Second, the refined structures do not precisely capture the actual protein structure, as suggested by \( R_{\text{free}} \) values ranging from 0.174 to as high as 0.284, hence possibly creating artificial hydrophobicity differences between different mutants, or, conversely, underestimating them.

The resulting hydrophobicity estimates are shown in Fig. 5.1. All measurements suggest that the entire protein, the N-terminus, and Patch 4 are overall hydrophilic, which is consistent with the fact that the protein is soluble in water. A more careful comparison is thus needed in order to determine whether certain substructures are more hydrophobic than others. We first compare the DBN and DBI structures, which are obtained from the same double mutant, R36S+P23T, and which are structurally very similar [98]. As expected, nearly all measurements for DBN and DBI overlap within their 95% confidence intervals. The only exceptions are the hydrophobicities of Patch 4 measured by the GRAVY and \( mf \) scales. The latter likely results from \( mf \) uniquely classifying prolines as hydrophobic. This discrepancy could then amplify the minute difference in surface exposure of Patch 4 prolines between DBI and DBN. A similar argument could be made about GRAVY, as tyrosine is considered to be the most hydrophilic residue on the GRAVY scale.

Overall, the N-terminus is the most hydrophobic region in nearly all scales and for all structures. However, other observations are not consistent across scales. In particular, a number of nonmonotonicities can be observed. For instance, Patch 4 is more hydrophobic in DBI than in WT for the GRAVY, \( hh \), and \( mf \) scales, but the \( ww \) and \( tt \) scales present no discernible difference. Similarly, Patch 4 is more hydrophobic in R36S than in WT for the \( mf \) scale, but the reverse is true for \( hh \). These discrepancies reflect the different ordering of residues on different scales. For instance, GRAVY, which is calculated from experimental measurements of transfer
free energies from water to water vapor, tends to assign aromatic side chains lower hydrophobicities than the other four scales, which instead consider the tendency of residues to transfer from bulk water to a lipid bilayer, a measurement prone to higher experimental uncertainty [199].

Interestingly, the N-terminus of the P23T mutant is the least hydrophobic structure for the GRAVY and \( \text{mf} \) scales. This trend, however, disappears when only Patch 4 residues are considered. Patch 4, which controls solubility inversion, is actually less hydrophobic than the overall N-terminus or the entire protein, except on the \( \text{mf} \) scale. Only for this last scale is Patch 4 clearly more hydrophobic. A similar inconsistency is observed for Patch 4 of DBI, which is more hydrophobic than the other proteins for GRAVY and \( \text{mf} \), but for these two scales P23T and DBN are not discernibly more hydrophobic than the structures without the mutation in the 23rd residue.

In summary, in none of the hydrophobicity scales do the structures with the (solubility-inverting) P23T mutation have a statistically and consistently higher hydrophobicity than those without. P23T mutations even result in lower hydrophobicity estimates on some scales. While these results are subject to errors from the crystal structure accuracy, as well as the imperfections of the hydrophobicity scales themselves, a microscopic change to the protein surface that could putatively underlie the inversion of solubility nonetheless remains elusive. Because hydrophobicity scales are but an indirect measure of protein-water interactions (and thus of protein-protein interactions), however, other, more detailed approaches could be more revealing.

5.3 Theory and Computational Details

Because a clear enhancement of hydrophobicity cannot be detected directly in mutants with inverted solubility, we next consider the thermodynamics of patchy models that incorporate various temperature-dependent patch energies. A schematic model
Figure 5.1: Hydrophobicity estimates for different crystal structures of single and double mutants of human $\gamma$D-crystallin. Proteins to the left of the black vertical line exhibit normal solubility, and those to the right exhibit inverted solubility. Error bars denote 95% confidence intervals. Lines connecting the points are but a guide to the eye. Note the flipped scales for the hh and mf scales, in which lower values denote higher hydrophobicity, by contrast to the other scales. Structures with the P23T mutation do not systematically present a higher hydrophobicity, which is inconsistent with the hydrophobicity scenario.

of the double mutant of human $\gamma$D-crystallin was previously studied in Ref. [98], and is here first modified to consider the hydrophobic scenario and then perturbed to evaluate the robustness of its inverted solubility regime.

5.3.1 Patchy Model

The schematic model consists of hard particles with attractive patches

$$u(r_{ij}, \Omega_i, \Omega_j) = u_{HS}(r_{ij}) + \sum_{a,b} u_{ab}(r_{ij}, \Omega_i, \Omega_j),$$

(5.2)

where $r_{ij}$ is the distance between particles $i$ and $j$, $\Omega$ denotes the particle orientation, and $u_{HS}(r_{ij})$ is the hard sphere potential for particles of diameter $\sigma$. The sum runs
over all patch pairs, with $n$ the total number of patches. The second contribution, $u_{ab}$, is further broken down into radial and orientational parts

$$u_{ab} = v_{ab}(r_{ij})f_{ab}(\Omega_i, \Omega_j). \quad (5.3)$$

The radial part, $v_{ab}$, is a square-well interaction

$$v_{ab}(r_{ij}) = \begin{cases} -\varepsilon_{ab}(T), & \sigma < r_{ij} < \lambda_a + \lambda_b \\ 0, & \text{otherwise} \end{cases}, \quad (5.4)$$

with interaction ranges $\lambda_a$ and $\lambda_b$ of patches $a$ and $b$, respectively, and with either constant or temperature-dependent patch energy $-\varepsilon_{ab}(T)$. The orientational part

$$f_{ab} = \begin{cases} 1, & \theta_{a,ij} \leq \delta_a \text{ and } \theta_{b,ij} \leq \delta_b \\ 0, & \text{otherwise} \end{cases} \times \begin{cases} 1, & \psi_{ij} \in [\varphi_{ab} - \Delta \varphi_{ab}, \varphi_{ab} + \Delta \varphi_{ab}] \\ 0, & \text{otherwise} \end{cases} \quad (5.5)$$

contains two contributions. The first ensures that the relative particle orientation enables them to interact with $\delta_a$ and $\delta_b$ the angular width for patches $a$ and $b$, respectively (Fig. 5.2a). The second limits the range $\varphi_{ab} \pm \Delta \varphi_{ab}$ of dihedral angles $\psi_{ij}$ allowed for each pair (Fig. 5.2b), with $\theta_{a,ij}$ the angle between the vector defining the location of patch $a$ and the vector that connects the centers of particles $i$ and $j$, and $\theta_{b,ij}$ similarly for patch $b$.

This model is parameterized such that each patch corresponds to a crystal contact in either the DBI or DBN crystal structure. This choice assumes that these surface patches are most chemically relevant for crystal formation, which is reasonable for such a small protein and is consistent with earlier studies of protein crystallization [65, 71]. We then obtain five patches for DBI – labeled with Arabic numerals – and five patches for DBN – labeled with Roman numerals. Because Patch 4 of DBI contains the 23rd residue, which is associated with the inverted solubility regime,
Figure 5.2: For two patches to interact, the relative particle orientation should satisfy the following. (a) The angle between the vector joining particles $i$ and $j$, $\mathbf{r}_{ij}$, and the patch vectors $\hat{\mathbf{e}}_\alpha$ and $\hat{\mathbf{e}}_\beta$ should be less than $\delta_\alpha$ and $\delta_\beta$, respectively. (b) The dihedral angle between two particles, which is defined as the angle between two planes defined by the vectors $(\mathbf{z}_i, \mathbf{r}_{ij})$ and $(\mathbf{z}_j, -\mathbf{r}_{ij})$, should be within the range $\varphi_{ab} \pm \Delta \varphi_{ab}$. The reference vector $\mathbf{z}$ is chosen such that its orientation relative to the patches is identical for all particles.

This patch is taken to be temperature dependent (see Sec. 5.3.2); other patches are assigned a constant energy. Patch energies and interaction ranges were previously extracted from all-atom molecular dynamics simulations [21], using umbrella sampling [95]. The resulting patchy particle model is sketched in Fig. 5.3 and the geometry details are given in the Supporting Information (Sec. S2). Note that the resulting effective single-component system model coarse-grains the role of solvent and ions in the crystallization cocktail. In what follows, unless otherwise specified, energies are reported in units of $k_B T_{\text{ref}}$, where $T_{\text{ref}} = 277\text{K}$ is the temperature at which DBN was crystallized experimentally, and distances are reported in units of the particle diameter $\sigma$, which here is taken to be 2.54 nm.

It is important to highlight that this protocol presents a number of limitations, including inaccuracies of the protein force field [197] and of the water model [8], as
5.3.2 Inverted Solubility Models

In order to represent the microscopic origin of the inverted solubility, we consider three models for the temperature-dependence of the Patch 4 interaction energy: the MLG model, the Wentzel-Gunton model, and the temperature-(de)activated patchy model. Note that the parameters explicitly defined in these models are discussed in this section, while the free parameters are left for Sec. 5.4.1.

**MLG model**– In this model, each of four water states is assigned a (relative) degeneracy, $q$, and an energy $E$. Degeneracies are ordered $q_{ds} > q_{db} > q_{ob} > q_{os}$. The last inequality follows from the hydration shell allowing the formation of only hydrogen bonds between water molecules and not to the hydrophobic solute. The higher degeneracy of the disordered shell compared to the disordered bulk follows
from the additional orientational constraints in the former compared to the latter. Because only relative information about the degeneracies is needed, the estimates proposed in Ref. [167] here suffice: $q_{ob} = 1.5$, $q_{db} = 30$, $q_{os} = 1$, and $q_{ds} = 48$. Meanwhile, the energies are ordered $E_{ds} > E_{db} > E_{ob} > E_{os}$. The ordered shell is expected to have a lower energy than the ordered bulk state, because hydrogen bonds that form via tangentially oriented water molecules tend to be stronger than radially oriented ones; the disordered shell is expected to have a higher energy than the disordered bulk because replacing the solute with water molecules slightly increases the number of hydrogen bonds. Because energy values used by Ref. [125] are on an arbitrary scale, which is incompatible with the specific energy scale of our patchy model, we here use the values reported by Silverstein et al. for the Mercedes-Benz model of water [166, 167]. Posing that the energy of the ordered bulk is about one hydrogen bond, $E_{ob} = -5.82 \ k_B T_{ref}$ [56], the other three states have: $E_{db} = -1.69 \ k_B T_{ref}$, $E_{os} = -5.90 \ k_B T_{ref}$, and $E_{ds} = -0.56 \ k_B T_{ref}$.

The energy and entropy per water molecule in the shell and are then given as [165]

$$E_s = \frac{E_{os} + E_{ds} e^{-\beta(E_{ds} - E_{os})}}{1 + e^{-\beta(E_{ds} - E_{os})}}$$  \hspace{1cm} (5.6)$$

$$E_b = \frac{E_{ob} + E_{db} e^{-\beta(E_{db} - E_{ob})}}{1 + e^{-\beta(E_{db} - E_{ob})}}$$  \hspace{1cm} (5.7)$$

and

$$s_s / k_B = \log \left( \frac{q_{os} + q_{ds} e^{-\beta(E_{ds} - E_{os})}}{1 + e^{-\beta(E_{ds} - E_{os})}} \right)$$  \hspace{1cm} (5.8)$$

$$s_b / k_B = \log \left( \frac{q_{ob} + q_{db} e^{-\beta(E_{db} - E_{ob})}}{1 + e^{-\beta(E_{db} - E_{ob})}} \right)$$  \hspace{1cm} (5.9)$$

The change in energy and entropy upon moving one water molecule from the bulk to the solvation shell of the protein are then simply $\varepsilon_w = E_s - E_b$ and $\Delta s_w = s_s - s_b$, respectively.
With this formulation the energy of Patch 4 is given by

$$\varepsilon'_4 = \varepsilon_4 + n_w \Delta \varepsilon(\beta) \tag{5.10}$$

where we have defined \(\Delta \varepsilon(\beta) = 2(\varepsilon_w - \Delta s_w / \beta)\), and \(n_w\) is the number of water molecules in the solvation shell around contact \(i\). Note that because patch parameters are measured at \(\beta_{ref} = 1\), parameters need to be tuned such that \(\varepsilon'_4(\beta = 1) = \varepsilon_4\), and hence \(\varepsilon'_4 = (\varepsilon_4 - \Delta \varepsilon(1)n_w) + \Delta \varepsilon(\beta)n_w\). Note also that the temperature scale for the MLG model cannot be changed arbitrarily by changing \(\beta_{ref}\), because its parameters already set the range of temperatures within which the hydrophobic effect changes the free energy of crystallization.

**Wentzel-Gunton Model**— Wentzel and Gunton proposed a simplified version of the MLG model in order to consider the phase behavior of particles with anisotropic interactions using Wertheim’s theory [195, 196, 164, 194]. This simple model assigns a linear temperature dependence for the patch energies

$$-\varepsilon'_4 = -\varepsilon_4 - 2\varepsilon_w + \frac{2}{\beta} \Delta s_w, \tag{5.11}$$

where \(-\varepsilon_w\) and \(-\Delta s_w\) are free parameters that account for the change in energy and in entropy, respectively, due to the displacement of water upon contact association. Patch energies should equal those of the original model at \(\beta_{ref}\), where the model was parameterized. This choice here suffices to set the overall temperature scale, because \(\varepsilon_w\) and \(\Delta s_w\) are arbitrary. Fixing \(\varepsilon_w\), such that \(\varepsilon'_4(\beta = \beta_{ref}) = \varepsilon_4\), thus results in \(\varepsilon'_4 = \varepsilon_4 + 2\Delta s_w(\frac{1}{\beta_{ref}} - \frac{1}{\beta})\).

**Temperature-(de)activated Patchy Model**— de Las Heras and de Gama [48] proposed a model for patch (de)activation with temperature inspired by DNA-grafted colloids, which lose their attractive patches above the DNA melting temperature [68],

94
Although this model does not correspond to a specific microscopic scenario in proteins, it can nevertheless be construed as a simple and elegant way to (de)activate a patch. The temperature dependence of the interaction is then

$$\varepsilon_4'(T) = \frac{\varepsilon_4}{2} \left[ 1 + \tanh \left( \frac{T - T_a}{\tau} \right) \right],$$

(5.12)

where $T_a$ is the deactivation temperature, $\tau$ controls the sharpness of that deactivation. For this model, Patch 4 is deactivated below $T_a$. 

5.3.3 Crystal Solubility Determination

Solubility lines are determined by first calculating the fluid and crystal chemical potentials, and then identifying the coexistence points at the intersection of these curves at fixed temperature and pressure. For both DBI and DBN, experimental solubilities correspond to protein volume fractions of $\phi = 10^{-3}$ or lower [89].

At such low densities simple local Monte Carlo (or molecular dynamics) sampling of the fluid phase is computationally inefficient, because transport is relatively slow. While this problem can be alleviated with advanced sampling methods such as aggregation volume bias Monte Carlo [42] and event chain Monte Carlo [25], we here instead estimate the fluid properties from the second virial coefficient, $B_2$, which is calculated as in Ref. [96] (see SI Sec. S3 [63]). Because the patch energies are high, $B_2$ can become very large and negative at low temperatures, but the protein density remains sufficiently low for $|B_2\rho| \ll 1$ in the regime of interest. In order to confirm that the third virial coefficient, $B_3$, can safely be neglected, we bound its value by noting that triply-bonded triplets of particles cannot form. The dominant contribution to $B_3$ thus comes from doubly-bonded triplets and scales as $B_2^2$, hence $|B_3\rho| \lesssim |B_2|^2\rho^2 \ll 1$ in the regime of interest. Its contribution to the fluid chemical potential, $\mu_f$, is therefore negligible, and so are higher-order corrections, thus justi-
fying this theoretical expediency. The fluid equation of state and chemical potential can then be written as

\[ \frac{\beta p}{\rho} = 1 + B_2 \rho, \]  

(5.13)

\[ \beta \mu_f = \beta \mu^{\text{id}} + 2B_2 \rho = \log \Lambda^3 \rho + 2B_2 \rho, \]  

(5.14)

where \( \beta \mu^{\text{id}} = \log \Lambda^3 \rho \) is the chemical potential of the ideal gas, and the thermal de Broglie wavelength \( \Lambda \) is set to unity, without loss of generality. With this formulation, we have

\[ \rho = \frac{-1 + \sqrt{1 + 4B_2 \beta p}}{2B_2}. \]  

(5.15)

Note that by thermodynamical stability, \( \beta \mu_f \) must decrease with decreasing pressure, and our estimate is consistent in this respect. If \( B_2 \) is positive, \( \partial \beta \mu_f / \partial \rho \) is also positive; if \( B_2 \) is negative, \( \partial \beta \mu_f / \partial \rho > 0 \) for \( \rho < -B_2/2 \), which is always true.

The crystal free energy at a given pressure and temperature is calculated using numerical simulations (see SI Sec. S3) with the Frenkel-Ladd method [60], which involves thermodynamically integrating from an Einstein crystal. From this reference free energy, thermodynamic integration along an isobar provides the crystal chemical potential, \( \mu_x \), at different temperatures,

\[ \beta \mu_x(\beta, p) = \beta_0 \mu_x(\beta_0, p) + \int_{\beta_0}^{\beta} \frac{\langle H(\beta') \rangle}{N} d\beta' + \int_{\beta_0}^{\beta} \beta' \frac{\langle dU/d\beta' \rangle}{N} d\beta', \]  

(5.16)

where \( \langle H \rangle = p\langle V \rangle + \langle U \rangle \) is the enthalpy and \( \langle \cdot \rangle \) denotes thermal averaging. Because of the highly constrained geometry of the patchy models, both crystals are almost
incompressible. As a result, $\langle V \rangle$ is essentially independent of temperature. To high accuracy, we can thus write

$$ \frac{1}{N} \int_{\beta_0}^{\beta} \frac{\langle H(\beta') \rangle}{N} d\beta' \approx \frac{1}{N} \int_{\beta_0}^{\beta} \frac{\langle U(\beta') \rangle}{N} d\beta' + \frac{p}{\rho} (\beta - \beta_0), \quad (5.17) $$

and at sufficiently low pressures the second term on the right hand side is also negligible.

We further approximate that all the crystal bonds are active, and hence $\langle U(\beta) \rangle \approx U_0(\beta)$, where $U_0(\beta)$ is the ground state energy, and $\langle dU/d\beta \rangle \approx dU_0/d\beta$. While this last approximation is generally quite good, it is overly crude in the patch deactivation regime, where the patch energy decreases rapidly around $\beta_\alpha$, and vanishes when temperature is reduced further. As a result, $\langle dU/d\beta \rangle \ll dU_0/d\beta$, which can result in a significant correction to $\beta \mu_x$ (see Fig. 5.4a). In the Wentzel-Gunton model, the patch similarly becomes non-attractive for $\beta > \beta_{\text{ref}}$, and upon further lowering the temperature, it eventually becomes repulsive. The topology of the DBI crystal then changes and the energy of the crystal once again becomes temperature-independent, which leads to a bending of the evolution of the chemical potential with temperature (Fig. 5.4b). In both cases, however, the DBI solubility curve is unaffected, because these changes occur in a region where DBI is metastable with respect to DBN. Taking $\langle dU/d\beta \rangle \approx dU_0/d\beta$ is thus reasonable for our purposes.

Under these two approximations, the crystal chemical potential for the MLG model can be written as

$$ \beta \mu_x(\beta, p) \approx \beta_0 \mu_x(\beta_0, p) + \frac{p}{\rho} (\beta - \beta_0) + \xi(\beta), \quad (5.18) $$

where
The simulation data (black) fully matches the individual Einstein crystal simulations (red data points). Estimates of $\beta \mu_x$ (blue) become significantly flawed at low temperatures, but because this regime is beyond the triple point, $\beta_{tp}$ (dashed line), the DBI solubility line is unaffected. (a) Patch 4 is deactivated below $T_a$, with $\tau = 0.05$, and (b) Patch 4 energy follows the Wentzel-Gunton model with $\Delta s_w = -50$.

\[
\xi(\beta) = \int_{\beta_0}^{\beta} \left[ -\varepsilon_{\text{tot}} + \Delta \varepsilon(1) n_w - 2 n_w \varepsilon_w + 2 n_w \left( \frac{d\Delta s_w}{d\beta'} - \beta' \frac{d\varepsilon_w}{d\beta'} \right) \right] d\beta' 
\]

\[
= (\Delta \varepsilon(1) n_w - \varepsilon_{\text{tot}})(\beta - \beta_0) - 2 n_w \left( \varepsilon_w \beta - \varepsilon_w \beta_0 \right) + 2 n_w \left( \beta \varepsilon_w \beta_0 \right) + 2 n_w \left( \Delta s_w(\beta) - \Delta s_w(\beta_0) \right)
\]

\[
= (\Delta \varepsilon(1) n_w - \varepsilon_{\text{tot}})(\beta - \beta_0) + 2 n_w \left( \Delta s_w(\beta) - \Delta s_w(\beta_0) \right)
\]

and where $-\varepsilon_{\text{tot}} = -\sum_{i=1}^{5} \varepsilon_i$ is the temperature independent portion of the crystal ground state energy per particle, i.e., $U_0/N = -\varepsilon_{\text{tot}} - n_w \Delta \varepsilon(\beta)$, for the MLG model. We thus have

\[
\frac{\beta \mu_x}{d\beta} \approx (\Delta \varepsilon(1) n_w - \varepsilon_{\text{tot}}) + 2 n_w \left( \frac{d\Delta s_w}{d\beta} - \varepsilon_w - \beta \frac{d\varepsilon_w}{d\beta} \right)
\]
which has a minimum when

\[ \Gamma(\beta) \equiv \frac{d\Delta s_w}{d\beta} - \varepsilon_w - \beta \frac{d\varepsilon_w}{d\beta} = \frac{\varepsilon_{\text{tot}}}{2n_w} - \frac{\Delta \varepsilon(1)}{2}. \]  

(5.23)

As noted above, \( \beta\mu_f \) decreases with decreasing pressure, and because by thermodynamic stability so does \( \rho \), an inverted solubility regime is only obtained when the slope of \( \beta\mu_x \) with respect to \( \beta \) is positive. For \( \Gamma(\beta) > \varepsilon_{\text{tot}}/(2n_w) - \Delta \varepsilon(1)/2 \), the slope of \( \beta\mu_x \) is positive, hence inverted solubility is observed.

For the Wentzel-Gunton model, the change in \( \beta\mu_x \) with temperature can be similarly estimated. We can write the energy per particle in the crystal as

\[ U_0(\beta)/N = -\varepsilon_{\text{tot}} - 2\Delta s_w \left( \frac{1}{\beta_{\text{ref}}} - \frac{1}{\beta} \right). \]  

(5.24)

and hence, following Eq. (5.16),

\[ \beta\mu_x(\beta, p) = \beta\mu_x(\beta_0, p) + \left( -\varepsilon_{\text{tot}} - 2\Delta s_w \frac{2}{\beta_{\text{ref}}} - \frac{p}{\rho} \right) (\beta - \beta_0). \]  

(5.25)

The slope of \( \beta\mu_x \) with respect to \( \beta \) is positive when \( -\varepsilon_{\text{tot}} - p/\rho > 2\Delta s_w/\beta_{\text{ref}} \), thus resulting in an inverted solubility regime.

Writing \( \beta\mu_x \) in a compact form for the temperature-(de)activated patchy model is not possible –the associated integrals need to be evaluated numerically, but the phenomenology is similar. Solubility is inverted in the region where \( \beta\mu_x \) has a positive slope, i.e., around \( T_a \), as can be seen in Fig. [5.4a].

If patch energies are modified by either randomly perturbing them or by scaling them by a constant factor, the free energy of the altered model can be estimated from the original model, assuming that the crystal free energy can be expressed as

\[ \beta A' = \beta A - \beta U_0/N + \beta U'_0/N, \]  

(5.26)
where $A'$ is the Helmholtz free energy and $U'_0$ is the ground state crystal energy for the altered model. This treatment amounts to neglecting the change in crystal entropy upon weakening or strengthening the patches, which is but a small contribution in this temperature regime. We separately verify that the crystal remains stable at the temperatures of interest.

Put together, various approximations described above allow for the expedited consideration of coexistence points that constitute the solubility curves by generating $\beta \mu_f$ and $\beta \mu_x$ curves as functions of temperature and pressure.

5.4 Results and Discussion

5.4.1 Inverted Solubility from Hydrophobicity Models

In order for the microscopic hydrophobicity models described in Sec. 5.3.2 to give rise to solubility inversion, a sufficiently large number of water molecules must be involved. In this section we first consider physical bounds on that number, and then consider how the corresponding crystal solubility lines are affected.

Effect of Parameters on Solubility Lines

The key free parameter in hydrophobicity models is the number of water molecules, $n_w$, solvating the hydrophobic patch. We first estimate the number of water molecules potentially available around Patch 4 by calculating the SASA for the participating residues \[98\] and then computing

$$n_w = A_4 \rho_w \int_{3\AA}^{4.5\AA} g_C(r) dr,$$

where $A_4$ is the solvent accessible surface area of Patch 4, $\rho_w = 3.3 \times 10^{-2} \text{Å}^{-3}$ is the number density of bulk water at room temperature, and $g_C(r)$ is the radial
distribution function of water around carbon atoms determined in Ref. Altan:2018. This estimate thus assumes that (i) the solvent has a radius of 1.4Å (the SASA definition), (ii) the average van der Waals radii of protein heavy atoms is \( \sim 1.6\text{Å} \), and (iii) the first solvation shell ends with the first peak of \( g(r) \) at 4.5Å. We also assume that the measured surface is flat, which is here but a small correction. If we further assume that all residues contributing to Patch 4 are hydrophobic, then \( n_w = 133 - 140 \) for all six protein structures. However, because Patch 4 contains only a handful of hydrophobic residues a more realistic estimate should use a smaller \( A_4 \).

Taking a residue as hydrophobic if it is labeled as such in any of the hydrophobicity scales considered in Sec. 5.2 gives instead \( n_w = 43 - 48 \). Because the hydrophobic residues within Patch 4 are not contiguous, the solvating water molecules are affected by the presence of hydrophilic surface residues nearby. This estimate for \( n_w \) should thus be treated as an upper bound.

We also consider the number of water molecules needed for Patch 4 to have its measured bonding strength. In particular, if we attribute the entire Patch 4 energy to the change in free energy upon moving solvating water molecules to the bulk, then the MLG model gives \( \varepsilon_4 = n_w \Delta \varepsilon(1) \), and thus \( n_w \approx 23 \). Because multiple hydrogen bonds also contribute to the patch energy, however, this number should also be treated as an upper bound.

We can now contrast these bounds with the minimum number of water molecules, \( n_w^* \), that need to be displaced to invert solubility. For the MLG model, we use Eq. (5.23) and the sum of DBI patch energies, \( \varepsilon_\text{tot} = 60 \), to estimate \( n_w^* \): it must be such that \( \xi(\beta_{\text{min}}) \) is a minimum, i.e., \( \Gamma(\beta) > \varepsilon_\text{tot} / (2n_w) - \Delta \varepsilon(1) / n_w \). In other words, solubility is inverted if \( \beta > \beta_{\text{min}} \). The numerical solution in Fig. 5.5 shows that \( n_w^* \gtrsim 71 \), and the corresponding change in \( \xi(\beta) \) is given in the inset. It should be noted, however, that \( n_w^* \) depends strongly on the MLG model parameters. For the multiplicities proposed by Shiryayev et al., for instance, inverted solubility is
Figure 5.5: The minimum of $\xi(\beta)$, and hence of $\beta\mu_x$, is obtained by the intersection of $n_w$ values (black lines) with $\Gamma(\beta)$ (the blue curve) as given in Eq. (5.23). The inset shows the corresponding $\xi(\beta)$, i.e. the temperature-dependent part of $\beta\mu_x$ for various $n_w$ using model parameters reported by Silverstein et al. Here, $n_w = 30$ is insufficient to invert solubility, but $n_w^* \geq 71$ is.

Possible with a mere $n_w^* \approx 18$. This second choice of degeneracies, however, seems unphysical [125]; taking $q_{ob} = 10$ and $q_{os} = 1$, indeed suggests that the ordered bulk degeneracy is an order of magnitude larger than that of the ordered shell. We thus expect $n_w^* \gtrsim 71$ to be a physically more reasonable estimate.

That said, however, because $\varepsilon_{tot} = 60$ results in room temperature solubilities that are orders of magnitude lower than experimental observations, and in light of the various sources error in patch energy determination (Sec. 5.3.1), Ref. [98] proposed to halve patch energies in order to reduce the discrepancies. For $\varepsilon_{tot} = 30$, $n_w^* \sim 35$ (Fig. 5.6), which is less than the 40 or so water molecules solvating hydrophobic residues in Patch 4, but more than the energy-based estimate.

In light of the many estimates involved in the above analysis, the hydrophobic effect as a cause of inverted solubility, although weakly supported, cannot be eliminated outright. Even if the hydrophobicity model parameters are kept constant, a possible resolution could be for Patch 4 to be stronger than estimated and the other patches weaker. The hydrophobicity scenario, however, does severely constrain the patch model parameters. It should further be noted that the P23T mutation is not
Figure 5.6: Solubility lines corresponding to different values of $n_w$ for $\varepsilon'_\text{tot} = 30$, for which $n^*_w \geq 35$.

associated with a systematic change in $A_4$ (and thus $n_w$), which suggests that a conformational change in the solvating water molecules should accompany the mutation.

*Solubility Lines for Models of Hydrophobicity*

In Sec. 5.4.1 we determined that the hydrophobicity scenario for inverting solubility requires a fine balance between the protein-protein patch energies, the size of the hydrophobic patch, and the number of water molecules solvating it. While this rare confluence of factors could explain why inverted solubility is not common among proteins, it is natural to wonder whether the presence of weak hydrophobic patches, which are ubiquitous in proteins, affects solubility lines without engendering a regime of inverted solubility. In this section, we study the Wentzel-Gunton model in order to explore this possibility.

The solubility lines for the Wentzel-Gunton model in Fig. 5.7 are specifically obtained for $\beta_{\text{ref}} = 0.5$, to match the experimental solubility as in Ref. [98], but our observations are qualitatively independent of this choice. Setting $\Delta s_w = -10$, which is here akin to $n_w \approx 20$ (assuming that the temperature-dependent energy in the MLG model scales as $n_w$), results in normal solubility, but the steepness of the solubility curve changes markedly compared to $\Delta s_w = 0$. Setting $\Delta s_w = -15$
Figure 5.7: Larger magnitudes of $\Delta s_w$ invert solubility, whereas $\Delta s_w = -15$ results in solubility that only weakly depends on temperature, and $\Delta s_w = -10$ (light blue) results in normal solubility. Note that, for the latter case, even though the solubility is not inverted, the solubility line is markedly altered compared to the $\Delta s_w = 0$ case (dark blue).

$(n_w \approx 30)$ results in the DBI solubility being almost independent of temperature and in DBN being more stable than DBI at $T < T_{tp} \approx 1.7$. Further reducing $\Delta s_w$ gives rise to an inverted solubility regime. The solubility curve then flattens below $T \approx 2$ and $\phi_{tp}$ moves to higher packing fractions. These observations thus emphasize that the presence of an inverted solubility regime is the limit case of a continuum of how hydrophobicity impacts the solubility line.

5.4.2 Solubility Lines for Temperature Deactivated Patches

Absent definitive microscopic evidence for the hydrophobic effect, we finally consider a generic model for patch deactivation. The temperature-deactivated patchy model, which was used to successfully capture the inverted solubility of DBI [98], stabilizes the crystal with increasing temperature without referring to any specific microscopic mechanism. In this section, we first discuss the physical constraints on the model parameters and then consider how solubility lines change with model parameters, paying particular attention to the robustness of the inverted solubility regime. We also estimate the binodal and the critical temperature, which have been experimen-
tally determined for certain human γD-crystallin mutants [121].

Parameter Estimates

Despite the absence of an explicit microscopic interpretation for the (de)activation model, one can still place some reasonably solid physical constraints its tuning parameters. First, the (de)activation temperature $T_a$ must lie in the vicinity of the triple point, and thus $T_a \approx T_{tp}$. For our model, the choice $T_a = 1.9$ ensures that the deactivation of Patch 4 makes DBI metastable with respect to DBN for $T < T_{tp}$. Second, $\tau$, which sets the temperature range over which (de)activation takes place, ought to capture the degree of cooperativity of the underlying microscopic process. It cannot be arbitrarily small, as it would be at a thermodynamic phase transition, because a macroscopic number of components would then need to be involved. It also cannot be arbitrarily large, because the inverted solubility regime then vanishes. For reference, recall that denaturing a protein takes place over a few degrees, and any smaller scale rearrangement that involves tens to hundreds of atoms should spread over at least $\gtrsim 10K$. We thus here consider a temperature range of $\sim 10K$, which corresponds to setting $\tau = 0.05$.

We first investigate how varying patch energies impacts the phase diagram, keeping $T_a = 1.9$ and $\tau = 0.05$ constant. As previously reported [98], the resulting phase diagram (Fig. 5.8a) exhibits a re-entrance regime bounded by the DBI solubility line, as well as a triple point between the fluid and the two crystal forms. The solubility lines that result from perturbing the patch energies by 5% and 10% shift to substantially lower or higher densities, but the existence of an inverted solubility regime is robust. As expected, the errors inherent to the overall parameterization of the model are therefore qualitatively benign.

We then investigate the robustness of the results with respect to the relative strength of the temperature-deactivated fourth patch, $\varepsilon_4$. This question is of interest
Figure 5.8: (a) Average solubility lines for perturbed parameters. Dashed lines denote 95% confidence intervals. DBI solubility line for 10% error (black) and 5% error (red), as well as DBN solubility line for 10% error (green) and 5% error (blue) are shown. Higher error levels increase the uncertainty in $\phi_{tp}$, as well as the minimum solubility observed for DBI, but inverted solubility is maintained. (b) The effect of changing the energy of the temperature-deactivated patch, such that $\varepsilon'_4 = f \varepsilon_4$. (c) The effect of changing $\varepsilon_4$ but keeping the total patch energies of DBI constant.

for two main reasons: (i) the strength and robustness of solubility inversion depend sensitively on the strength of that patch; and (ii) the ordering of the single mutant solubilities directly correlates with their respective Patch 4 energies.

The impact of multiplying $\varepsilon_4$ by $f \in \{0.9, 1.0, 1.5\}$, while keeping the other patch parameters constant is shown in Fig. 5.8b. Increasing the strength of Patch 4 systematically decreases the solubility of DBI and lowers $\phi_{tp}$. Interestingly, the decrease in solubility with increasing $f$ is consistent with the experimental observations for the single mutants, P23T, P23S, and P23V [12]. Because a stronger Patch 4 decreases the DBI solubility (R36S+P23T double mutant), assuming that the difference between crystals arises due to Patch 4 only, we speculate that if two other double mutants, R36S+P23S and R36S+R23V, were crystallized with similar crystal contacts, then their inverted solubility would order similarly.

The impact of changing the energy of the temperature-dependent patch while keeping the total energy of DBI patches constant, i.e., $\varepsilon_{tot} = f_1 \varepsilon_4 + f_2 (\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_5)$ as shown in Fig. 5.8c. (Because the second patch corresponds to a shared contact between DBI and DBN, the DBN solubility is then also slightly perturbed.) Here
Figure 5.9: (a) As $\tau$ is increased, the DBI solubility line becomes less flat, and eventually inverted solubility is lost (e.g. $\tau = 0.35$). (b) Manipulating the sum of DBI patch energies and $\tau$, one can obtain a temperature range over which the solubility is almost temperature-independent.

Again, the inverted solubility regime vanishes upon markedly reducing the strength of Patch 4. The difference with the first case is that DBN is now metastable with respect to DBI within the probed temperature range, while DBI becomes metastable with respect to DBN otherwise. For $f_1 = 0.4$, DBN is still metastable with respect to DBI, but inverted solubility is observed only over a narrow range of density.

We also investigate the robustness of the phenomenology with respect to changes in $\tau$. Decreasing $\tau$ corresponds to a faster temperature (de)activation of the patch, which flattens the inverted solubility region and results in $T_{tp} \to T_a$ as $\tau \to 0$ (Fig. 5.9a). $\phi_{tp}$ similarly gets pushed to higher packing fractions, suggesting that a protein solution prepared very near $T_a$ could reach remarkably high concentrations compared to solutions prepared at surrounding temperatures. However, as argued above, very small values of $\tau$ are physically unreasonable. Conversely, increasing $\tau$ weakens this transition and eventually eliminates the inverted solubility regime. Interestingly, a specific choice of $\tau$, with a minor tweak to patch energies ($\tau = 0.28$, $\varepsilon_i' = 1.1\varepsilon_i$), gives rise to a nearly vertical solubility curve (Fig. 5.9b), similar to the temperature-independent solubility of apoferritin [184].
Figure 5.10: The liquid-liquid binodal regions for the temperature-deactivated patch model for various Patch 4 energies, $\varepsilon_4$. For the reference model patch energy, $\varepsilon_4 = 16$ ($\tau = 0.05, T_a = 1.9$), there is no closed-loop binodal. Only if $\varepsilon_4 > 36$, here $\varepsilon_4 = 40$ is shown (red), are multiple critical points and a closed-loop binodal obtained.

Estimation of the Critical Temperature

Although various theoretical results suggest that a closed-loop binodal with multiple critical points is possible upon introducing temperature-dependent binding energies [165, 194, 48], no experimental evidence of such a binodal has been found for any human $\gamma$D-crystallin mutant. In addition, experiments find that the P23V mutation, which also inverts solubility, has a binodal that is indistinguishable from that of the wild type [121]. Here we use Wertheim’s perturbation theory [195, 196], which provides quantitatively good estimates of the binodals in patchy models to explore this question (see SI Sec. S4 [112]). In order to determine whether our model is consistent with these observations, we estimate the liquid-liquid binodal and the associated critical temperature, $T_c$.

Choosing $\tau = 0.05$ and $T_a = 1.9$, as above, results in a typical binodal with a single critical point at $T_c = 1.85$ (Fig. 5.10). Hence, without altering patch energies, our model does not give rise to a closed-loop binodal (Fig. 5.10), consistently with experiments [121]. In order to determine how far our model is from exhibiting a
closed-loop binodal, we systematically increase the energy of Patch 4. Only when 
$\varepsilon_4$ more than doubles does a closed-loop binodal appear. (See Fig. 5.10). This 
perturbation, however, falls far outside of the error estimates of the patch energies, 
which further supports the qualitative robustness of our model prediction.

5.5 Conclusion

In this article, we have attempted to rationalize the inverted solubility of certain 
mutants of $\gamma$D-crystallin based on microscopic models of protein-protein interactions 
and their temperature dependence. We have paid particular attention to the putative 
role of hydrophobicity. Estimating surface hydrophobicity using different scales 
did not reveal the presence of any pertinent surface feature, but microscopic models 
of hydrophobicity suggest that the number of available surrounding water molecules 
might suffice. Although our analysis falls short of conclusively determining whether 
hydrophobicity plays a determining role or not in this protein, this scenario nonetheless 
seems a bit far fetched due to the lack of structural evidence, as well as the 
thermodynamic constraints on the patch energies for such a hydrophobic scenario 
to occur. (By the same token, however, this analysis provides an explanation as to 
why inverted solubility is far less common than proteins with hydrophobic patches on 
their surface.) A more conclusive determination would require for the water structure 
around the region of interest to be more specifically probed. Because standard water 
models are insufficiently sensitive to this feature [8], simulations with more sophisti-
cated water models [97, 175, 110, 12], and neutron diffraction or hydrogen-deuterium 
exchange experiments might be more productive avenues. For the former, however, 
extensive testing, including comparison with experimental structures, of the ability 
of these water models in capturing protein-water interactions would first be needed.

Even though the microscopic origin of inverted solubility in human $\gamma$D-crystallin 
still remains somewhat elusive, additional insight from the crystallization of other
double mutants, such as R36S+P23V and R36S+P23S, might be helpful in identifying generic features that might have eluded the analysis thus far. Repeating the above structural and thermodynamical for these mutants could help tease out more subtle features that might be at play.
This Chapter presents methods developed to properly sample a two-dimensional crystalline sheet of proteins.

6.1 Introduction

Self-assembly into single-layer sheets typically proceeds through the support of a surface or an interface \[30, 53, 117\]. Without such support, the thickness, \( h \), of membranes, such as lipid bilayers, diverges with system size, \( L \), as \( \langle h^2 \rangle \sim L^2 \) \[114\], and hence the resulting membranes are not thermodynamically stable. A few systems, however, appear to violate this physical argument, such as protein S-layers \[146\] and CdTe nanocrystals \[176\], both of which can form macroscopic sheets in solution, without any support. More recently, a protein has even been engineered to self-assemble into micron-sized, defect-free two-dimensional crystals \[173\]. At first sight, this result further seems to violate the Mermin-Wagner theorem, which states two-dimensional perfect crystals are unstable to topological defects. A class of crys-
talline (or tethered) membranes nonetheless has been shown to display long-range orientational order in their flat phase thanks to the coupling of in- and out-of-plane fluctuations. The correlation of surface normals, $G(r)$, is then predicted to scale as (at zero surface tension)

$$G(r) = C + \frac{b}{r^\eta}, \quad (6.1)$$

where $C$ and $b$ are materials constants, and $\eta$ is a universal critical exponent. Recent theoretical estimates give $\eta = 0.85 \[105\]$, which has been verified in molecular simulations of graphene \[114\]. In Fourier space, the same relationship can be written as

$$G_q = \frac{1}{A\beta \kappa q^{2-\eta}}, \quad (6.2)$$

where $q$ is the magnitude of the wave vector, $|q|$, $\kappa$ is the bending rigidity, and $A$ is a normalization factor. This quantity is related to the out-of-plane fluctuations in Fourier space, $\langle h_q \rangle$, as

$$G_q = q^2 \langle |h_q|^2 \rangle. \quad (6.3)$$

By contrast, a non-crystalline membrane without such a long-range order (and for the general case of a non-zero surface tension, $\gamma$, has

$$G_q = \frac{1}{A\beta (\gamma + \kappa q^2)}. \quad (6.4)$$

Our aim is to investigate the phase behavior of a protein engineered to assemble into a two-dimensional crystal \[173\]. The plan for this chapter is as follows. We first introduce the patchy model for this protein, as well as the specialized Monte Carlo
The protein has $C_4$ symmetry, and cysteines at each of the four “corners”. This allows for self-assembly into a square lattice. We construct a patchy particle for this protein such that there are four equally-spaced patches arranged around the equator.

algorithm we use. We then introduce a benchmarking model for a crystalline membrane and present validation results for these models. We conclude by identifying the next steps for describing the phase behavior of this system.

6.2 Methods

In this section, we describe the test systems and the specialized Monte Carlo algorithm.

6.2.1 Patchy Model and Self-Assembly

Because the protein has $C_4$ symmetry and assembles through co-planar cysteines, we model this protein using a patchy particle with four patches along the equator, such that the patch vectors are given by $\mathbf{p} = (\pm 1, \pm 1, 0)$ (Fig. 6.1). Unless otherwise noted, the range of the square well attraction is $\lambda = 1.1\sigma$, where $\sigma$ is the particle diameter. The patch width is $\theta = \cos^{-1}(0.95)$, where $2\theta$ is the apex angle of the cone that defines the patch width.
6.2.2 Specialized Monte Carlo Moves

In the following, we describe the specialized Monte Carlo moves we use for this study. The system setup is such that the membrane is aligned to be parallel to the $xy$-plane of an $L_x \times L_y \times L_z$ box, with height, $L_z \gg L_x, L_y$, so that out-of-plane fluctuations are not affected by their periodic image in the $z$-direction.

**Area Moves**

In order to work in the constant surface tension, $\gamma$, ensemble, we implement constant area moves. Two variants are possible: keeping the total volume constant, or keeping the box height constant. The area is changed following a random walk in $\log A$, which corresponds to randomly selecting $\Delta \log A \in (-\Delta \log A_{\text{max}}, \Delta \log A_{\text{max}})$. Because the system is expected to preserve its aspect ratio, isotropic area moves, i.e., $L_x$ and $L_y$ are changed by the same percentage.

For volume-conserving area moves \[55\], the box size is changed such that

\begin{align*}
L'_x &= f L_x \\
L'_y &= f L_y \\
L'_z &= L_x L_y L_z / (L'_x L'_y),
\end{align*}

where $f = \exp(0.5 \times \Delta \log A)$. The position of particle $i$ is also scaled accordingly:

\begin{align*}
x'_i &= f x_i \\
y'_i &= f y_i \\
z'_i &= L_x L_y / (L'_x L'_y) z_i.
\end{align*}

The acceptance rule then becomes

\[\min(1, e^{-\beta \Delta U + \beta \Sigma \Delta A + \Delta \log A}),\]
where $\Sigma$ is the tension imposed on the system.

For constant height area moves $[191]$, $L_z$ and $z_i$ are kept constant and $L_x$, $L_y$, $x_i$, and $y_i$ are changed the same way as in the constant volume area moves. The acceptance rule for this case is then

$$\min(1, e^{-\beta \Delta U} e^{\beta \Sigma A + (N+1) \Delta \log A}). \tag{6.12}$$

**Wave Moves**

In order to measure $\gamma$ and $\kappa$, as well as to determine if the crystalline membrane is in the flat phase, it is imperative to sample small $q$ surface fluctuations efficiently. Because such long wavelength fluctuations involve a collective reorganization of particles, sampling with single particle moves alone becomes very inefficient as the system size grows. We therefore implement wave moves as described in Ref. [114]. We choose a cutoff wavevector magnitude, $q_{\text{max}}$, and apply wave moves for smaller $q$. Random positive integers $n_x$ and $n_y$ are chosen such that the components of the wave vector, $q_x = 2\pi n_x / L_x$ and $q_y = 2\pi n_y / L_y$, result in $q < q_{\text{max}}$. We also choose random phase shifts $\varphi_x$ and $\varphi_y$, between 0 and $L_x$ and $L_y$, respectively. The random amplitude $A_q \in (-A/q, A/q)$ is chosen to account for the energy scaling of these excitations, $\sim A^2 q^2$. The displacement of the $z$-coordinate of particle $i$, $\Delta z_i$, is then given by

$$\Delta z_i = A_q \sin(q_x x_i + \varphi_x) \sin(q_y y_i + \varphi_y), \tag{6.13}$$

and the acceptance rule is simply

$$\min(1, e^{-\beta \Delta U}). \tag{6.14}$$

**6.2.3 The Simplest Model for a Crystalline Membrane**

The simplest possible model for a crystalline membrane consists of a triangular mesh $[33]$ with the following Hamiltonian
\[ H = \sum_{e \in \text{edges}} l_e^2 + \kappa \sum_{a,b} (1 - n_a \cdot n_b), \]  

(6.15)

where \( l_e \) is the length of an edge \( e \) in the mesh, \( a \) and \( b \) are neighboring triangular facets, and \( n_a \) and \( n_b \) are their normal vectors. Note that this model does not have any self-avoidance, and thus the surface becomes crumpled below a critical bending rigidity, \( \kappa_c \). Previous studies involving this model had free boundary conditions, and did not consider wave moves [33].

6.3 Results

In this section, we present results of consistency checks for the Monte Carlo algorithm, and determine the crumpling transition for the simple crystalline membrane model.

6.3.1 Performance Increase due to Wave Moves

In order to compare the efficiency of various simulation settings, we compute the autocorrelation functions of \( |h_q|^2 \) for low \( q \). We simulate strips of membranes containing \( N = 5 \times N_y \) particles, separated by 1.05\( \sigma \) at \( \beta = 10 \). Introducing wave moves reduces the relaxation timescale for small \( q \) by up to an order of magnitude. For instance, \( N_y = 500 \) for \( q = 0.0120 \) (the smallest nonzero \( q \) for this system size) has \( \tau_\alpha = 1.4 \times 10^5 \) when no wave moves are performed, but \( \tau_\alpha = 3.4 \times 10^4 \) when ten wave moves are performed per MC sweep. \( \tau_\alpha \) for a given \( q \) increases with system size (Fig. 6.2), as well as for a given system size with decreasing \( q \). Although increasing the number of wave moves per MC sweep reduces \( \tau_\alpha \) slightly, it also slows down the overall simulation because the time per MC sweep increases more quickly than \( \tau_\alpha \) is reduced. Therefore, \( O(1) \) wave moves per MC sweep seems to be the most efficient choice. With this setup, \( \tau_\alpha \sim 10^4 \) for out-of-plane displacements.
Figure 6.2: Autocorrelation function, $C(t)$, for $q = 0.0598$ for three different system sizes. $\tau_\alpha$ for a given $q$ increases with system size, and is drastically larger for simulations with no wave moves (blue curve). The black line shows $C = 1/e$.

Figure 6.3: (a) Imposed frame tension, $\Sigma$ vs. measured surface tension, $\gamma$, and bending rigidity $\kappa$, for $N = 35^2$ and $N = 100^2$. The results are given for constant volume area moves. Error bars denote the 95% confidence interval, and the $\Sigma = \gamma$ line is shown in black. Fits to the linearized form of Eq. 6.4 for (b) $N=100^2$ and (c) $N=35^2$.

6.3.2 Consistency Check for Area Moves

We check if the measured $\gamma$ matches the imposed frame tension, $\Sigma$. (Results for $N = 35^2$ and $N = 100^2$ are as shown in Fig. 6.3a). We find that the trend in $\langle |h_q|^2 \rangle$ is well explained by Eq. 6.4 (Fig. 6.3a and b) and thus extract the surface tension and bending rigidity by fitting this equation.
6.3.3 Crumpling Transition in the Triangular Mesh Model

Bowick et al. estimated that the crumpling transition of the mesh model takes place at $\kappa_c = 0.79$ \[33\]. We verify this measurement using both membranes with free boundaries as well as with periodic boundary conditions and constant volume area moves and wave moves. Fig. 6.4 presents how system size and bending rigidity affects the radius of gyration, $R_g$, as a function of $\kappa$ for various system sizes, where $R_g$ is defined as

$$R_g = \langle \frac{1}{3N} \sum_i r_i^2 \rangle,$$  \hspace{1cm} (6.16)

where $r_i$ is the distance of the $i^{th}$ particle from the membrane center of mass.

We find that below $\kappa_c = 0.8 \pm 0.1$, the membrane crumples, resulting in drastically smaller $R_g$ values. The system size dependence of $R_g$ well below and above $\kappa_c$ is also consistent with theoretical predictions (Fig. 6.5). For both the free and periodic boundary cases, $R_g \sim \log(N)$ at $\kappa = 0$. For $\kappa = 2$, $R_g \sim N^{\nu/2}$, with $\nu = 1 \pm 0.02$ for the free boundary case and $\nu = 0.986 \pm 0.004$ for the periodic boundary case, consistent with the theoretical prediction, $\nu = 1$ \[33\].

6.4 Outlook

We have implemented and tested wave and area moves, and showed that our simulation setup captures the crumpling transition of a simple crystalline membrane model. More tests are required before we can investigate the phase behavior of the protein that assembles into sheets. First, more systematic testing of the area moves, also using the constant height scheme is required. An area move that preserves the box height, but scales the $z$-coordinates of particles should also be considered. Second, attempts to extract $\eta$ from various observables calculated for the triangular mesh
should be made. Once the properties of this system are reproduced using our novel methodology, the patchy protein model will be considered and the model parameters will be optimized to match with experimental systems that form a stable crystalline membrane.

**Figure 6.4:** Crumpling transition, measured in terms of the radius of gyration for (a) free boundaries and (b) periodic boundary conditions.

**Figure 6.5:** $R_g$ scales as (a) $\log(N)$ at $\kappa = 0$ and as (b) $N^{\nu/2}$ at $\kappa = 2$, for both free boundaries and periodic boundary conditions (PBC).
In this dissertation, we tackled different problems associated with protein crystallization. First, we studied the effect of biomolecular solvation on the amount and accuracy of structural information that can be obtained from X-ray diffraction data. Testing classical MD water models commonly used in simulations of biomolecules in water revealed that a lot of these models fall short of reproducing the structure of water inside protein crystals with satisfactory accuracy for enhancing structural refinement. Nevertheless, MD simulations produce high-intensity peaks in water density that accurately predict crystal waters assigned during protein refinement, which allows for the detection of the protonation state of sufficiently solvent-exposed residues. The framework we developed can be used for testing additional water models and proteins, which provides a way of determining if new and improved water models can capture biomolecular solvation adequately. If a satisfactory model were identified, our methodology would further allow for incorporating MD-generated water densities into the protein model during protein structure refinement.

Second, we investigated the microscopic and thermodynamic causes for the inverted solubility of a human γD-crystallin mutant. Developing and parameterizing
a patchy particle model with temperature-dependent interactions allowed for recapitulating the protein solubility lines. Next, we tested the hypothesis that increased surface hydrophobicity upon mutation leads to that inverted solubility. We investigated the crystal structures of six proteins with and without the solubility inverting mutation and did not find any discernible increase in hydrophobicity. Nevertheless, we were able to show that the hydrophobic scenario could thermodynamically give rise to inverted solubility. It is not clear, however, whether hydrophobic patches on the surface of proteins can be treated similarly as entirely hydrophobic solutes.

One common finding of Chapters 3, 4, and 5 is that water models commonly used in MD simulations fall short of describing hydrophobic solvation. This shortcoming severely limits the accuracy with which the structure of water around proteins is captured, particularly their orientation. This failure is most likely because water models are parameterized to reproduce the thermodynamic properties of bulk water, which offers no immediate basis for protein-water interactions. While considering polarizable water models, or models that take into account many-body effects, could alleviate this problem, they would first need to be tested extensively, using a framework akin to that developed in Chapter 3. Only then, an accurate *in silico* investigation of water orientation around hydrophobic residues can be performed. Testing the accuracy of the microscopic arguments used for developing the hydrophobicity model we use in Chapter 5 would then also be possible.

Finally, we investigated a protein that assembles into a two-dimensional crystal. Because this crystal is thermodynamically stable, it is likely a member of a class of materials called crystalline membranes. We developed the tools required to sample the membrane properties adequately. These will help determine the protein properties that allow for a stable two-dimensional crystal to form, and possibly guide future efforts at designing such biomaterials.
Bibliography


132


134


Biography

Irem Altan attended Koc University between 2009 and 2014, where she obtained two BSc degrees, one in Chemistry and the other in Physics. In 2014, she moved to Durham, North Carolina for her doctoral studies in the Department of Chemistry at Duke University, under the supervision of Dr. Patrick Charbonneau. While at Duke, she was awarded the Charles K. Bradsher and Kathleen Zielik Departmental fellowships.

Publications in reverse chronological order:


