UNIVERSITY OF CALIFORNIA,
IRVINE

# Modeling the structure and dynamics of gamma-crystallins and their cataract-related variants

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Chemistry

by

Eric K. Wong

Dissertation Committee:
Professor Douglas J. Tobias, Chair
Professor Rachel W. Martin
Professor Craig C. Martens

2018

# DEDICATION

To my parents, Dr. Marston and Arleen Wong, for their continual love and support.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

Page

# ACKNOWLEDGMENTS

First and foremost, I would like to thank my research advisor, Dr. Douglas J. Tobias. Leading a research group with many members across many fields, he has continually strove to provide for a diverse set of projects and collaborators to each member. His enthusiasm for scientific research as well as teaching has left a long standing impression upon me, which I will carry through to my future endeavors. To Dr. Rachel W. Martin, my collaborating research advisor, I thank her for providing me academic support as well as the interdisciplinary research environment that is essential for computational research. I also want to Dr. Juan Alfredo Freites, who has served as a collaborator as well as a PI to several projects. His expansive and in depth knowledge across many fields has been of great help to me, co-workers, and members from several other labs.

To my experimental collaborators, Dr. Domarin Khago, Dr. Carolyn Kingsley, and Jan Bierma, I would like to thank them for all of their hard work on $\gamma$S-crystallin experiments. I would also like to thank visiting members of the lab: Matt McCummins, David Wych, Also, I want to extend thanks to Joseph Farran and the rest of the HPC support staff for providing and maintaining such an essential computational resource. A special thanks goes to my friends from graduate school: Mya Le Thai, Paolo Reyes, Jerry Guo, Suvrajit Sengupta, Raul Ocampo, and Julie Hsu for the good food and laughter after hours.

I would also like to thank Jerry Hu and Dr. Ben Burke, my employers and mentors before I joined graduate school. Their mentorship was the first to inspire me to pursue a career in research. I cannot thank them enough for affording me the opportunity to work and learn from them at such an early point in my career.

Finally, I would like express deep gratitude to my family. Their support and motivation has been unending during my years spent obtaining my graduate degrees. My sister has a spontaneity and vitality that she puts into everything that she does. From her I learned the importance of working with passion and enthusiasm. Last but not least, I would like to thank my parents for their hard work and devotion to me and my sister. They have taught me the importance of a strong work ethic, while at the same time, allowing me the freedom to pursue my own path. They have led fulfilling lives together starting from humble beginnings, to traveling the world serving their country, to starting their own business while raising two troublemakers. I feel forever blessed to have them as both my parents and my role models. Thank you.

Computational resources were provided by the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation grant ACI-1053575. Anton/Anton 2 computer time was provided by the Pittsburgh Supercomputing Center (PSC) through Grant R01GM116961 from the National Institutes of Health. Financial support was provided under the following grants: UCCRCC-103783, DMR-1410415, and NIH R01 EY025328-01A1.

The text of chapter 2 in this thesis/dissertation is a reprint of the material as it appears in

"Increased hydrophobic surface exposure in the cataract-related G18V variant of human $\gamma$S-crystallin". The co-author listed in this publication directed and supervised research which forms the basis for this thesis.

# CURRICULUM VITAE

## Eric K. Wong

### EDUCATION

**Doctor of Philosophy in Chemistry** **2018**
University of California, Irvine *Irvine, CA*

**Masters of Science in Chemistry** **2017**
University of California, Irvine *Irvine, CA*

**Bachelor of Science in Chemical Engineering** **2010**
University of California, San Diego *La Jolla, CA*

### RESEARCH EXPERIENCE

**Graduate Research Assistant** **2011–2017**
University of California, Irvine *Irvine, CA*

**Computational Chemistry Intern** **2009–2011**
Pfizer Inc. *La Jolla, CA*

### TEACHING EXPERIENCE

**Teaching Assistant** **2011–2015**
University of California, Irvine *Irvine, CA*

- Computational Methods (Chem/Phys 229A)
- Chemical Kinetics (Chem 213)
- Analytical Chemistry (Chem 151)
- Scientific Computing Skills (Chem 5)
- Honors/Majors General Chemistry Lab (Chem H2LA/B/C)
- General Chemistry Lab (Chem 1LC/D/E)

## PUBLICATIONS

- **Increased hydrophobic surface exposure in the cataract-related G18V variant of human S-crystallin.** Khago, D.*; **Wong, E. K.***; Kingsley, C. N.; Freites, J. A.; Tobias, D. J.; Martin, R. W. Biochimica et Biophysica Acta (2016) 1860, 325–332.
- **Cataract-related W42R $\gamma$D-crystallins show spontaneous domain separation and increased propensity for interprotein interaction at high concentrations.** **Wong, E. K.***; Prytkov, V*; Freites, J. A.; Tobias, D. J.*In preparation*
- **Conservation of structure and dynamics in mesophilic and psychrophilic $\gamma$S-crystallin: thermal adaptation of structural eye lens proteins.** **Wong, E. K.**; Freites, J. A.; Martin, R. W.; Tobias, D. J. *In preparation*
- **Atomistic simulations show anomalous subdiffusion and ergodicity breaking resulting from macromolecular crowding.** **Wong, E. K.**; Freites, J. A.; Tobias, D. J. *In preparation*
- **k-Cores in Dynamic Networks as a Tool for Analysis of Protein Structure Cohesiveness** Freites, J. A.; Zhang, X.; **Wong, E. K.**; Martin, R. W.; Tobias, D. J.; Butts, C. T. *In preparation*
- **Microsecond simulations of an open state model of the human Hv1 proton channel.** Geragotelis, A.; Wood, M. L.; Goeddeke, H.; Hong, L.; **Wong, E. K.**; Freites, J. A.; Tombola, F.; Tobias, D. J. *In preparation*

* = Equal contributions

## POSTER PRESENTATIONS

- **Computational Study of Anthracycline Interactions with Membrane-embedded P-Glycoprotein.** **Wong, E. K.**; Freites, J. A.; Tobias, J. D. *Biophysical Society Meeting.* Los Angeles, CA 2016
- **Modeling Interprotein Interactions in Concentrated Solutions of Wild-Type and Cataract-Related Variants of $\gamma$D- and $\gamma$S-Crystallins.** Pytkova, V. D.; Heyden, M. B.; **Wong, E.**; Freites, J. A.; Tobias, D. J. *Biophysical Society Meeting.* Los Angeles, CA 2016
- **Molecular Dynamics Simulations of $\gamma$S-Crystallin.** Freites, J. A.; **Wong, E. K.**; Brubaker, W. D.; Kingsley, C. N; Brindley, A.; Martin, R. W.; Tobias, D. J. *Biophysical Society Meeting.* San Francisco, CA. 2013

- **Novel Application of Cytochrome P450 Tools to Impact Drug Design and Evaluate Drug Interactions** Hu, J.; Burke, B.; **Wong, E.** *Pfizer Internal CYPTech Convention.* La Jolla, CA. 2009

# ABSTRACT OF THE DISSERTATION

Modeling the structure and dynamics of gamma-crystallins and their cataract-related variants

By

Eric K. Wong

Doctor of Philosophy in Chemistry

University of California, Irvine, 2018

Professor Douglas J. Tobias, Chair

$\gamma$-crystallins are structural eye lens proteins responsible for focusing light into the retina. These proteins are highly stable, being capable of remaining soluble at concentratins exceeding 300 g/L for an entire lifetime. Upon a loss in solubility, either by mutation or chemical damage to the protein structure, opaque aggregates form in the eye lens, known as cataract. Understanding the conformations and interactions of the $\gamma$-crystallins in their aggregated state will provide a deeper understanding of the mechanisms behind cataract formation.The work presented in this dissertation will investigate cataract-related sequences of $\gamma$S- and $\gamma$D-crystallins ($\gamma$S-WT and $\gamma$D-WT, respectively). Using molecular dynamics and other modeling techniques, I investigate potential sites for interprotein interaction, often through the exposure of hydrophobic residues. In the congenital cataract-related G18V variant of $\gamma$S-crystallin ($\gamma$S-G18V), the exposure of hydrophobic patches are identified in a relatively folded protein. Simulated protein-ligand conformations of a hydrophobic probe (ANS) identify hydrophobic sites both local and allosteric to the site of mutation. In the W42R variant of $\gamma$D-crystallin ($\gamma$D-W42R), a 17 $\mu$s molecular dynamics simulations shows the spontaneous separation of the N- and C-terminal domains (whereas $\gamma$S-WT remains stable for 50 $\mu$s). Two protein simulations show that the hydrophobic interdomain interface becomes the main site for interprotein interaction, providing strong support for a domain swapping aggregation pathway at physiological conditions. The thermal stability is analyzed for the $\gamma$S-crystallins of the human and Antarctic toothfish. The less thermally stable toothfish proteins show correlated increases in backbone flexibility and decreases in the packing of the hydrophobic core, yet maintain similar structure and dynamics at their native temperature. Finally, an analysis of the domain-domain motions in $\gamma$D-WT fluctuations is performed in the presence and absence of macromolecular crowding. The autocorrelation function of the interdomain distance ages over 50 $\mu$s of simulation, showing non-convergent dynamics even after significant computational sampling.

# Chapter 1

# Introduction

Cataract is the leading cause of blindness worldwide, accounting for more than 50% of cases of blindness, globally[1]. Cataract is the opacification of the eye lens caused by a loss in solubility and aggregation of structural eye lens proteins called crystallins. The cloudiness formed in cataract is a result of the scattering of visible light from large aggregates of the crystallin proteins[2]. The cataracterous lenses can be replaced surgically, however the procedure can be expensive and the cataract surgical rates are reported to be low in some regions such as sub-Saharan Africa and southeast Asia[3, 4]. This necessitates the development of non-surgical methods for cataract treatment. A deeper understanding of the molecular mechanisms behind cataract formation is essential for the development of novel treatments for regions without access to cataract surgery.

The human eye lens contains some of the oldest cells in the human body[5]. Nearly 99% of the eye lens core[6] is composed of bundles of elongated, transparent cells called lens fiber cells. Upon cellular differentiation, all lens fiber cell organelles are lost, leaving high concentrations (90% by mass[7]) of tightly packed crystallin proteins. These crystallin proteins provide the refractive medium to focus light towards the retina[8], and, since lens fiber cells lack any protein expression mechanisms, must remain soluble for an entire lifetime. Crystallins are divided into 2 families: $\alpha$- and $\beta\gamma$-crystallin. $\alpha$-crystallins are large chaperone complexes that protect the eye lens from large scale aggregation by binding misfolded proteins[9, 10]. The $\beta\gamma$-crystallins are the structural eye lens proteins where $\beta$-crystallins exist as dimers and $\gamma$-crystallins exist as monomers. In this dissertation, I will be focusing on structure and dynamics related to the monomeric $\gamma$-crystallin.

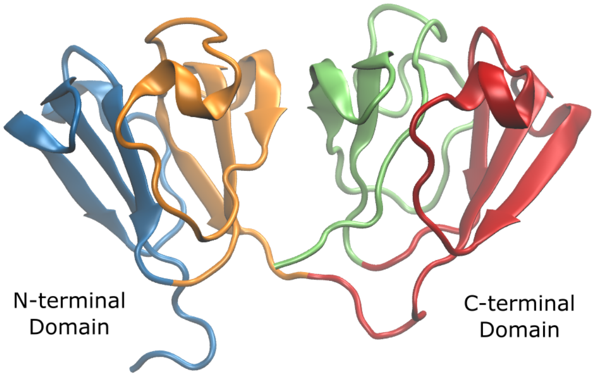

Figure 1.1: Cartoon representation of the NMR structure of human $\gamma$S-crystallin[10]. Each Greek key motif is colored separately, and the protein is oriented such that the N-terminal domain is on the left and the C-terminal domain is on the right.

$\gamma$-crystallins exist as a 21 kDa, two domain protein, where each domain contains two Greek key motifs, composed of four anti-parallel $\beta$-strands in each motif (Figure 1.1). Each of the four Greek keys are structurally homologous, yet non-identical. The N-terminal (NTD) and C-terminal (CTD) domains are joined by a compact hydrophobic interface, resulting in a globular protein about 5 nm in diameter. $\gamma$-crystallins are highly stable proteins, capable of remaining soluble at high concentrations (450-1000 g/L)[11, 12] in the eye lens. However, congenital or post-translational modifications to the protein sequence can result in a loss in solubility and aggregation, resulting in a loss of lens opacity. The cloudiness that is characteristic of lens cataract is a result of the scattering of visible light from these insoluble crystallin aggregates.

The crystallin aggregation pathway has been shown to be highly dependent on both the protein environment[13, 14, 15] and alterations to the protein sequence[10, 16, 17]. The current hypothesized aggregation mechanisms involve different degrees of misfolding[18]. The condensation mechanism details the aggregation of mostly folded proteins containing altered interprotein interactions[19, 20, 21]. Partial unfolding, usually involving intact Greek key domains, can result in the swapping of $\gamma$-crystallin domains[22, 23]. Understanding the structure and dynamics behind these aggregation prone states can give useful insights into the mechanisms behind aggregation and can inform strategies for its prevention.

Several structures of $\gamma$-crystallin and their cataract-causing variants have been reported using solution-state NMR spectroscopy[10, 24, 25] and x-ray crystallography[26, 27, 28]. Despite the variants being capable of rapid aggregation, many of these structures closely resemble their wild-type counterparts. However, reports of changes in structural stability[29, 21], exposed hydrophobicity[20, 30], and the presence of partially unfolded states[27] suggest that further insights can be obtained beyond the average structures. In fact, if such subpopulations of unfolded intermediates exist, computational analysis of the protein structure and stability may be able give predictions on how these aggregation-prone states arise.

In this dissertation, I will computationally model $\gamma$-crystallins to provide examples of:

1. Altered surface hydrophobicity in $\gamma$S-G18V, a cataract-related variant with minimal unfolding
2. Changes in tertiary structure in $\gamma$D-W42R, a protein with a known partially unfolded intermediate
3. Conservation of structure and dynamics in the cold adapted toothfish $\gamma$S-crystallins
4. Interdomain dynamics that are non-stationary in the microsecond timescale

To approach these questions, I will present results from atomistic molecular dynamics (MD) simulations. Given a reasonable estimate of a protein conformation, MD simulations produce a fully atomistic description of protein motions in a solvated environment. Much of the accuracy of MD simulations is owed to the close curation and validation of the MD force field parameters[31, 32], the set of force constants needed to define the bonded and nonbonded interaction potentials between all atoms. By integrating through time using Newton's equations of motion, protein-solvent trajectories can be generated and analyzed for their interactions and dynamics. The data from these atomistic models have potential for guiding new experiments as well as providing insights into protein function.

However, sufficient sampling of relevant protein motions can be computationally expensive. Since the MD simulations must integrate Newton's equations of motion with a femtosecond ($10^{-15}$s) timestep, protein dynamics simulations require $> 10^9$ timesteps to sample the larger protein motions residing in the microsecond timescale. Currently, most high-performance computing clusters are capable of sampling about tens of nanoseconds in a day. This is sufficient to sample local fluctuations, yet larger domain motions from the microsecond to millisecond timescales[33] remain inaccessible. Recently, the availablility of Anton 2, a special purpose supercomputer for MD simulations[34], has made microsecond to millisecond MD simulations computationally feasible. With this, the larger, more collective protein motions can be investigated for applications in protein folding and conformational change.

Chapter 2 will characterize changes in the hydrophobic surface of $\gamma$S-G18V, a $\gamma$S-crystallin variant that readily aggregates at low concentrations. Despite being aggregation-prone, the solution-state structures show only minimal unfolding and is more thermodynamically stable than other non-aggregating variants[21]. Using a combination of molecular docking and NMR chemical perturbation experiments performed by Domarin Khago, we identify regions of exposed hydrophobicity related to the local unfolding of the protein. This demonstrates the exposure of sites for interprotein interaction with only local unfolding on the protein surface.

Chapter 3 is a working manuscript presenting results from microsecond timescale MD simulations of the cataract-related W42R variant of $\gamma$D-crystallin ($\gamma$D-W42R). The crystal structure of this protein closely resembles the wild-type structure[27]. However, biophysical experiments indicate changes in tertiary structure[35] as well as small populations of partially unfolded proteins[27]. As a part of this work, MD simulations show that, when modeled under physiological temperature and pH, the two domains separate, resulting in an open conformation that is prone to increased interprotein interaction. Previous works have hypothesized a similar domain-swapping pathway through the use of high temperatures, strong denaturing conditions[22], and atomic force microscopy experiments[23]. This new conformation provides support for the domain-swapping hypothesis under physiological temperature and pH.

Chapter 4 will highlight the functional importance of protein dynamics in $\gamma$S-crystallin as structural proteins. The psychrophilic, or cold-adapted, toothfish $\gamma$-crystallins are resistant to cold-cataract formation yet have a decreased thermal stability. MD simulations comparing the human and toothfish crystallins show a loss in thermostabilizing salt bridges, yet have a similar structure and dynamics at their respective environmental temperatures (277 K and 300 K for toothfish and human, respectively). The conservation of dynamics in temperature adaptation highlights the importance of protein flexibility in the functional role of $\gamma$S-crystallins as eye lens proteins.

Chapter 5 will characterize the dynamics of the interdomain fluctuations of wild-type human $\gamma$D-crystallin (H$\gamma$D). A 50 $\mu$s MD simulation of H$\gamma$D shows that the interdomain motions are subdiffusive with fluctuations exhibiting a non-exponential decay in the two-time autocorrelation function. Most notably, the effective relaxation time for domain fluctuations is dependent on observation time. This observation time dependence is evidence of a non-ergodic or ageing process, where dynamics remains non-convergent well into the microsecond timescale. This non-ergodic dynamics shows that $\gamma$-crystallins contain internal motions that show non-convergent behavior well into the microsecond timescale.

# Chapter 2

# ANS Docking to the G18V variant of human $\gamma$S-crystallin

## 2.1 Introduction

High concentrations of closely packed crystallin proteins are necessary for maintaining the transparency and refractive index gradient of the eye lens. The human lens has several structural crystallins that are found with different radial distributions; the focus of this study is $\gamma$S-crystallin, which is preferentially located in the lens cortex (periphery) [36, 37]. The solution-state NMR structure of wild-type $\gamma$S-crystallin has been determined [10], revealing a double Greek key architecture for each of the two domains, consistent with the structures of other $\beta\gamma$-crystallins. The childhood-onset cataract variant G18V ($\gamma$S-G18V) is structurally similar to $\gamma$S-WT, but it has dramatically lower thermal stability and solublity [38, 21], as well as strong, specific interactions with $\alpha$B-crystallin, the holdase chaperone of the lens [10]. Despite the well-documented aggregation propensity and reduced stability of $\gamma$S-G18V, the particular intermolecular interactions leading to its aggregation are as yet unknown.

Protein self-aggregation leading to cataract can occur due to an increase in net hydrophobic interactions, as previously shown in the congenital Coppock-type cataract variant D26G $\gamma$S-crystallin [39], the cerulean cataract variant P23T $\gamma$D-crystallin [20], acetylation of G1 and K2 residues in $\gamma$D-crystallin [40], and the lamellar cataract variant D140N $\alpha$B-crystallin [41]. All of these mutations introduce altered conformations that produce lowered solubility by exposure of hydrophobic patches on the surface, even though the structural differences from their wild-type counterparts are relatively subtle. $\gamma$S-G18V is no exception; the mutation does not cause large-scale unfolding or rearrangement into a misfolded conformation, but rather produces altered intermolecular interactions with itself and with $\alpha$B-crystallin [10].

The fluorescent probe 1-anilinonaphthanlene-8-sulfonate (ANS), which has both negatively charged and hydrophobic moieties, is often used to quantify exposed hydrophobic surface in binding to hydrophobic surface patches in proteins [20, 42, 43]. Two types of protein-ANS interactions are required for fluorescence enhancement: hydrophobic interactions between the conjugated ring system of ANS and the protein surface [44], and electrostatically between the sulfonate group and positively charged side chains at the binding site [45]. An increase in fluorescence intensity indicates that either more ANS is binding to the protein surface, or that it is bound more tightly, correlating with higher surface hydrophobicity. This method has been used to characterize exposed hydrophobic surface in a number of protein systems, including the mitochondiral chaperone protein Atp11p, which recognizes its client proteins via hydrophobic interactions [46], and aggregation-prone variants of superoxide dismutase-1 (SOD1), an essential cellular enzyme whose aggregation is associated with amyotrophic lateral sclerosis (ALS) [47, 48]. Despite the utility of ANS binding as a probe of hydrophobic surface exposure, and the sensitivity afforded by using fluorescence as a reporter, this assay is limited by the lack of detailed information about which amino acid residues, or even general regions of the protein, are taking part in the dye-binding interaction. NMR chemical shift perturbation (CSP) mapping can forge a link between fluorescence enhancement upon dye binding and the corresponding changes in the local chemical environment of specific residues in the protein. Comparisons between wild-type and variant proteins can then be used to compare differences in exposure of hydrophobic residues on the surface under particular solution conditions. CSP mapping is a commonly used technique for investigating protein-protein or protein-ligand interactions and interfaces [49], and is the basis of the "SAR by NMR" methodology that is indispensable in the identification of active pharmaceutical agents [50].

Molecular docking, a computational technique widely used to model the conformation of protein-ligand complexes, enables experimental perturbations to be analyzed in atomistic detail. Bound ligand conformations, or poses, are ranked using an empirical scoring function designed to evaluate intermolecular interactions using minimal computational time. Conventionally, knowledge of the active site is used to guide the pose generation, often in the context of screening large libraries of compounds against known protein structures [51, 52, 53, 54]. However, docking protocols without prior knowledge of the active site (blind docking) [55], have successfully identified putative allosteric binding sites of drugs, leading to the design of novel allosteric modulators [56], and fluorescent dyes [57, 42]. Bis-ANS binding sites found by docking, validated with steady-state and time-resolved fluorescence assays, have been used to identify hydrophobic patches in a lipase from *Bacillus subtilis*[58].

## 2.2 Materials and Methods

### 2.2.1 ANS fluorescence assay

Wild-type and G18V $\gamma$S-crystallins were expressed and purified as previously described [21]. Fluorescence spectra were collected as a function of ANS binding for $\gamma$S-WT and $\gamma$S-G18V with a F4500 Hitachi fluorescence spectrophotometer. The excitation and emission wavelengths were 390 nm and 500 nm, respectively, with slits set to 5 nm. Protein concentrations

| | | | |
|---|---|---|---|
| Center | $^{1}$H: 799.8056964 MHz | $^{13}$C: 201.1282461 MHz | $^{15}$N: 81.0504078 MHz |
| Offset | $^{1}$H: -294.932 Hz (4.8 ppm) | $^{13}$C: -9863.17 Hz (43 ppm) | $^{15}$N: 2400 Hz (116.7 ppm) |

Table 2.1: Final concentrations of $\gamma$S-WT and $\gamma$S-G18V

for both $\gamma$S-WT and $\gamma$S-G18V were approximately 1 mg/mL in 10 mM sodium phosphate buffer and 0.05% sodium azide at pH 6.9. ANS concentrations ranged from 5 $\mu$M to 2 mM were measured using $\epsilon$= 4.95 mM$^{-1}$ cm$^{-1}$ at 350 nm [59].

### 2.2.2 NMR sample preparation

Purified protein with the 6x-His tag removed was concentrated and supplemented with 2 mM TMSP, 10% D2O, and 0.05% sodium azide. The final concentration of all $\gamma$S-WT and $\gamma$S-G18V samples was 0.3 mM. ANS was titrated into the protein samples to give final molar ratios of 1:0, 1:0.5, 1:1, and 1:2 of $\gamma$S:ANS. Spectra were acquired at 25 °C.

### 2.2.3 NMR experiments

Experiments were performed on a Varian $^{\text{Unity}}$INOVA spectrometer (Agilent Technologies) operating at 800 MHz and equipped with a $^{1}$H–$^{13}$C–$^{15}$N 5 mm tri-axis PFG triple-resonance probe, using an 18.8 Tesla superconducting electromagnet (Oxford instruments). Decoupling of $^{15}$N nuclei was performed using the GARP sequence [60]. $^{1}$H chemical shifts were referenced to TMSP, and $^{15}$N shifts were referenced indirectly to TMSP. NMR data were processed using NMRPipe [61] and analyzed using CcpNMR Analysis [62]. Center operating frequencies and (unless otherwise stated) center frequency offsets were as follows:

### 2.2.4 Calculation of chemical shift perturbations

$^1$H-$^{15}$N HSQC spectra of $\gamma$S-WT and $\gamma$S-G18V were collected in the presence and absence of ANS at concentration ratios of 1:0, 1:0.5, 1:1, and 1:2 of $\gamma$S:ANS, and resonances were identified and assigned based on chemical shift data previously collected by our group. Resonances showed perturbations that are indicative of ANS binding. The change in chemical shift for each peak in the 2D spectrum upon ANS binding was calculated using the following chemical shift perturbation (CSP) equation:

$$\Delta\delta_{avg} = \sqrt{\frac{(\Delta\delta_N/5)^2 + (\Delta\delta_H)^2}{2}} \tag{2.1}$$

A strong-binding threshold for each set of conditions was set at two times the root mean square (RMS) of the calculated CSP, while the weak-binding threshold was set at half the RMS to determine which residues had strong or weak binding with ANS. The values used for each threshold appear in Supplementary Table S1.

### 2.2.5 Binding site search by rigid receptor docking

Protein coordinates were obtained from the NMR structures of $\gamma$S-WT and $\gamma$S-G18V crystallins (PDB ID: 2M3T and 2M3U) [10]. Autodock Tools [63] was used to prepare both the receptor (crystallin) and ligand (ANS) by merging non-polar hydrogens atoms into united heavy atoms. Gasteiger charges[64] were added to each atom. The sulfonic acid group of ANS was deprotonated before processing by Autodock Tools. Molecular docking was performed using Autodock Vina [65]. In order to ensure good coverage of the protein binding surface, 27 search spaces were placed in an overlapping 3 x 3 x 3 grid around the protein (Supplementary Figure A.1). Since Autodock Vina works optimally with search spaces with at most a 27,000 Å$^3$ volume, a 30 x 30 x 30 Å search space was chosen. The exhaustiveness parameter was set to 20 (over the default value of 8) in order to ensure an extensive search of the protein surface. Docking was performed over each one of twenty solution-state NMR conformations for either $\gamma$S-WT or $\gamma$S-G18V. The resulting poses were screened to ensure that both electrostatic and hydrophobic interactions required for ANS fluorescence enhancement upon binding were present. Docked poses that did not include both interactions within the first coordination shell of the ANS-protein radial distribution function were considered non-fluorescent and removed from the docked set. The screened docked set covers most of the protein surface (see Supplementary Figure A.2).

### 2.2.6 Calculation of residue contacts

To compare the screened docked set with the residue-based CSP data, ANS-residue contact frequencies were calculated by summing the Boltzmann weights of all the poses in contact with a given residue. The Boltzmann weight of a given docked pose was calculated according to

$$w_i = \frac{exp(-E_i/k_BT)}{\sum_i exp(-E_i/k_BT)} \tag{2.2}$$

where $i$ is the index of the docked pose, $E_i$ is the pose binding energy, k$_B$ is the Boltzmann constant, and $T$ is the absolute temperature. The residue contact frequencies for each protein are shown in Supplementary Figure S3. Following the CSP analysis, to determine which residues had strong or weak binding with ANS, a strong-binding threshold was set at two times the RMS of the calculated ANS-residue contact frequency, while the weak-binding threshold was set at the RMS value. The values used for each threshold appear in Supplementary Table S1.

### 2.2.7 Flexible refinement of binding sites

A flexible docking refinement was performed near all the highly perturbed residues accorindg to the strong-binding cutoff on the CSP data. Docking search spaces were defined by clustered conformations of ANS from the screened docked set used to calculate the ANS-residue contact frequencies. Using a root-mean-square deviation cutoff (RMSD) of 5.0 Å, clustered poses were grouped into potential binding sites near the experimentally perturbed residues (see Supplementary Figure A.2). Search spaces were defined as boxes surrounding the clustered ligands with an 8 Å padding. The padding was necessary to include flexible side chains within the search space. Residues with an experimental CSP above the low-binding cutoff were considered as flexible. A total of five potential binding sites were used to dock ANS to either flexible $\gamma$S-WT or $\gamma$S-G18V. The resulting poses were clustered again, and the location and interactions of each pose were compared visually.

## 2.3 Results and Discussion

### 2.3.1 ANS fluoresence indicates that the relative surface hydrophobicity of $\gamma$S-G18V is higher than that of $\gamma$S-WT

Dye-binding assays were performed on $\gamma$S-WT and the aggregation prone variant, $\gamma$S-G18V. The ANS fluroscence measurements for $\gamma$S-WT and $\gamma$S-G18V, shown in Figure 2.1, indicate more exposed hydrophobic surface in $\gamma$S-G18V compared to its wild type counterpart. These data also allow determination of the lowest ANS concentration required to produce the maximum emission before saturation, which was 1.5 mM for $\gamma$S-WT and 1 mM ANS for $\gamma$S-G18V. The lower concentration required to saturate $\gamma$S-G18V is consistent with the observation that it binds ANS more readily than wild-type.

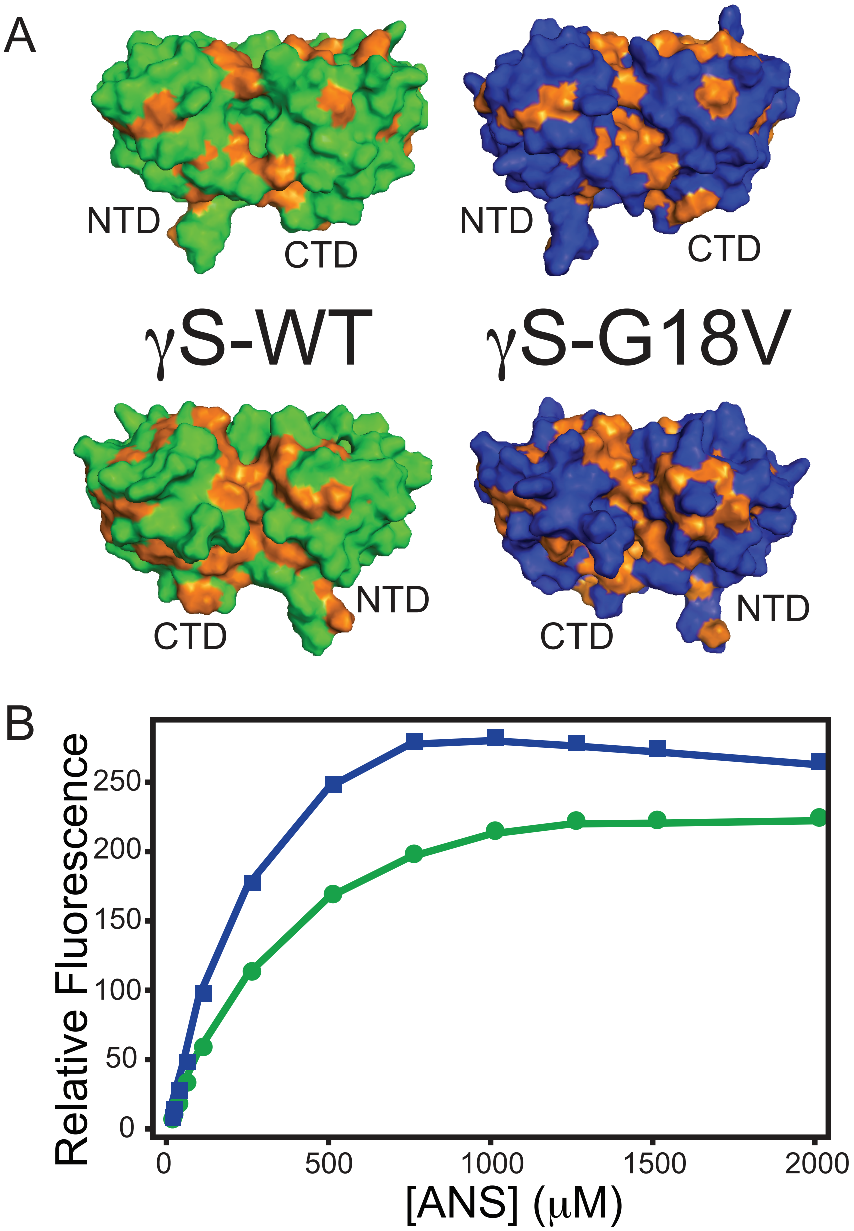

Figure 2.1: A. Molecular surface representation of $\gamma$S-WT (green) and $\gamma$S-G18V (blue) based on the solution-state NMR structures (PDB ID 2M3T and 2M3U, respectively). Hydrophobic residues are highlighted in orange. B. Fluorescence spectra representing ANS binding monitored at 500 nm using $\gamma$S-WT and $\gamma$S-G18V crystallins. Protein concentrations for both S-WT and S-G18V were approximately 1mg/mL. Saturation occurred at 1.5mM ANS for S-WT and 1 mM ANS for $\gamma$S-G18V. Higher emission was observed for $\gamma$S-G18V, indicating more hydrophobic surface area exposed than for $\gamma$S-WT.

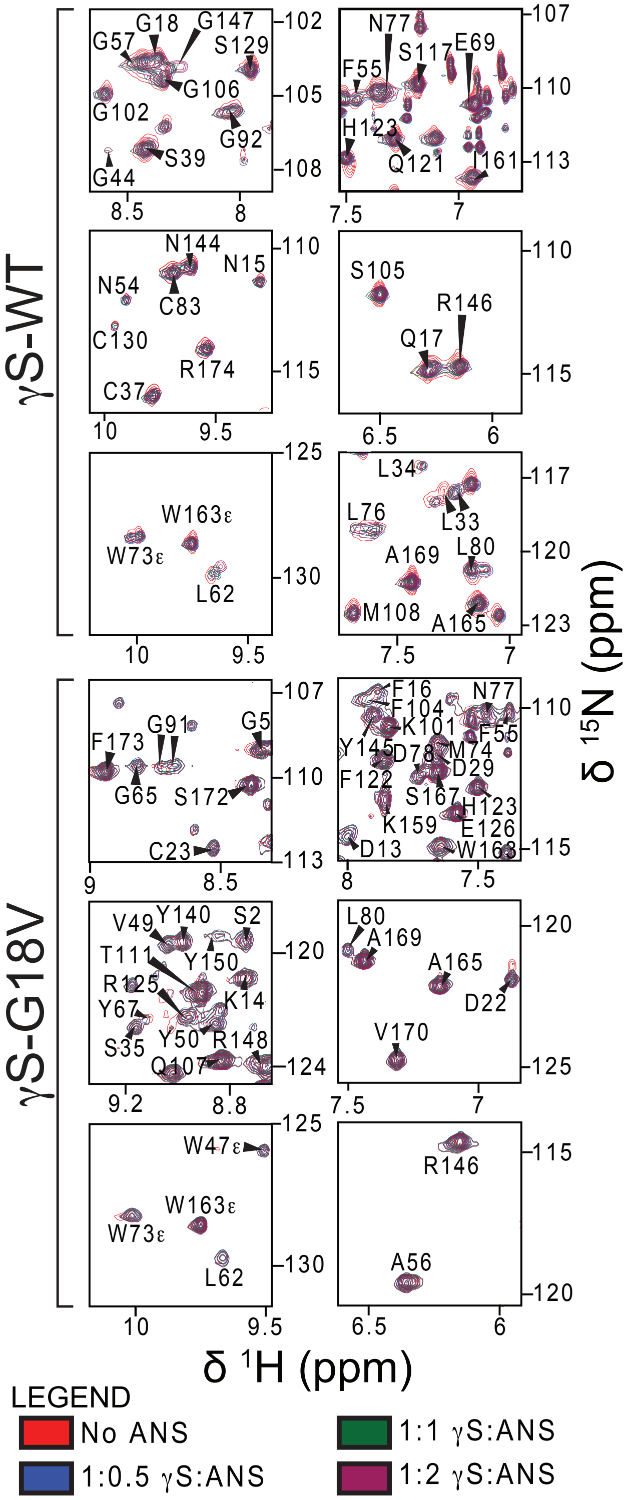

Figure 2.2: Selected portions of the $^{1}$H-$^{15}$N HSQC spectra of $\gamma$S-WT and $\gamma$S-G18V. Experiments were carried out using final ratios of 1:0, 1:0.5, 1:1, and 1:2 of $\gamma$S:ANS. Resonances indicative of a change in chemical shift are indicative of multiple ANS binding to specific residues. However the perturbations observed are small due to the low concentrations of $\gamma$S-crystallin and ANS used. Spectra were acquired at 25 °C with final concentrations of all $\gamma$S-WT and $\gamma$S-G18V samples at 0.3 mM.

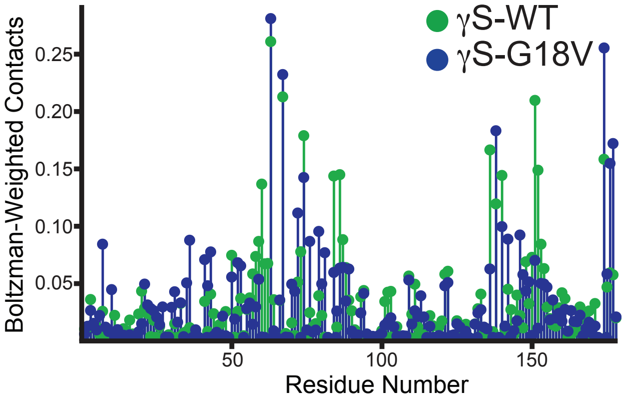

Figure 2.3: Average chemical shift perturbation (CSP) of $\gamma$S-WT (green) and $\gamma$S-G18V (blue). Nonspecific binding, with maximum perturbation in the N-terminal domain, is observed in both proteins. However, in $\gamma$S-G18V, more of the CSPs are localized to the N-terminal domain. Particularly between residues 15 to 50, in the cysteine loop near the mutation site. Inspection of the structures confirms that this region is exposed to solvent in $\gamma$S-G18V but not in $\gamma$S-WT

### 2.3.2 Chemical shift perturbation mapping reveals the residues involved in ANS binding and the relative strengths of the interactions

Binding interactions between ANS and $\gamma$S-WT or $\gamma$S-G18V were measured at concentration ratios of 1:0, 1:0.5, 1:1, and 1:2 of $\gamma$S:ANS, using CSP mapping via $^{1}$H-$^{15}$N HSQC spectra. Selected regions of the NMR spectra where resonances show perturbations indicative of ANS binding are shown in Figure 2.2. The full NMR spectra can be found in the Supplemental Information (Supplementary Figures S4 and S5). The change in chemical shift for each peak in the 2D spectrum upon ANS binding was calculated using Equation 1. Representative CSP data for $\gamma$S-WT and $\gamma$S-G18V upon 1:1 ANS binding is shown in graphical form in Figure 2.3. The complete set of calculated CSP data can be found in the Supplemental Information (Supplementary Figures S6 and S7). As shown in Figure 2.3, although nonspecific binding is observed throughout the surfaces of both proteins, $\gamma$S-G18V binds ANS more strongly in the N-terminal domain (approximately the first 100 residues). The maximum ANS binding occurs within residues 15 through 50, close to the mutation site. These observations are

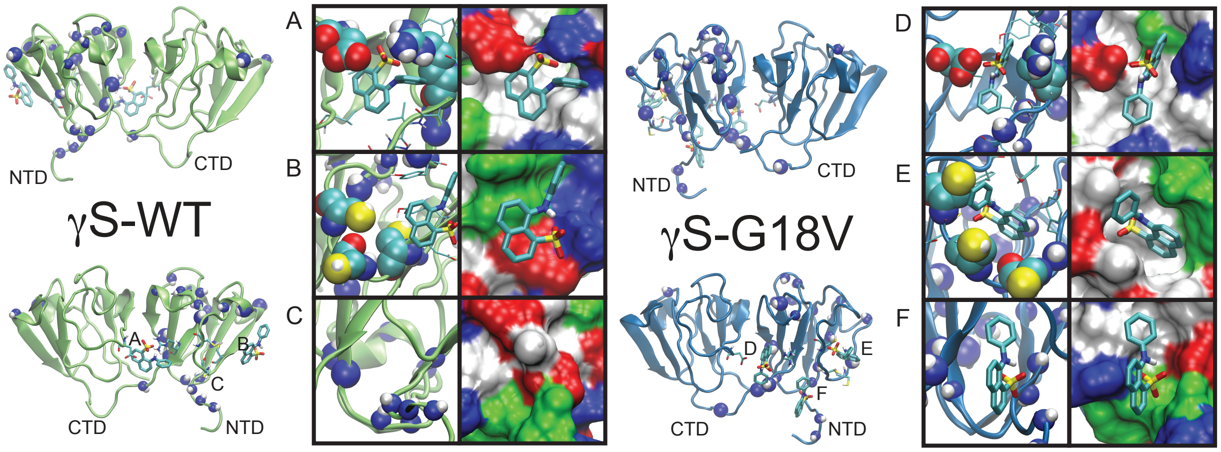

Figure 2.4: ANS interactions with $\gamma$S-WT and $\gamma$S-G18V. The strong-binding threshold and weak-binding threshold were defined as two times the RMS and half the RMS, respectively. Experimental CSP values indicate that ANS binding occurs throughout the N- and C-terminal domains for S-WT, (strong binding residues in green and weak binding residues in pale green), while in $\gamma$S-G18V ANS binding mainly occurs at the N-terminal domain (strong binding residues in blue and weak binding residues in pale blue). Some strong binding is observed in the N-terminal domain for both proteins near the mutation site, e.g.G18 in $\gamma$S-WT and D22 in $\gamma$S-G18V. However, $\gamma$S-G18V displays more ANS binding (both strong and weak) overall in the N-terminal domain. Strong binding is also observed in the interdomain interface of $\gamma$S-WT, residues L62, S82, and H123, and $\gamma$S-G18V, residues L62,W73, H87, L88, and G91. G18V exhibits more binding (strong and weak) within that interdomain interface suggesting that this variant has higher surface hydrophobicity localized to the N-terminal domain near the mutation site and the interdomain interface. Coverage of both strong and weak binding residues are nearly identical between experimental and docking results, highlighted in dark green for $\gamma$S-WT and dark blue for $\gamma$S-G18V, indicating that the docking results are in good agreement with the experimental data.

mapped onto the protein structures in Figure 4 (left panel) where the residues having strong and weak ANS binding are highlighted. A strong-binding threshold (two times the RMS) and a weak-binding threshold (half the RMS) were set for each condition. For $\gamma$S-WT, strong binding residues are highlighted in bright green and weak binding residues in pale green. For $\gamma$S-G18V, strong binding residues are highlighted in dark blue and weak binding residues in pale blue. The CSP data indicate that although both proteins display ANS binding throughout both the N- and C-terminal domains, $\gamma$S-G18V has additional ANS binding residues, mostly in the N-terminal domain. Although some strong binding residues exhibited in the N-terminal domain for both proteins near the mutation site, (e.g. G18 in $\gamma$S-WT and D22 in $\gamma$S-G18V), G18V displays more ANS binding, both strong and weak, in the N-terminal domain. Strong binding is also seen in the interdomain interface of $\gamma$S-WT, (residues L62, S82, and H123), and $\gamma$S-G18V (residues L62, W73, H87, L88, and G91). Similar perturbations at the interdomain interface (residues G65, Y67, S82, S85, and G91) were shown for $\gamma$S-G18V in the presence of $\alpha$B-crystallin, the primary holdase chaperone protein of the eye lens [10]. $\alpha$B-crystallin only weakly interacts with $\gamma$S-WT at the surface of the protein (residues S35, W47, E66, G92, F122, and H123) [10]. The full list of residues interacting with both ANS and $\alpha$B-crystallin for both $\gamma$S-WT and $\gamma$S-G18V is tabulated in Table 2 for comparison. Notably, $\alpha$B-crystallin strongly binds near the interdomain interface in $\gamma$S-G18V in but not $\gamma$S-WT, consistent with the hypothesis that the chaperone is recognizing an exposed hydrophobic patch in this region of $\gamma$S-G18V.

ANS is a small molecule probe that reports hydrophobic exposure due to its propensity to bind to hydrophobic patches. Due to its small size, ANS will perturb residues local to the hydrophobic patch without largely affecting the receptor conformation. These perturbed residues reveal hydrophobic patches on the apo form of $\gamma$S-G18V that are local to residues implicated in $\gamma$S-G18V interaction with $\alpha$B-crystallin. This is consistent with the hypothesis that the chaperone is recognizing an exposed hydrophobic patch in $\gamma$S-G18V. The larger set of perturbed residues in $\alpha$B-crystallin can be attributed to size of $\alpha$B-crystallin relative to

ANS, resulting in a much larger interaction interface.

The concentrations used for these NMR studies of both $\gamma$S-WT and $\gamma$S-G18V were below concentrations used in previous studies, where dynamic light scattering (DLS) data observed monomeric conditions for both proteins [10]. Another indication that monomeric conditions were used was based on the observed line widths. Table 3 in the Supplemental Information reports line widths of $\gamma$S-WT and $\gamma$S-G18V for several residues from the 1:1 $\gamma$S:ANS mixtures. The line widths reported for WT and G18V are comparable to one another, unlike previous studies were crosspeaks were broadened out below the noise threshold due to the presence of large complexes in solution resulting in the use of transverse-relaxation optimized spectroscopy (TROSY)-HSQC [10].

### 2.3.3 Docking of ANS on the protein surface predicts more binding sites on $\gamma$S-G18V than $\gamma$S-WT and allows interpretation of the CSP data

Rigid receptor docking resulted in a total of 4860 docked poses (27 search spaces × 20 NMR conformations × 9 poses/search space). After screening for poses consistent with ANS fluorescence enhancement upon binding, 3423 poses and 3367 poses remained for $\gamma$S-WT and $\gamma$S-G18V, respectively (see Supplementary Figure S2A). Filtered poses covered nearly the entire surface of the protein and exhibit a broad range of scores (from -2 kcal/mol to -7 kcal/mol, with a mean of -4.5 kcal/mol). Due to the pocket-like shape of the interdomain interface, ANS preferentially bound to the large hydrophobic pocket between the N- and C-terminal domains. However, sites were identified near all highly perturbed residues with comparable binding scores (see Supplementary Figure S3). Flexible docking poses located near the highly perturbed residues according to the CSP data had binding scores between -4.5 kcal/mol and -6.0 kcal/mol, consistent with a stronger preference for ANS to bind near the perturbed residues. A total of ten binding sites were found for $\gamma$S-G18V and nine binding sites for $\gamma$S-WT using flexible docking. Most of these binding sites were found to be very similar in both $\gamma$S-WT and $\gamma$S-G18V. However, three binding modes were found to be unique to $\gamma$S-G18V. The first and most populated binding mode is located in the hydrophobic cavity at the interface between the N- and C-terminal domains, shown in Figure 5A and 5D [66, 29]. Although this binding site was found in both $\gamma$S-WT and $\gamma$S-G18V, the presence of the R84-D153 salt-bridge blocks the exposure of the hydrophobic surface in $\gamma$S-WT. In contrast, $\gamma$S-G18V lacks this salt-bridge interaction, which exposes the interdomain hydrophobic cavity and allows the entry of ANS into the interdomain binding site. This is consistent with the experimental NMR data, which indicate that chemically perturbed residues, H87 and L88, located near the interdomain pose, interact strongly with ANS only in $\gamma$S-G18V (Figure 5). The second and third binding sites are located close to residues 20 - 30, which includes a loop region containing three cysteine residues (C23, C25, and C27). As a result of the G18V mutation, C23 and C27 become solvent exposed, suggesting possible formation of intermolecular disulfide bridges, consistent with the observation that an excess of reducing agents abrogates the formation of small oligomers [67]. Previous studies suggested that the exposure of these cysteines results from a disruption in secondary structure due to the burial of V18 side chain [10]. As a result of this cysteine exposure and concomitant structural changes, a new hydrophobic pocket is uncovered as the second ANS binding site. Although ANS binds this Cys loop in $\gamma$S-WT after flexible docking refinement, it is not in direct contact with any hydrophobic surface, suggesting that the pose may not be consistent with and enhancement in ANS fluorescence (Figure 5B). In contrast, when the hydrophobic pocket is exposed, as it is in $\gamma$S-G18V, ANS becomes buried deep within the pocket (Figure 5E). This conformation provides both the hydrophobic interactions necessary for fluorescence as well as reduced quenching due to water exposure [44]. In addition to cysteine exposure, the third binding site shows additional hydrophobic surface exposure due to the cysteine loop separating from the main Greek key motif. This binding site is not found in $\gamma$S-WT using the same docking search space, indicating that this hydrophobic patch is a unique characteristic of $\gamma$S-G18V (Figure 5F). Additionally, the CSP data shows local perturbation of the backbone amides of the residues involved in these three binding sites only in for $\gamma$S-G18V. The presence of these new binding sites can explain the higher ANS fluorescence intensity of $\gamma$S-G18V over WT, and they also identify exposed hydrophobic patches which may potentially serve as protein-protein interfaces in crystallin aggregates, and which can be targeted in future mutagenesis studies.

The CSP data and the ANS-residue contact data from the docking simulations show generally good agreement in that the same protein regions were observed to bind ANS (see Figure 4). In some cases, the specific residues classified as strong binding vary between experimental and docking results, but coverage of both strong- and weak-binding residues are nearly identical (highlighted in dark green for $\gamma$S-WT, and dark blue $\gamma$S-G18V in the right panel of Figure 4). This outcome is to be expected because the docking scoring function is more effective at identifying binding sites than distinguishing more subtle changes in binding energy: the standard error of the Autodock Vina scoring function [65] is larger than the variation among scored poses. The agreement between rigid protein docking results and experimental ANS binding results suggests that there is no major change in protein conformation upon binding of ANS, supporting the hypothesis that hydrophobic patches on the surface are involved in intermolecular interactions. Good agreement between the experimental and docking results further confirms that ANS binding is localized near the mutation site in the N-terminal domain for $\gamma$S-G18V, consistent with the CSP data. Experimental and simulation results are also consistent on the binding of ANS to the exposed interdomain hydrophobic surface located in the interdomain interface between the two domains due to the breaking of the R84-D153 salt bridge in $\gamma$S-G18V. Exposure of this hydrophobic patch facilitates ANS binding and may be involved in hydrophobic protein-protein interactions.

# Chapter 3

# MD simulations of the W42R variant of human $\gamma$D-crystallin

## 3.1 Introduction

Cataract, the opacification of the eye lens, is the leading cause of blindness in many developing countries[3, 4]. This opacification is caused by the aggregation of a family of proteins in the eye lens called crystallins[68]. These proteins make up 90% of the protein content in the eye lens fiber cells. Upon cell differentiation, these cells are denucleated, resulting in the loss of protein turnover for all eye lens proteins. Thus, in order for the eye lens to maintain its function, crystallin proteins must remain soluble at concentrations exceeding 400 g/L for an entire lifetime[69]. However, congenital defects and post-translational modifications, such as UV photo-oxidation, deamidation, and truncation, can result in the formation of large protein aggregates that diffract light[17, 70, 71]. These light diffracting aggregates make up the cloudiness that is characteristic of nuclear cataract.

Crystallin proteins are divided into 3 families: $\alpha$-, $\beta$-, and $\gamma$-crystallins. $\alpha$-crystallins are heat shock proteins that serve as holdase chaperones. These chaperones bind misfolded $\beta/\gamma$-crystallins, but do not refold them, preventing further aggregation[10, 9]. $\beta$- and $\gamma$-crystallins are dimeric and monomeric structural proteins, respectively. These structural proteins allow the eye lens to modulate the index of refraction while maintaining its lens transparency. Many congenital[21, 27, 28]) and post-translational modifications[72, 73] have been linked to cataract formation, often through different mechanisms of structural change. In this paper, we focus on the aggregation-related properties of the monomeric human $\gamma$D-crystallin ($\gamma$D-WT) and its cataract-related W42R variant ($\gamma$D-W42R).

$\gamma$D-crystallin is a 173 residue,  21 kDa, monomeric protein comprised of primarily -sheets organized into two Greek key domains. The congenital cataract-related W42R variant was reported to have changes in tertiary structure and a reduced thermal stability[35]. However, the reported crystal structure of $\gamma$D-W42R is near identical to that of wild-type[27]. Recent experiments have identified the formation of internal disulfide cross linkages upon chemical denaturation of the N-terminal domain[74], suggesting the existence of a small population of partially unfolded intermediates. However, structural details of this intermediate state and its aggregation pathway at physiological conditions remain uncertain.

Several mechanisms of aggregation have been reported, often involving different degrees of unfolding. Large-scale denaturation of the Greek key domain structure can lead to the formation of amyloid fibrils, characterized by fibrillar aggregates containing intermolecular $\beta$-sheets[13, 15]. Moderate unfolding, resulting in the separation of the N- and C-terminal domains, can lead to domain-swapped aggregates, where the separated domains reform new inter-protein domain interfaces[22]. Lastly, minimal unfolding can alter interprotein interactions such that the proteins aggregate in a native-like state[19, 20]. It is important to note, that the aggregation pathway is dependent on the type of stress (chemical, pH, thermal, etc.) put on the protein. pH stress on $\gamma$-crystallins has been shown to form amyloid fibrillar aggregates that differ in morphology from aggregates formed under physiological conditions[13, 75]. Furthermore, some cataract-forming crystallin variants have been shown to be more thermally stable than non-cataract forming variants[21]. Therefore, it is important to characterize $\gamma$-crystallins variants at physiological conditions to understand the pathway to cataract formation.

To computationally investigate the cataract-related conformations and interactions of $\gamma$D-W42R and its aggregates, we use a combination of microsecond-scale molecular dynamics (MD) simulations and Multi-Conformation Monte-Carlo (mcMC) simulations to model the single protein dynamics and interactions at high concentration, respectively. Additionally, we use network analysis to investigate the morphologies of the aggregates. We show that the N- and C-terminal domains (NTD and CTD, respectively) of $\gamma$D-W42R spontaneously separate in absence of thermal or chemical denaturation. The resulting domain-separated conformations contain patches of exposed hydrophobic residues that, in turn, become the primary sites of interprotein interaction in Monte Carlo simulations of $\gamma$D-W42R at high concentrations. These domain-separated conformations of $\gamma$D-W42R show a propensity to form higher order aggregates than $\gamma$D-WT in mcMC simulations. Based on our overall results, we provide atomistic computational evidence of significant conformational changes in $\gamma$D-W42R that result in large-scale aggregation.

## 3.2 Results

### 3.2.1 W42R human $\gamma$D-crystallin domains separate after salt-bridge interaction

To investigate the conformational dynamics of wild-type $\gamma$D-crystallin and its cataract-related W42R variant, single protein MD simulations of $\gamma$D-WT and $\gamma$D-W42R were run for 50 $\mu$s and 17 $\mu$s, respectively. Although the crystal structure of $\gamma$D-W42R contains the

W42R point mutation, the mutant structure is strikingly similar to that of the wild-type protein (backbone RMSD of 0.795 Å)[27]. In the single protein MD simulation of $\gamma$D-W42R, the protein structure remains similar to its initial structure for several microseconds. Over the course of 6 $\mu$s, the residue R42 gradually becomes solvent exposed and ultimately forms a salt-bridge with the C-terminal carboxyl group (Figure 3.1B). Upon salt-bridge formation, the N- and C-terminal domains separate, resulting in a $> 10$ Å increase in the $\alpha$C-RMSD for $\gamma$D-W42R (Figure 3.1B). As the domains separate, the C-terminal domain rotates such that the set of hydrophobic interdomain residues become solvent exposed for both domains (Figure 3.2B). Interestingly, the internal fluctuations of the N- and C-terminal domains of the domain-separated $\gamma$D-W42R is similar to that of $\gamma$D-WT (Figure 3.2C). This gives no indication for further unfolding of the Greek key domains after domain separation in $\gamma$D-W42R from our observable timescale. In the case of the wild-type protein, the protein maintains its native conformation over the course of 50 $\mu$s of simulation (Figure 3.1A).

To identify the exposure of new potential interprotein contacts, the relative solvent accessibilities (RSA) of $\gamma$D-WT and $\gamma$D-W42R were calculated for each clustered conformation prepared for the mcMC simulations (Figure 2D). The RSA represent the residue solvent accessibility normalized over the theoretical maximum solvent accessibility for each type of residue (Tien et al., 2013). Residues with a change in RSA $> 0.13$ are all located at the interdomain interface in $\gamma$D-WT (shown as VDW spheres in Figure 2A & B). These are primarily patches of hydrophobic residues that form the hydrophobic core of the N- and C-terminal domains. These exposed hydrophobic patches give opportunities for the formation of new interprotein contacts. To test the role of these residues in interprotein interaction, we performed multi-conformation Monte-Carlo simulations of $\gamma$D-WT and $\gamma$D-W42R at high concentrations. Clustered conformations were obtained for $\gamma$D-WT and the open conformation of $\gamma$D-W42R. The top 50 most populated clusters were used as input for mcMC simulations of $\gamma$D-WT and $\gamma$D-W42R at 200 g/L.

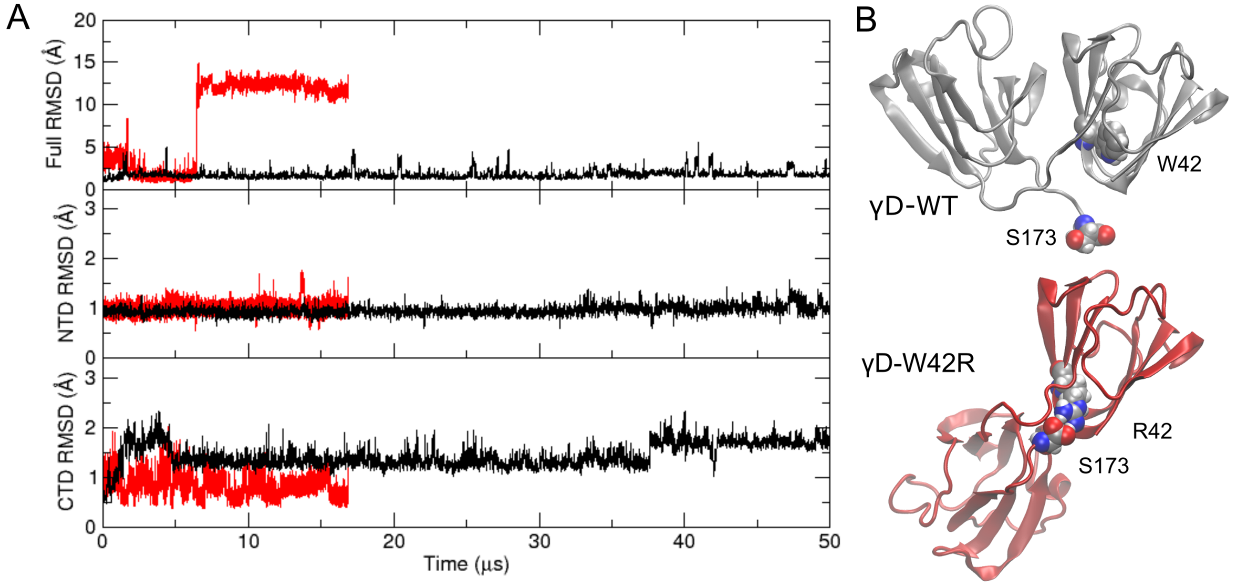

Figure 3.1: (A) Root mean square deviation (RMSD) of the backbone alpha carbons plotted vs. time. RMSD evolutions were reported after fitting and measuring the full backbone (full RMSD), the N-terminal domain (NTD RMSD), and the C-terminal domain (CTD RMSD). (B) Closed conformation of $\gamma$D-WT and the open conformation of $\gamma$D-W42R are shown in grey and red cartoon representation, respectively. Snapshots of $\gamma$D-WT and $\gamma$D-W42R were taken after 50 $\mu$s and 10 $\mu$s of MD simulation, respectively. The S173-R42 salt bridge and its analogous residues are shown in VDW representation.

### 3.2.2 Monomer mcMC simulation results

To analyze the many body interactions of $\gamma$D-crystallin, mcMC simulations of $\gamma$D-WT and $\gamma$D-W42R (named as monomer mcMC simulations) were performed and analyzed by Vera Prytkova. After performing Monte Carlo simulations of $\gamma$D-WT and $\gamma$D-W42R at 200 g/L for $2 \times 10^5$ MC steps, domain-based radial distribution functions were computed. The domain-based radial distribution functions (Figure 3.3) were considered instead of the center of mass based radial distribution function due to the shape of the protein. Since the $\gamma$-crystallins are composed of two domains – NTD and CTD, the entire protein has an elongated structure. Considering the radial distribution function of the centers of mass of each domain allows us to see which domain has the strongest preference for interaction.

According to domain-based radial distribution functions, the NTD of $\gamma$D-W42R has the

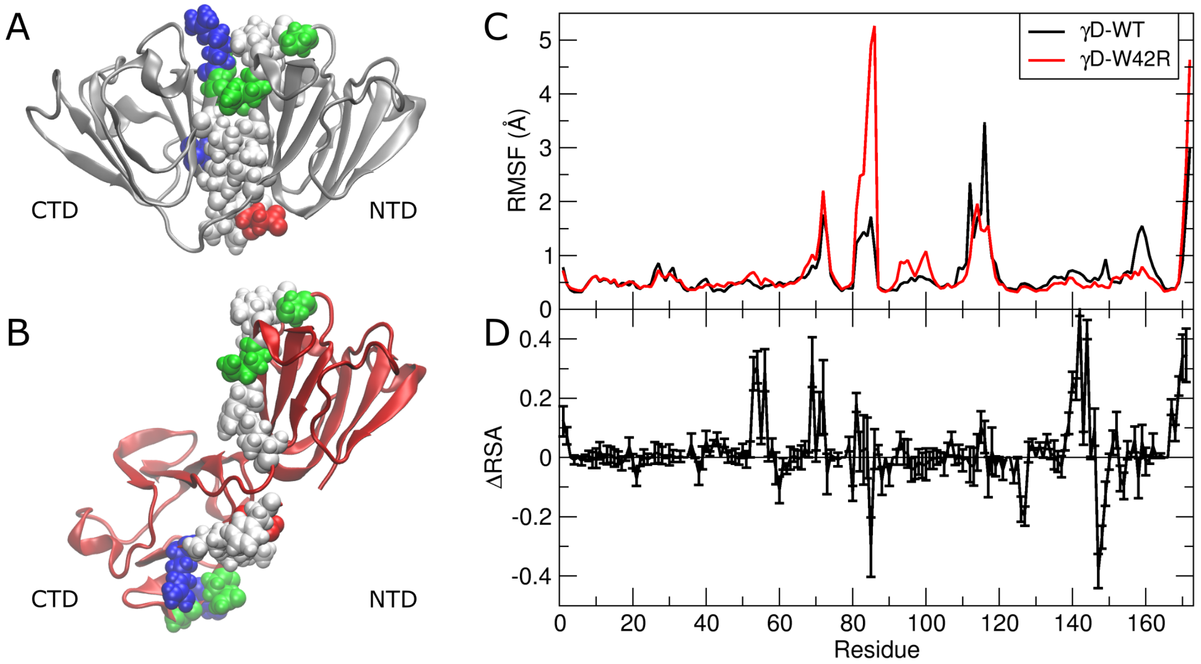

Figure 3.2: Representative conformations from highest populated clusters obtained from RMSD clustering of the MD trajectories of (A) $\gamma$D-WT and (B) $\gamma$D-W42R. Residues with a $> 0.13$ (1 standard deviation above the mean) increase in relative solvent exposure ($\Delta$RSA) from the W42R point mutation are represented in VDW spheres on both wild-type and mutant proteins. Residues are colored by residue type. Non-polar residues are colored white, polar residues are colored green, basic residues are colored blue, and acidic residues are colored red. (C) Backbone alpha carbon root mean square fluctuations are shown in black and red, respectively. Interdomain fluctuations removed by superimposing the individual domains before calculating the RMSF. (D) Difference in RSA between $\gamma$D-W42R and $\gamma$D-WT. Positive $\Delta$RSA values correspond to an increase in residue exposure for $\gamma$D-W42R. The error bars represent the standard error of the mean.

strongest preference for interaction with the NTD of other proteins comparing to that of the wild-type. NTD-CTD interactions are less strong, but also more prevalent in the W42R variant, while CTD-CTD interactions are very similar in $\gamma$D-W42R and $\gamma$D-WT.

To examine the origin of this preferential interaction in $\gamma$D-W42R all protein pairs contributing to the radial distribution function were selected and the contacts between them were analyzed. The protein pairs were selected based on the distance between the centers of mass of protein domains. Two residues are said to be in contact if any two heavy atoms are within 3.5 Ådistance of each other. Based on these criteria, the total number of contacts

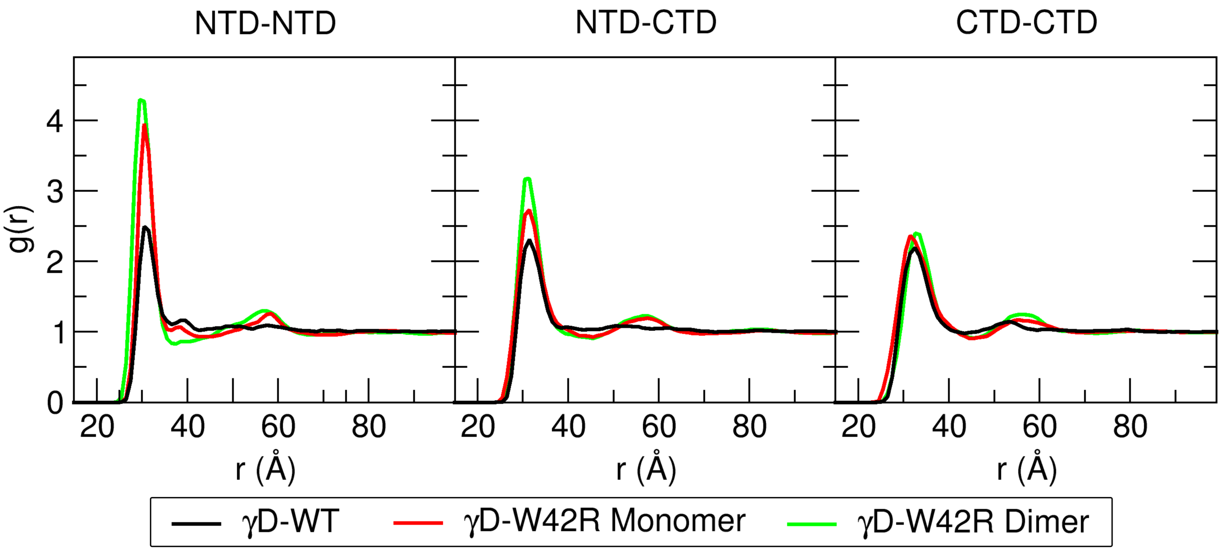

Figure 3.3: Radial distribution functions of domain centers of mass obtained from mcMC simulations of $\gamma$D-WT and $\gamma$D-W42R. $\gamma$D-W42R "Monomer" and "Dimer" corresponds to mcMC simulations using conformational libraries obtained from single protein and two protein MD simulations of $\gamma$D-W42R, respectively.

found in the $\gamma$D-WT simulation is 24,991, and for the $\gamma$D-W42R simulation – 47,887, which is almost twice as many as that for $\gamma$D-WT. Figure 3.4 shows the total number of contact found for each residue. $\gamma$D-W42R has slightly more CTD-CTD contacts than the wild type. Specifically, residue L144 is responsible for the increased CTD-CTD interactions. The NTD interacts more with both N- and C- terminal domains of other proteins through residues L53, M69, and L71. As can be observed in the contact analysis plot, these hydrophobic residues dont make a single specific contact, but instead contact many other residues. If we compare these residues to changes in solvent exposure of the fifty mcMC conformations (Figure 3.2D), we can see that these residues indeed belong to those patches of the protein that have significantly increased exposure in $\gamma$D-W42R relative to $\gamma$D-WT.

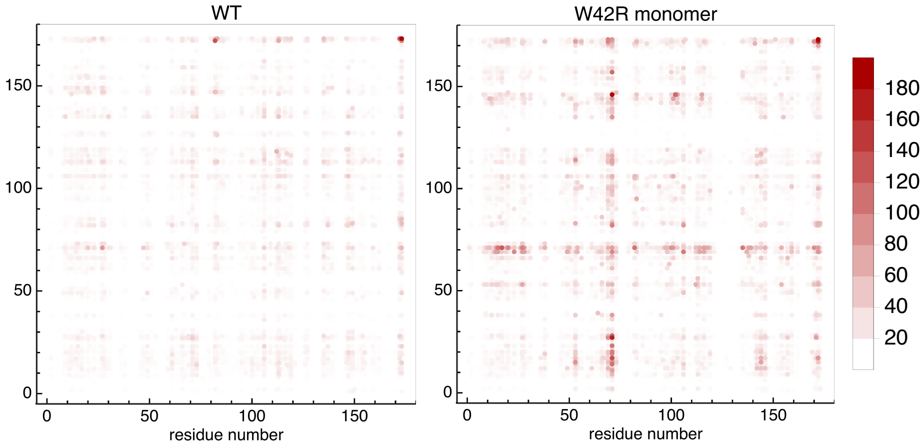

Figure 3.4: The total number of contacts of $\gamma$D and its W42R variant protein residues with other protein residues in mcMC simulations. No preferential contacts are found between wild-type proteins. The darkest points relate to contacts of the flexible end on C-terminal. However, W42R variant contacts multiple residues of other proteins with residues L53, M69, L71.

### 3.2.3 Protein-protein interaction results in additional domain separation

As the increased interaction of the W42R variant is apparent both from the radial distribution functions and contact analysis, an MD simulation of a two protein system of $\gamma$D-W42R was started using the final structures of the single protein simulations. This allows us to see whether there is any further structural change, as well as to expand the ensemble of input structures for mcMC simulations. An MD simulation of $\gamma$D-W42R was prepared with two copies of $\gamma$D-W42R and their solvation shells (Appendix Figure C.1). The two protein copies were placed such that no protein heavy atoms are within 12 Å of each other. Over the course of the 7 $\mu$s MD simulation, the two proteins diffuse together and interact such that the exposed hydrophobic interdomain residues (specifically, L53, F56, M69, and L71) from the NTD of both proteins form a new interprotein hydrophobic core. This new co-conformation forms after 2.4 $\mu$s and persists for the remainder of the simulation. Apart from L53, the four mentioned residues are the same ones with strong contacts observed in the monomer mcMC simulations. The contact matrix for the two protein $\gamma$D-W42R simulation (Figure C.2) shows that the interprotein contacts are rather sparse. However, these few contacts are sufficient to keep the two proteins associated for the duration of the simulations. When the two proteins interact, the CTD separates further from the NTD resulting in a further exposed interdomain interface for both proteins (Figure 3.6). This results in residue F56, located deeper in the interdomain interface, becoming further exposed to protein interaction (Figure 3.5).

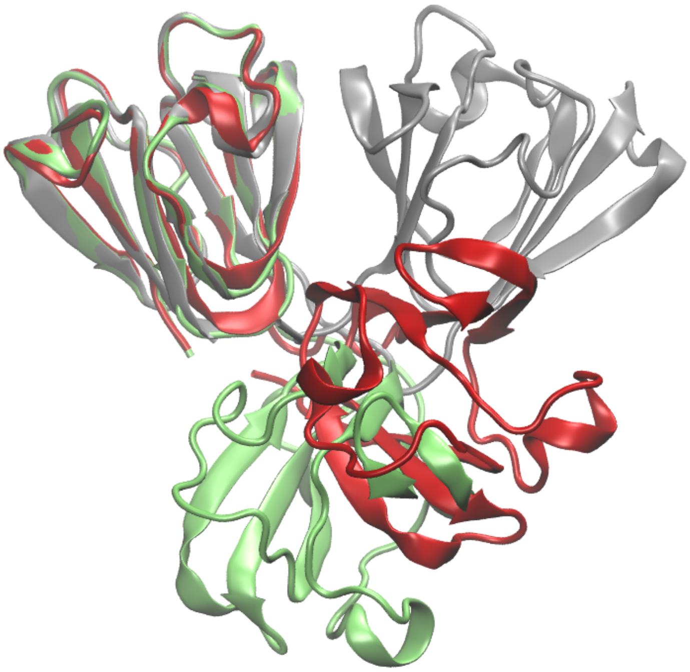

Figure 3.5: Snapshots of the highest probability structures used in mcMC simulations. In gray is the wild type structure, in red – the $\gamma$D-W42R at infinite dilution, and in green – $\gamma$D-W42R at 150 g/L. The relative angle of domain centers-of-mass show that the C-terminal domain is tilted by 45.2 and 65.6 degrees in monomer and dimer structures, respectively, relative to $\gamma$D-WT.

### 3.2.4 Dimer mcMC Simulation results

mcMC simulations using conformations from the two protein MD simulations were perfomed by Vera Prytkova. The two protein MD trajectory was clustered once again based on the sidechain RMSD to create a new ensemble of fifty conformations for mcMC simulations. We

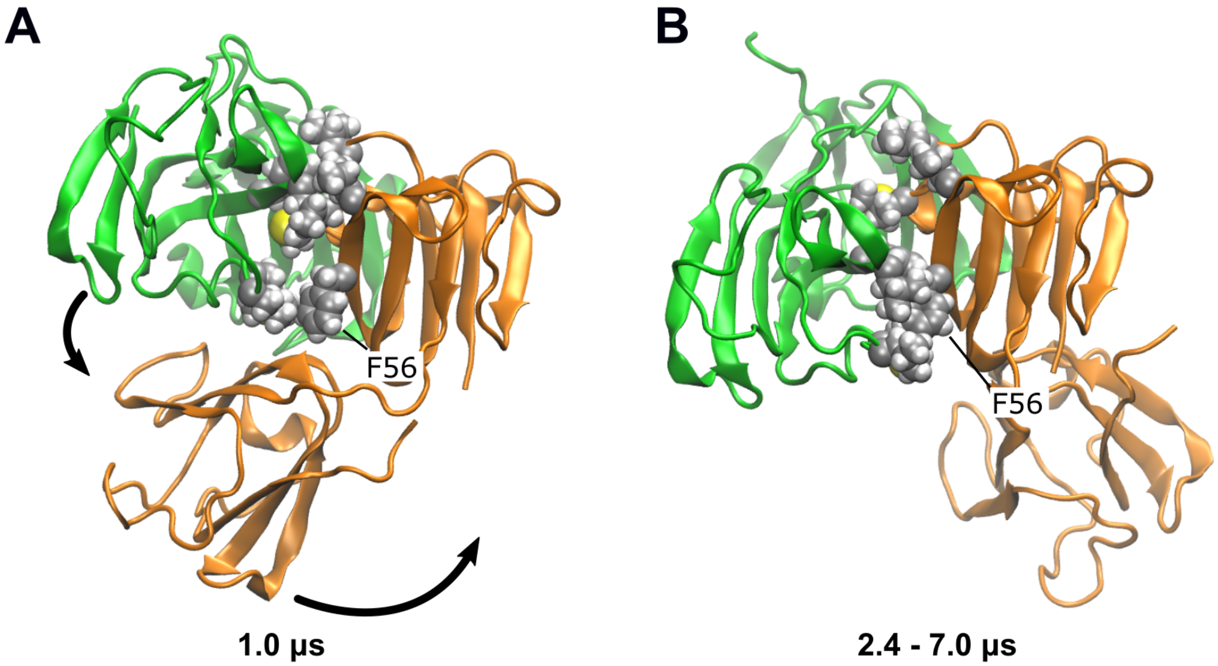

Figure 3.6: Snapshots of the two-protein MD simulation of $\gamma$D-W42R after (A) initial binding and (B) after binding-induced conformational change. Residues L53, F56, M69, and L71 (strong binding residues identified from mcMC simulations) are shown as VDW spheres in both proteins. After 1 $\mu$s, the four strong binding residues from both proteins come into contact. After 2.4 $\mu$s, the C-terminal domain of one protein is displaced, and the interprotein interface shifts towards F56. mcMC simulations using these new conformations show stronger contacts at F56.

further denote this ensemble as the dimer simulation conformational ensemble. In Figure 3.3, we consider the domain-based radial distribution function of the dimer ensemble of W42R variant. It can be observed that the N-terminal domain has an even higher propensity to interact in the dimer ensemble of the W42R than in the monomer ensemble.

The total number of protein pairs found in the contact analysis for the dimer ensemble is even higher – 67,415. The comparison of the contact analysis between two ensembles of the W42R variant  the one generated from the monomer MD simulation and from the dimer MD simulation – is presented on the plot in Figure 3.7. The main features of interaction are preserved, the same residues are participating in contact, however the NTD interacts even more in the dimer ensemble.

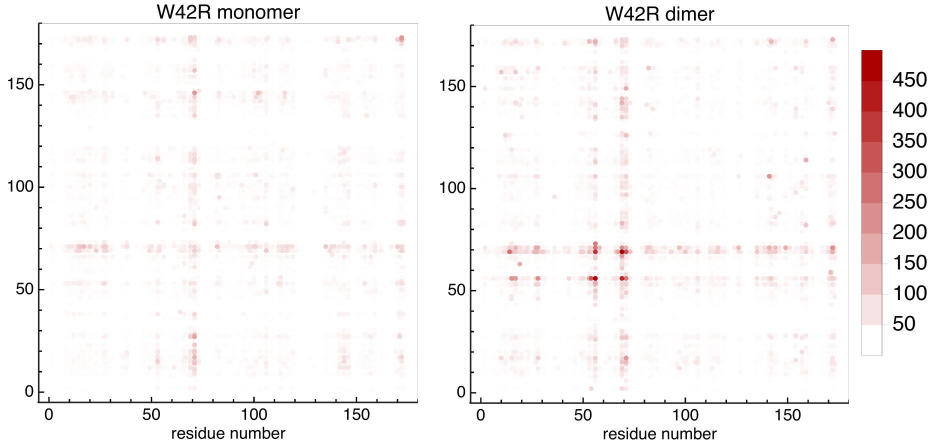

Figure 3.7: The total number of contacts for each residue of W42R variant from monomer and dimer conformations with other protein residues in mcMC simulations. Dimer conformation of W42R variant have a higher number of contacts than the monomer conformation for residues 53, 56, 69, 71.

### 3.2.5 The open conformation of W42R variant of human $\gamma$D-crystallins forms larger sized aggregates relative to wild-type

To investigate the size and morphologies of the $\gamma$D-crystallin aggregates, a network analysis of the aggregates was performed by J. Alfredo Freites. mcMC simulations indicate that increased exposure of the interdomain interface leads to a larger number of contacts between proteins in a concentrated system of W42R $\gamma$D-crystallins. To analyze how this increased propensity to make contacts leads to aggregate formation, two proteins are said to be in contact with one another if the distance between the domain centers of mass is within 31 Å. Such distance is chosen from to the position of the maximum of the domain-based radial distribution function. Protein aggregates chosen by the above criteria are now analyzed. In bottom plot of Figure 3.8, the probability density of cluster size distribution is shown for the $\gamma$D-WT simulations as well as monomer and dimer ensembles of $\gamma$D-W42R. The cluster size distributions of the monomer and dimer simulations of $\gamma$D-W42R show a significant increase in cluster size over $\gamma$D-WT. Though $\gamma$D-WT shows some propensity to form small clusters, the $\gamma$D-W42R clusters can be composed of more than half of the proteins present in the simulation (375 proteins total). When the domains are further separated (represented in the dimer simulation of $\gamma$D-W42R), the distribution of clusters becomes much broader, signifying a large proportion of higher order aggregates.

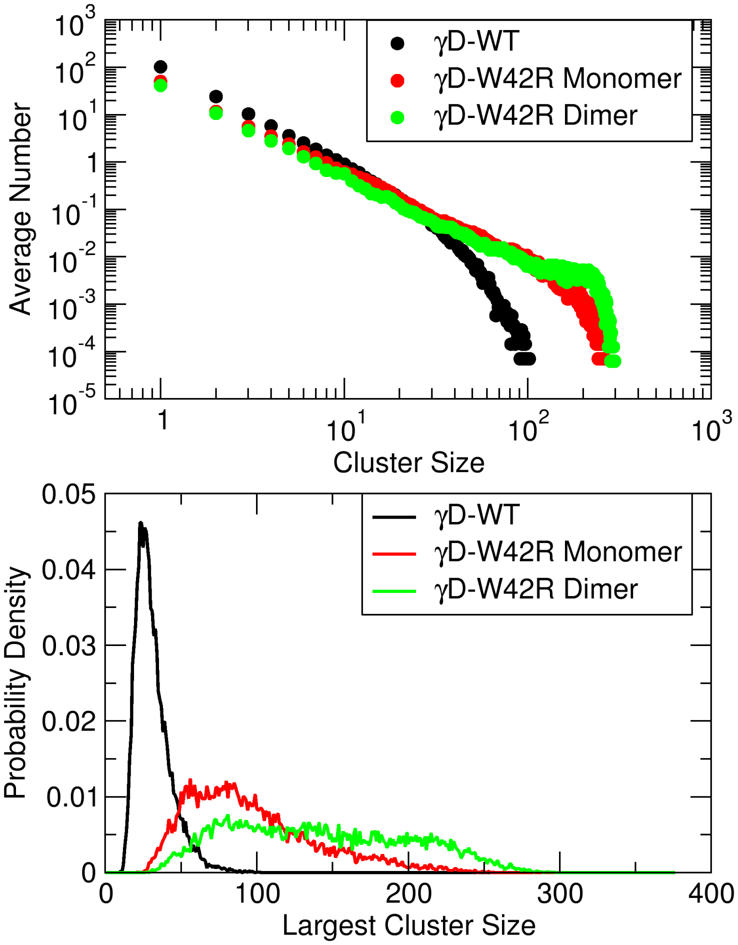

Figure 3.8: (Top) Distribution of the size of all clusters present in Monte-Carlo simulations of 375 proteins (200 g/L). (Bottom) Probability density of the largest cluster sizes from Monte-Carlo simulations of 375 proteins (200 g/L).

Visually inspecting the morphologies of the isolated clusters (Figure 3.9), there is an amorphous structure in both mutant and wild-type aggregates. However, the clusters of the mutant protein have a much larger apparent size, likely resulting from the increased propensity for interprotein interaction. Additionally, the newly exposed hydrophobic interfacial residues in $\gamma$D-W42R introduce a new interprotein interaction surface, allowing for a single

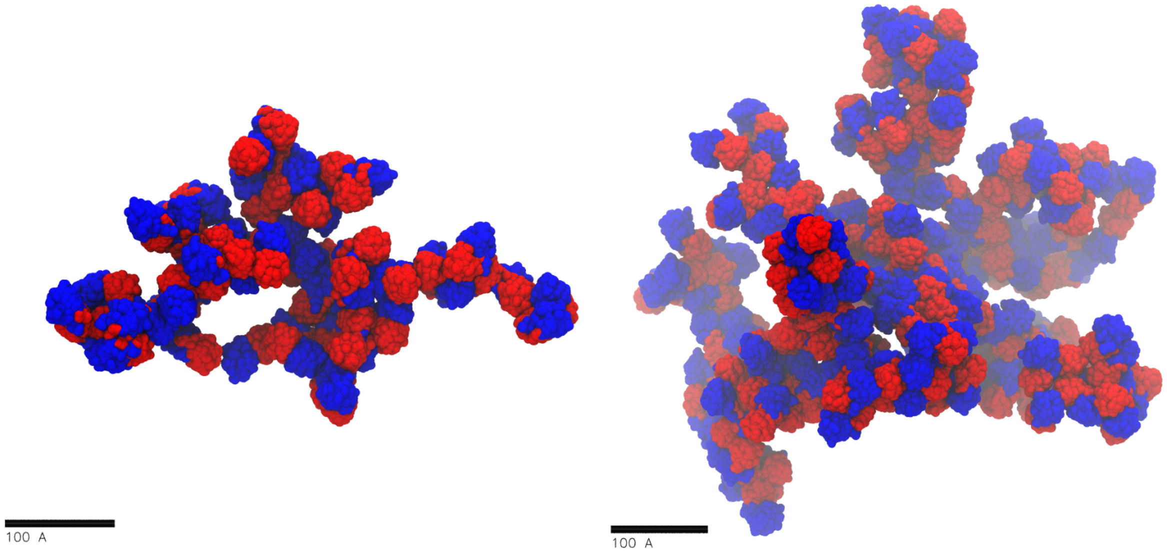

Figure 3.9: Isolated cluster conformations taken from the 95th percentile of the cluster size distribution. Clusters were formed in mcMC simulations using conformations of $\gamma$D-WT (left) and the $\gamma$D-W42R monomer (right) at 200 g/L. The N- and C- terminal domains of the proteins are colored in red and blue, respectively. Cluster sizes are 54 and 181 proteins for $\gamma$D-WT and $\gamma$D-W42R, respectively.

protein to interact with several more proteins. To investigate the statistics of the cluster-forming proteins, we analyize the portions of the aggregates where proteins form interactions with more than one other protein. By representing the aggregates as a network of domain-domain interactions, the $\gamma$D-W42R conformation is shown to be capable of a higher degree of interactions (Figure 3.10), particularly involving the N-terminal domain.

## 3.3 Discussion

We compare the conformational change of $\gamma$D-W42R in a long MD simulation to the wild type protein. The simulation of $\gamma$D-WT was conducted for 50 $\mu$s and no structural change was observed. However, after 6 $\mu$s of $\gamma$D-W42R simulation the protein undergoes structural change due to solvent exposure of residue R42 and formation of a salt bridge between this residue and C-terminal carboxyl group. This leads to separation of two domains while the

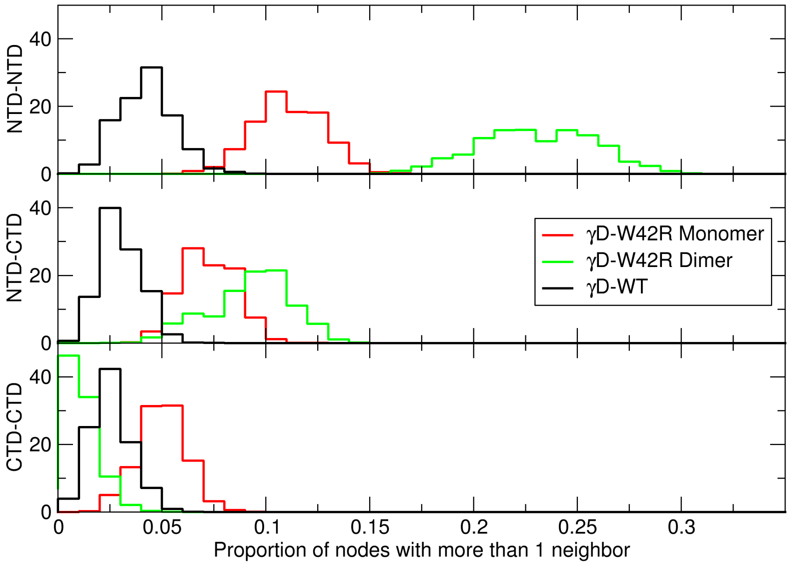

Figure 3.10: Histogram of the interprotein domain interactions involving more than one neighbor.

structure of separate domains stays intact. This structural change exposes several hydrophobic residues on the interdomain interface of both domains that were buried in the $\gamma$D-WT. No other significant conformational changes occur for the remaining 11 $\mu$s of simulation, suggesting that this conformation of the protein is stable in dilute protein concentrations. The two protein MD simulations of $\gamma$D-W42R reveals that the hydrophobic interdomain residues participate in protein-protein interaction, and two domains separate even further as a result of protein-protein interaction, fully exposing the interdomain interface.

Early experiments by Wang et al.[35] report an initial comparison of the hexahistidine-tagged $\gamma$D-WT and $\gamma$D-W42R. They report similar secondary structure between $\gamma$D-W42R and $\gamma$D-WT through far UV CD spectroscopy, indicating that much of the $\beta$-sheet content is maintained. However, ANS fluorescence experiments show a significant increase in hydrophobic exposure in $\gamma$D-W42R compared to $\gamma$D-WT. They suggest that this hydrophobic exposure is a result of a change in tertiary structure in W42R, consistent with the separation of intact Greek key domains observed in our MD simulations.

Ji et. al.[27] report shows the crystal structure of W42R variant has two domains tilted only by 9 degrees comparing to the wild type while structurally each domain remains intact. They find that wild type protein and W42R variant possess almost identical solvent accessible surface areas (8,894.32 $\AA^2$ and 8,546.23 $\AA^2$, respectively, as calculated by VMD). However, the unfolding curve of W42R variant under close to physiological conditions (37 °C and 7 pH) shows two-step unfolding, indicating that an unfolding intermediate exists and that W42R variant has a lower chemical stability than $\gamma$D-WT. A domain-separated conformation has been observed in annealing simulations of wild-type $\gamma$D-crystallin[22]. They proceed to observe domain swapping interactions with the CTD. It is worth noting, however, that the unfolding simulations involved simultaneous thermal unfolding and chemical denaturation with urea, resulting in an unfolded NTD. We observe a separation of intact domains in absence of thermal and chemical denaturation. Our MD simulations under physiological conditions, and therefore close to eye lens conditions, reveal that without unfolding the structure of separate domains, just by further tilting the angle between them, solvent accessible surface area increases from 8,991.08 $\AA^2$ in $\gamma$D-WT to 9,615.64 $\AA^2$ in $\gamma$D-W42R monomer conformation to 9,784.23 $\AA^2$ in W42R dimer conformation.

Serebryany et al.[74] also report that tryptophan fluorescence spectra, an evidence of conformational change in $\gamma$D, show no difference between $\gamma$D-WT and $\gamma$D-W42R. They observe that, in oxidizing conditions, the $\gamma$D-W42R spectra becomes red-shifted indicating the process of unfolding. The spectral change was observed over the course of 60 minutes, a timescale yet inaccessible to MD simulations. However, the presence of unfolded proteins was either observed or indicated to exist in conditions that were far from physiological – in the presence of oxidizing agents and at very low protein concentration. It is possible that if we had the time scale of minutes or hours, more conformational changes would take place. However, $\gamma$D-crystallin exists in the eye lens at concentrations exceeding 400 g/L, and even small structural changes associated with point mutations may lead to enhanced local interaction and aggregation while full protein unfolding may become impossible under crowded conditions.

High concentration (200 g/L) mcMC simulation of 375 proteins with conformations extracted from the monomer MD simulations show that $\gamma$D-W42R has a higher propensity to aggregate than the $\gamma$D-WT protein. Dimer ensemble of W42R is even more likely to aggregate than the monomer ensemble since further conformational change of the protein occurs due to protein-protein contact. The radial distribution function indicates that N-terminal domain of $\gamma$D-W42R interacts with both N-terminal and C-terminal domains of other proteins, while C-terminal to C-terminal interaction in $\gamma$D-W42R are very similar to those in wild type.

Hydrophobic residues, specifically residues L53, M69, and L71, located at the interdomain interface of the N-terminal domain, come in contact with many other residues of other proteins. Therefore, many orientations of proteins with respect to one another are possible during aggregation, as long as the residues mentioned above are participating in contacts. Those contacts are successful in creating clusters of W42R variant much larger than clusters of WT protein. The isolated aggregates are amorphous and stringy in shape since each protein may have only a few neighbors. Aggregates of similar shape were recently detected by Boatz et al.[13] in negative-stain TEM images for another $\gamma$D variant – P23T. In fact, the shape of the aggregates is sensitive to aggregation conditions – when the pH of solution is decreased to 3, the proteins form amyloid fibrils containing interprotein $\beta$-sheets. Solid state NMR spectroscopy did not detect any structural change of the Greek key domains in the P23T variant at neutral pH. This indicates that at high concentrations of $\gamma$D-crystallin, small changes of protein surface charge of structure disrupt the careful balance of protein solubility and aggregates form.

## 3.4 Methods

### 3.4.1 Single Protein Molecular Dynamics Simulation System Preparation and Equilibration

The initial protein coordinates of $\gamma$D-WT and $\gamma$D-W42R were built from the crystal structures deposited into the Protein Data Bank (PDB ID code 1HK0 for $\gamma$D-WT and 4GR7 for $\gamma$D-W42R)[26, 27]. Histidine protonation states were set to be the same as those published in the solution state NMR structure of the P23T variant of $\gamma$D-crystallin (PDB ID code 2KFB)[76]. Protein atoms were parameterized with the CHARMM36 force field[31]. The crystal structure waters were kept and the proteins were solvated in a cubic TIP3P[77] water box measuring 80 Åon a side. The system was neutralized with chloride counterions. The single protein systems contained 48,309 atoms and 48,367 atoms for $\gamma$D-WT and $\gamma$D-W42R, respectively. All system preparation was performed using the VMD 1.9.1 software package[78].

A 20 ns pre-production simulation equilibration was performed with NAMD 2.9[79]. The prepared systems were minimized for 10,000 steps in the NPT ensemble at 310 K and 1 atm. Protein heavy atoms were restrained with harmonic positional restraints and were gradually relaxed over 200 ps. NAMD was parameterized with the smooth particle mesh Ewald method[80, 81] for long-range electrostatic interactions, a real space interaction cutoff at 11 Å, and an integration time step of 2 fs/timestep. The RESPA algorithm[82] was used with a timestep of 4 fs for electrostatic forces, 2 fs for nonbonded forces, and 1 fs for bonded forces. Hydrogen covalent bonds were held fixed using the SHAKE[83] and SETTLE[84] algorithms. Constant temperature and pressure was maintained using a Langevin thermostat and a Nos-Hoover-Langevin piston[85, 86].

### 3.4.2 Microsecond Time Scale Molecular Dynamics Simulations

Production simulations were performed on the Anton 2 supercomputer, a special-purpose computer for molecular dynamics simulations of biomolecules[34]. Protein and solvent atoms were parameterized with the CHARMM36[31] and TIP3P[77] forcefields, respectively. The multigrator scheme[87] was used to integrate Newtons equation of motion at 2.5 fs/timestep. Using the RESPA algorithm (Grubmuller et al., 1991), long-range nonbonded, short-range nonbonded, and bonded forces were calculated at a timestep of 7.5 fs, 2.5 fs, and 2.5 fs, respectively. Long-range electrostatic forces were calculated using the k-Gaussian split Ewald method[88]. Hydrogen covalent bonds were held fixed using the SHAKE[83] algorithm. Constant temperature and pressure was maintained using Nose-Hoover chains[89] and the Martyna-Tobias-Klein barostat[86], respectively. Single protein simulations of $\gamma$D-WT and $\gamma$D-W42R were run for a total of 50 $\mu$s and 17 $\mu$s of production simulation, respectively.

### 3.4.3 W42R human $\gamma$D-crystallin Dimer System Preparation and Simulation

To prepare the two protein MD simulation of $\gamma$D-W42R, the protein and its solvation shell (waters within 6 Åof the protein) were extracted from the last frame of the 17 $\mu$s single protein MD simulation. Two copies of the proteins were placed such that the proteins are separated by at least 9 Åand the solvating waters do not overlap (Figure C.1). The cubic water box is parameterized, solvated, neutralized, and equilibrated as previously described. The resulting system contains a total of 46,655 atoms contained in an 80 Å× 80 Å× 80 Åperiodic cell, the same dimensions as the single protein simulation. The two-protein simulation of $\gamma$D-W42R was run for a total of 7 $\mu$s on the Anton 2 supercomputer using the same parameters as the single protein simulation.

### 3.4.4 mcMC simulation protein-protein interaction potential

Our mcMC simulation employ protein-protein interaction potential developed by Mereghetti et al.[90] for Brownian dynamics simulations in the SDAMM software package[91]. Two proteins interact through the following potential function:

$$\begin{aligned}\Delta U = \frac{1}{2}\sum_{i_2}\Phi_{el_1}(r_{i_2})\cdot q_{i_2} + \frac{1}{2}\sum_{j_1}\Phi_{el_2}(r_{j_1})\cdot q_{j_1} \ + \sum_{i_2}\Phi_{el_1}(r_{i_2})\cdot q_{i_2}^2 \\ + \sum_{j_1}\Phi_{el_1}(r_{j_1})\cdot q_{j_1}^2 \ + \sum_{m_2}\Phi_{ND_1}(r_{m_2})\cdot SASA_{m_2} + \sum_{n_1}\Phi_{ND_2}(r_{n_1})\cdot SASA_{n_1} \\ + \sum_{m_2}E_{softcore_1}(r_{m_2}) + \sum_{n_1}E_{softcore_2}(r_{n_1})\end{aligned} \tag{3.1}$$

The first two terms denote the interaction of electrostatic potential of one of the proteins with the charges of another protein[92]. The charges are computed through the effective charge approximation implemented in SDAMM software package. The second two terms refer to electrostatic desolvation penalty that appears due to location of solvated polar groups of one protein in proximity of the low dielectric environment of another protein and consequential simultaneous loss of solvation shell[93]. Terms five and six correspond to an attractive short-range non-polar desolvation interaction between two proteins that appears when solvent exposed hydrophobic atoms of one protein are buried by another protein. This interaction can be scaled by modifying a prefactor $\beta$ used to convert the buried area of the protein surface into a desolvation energy. The value used in our simulation is -9 cal $mol^{-1}\mathring{A}^{-2}$. Seventh and eighth terms denote the softcore repulsive interaction energy terms. The interaction potential terms were computed prior to simulations on 200 Åx 200 Å× 200 Ågrids with the grid spacing of 1 Å. The electrostatic potential grids were computed at 50 mM ionic strength according to OPLS force field[94] by finite difference solution of the linearized Poisson-Boltzmann equation using the UHBD[95] software package.

### 3.4.5 Multiple conformation Monte Carlo simulations

A multiple conformation Monte Carlo (mcMC) algorithm[96] employs translational and rotational moves on randomly selected proteins in combination with a conformational swap from a finite size library of structures. For rotational and translational moves a basic Metropolis scheme[97] is used and the size of moves is adjusted to provide a 50% acceptance ratio. The appearance of each conformation in the simulation is proportional to its probability of appearance in MD simulation from which it was extracted. The entire MD simulation is used for clustering based on sidechain orientation. Top fifty structures are selected for the mcMC conformational swap library. Each trial move is accepted according to the Metropolis criterion with acceptance probability:

$$P_acc = min(1, exp[\frac{-\Delta E}{k_B T}]) \tag{3.2}$$

Where ΔE is the difference between the energy of the system before and after the trial move, kB is the Boltzmann constant and T is the temperature. All simulations were performed with 375 proteins at 200 mg/mL protein concentration, which is approximately half the density of the eye lens. A total of $2\times10^5$ MC cycles at 310 K are performed for each protein type.

# Chapter 4

# Thermal stability of the Antarctic toothfish $\gamma$S-crystallins

## 4.1 Introduction

Protein function is closely connected to both the structure and dynamics of the three dimensional fold. Therefore, when a protein adapts to environmental stresses, dynamic properties must be preserved in order to maintain function[98]. In one example, families and superfamilies of enzymes have been shown to conserve their dominant normal modes of vibration[99] and backbone flexibility profiles[100]. For cold adaptation, increased flexibility is often linked to regions near the enzymatic active site. Several enzymes of different functions have independently adopted this strategy for cold adaptation, emphasizing the pervasiveness of this strategy[101]. Developing a better understanding of thermal adaptation can lead to improved strategies for the rational design of biocatalysts as well as a better understanding of the structural stability of proteins.

Psychrophiles, organisms that have adapted to low temperatures, have provided examples of cold stable enzymes capable of low temperature activity[102, 103] and high thermolability[104]. The current hypothesis behind thermal adaptation is the ”activity-stability-flexibility” relationship[105, 106], where increased flexibility at vital regions of an enzyme counteract the ”freezing” effects of low temperatures, thus promoting activity. Protein mutagenesis experiments[107, 108, 109] have provided support for the developing hypothesis that a cold-adaptation can be achieved by strategically removing the thermostabilizing interactions found in the thermophilic counterparts. However, reported cases[110] have shown that rigidity, or a loss thereof, is not sufficient to describe changes in thermal stability. Therefore, close attention should be paid to the individual factors affecting a protein's flexibility.

At the microscopic level, structural rigidity in thermostable enzymes can come from several factors, including: salt bridge interactions, disulfide bond formation, proline content, and packing of the hydrophobic core. In most cases, the predominant factor affecting thermal stability is the surface charge content of the protein[111, 112, 113]. Both crystallographic data and MD simulations provide evidence that thermal stability is maintained through clusters of salt bridging residues, typically at domain-domain or protein-protein interfaces[114, 115]. A secondary, but yet still important, factor is the packing of the hydrophobic core. Mutations as small as reducing the hydrophobic side chain length of a single residue can result in a loss of thermodynamic stability[116, 117]. Most of these studies have been performed on enzymes, where mutations focus on well defined active site. In this study, we investigate the thermal stability of a structural eye lens protein, a protein with no known catalytic activity.

In the vertebrate eye, there are three main classifications of eye lens structural proteins: $\alpha$-, $\beta$-, and $\gamma$-crystallin. $\alpha$-crystallins are large multimeric chaperone complexes that bind other misfolded crystallins, preventing further aggregation[9, 10]. The $\beta\gamma$-crystallins are structural proteins responsible for maintaining eye lens clarity; where the $\beta$-crystallins exist as dimers and the $\gamma$-crystallins exist as monomers. Their structure is characterized by two structurally homologous, yet non-identical, domains composed of two Greek key motifs in

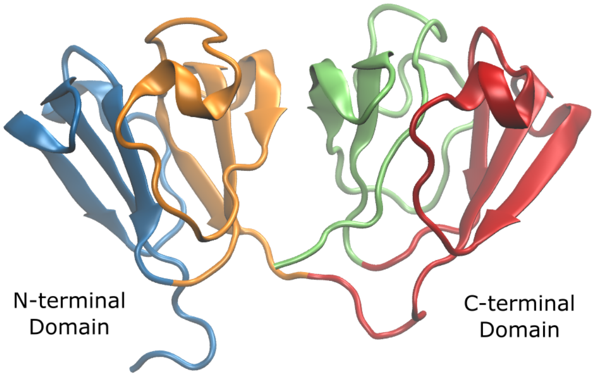

Figure 4.1: Cartoon representation of the NMR structure of human $\gamma$S-crystallin[10]. Each Greek key motif is colored separately, and the protein is oriented such that the N-terminal domain is on the left and the C-terminal domain is on the right.

each domain (Figure 4.1). These proteins are highly stable, capable of remaining soluble at high physiological concentrations (300-1000 g/L)[118, 12]. In the event of a loss in solubility, opaque aggregates, known as cataract, begin to form. In the case of near freezing temperatures, the mammalian and tropical fish $\gamma$-crystallins form light diffracting liquid-liquid phase separations composed of a protein-rich and protein-poor phase[11, 119], known as ”cold cataract”.

The Antarctic toothfish (*Dissostichus mawsoni*) lives in waters as cold as -2 °C, and have developed a resistance to cold cataract formation[120] and the loss of eye lens function associated with its formation. The two toothfish paralogs, $\gamma$S1- and $\gamma$S2-crystallin (T$\gamma$S1 and $\gamma$S2), are structurally homologous to human $\gamma$S-crystallin (H$\gamma$S). Yet, T$\gamma$S1 and T$\gamma$S2 have sequences identities of 57% and 53% with respect to H$\gamma$S. Biophysical characterization of the toothfish $\gamma$S-crystallins shows that T$\gamma$S1 is more susceptible to chemical denaturation, however T$\gamma$S2 is more susceptible to thermal unfolding. This presents an interesting distinction since structural stability is often measured interchangeably with thermal and chemical unfolding. This distinction poses the toothfish $\gamma$S-crystallins as models for identifying structural characteristics that contribute uniquely to thermal stability.

In this work, we model the structure and dynamics of H$\gamma$S, T$\gamma$S1, and T$\gamma$S2, each having varying levels of thermal stability. At room temperature, the less thermally stable proteins have a loss in cohesion of interactions at 3 regions in the C-terminal domain. These same regions are sensitive to thermal fluctuations as indicated by an increased backbone RMSF that correlate with decreasing thermal stability. In the mesophilic H$\gamma$S, these fluctuations are reduced by a set of salt bridging interactions that bridge between the two Greek key domains contained within the C-terminal domain. In the toothfish $\gamma$S-crystallins, substitution of these salt bridge interactions with weaker hydrogen bonds result in the aforementioned increased flexibility, yet at low temperatures a similar structure and flexibility is recovered. The similarity of protein structure and flexibility of the $\gamma$S-crystallins at their operating temperatures highlights the necessity of conserved dynamics for structural function at low temperatures. Additionally, we incorporate a k-core network analysis[121, 122] to provide a quantitative measure of the hydrophobic packing in the protein. We identify losses in cohesion of the hydrophobic contacts that correlate with the protein thermal stability. This network analysis allows a new method to potentially quantify hydrophobic packing and context of protein thermal stability.

## 4.2 Methods

### 4.2.1 Molecular Dynamics Simulations

The initial coordinates of H$\gamma$S were built from the lowest energy solution-state NMR conformation deposited in the Protein Data Bank (PDB ID code 2M3T)[10]. Protein coordinates for T$\gamma$S1 and T$\gamma$S2 were predicted using homology modeling with SwissModel by Kingsley et al[123] using the NMR solution-state structure of H$\gamma$S[10] as a template. The initial structures of each protein was superimposed and then solvated in a TIP3P[77] water box such that each protein atom is at least 15 Å from the edge of the periodic cell. Each system was then neutralized with chloride counter-ions. Each simulation had an approximate periodic cell size of 64 Å × 80 Å × 80 Å. Resulting in a total atom count of 2867, 2814, and 2727 atoms for H$\gamma$S, T$\gamma$S1, and T$\gamma$S2, respectively. Protein and ion atoms were parameterized using the CHARMM36 force field[31]. The three systems of H$\gamma$S, T$\gamma$S1, and $\gamma$S2 were duplicated for constant temperature simulations at 277 K and 300 K. All system preparation was performed using the VMD 1.9.1 software package[78].

Before production simulation, each system was equilibrated in the NPT ensemble using 20,000 steps of minimization. Harmonic positional restraints were placed on each protein heavy atom and were gradually relaxed over the course of 1 ns of simulation at 1 fs/timestep. Once the harmonic positional restraints were removed, the integration timestep was changed to 2 fs/timestep for production simulation. All MD simulations were performed using the NAMD 2.9[79] software package. Long-range electrostatic interactions were calculated using a smooth particle mesh Ewald method[80, 81], and a cutoff of 11 Å was used for short-range, real-space interactions. The RESPA algorithm[82] was used to integrate at multiple timesteps of 4fs for electrostatic forces, 2fs for nonbonded forces, and 1fs for bonded forces. The SHAKE[83] and SETTLE[84] were used to fix hydrogen covalent bonds. Constant temperature and pressure were maintained using a Langevin thermostat and a Nosè-Hoover-Langevin piston[85, 86]. Each system was run for approximately 1.5 $\mu$s and the last 400 ns was extracted for analysis.

### 4.2.2 Chemical group graph representation

MD conformations were reduced into a chemical group graph representation using a chemical group scheme developed by Benson et al.[124]. In this scheme, each of the 20 amino acids are reduced into sets of moieties and classified as a polar, non-polar, positive, or negative node. For example, an arginine would be composed of one positive node for the guanidine group, two non-polar nodes for the 4 carbon side chain, and a dipolar node for the polar backbone carboxamide. A complete description of the chemical groups for each amino acid is detailed by Benson et al.[124]. In the graph space, an edge connects two nodes if at least one atom-atom contact between the two nodes. A contact is defined by an interaction cutoff of 4.6 Å or 5.6 Å if both of the contacting atoms are carbons. The resulting nodes/edges constitute a graph that represents the network of non-covalent interactions within the protein. Atomistic protein conformations were extracted every 200 ps was extracted from the 400 ns MD trajectory and converted to the chemical group graph representation. The k-core of each node was calculated using the ”sna” package for social network analysis[122, 121].

### 4.2.3 Protein sequence analysis

A library of homologous $\gamma$-crystallin sequences was generated using BLASTP[125]. The sequences of H$\gamma$S, T$\gamma$S1, and T$\gamma$S2 were used as query sequences, and the identified homologous sequences were grouped in a single library. Partial, synthetic, hypothetical, and predicted sequences were filtered from the set. Additionally, all non-$\gamma$-crystallin sequences were removed from the set (this included homologous $\alpha/\beta$-crystallins and absent in melanoma 1 (AIM1)). Entries with duplicate NCBI accession numbers were removed, and the remaining set was re-aligned using CLUSTALW[126].

Table 4.1: Summary of charge content in Human and Toothfish $\gamma$S-crystallins

|  | # Negative Residues | # Positive Residues | Net Charge | Melting Temperature[123] |
|---|---|---|---|---|
| H$\gamma$S | 24 | 23 | -1.0 | 72.0 °C |
| T$\gamma$S1 | 24 | 20 | -4.0 | 68.5 °C |
| T$\gamma$S2 | 23 | 17 | -6.0 | 58.0 °C |

## 4.3 Results

### 4.3.1 Toothfish $\gamma$S-crystallins have reduced basic residue content relative to human $\gamma$S-crystallin

A notable distinction between the human and toothfish $\gamma$S-crystallin sequences is their difference in charged residue content. Between H$\gamma$S, T$\gamma$S1, and T$\gamma$S2, each protein contains 23, 20, and 17 positively charged residues, while maintaining similar amounts of negatively charged residues (Table 4.1). This results in a decrease in net charge for the toothfish $\gamma$S-crystallins and a reduction in potential salt-bridging interactions. The aligned sequences of the three proteins (Figure 4.2) show that most of the basic residues lost in the toothfish crystallins are located in the C-terminal domain (residues 94-178), affecting a set of clustered salt bridges that bridge between the two Greek key motifs in the C-terminal domain. In the toothfish $\gamma$S-crystallins, the salt bridges remain much more sparse due to the lost basic residues (Figure 4.2).

### 4.3.2 Protein backbone flexibility correlate with the thermal unfolding experiments

The thermal unfolding temperatures of H$\gamma$S, T$\gamma$S1, and T$\gamma$S2 (Table 4.1) were reported from circular dichroism experiments[123]. T$\gamma$S1 has a $T_m$ of 68.5 °C, while T$\gamma$S2 has a

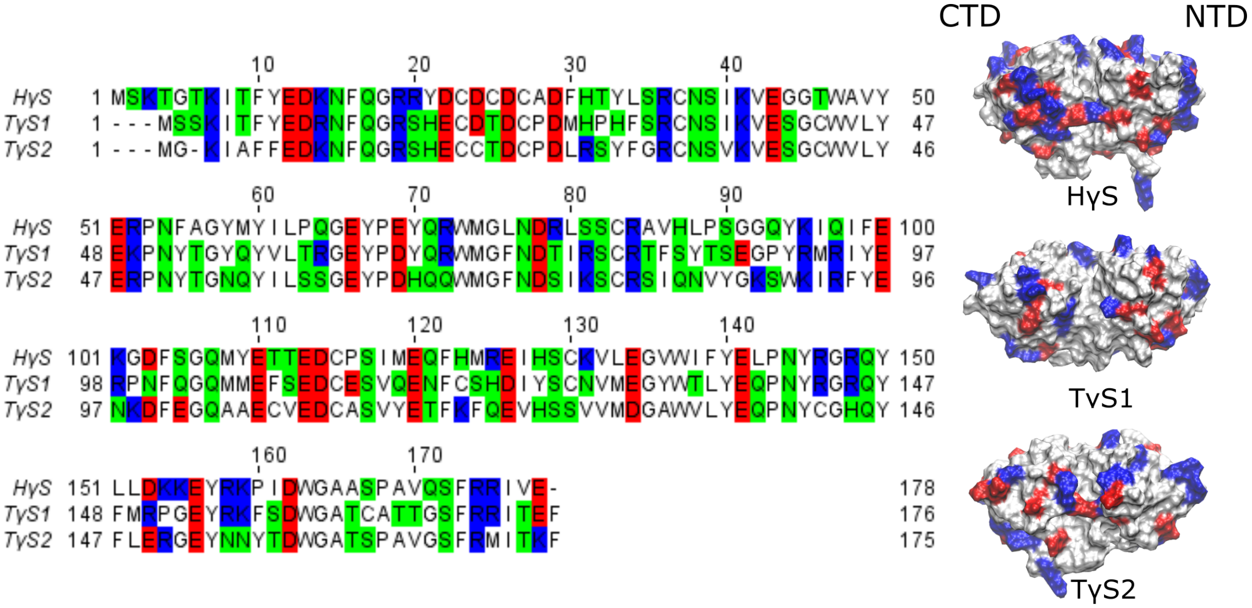
CTD
NTD
10
20
30
40
HγS 1 MSKTGTKITFYEDKNFQGRRYDCDCDCADFHTYLSRCNSIKVEGGTWAVY 50
TγS1 1 ---MSSKITFYEDRNFQGRSHECDTDCPDMHPHFSRCNSIKVESGCWVLY 47
TγS2 1 ---MG-KIAFFEDKNFQGRSHECCTDCPDLRSYFGRCNSVKVESGCWVLY 46
60
70
80
90
HγS 51 ERPNFAGYMYILPQGEYPEYQRWMGLNDRLSSCRAVHLPSGGQYKIQIFE 100
TγS1 48 EKPNYTGYQYVLTRGEYPDYQRWMGFNDTIRSCRTFSYTSEGPYRMRIYE 97
TγS2 47 ERPNYTGNQYILSSGEYPDHQQWMGFNDSIKSCRSIQNVYGKSWKIRFYE 96
110
120
130
140
HγS 101 KGDFSGQMYETTEDCPSIMEQFHMREIHSCKVLEGVWIFYELPNYRGRQY 150
TγS1 98 RPNFQGQMMEFSEDCESVQENFCSHDIYSCNVMEGYWTLYEQPNYRGRQY 147
TγS2 97 NKDFEGQAAECVEDCASVYETFKFQEVHSSVVMDGAWVLYEQPNYCGHQY 146
160
170
HγS 151 LLDKKEYRKPIDWGAASPAVQSFRRIVE- 178
TγS1 148 FMRPGEYRKFSDWGATCATTGSFRRITEF 176
TγS2 147 FLERGEYNNYTDWGATSPAVGSFRMITKF 175
HγS
TvS1
TγS2

Figure 4.2: (Left) Aligned sequences of human and toothfish $\gamma$S-crystallin. Amino acids are colored by residue type (Red: Acidic, Blue: Basic, Green: Polar, White: Hydrophobic). (Right) Molecular surface representation of H$\gamma$S, T$\gamma$S1, and T$\gamma$S2 taken from snapshots after 1.5 $\mu$s of MD simulation. Proteins are oriented such that the C-terminal domain (CTD) is positioned on the left and the N-terminal domain (NTD) is positioned on the right. Positive, negative, and neutral charged residues are colored blue, red, and white, respectively.

T$_m$ of 58.0 °C, and the mesostable H$\gamma$S has the highest unfolding temperature of 72.0 °C. These unfolding temperatures all correlate with the number of basic residues. To model the structure and dynamics of these proteins in response to thermal fluctuations, MD simulations were run at 300 K and 277 K, the temperatures at which human and toothfish crystallin optimally exist. MD simulations were run for at least 1.5 $\mu$s, to ensure any changes to the toothfish homology model structures has been equilibrated.

The internal fluctuations within the Greek key domains were measured using the backbone alpha carbon root mean squared fluctuations (RMSF). Each individual domain was fit before calculating the average fluctuations. This effectively measures the intradomain fluctuations of the N- and C-terminal domain by fitting each domain before calculating the fluctuations. At room temperature, T$\gamma$S1 and T$\gamma$S2 show much larger fluctuations in C-terminal domain that correlate in intensity with the melting temperatures of the proteins (Figure 4.3A). These

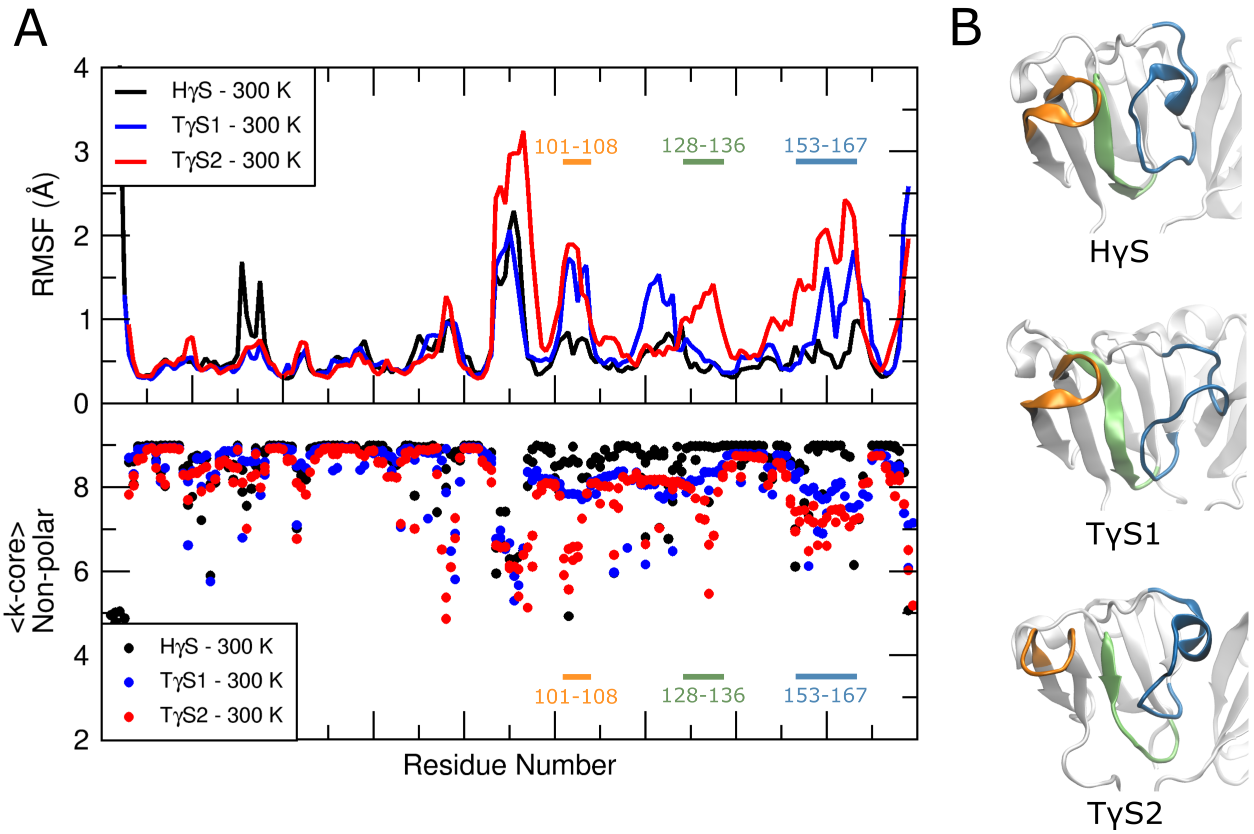

Figure 4.3: (A) Changes in protein dynamics and structure. The alpha carbon backbone RMSF is plotted at 300 K (top) and with toothfish at 277 K and human at 300 K (middle). The time-averaged k-core number of each non-polar node in the chemical group graph representation is plotted vs. residue number (bottom). Three regions with a significant response to heat at 300 K are highlighted with colored bars in orange, green and blue and labeled with their corresponding residue ranges. (B) MD snapshots of human and toothfish $\gamma$S-crystallin C-terminal domain. The backbone is colored corresponding to the three colored regions highlighted in the backbone RMSF and k-core plots.

fluctuations are mainly composed of the separation of a $\beta$-hairpin and a loop containing an $\alpha$-helix in the C-terminal domain (highlighted in Figure 4.3B). At 300 K, the loop separates from the $\beta$-hairpin in the toothfish proteins, whereas the fold remains intact in H$\gamma$S. Excluding the N-terminal strand and the interdomain linker, the average residue RMSF was calculated and summarized in Table 4.2. At 300 K, average RMSF increases with the loss in thermal stability. However, at low temperatures, the toothfish average RMSF recovers an average RMSF value much more similar to that of H$\gamma$S at 300 K.

Table 4.2: Average root mean square fluctuations of domain residues

| | 277 K | 300 K |
|---|---|---|
| H$\gamma$S | 0.5318 Å | 0.5417 Å |
| T$\gamma$S1 | 0.5395 Å | 0.6507 Å |
| T$\gamma$S2 | 0.4466 Å | 0.7695 Å |

### 4.3.3 Less thermally stable proteins show a loss in interactions at the hydrophobic core

In order to track the changes in interactions between human and toothfish proteins, the protein conformations from the MD simulations were reduced into networks of interactions between chemical groups. Using a chemical group graph representation[124], the atomic protein structure is parsed into moieties based on charge and the interactions between these chemical groups are calculated by a distance cutoff. The result is a graph representation of intra-protein interactions where chemical groups are represented as nodes and the interactions are represented as edges. A more in depth description of the network representation is described in the methods.

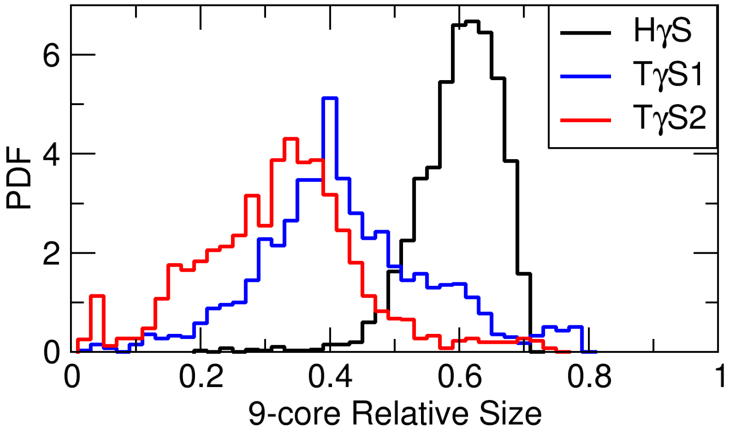

Figure 4.4: Histogram plotting the probability density function (PDF) of the relative 9-core size of H$\gamma$S, T$\gamma$S1, and T$\gamma$S2 at 300 K. H$\gamma$S, T$\gamma$S1, and T$\gamma$S2 are colored in black, blue, and red, respectively.

The network k-cores were calculated to measure the connectivity of the protein hydrophobic core. The k$^{th}$ core represents the sub-graph that remains when all nodes with a degree, or number of connections, less than k is removed. The MD simulations of all three proteins at 300 K have a deepest core of k=9. In H$\gamma$S, the 9-core composes 61.1% of the entire network, reflecting the well-connected internal structure of the protein. In the less thermally stable T$\gamma$S1 and T$\gamma$S2, the relative 9-core sizes drop to 43.3% and 33.4%, respectively (Figure 4.4). The total network size of H$\gamma$S, T$\gamma$S1, and T$\gamma$S2 is 572, 572, and 553 nodes, respectively. Much of the loss in k-core content is located at residues 101-108, 128-136, and 153-167 (Figure 4.3B), the same regions where increases in backbone flexibility is calculated. Combined, these regions have a decrease in average k-core and increase in backbone RMSF that correlate with the thermal stability of the protein(Figure 4.3).

### 4.3.4 Salt-bridge to hydrogen bond substitution results in increased flexibility at higher temperatures

As previously mentioned, increased fluctuations and loss in interaction connectivity were measured at residues 101-108, 128-136, and 153-167 (in the H$\gamma$S sequence). These regions correspond to an outer $\beta$-hairpin, a buried $\beta$-strand, and a unstructured loop located near the interdomain interface. In H$\gamma$S, these three regions are held by a cluster of salt bridges that prevent larger backbone fluctuations at higher temperatures. In the T$\gamma$S1 and T$\gamma$S2, two of the salt bridges are lost by substitution of the basic residues. The pair of residues is a D103-K149 salt bridge that link the disordered loop to the $\beta$-hairpin, and the second pair is a K131-E156 salt bridge that links the disordered loop to the buried $\beta$-strand (Figure 4.5).

For both toothfish $\gamma$S-crystallins, at least one of the charged residues is replaced with a polar asparagine, effectively substituting a salt bridge for a hydrogen-bonding interaction. In the case of T$\gamma$S2, K131 in H$\gamma$S is additionally replaced with a hydrophobic V126, resulting in a repulsive pairwise interaction with E151 (Figure 4.5A). In figure 4.5B, snapshots taken

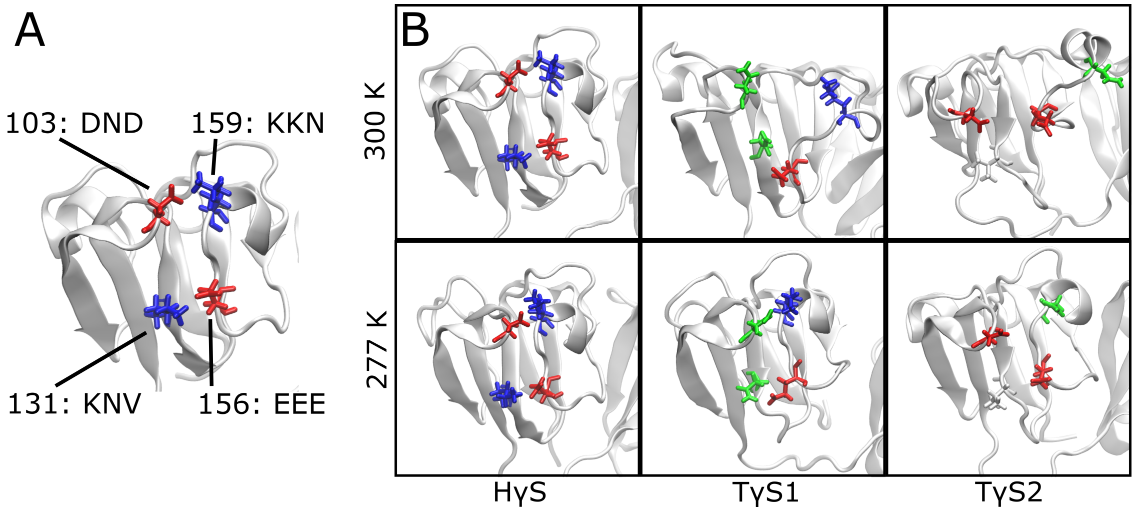

Figure 4.5: (A) Snapshot of the stabilizing salt bridges in the C-terminal domain of HγS. Each residue is labeled with the HγS residue number, and the residue name at that position for HγS, TγS1, and TγS2, respectively. (B) Snapshots of the homologous residues to the stabilizing salt-bridges in HγS. For both panels, residues are rendered in licorice representation and colored by residue type (Red: Acidic, Blue: Basic, Green: Polar, White: Hydrophobic). The protein backbone is rendered in cartoon representation and colored white

from MD simulations at 277 K and 300 K show the difference in backbone and salt-bridge conformation between the three proteins. At 277 K, both TγS1 and TγS2 maintain the native crystallin fold similar to HγS. However, at 300 K, the weaker hydrogen bonding interactions at the $\beta$-hairpin break, resulting in separation of the disordered loop region for both proteins. For TγS2 at 300 K, the disordered loop is separated even further near the K131 to V126 substitution. At 277 K, this separation is still visible, to a lesser degree, however a native-like fold is still observed. HγS, containing both salt-bridges intact, maintains its native structure at both 277 K and 300 K.

Similar salt-bridging interactions are also seen in the N-terminal domain, which show no change in dynamics across all three proteins. At analogous locations to the D103-K149 and K131-E156 salt bridges, two salt bridges are also seen bridging an unstructured loop to a nearby $\beta$-hairpin (K14-E69) and a $\beta$-sheet (K41-E66) (Figure D.1). Though there is some variation in amino acid, the charges and salt bridges are conserved across all three proteins. At 277 K and 300 K, these salt bridges remain intact and the N-terminal domain remains stable with no notable changes in flexibility or packing of the hydrophobic core.

To investigate the conservation of these mutations, the sequences of 667 $\gamma$-crystallins were analyzed. Each of the sequences were aligned with the sequence of H$\gamma$S, and the four salt bridging residues (D103, K159, K131, and E156) were analyzed for conservation (summarized in Figure D.2 and Table D.1). For positions at D103, K159 and K131, an asparagine is found in 40.6%, 22.2%, and 44.2% of aligned sequences, respectively. E156 is found in 66.9% of the aligned sequences as well as being conserved across H$\gamma$S, T$\gamma$S1, and T$\gamma$S2. The lysines (K131 and K159 in H$\gamma$S) have a relatively low frequency, however there is a large abundance of sequences with arginine (at the K159 position) and histidine (at the K131 position). The V126 substitution at K131 only appears in 15 of the 667 sequences, with no other frequent hydrophobic substitutions. These 15 sequences containing this valine all belong to the "M2-like" $\gamma$-crystallins of fish eye lenses. With the exception of V126, there is a frequent selection for polar or charged residues with long sidechains at the domain-domain interface.

## 4.4 Discussion

Protein thermal stability has been shown to be correlated with average body temperature using neutron scattering experiments[127]. For enzymes, this has been attributed to the role of structural flexibility required for the function of a catalytic active site at its operating temperature[128]. The $\gamma$S-crystallins from the human and toothfish eye lenses show a similar correlation with thermal stability and flexibility. This results in the toothfish $\gamma$S-crystallin proteins having similar cold temperature dynamics to that of human $\gamma$S-crystallin at room temperature. Though it is commonly accepted that the structure of $\gamma$S-crystallin is important for the proper function of the structural protein, the cold adapted toothfish $\gamma$S-crystallin provide an example where dynamics is preserved at operating temperature as well. Particularly, the toothfish $\gamma$S-crystallins counteract the "freezing" effects of low temperatures through a more flexible, yet more labile, protein backbone. This suggests that the $\gamma$S-crystallin flexibility holds an important role for its function as a structural eye lens protein.

Comparisons of MD simulations at 277 K and 300 K shows that increased lability in the toothfish crystallins are isolated to three regions in the C-terminal domain (highlighted in Figure 4.5). In H$\gamma$S, these regions are restrained by clusters of domain-domain salt bridges that cross between the two homologous Greek key motifs. These domain-domain salt bridge clusters are common to many thermophilic proteins[129, 115, 111, 113], and have been proposed as a strategy to rigidify labile regions without altering the core residues responsible for the protein fold[73]. In T$\gamma$S1 and T$\gamma$S2, key salt bridges are replaced with sidechain hydrogen bonds, resulting in increased backbone flexibility, while preserving the protein structure at lower temperatures. It is important to note that altered electrostatics have long-range effects, including altered inter-protein interaction. Therefore additional contributions to protein stability can come from interprotein interactions, especially when considering higher concentrations. However, the results from MD simulations (performed at infinite dilution) correlate well with the thermal unfolding experiments, which were performed at 0.25 g/L[123].

In addition to the correlation between protein stability and flexibility, packing of the hydrophobic core at 300 K correlates with protein stability as well. The burial and packing of protein hydrophobicity is an important factor for the thermodynamic stability[130, 131]. Since the protein core consists of a dense network of both polar and non-polar interactions, effectively parsing such interactions can be difficult. The chemical group graph representation coarse grains amino acids into moieties based on charge, allowing the protein structure to be analyzed as a network of charged, polar and nonpolar interactions. Additionally, hydrophobicity from non-hydrophobic residues, such as the carbon sidechain of a glutamine, are identified as part of the protein's hydrophobicity. This group representation provides an optimized scheme for network analysis of the hydrophobic core.

The K-core analysis of the human and toothfish $\gamma$S-crystallin hydrophobicity shows a correlation with thermal stability and the packing of the hydrophobic core. Both the size of the most dense (k=9) k-core and the profile of the average k-core (Figures 4.4 and 4.3) decrease in T$\gamma$S1 and, to a larger degree, in T$\gamma$S2. These metrics correspond to a decrease in cohesion of the collective hydrophobic interactions in more thermostable proteins. This correlation is consistent with previous observations of thermophilic proteins containing tightly packed hydrophobic cores[132, 117]. The k-core analysis shows that, when analyzing the core as a network of interactions, the tighter packing contains a larger degree of hydrophobic interactions at each site. This tighter cohesion creates an overall interaction network structure that is more resilient in the event of losses in contacts. For the case of thermal stability, a cohesive hydrophobic core would be more resistant to unfolding (loss in structure of the hydrophobic core) in the event of increased thermal fluctuations. The k-core analysis of the network of non-polar interactions shows a loss in cohesion that correlates with thermal stability, and may provide a useful tool for analyzing the dense hydrophobic core in thermostable proteins.

## 4.5 Conclusion

MD simulations of human and toothfish $\gamma$S-crystallins shows changes in flexibility that correlate with the protein thermal stability reported from circular dichroism experiments. Much of this increase in flexibility, in the toothfish $\gamma$S-crystallins, can be attributed to a loss in domain-domain salt bridges, a common trait of thermostable proteins. However, at lower temperatures, the toothfish $\gamma$S-crystallins recover a native-like structure and dynamics similar to that of H$\gamma$S. This shows that both structure and flexibility are conserved as part of the cold adaptation of the toothfish $\gamma$S-crystallins, and highlights the importance of flexibility in the functional role of the structural proteins. Additionally, network k-core analysis shows a loss in cohesion of the non-polar interactions at the densest portions of the hydrophobic core. This signifies a loss in hydrophobic packing that correlates with protein thermal stability as well. The chemical group graph representation combined with k-core analysis provides a potentially useful method for effectively measuring the packing of a protein's hydrophobic core.

# Chapter 5

# Interdomain dynamics of human $\gamma$D-crystallin

## 5.1 Introduction

One of the main principles important to the statistical analysis of MD simulations is ergodicity, the property by which a dynamical system at equilibrium should have a single particle time average that is equal to the average of the ensemble. This means that, given a sufficient amount of time, a single particle is capable of sampling all accessible regions in phase space. However, in the case where regions in phase space are restricted (e.g. long transition rates between states), the time average may not converge to the ensemble average until much longer timescales. This slow convergence where the time average does not represent the ensemble is known as weak ergodicity breaking. In this work, I will test the ergodicity of the diffusion and fluctuations of the interdomain motions of human $\gamma$D-crystallin (H$\gamma$D).

The diffusion process of H$\gamma$D is measured by the mean squared displacement (MSD) of the distance between the centers of mass of the N- and C-terminal domains. In an ensemble of

N particles, the ensemble averaged MSD (EA-MSD) is expressed as:

$$\langle x^2(t)\rangle = \frac{1}{N}\sum_{i=1}^{N} x_i^2(t) \tag{5.1}$$

where $x_i$ is the distance between centers of mass of domains for the $i^{th}$ replicate of trajectory. In the case a single trajectory, the time averaged MSD (TA-MSD) can be calculated by averaging displacements over a sliding time window over multiple time origins:

$$\overline{\delta^2(t)} = \frac{1}{T-t}\int_0^{T-t} (x(t'+t) - x(t'))^2 dt' \tag{5.2}$$

where T is the trajectory length, t is the lag time, and x(t') is the interdomain distance at time t'. When observing systems with anomalous diffusion, the MSD exhibits nonlinear scaling with time. This work will cover a subdiffusive MSD, which is expressed as $\langle x^2(t)\rangle \propto t^{\alpha}$, where $\alpha$ is the anomalous diffusion coefficient ($0 < \alpha < 1$). Single particle tracking experiments have shown that ergodicity breaking often shows different power law scaling behavior of the MSD between temporal and ensemble averaging[133, 134]. Ergodicity breaking in a diffusive process is a strong indication that the diffusion process follow (or contains) a continuous time random walk (CTRW) model consisting of instantaneous jumps of variable after random waiting times. The CTRW model can be likened to a dynamical system with an energy landscape containing wells of variable depths. However, in the case of non-ergodicity, exceptionally deep wells can essentially ”freeze” a system into a particular state due to the slow kinetics to crossing the energy barrier[135]. Discrepencies in the temporal and ensemble averages will begin to show when these long waiting times are comparable to the observation time of the system. Examples of these processes containing long waiting time distributions have been found in biological systems, such as trapping of ion channels to actin network on a cellular membrane[133] and the domain-domain motions of biological enzymes[136].

In addition to testing for ergodicity breaking in diffuion, a non-stationarity of the interdomain fluctuations, measured by the two-time correlation function, is identified and analyzed. This observation time dependence of the correlation function is known as aging. In this work, the interdomain distance fluctuations are calculated from the two time correlation function, C'(t;T), of the interdomain distance centered at the mean:

$$C'(t;T) = \frac{1}{T-t}\int_0^{T-t}[x(t'+t)-\overline{x}][x(t')-\overline{x}]dt' \tag{5.3}$$

$$C(t;T) = C'(t;T)/C'(0;t) \tag{5.4}$$

where T is the trajectory length, t is the lag time, $\overline{x}$ is the average distance between centers of mass of the domains, and C(t;T) is the normalized interdomain distance autocorrelation function (ACF). Relatively recent investigations on the domain-domain and sidechain-sidechain distance fluctuations of several enzymes report a characteristic relaxation time that is observation time dependent well into the tens of microseconds of MD simulations[136]. The characteristic relaxation time dependence on the observation time follows a power law relation that can be correlated with experimental measurements on the timescale of minutes[137, 138]. Should aging indeed last into the minutes timescale and beyond, this could mean that this non-ergodic behavior could persist throughout the *in vivo* lifetime of a protein.

In this work, single protein atomistic molecular dynamics simulations of H$\gamma$D were performed at three different timescales: 10 ns, 2 $\mu$s, and 44 $\mu$s. Similar to results presented by Hu et al.[136], interdomain center of mass motions are subdiffusive and show aging fluctuations over the three observed timescales. However, at 44 $\mu$s, the ACF begins to converge, resulting in the breaking of the power law relation between the characteristic relaxation time and observation time and the end of aging. In contrast, macromolecular crowding is introduced by modeling four proteins in a periodic cell, resulting in prolonged aging of the ACF to 44 s with no sign of convergence. Characterization of the dynamics of the dilute (single protein) and crowded (four protein) simulations shows seemingly ergodic diffusion in a geometrically confined landscape despite clear signs of aging. Further work on identifying the underlying aging process will more clearly identify the physical origins of this non-ergodicity and its relation with crowding.

## 5.2 Results and Discussion

### 5.2.1 Interdomain motions are subdiffusive with a non-stationary time correlation function

To analyze the domain-domain dynamics of $\gamma$D-crystallin, the distances between centers of mass of the two domains were tracked over time. For each protein in each simulation, the TA-MSD and TA-ACF were calculated over three different trajectory lengths: 10 ns, 2 $\mu$s, and 44 $\mu$s (simulation details are outlined in the methods). With the exception of trajectory with lengths of T = 44 $\mu$s, TA-MSDs were averaged over 12 replicates of simulation. The TA-MSD (shown in Figure 5.1A) three regimes of diffusive motion. In the sub-ps timescale, the dynamics is super-diffusive, where the TA-MSD scales as $t^{\alpha}$ where the average subdiffusive exponent, $\alpha$, is 1.355 with a standard error of 0.002. From ∼100 ps, the MSD shifts to $\alpha = 0.159 \pm 0.013$ for the dilute simulation and a lesser $\alpha$=0.106±0.012 for crowded simulations. At ∼300 ns, the MSD plateaus, suggesting that the limits of confinement has been reached. In the crowded simulations, a similar plateau is observed at a similar displacement, yet is not reached until well into the microsecond timescale (orange line in Figure 5.1A).

Similar to the results reported by Hu et al[136], there is a clear observation time dependence in the time correlation function of interdomain motions. From Figure 5.1B, doth dilute and crowded simulations show that the effective decay rate, $\tau_e$ (as indicated by the 1/e cutoff)

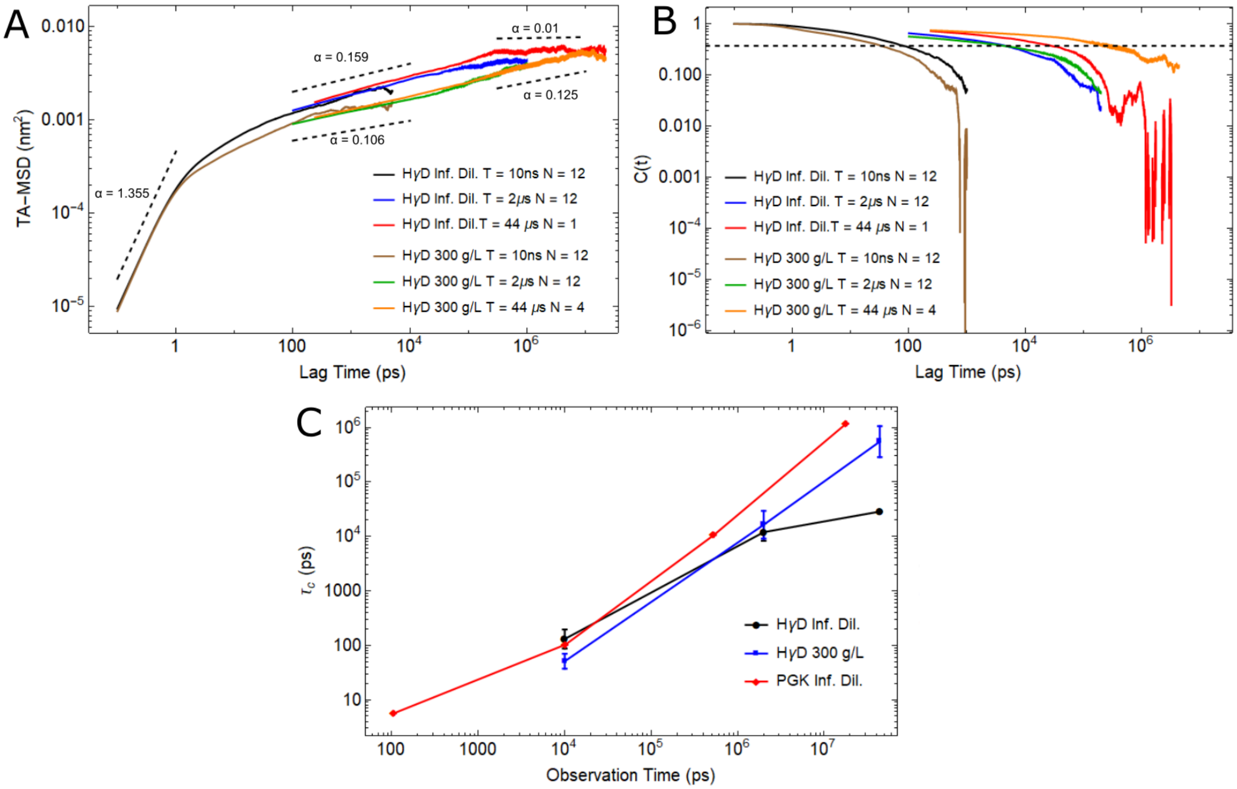

Figure 5.1: (A) Time averaged MSD (TA-MSD) and (B) Time averaged autocorrelation function (C(t)) from simulations of H$\gamma$D in crowded and dilute conditions. The total trajectory length (T) and number of replicates (N) are indicated in both legends. The four proteins in the 300 g/L simulations were considered as replicates. The 1/e cutoff used to calculate the effective decay rate $\tau_e$ is shown as a dotted line in the time correlation plot. (C) Scatter plot of the effective decay rate ($\tau_e$ vs. observation time or trajectory length). Data points from domain fluctuations in PGK from Hu et al.[136] are plotted in red. Error bars are reported as the standard error of the mean.

increases with respect to the trajectory length via a power law relation(Figure 5.1C). Under crowded conditions, the power law relation is maintained for the full 44 $\mu$s. Under dilute conditions, however, the characteristic relaxation times lose their power law dependence with observation times at 44 $\mu$s and seems to converge to a stationary value

### 5.2.2 Dynamics characterization workflow

Interdomain motions were characterized using a workflow developed by Meroz and Sokolov[139, 140]. This workflow deduces a known model of subdiffusion from the basic characteristics of the protein motion. In this case, the motion investigated is the distance between the centers-of-mass of the N- and C-terminal. A diagram outlining the characterization workflow is shown in Figure 5.2. The following sections will cover individual steps in the characterization workflow.

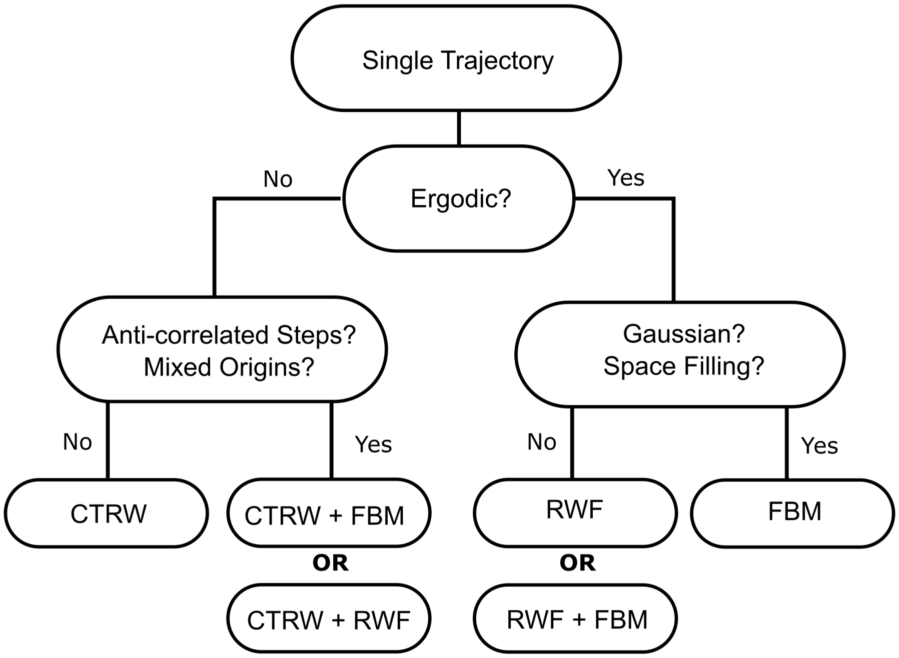

Figure 5.2: Flowchart diagram reproduced from work published by Meroz et al.[140]. This flowchart outlines methodology to characterize subdiffusive motion into one or a combination of three subdiffusive models: Continuous-time random walk (CTRW), Fractional Brownian motion (FBM), and Random walk on a fractal (RWF).

### 5.2.3 $\gamma$D-crystallin exhibits ergodic interdomain diffusion

The first distinguishing factor between subdiffusive mechanisms is ergodicity, or, particularly, whether the statistical time average of an observable is equal to the average of the ensemble.

To check for ergodicity in diffusion, the first test would be to check whether the time averaged MSD (averaging over multiple time origins) is consistent with the ensemble average over multiple independent trajectories. In examples of ergodicity breaking[134, 141, 140], the TA-MSD and EA-MSD show different behavior in terms of their power-law scaling.

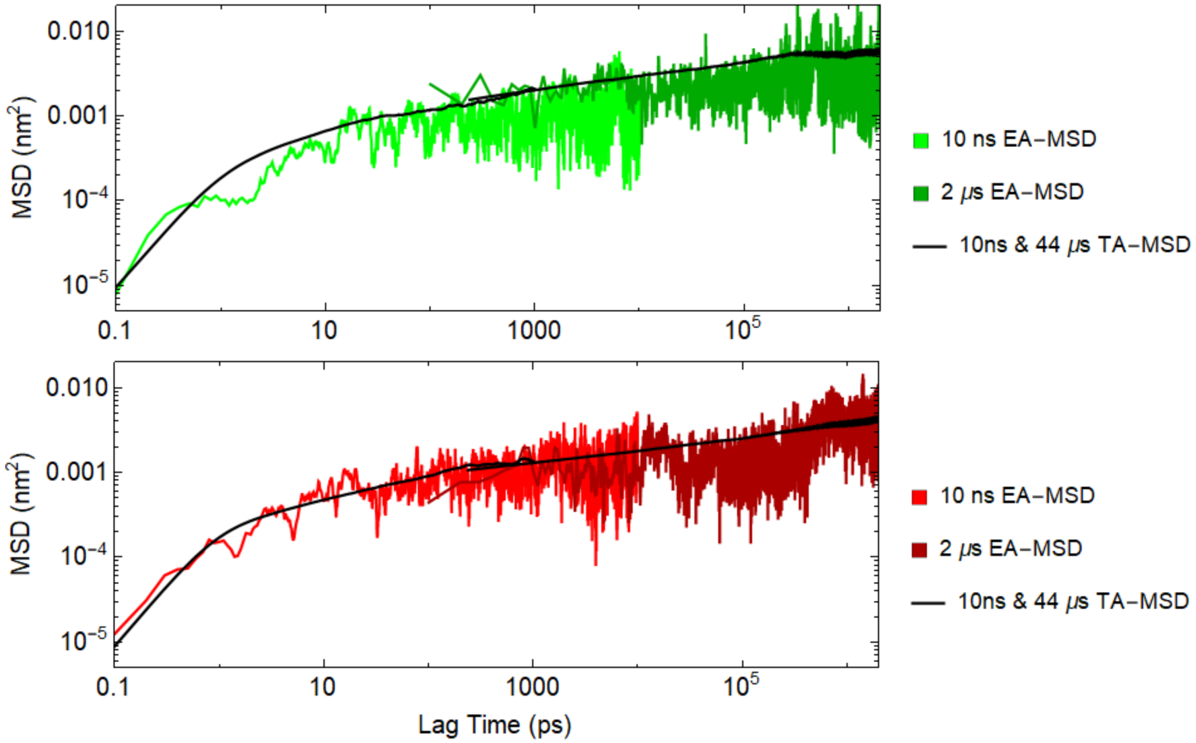

Figure 5.3: Comparisons of temporal and ensemble averaging of the MSD. The time averaged MSD (TA-MSD) values are measured from the 44 $\mu$s trajectory and a single 10 ns trajectory for lag times $<$ 100 ps. EA-MSD values were averaged over 12 replicates of 10 ns and 2 s trajectories.

Shown in Figure 5.3 is an overlay of the TA-MSD and the EA-MSD from both dilute and crowded simulations. 12 replicates of 2 $\mu$s trajectories were used for ensemble averaging. The TA-MSD overlaps within the dispersion of the EA-MSD, and there is no discernible difference in the power-law scaling. Similar ensemble and temporal profiles were seen using single particle tracking and optical tweezer experiments[142], leading to the conclusion of confined ergodic motion. This suggests that the diffusion process is ergodic, though the dispersion in the EA-MSD indicates that additional replicates are required to more definitely distinguish both MSD profiles.

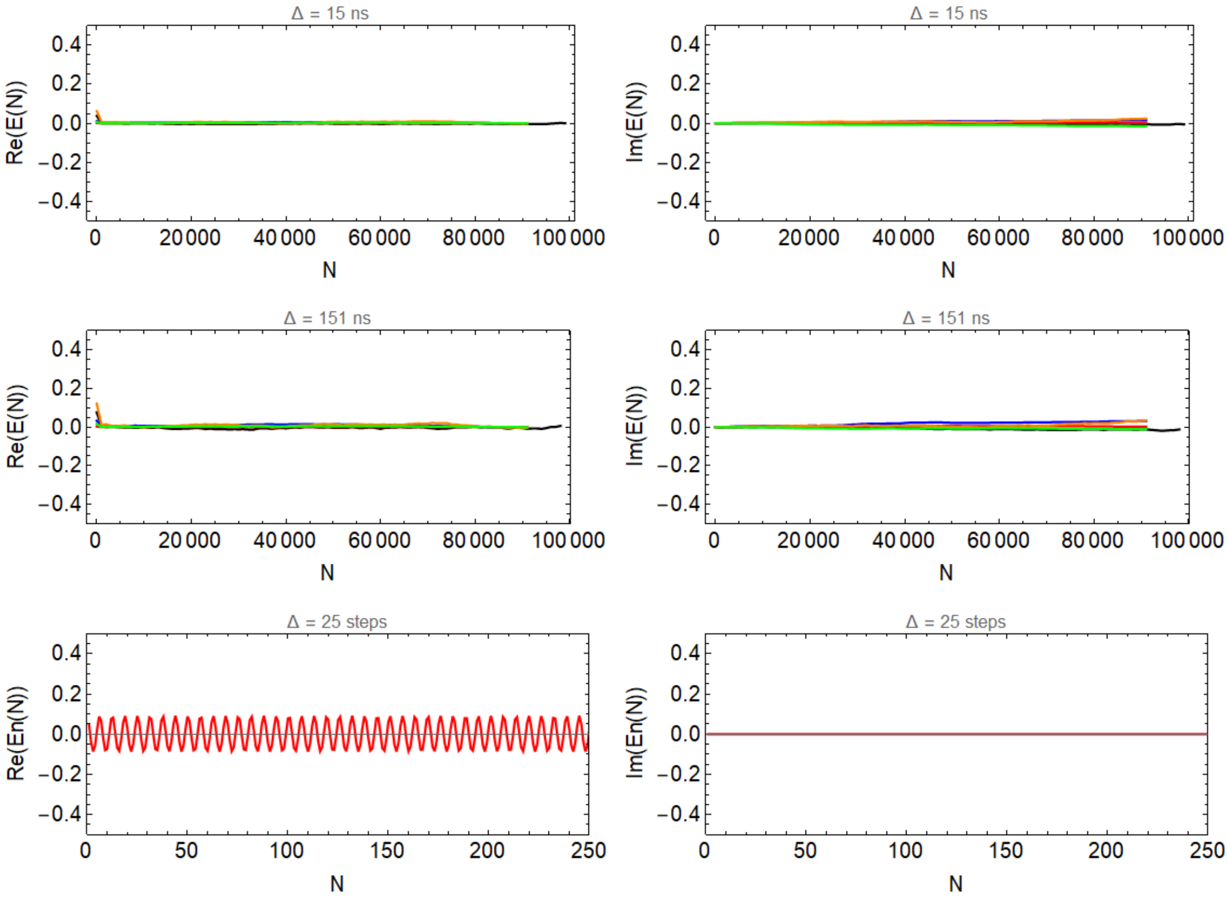

Figure 5.4: Real and imaginary parts of E(N) from the mixing test[143] plotted for (top) H$\gamma$D simulations at a lagtime of 15ns, (middle) H$\gamma$D simulations at a lagtime of 151 ns and (bottom) a non-mixing, stationary stable harmonizable process described by Magdiarz et al.[143]. For the simulations of H$\gamma$D, the infinite dilution simulation is plotted in black, while the four proteins from the crowded simulations are plotted in blue, red, green and orange. Convergence of real and imaginary parts of E(N) at large values of N is a sign of mixing, and furthermore ergodicity.

A second test for ergodicity is a correlation test for mixing, the asymptotic statistical independence of a random process at infinite time separation. Since mixing is a stronger property than ergodicity (all mixing processes are ergodic)[144], proving a process to be mixing is sufficient to proving ergodicity. For this case, the interdomain displacements of H$\gamma$D is tested at several lag times using the mixing test developed by Magdziarz et al.[143]. This method tests if the joint probability density function (PDF) of the displacements Y(0) and Y(N), where Y(N) is the displacement at time N, can be expressed as the product of the marginal PDF. Magdziarz et al. introduces the measure E(N):

$$E(N) = D(N) - |\langle \exp\{iY(0)\}\rangle|^2 \tag{5.5}$$

where $D(N) = \langle \exp\{i[Y(N) - Y(0)]\}\rangle$. Under the condition of ergodicity, the two terms in E(N) cancel and E(N) vanishes as N → ∞. In Figure 5.4, the real and imaginary parts of E(N) asymptotically converge to zero for H$\gamma$D in dilute conditions. At crowded conditions, E(N) quickly drops to zero, but very slightly deviates from zero at large values of N. Additionally, a non-ergodic, non-mixing process was tested to show non-ergodic behavior in the mixing test. For this case, I used the stable harmonizable process of the form $Y(t) = A^{1/2}[G_1 cos(t) + G_2 sin(t)]$, where $G_1$ and $G_2$ are standard normal variables and A $> 0$ is sampled from a one-sided $\alpha$-stable distribution[143]. Like the results published in the original paper, the real part of E(n) oscillates indefinitely, confirming the process to be non-mixing.

Though both equivalence of ensemble and temporal averaging as well as properties of mixing make a strong case for ergodicity, a previous report of a CTRW model with Gaussian noise shows deceptively ergodic behavior, despite having an underlying aging CTRW[134]. This includes, at high magnitudes of noise, correspondence between TA-MSD and EA-MSD profiles. Since motions in biological systems are rife with Brownian noise, this highlights the need to separate ergodic fluctuations from the aging process to effectively characterize aging. There have been coarse-graining methods[136, 145], proposed to smooth out fluctuations in proteins. Partial analysis of the coarse-grained dynamics of H$\gamma$D using the conformational cluster transition network (CCTN) method is outlined in Appendix E.

### 5.2.4 Interdomain motions resemble a random walk on a confined landscape

The previous tests identify ergodic behavior in the interdomain fluctuations of H$\gamma$D. The next step is to determined if there is an underlying structure that confines motion. The workflow previously described (Figure 5.2), reduces ergodic motion to two popular subdiffusive models: fractional Brownian motion (FBM) and random walk on a fractal (RWF). Both models are ergodic and subdiffusive, yet result from different environments. FBM subdiffusion is seen in motions restricted in a viscoelastic medium[146, 142], while RWF has been reported in environments that are geometrically constricted[133].

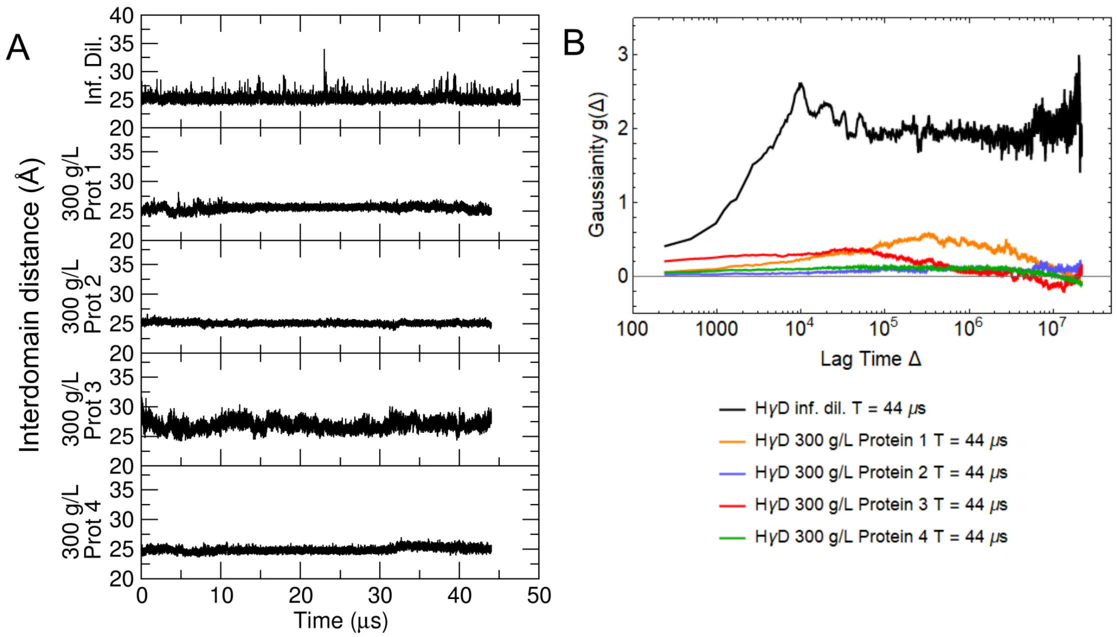

Figure 5.5: (A) Time evolution plots of the H$\gamma$D interdomain distance (plotted in Angstroms) under dilute and crowded conditions. The interdomain plots from the four proteins of the crowded simulations (300 g/L) are plotted individually. (B) Plots of the Gaussian parameter (g(t)) with respect to lag time from 44 $\mu$s trajectories of the dilute and crowded simulations of H$\gamma$D.

The main distinction between FBM and RWF can be found in the displacements. Driven by Gaussian noise, FBM should have a normally distributed displacement PDF, while the

PDF from RWF processes should deviate from a Gaussian distribution due to the underlying geometry[147]. One can use the Gaussian parameter (for one-dimensional motion)[148]:

$$g(t) = \frac{\langle dr^4 \rangle}{3\langle dr^2 \rangle^2} - 1 \tag{5.6}$$

where $dr$ is the displacement at lagtime $t$. This parameter represents any deviation from a Gaussian distribution, where $g(t) = 0$ represents a normally distributed PDF. The time evolution of the interdomain distance (Figure 5.5A) shows that, under dilute conditions, the domains separate intermittently, resulting in a slightly tailed displacement PDF. In concentrated conditions, the molecular packing restricts these intermittent domain separations, resulting in a narrower but Gaussian displacement PDF. This difference is reflected in the Gaussianity, where the value is non-zero only under dilute conditions (Figure 5.5B). This comes from the heavy tails of the dilute PDF resulting in an increased fourth moment.

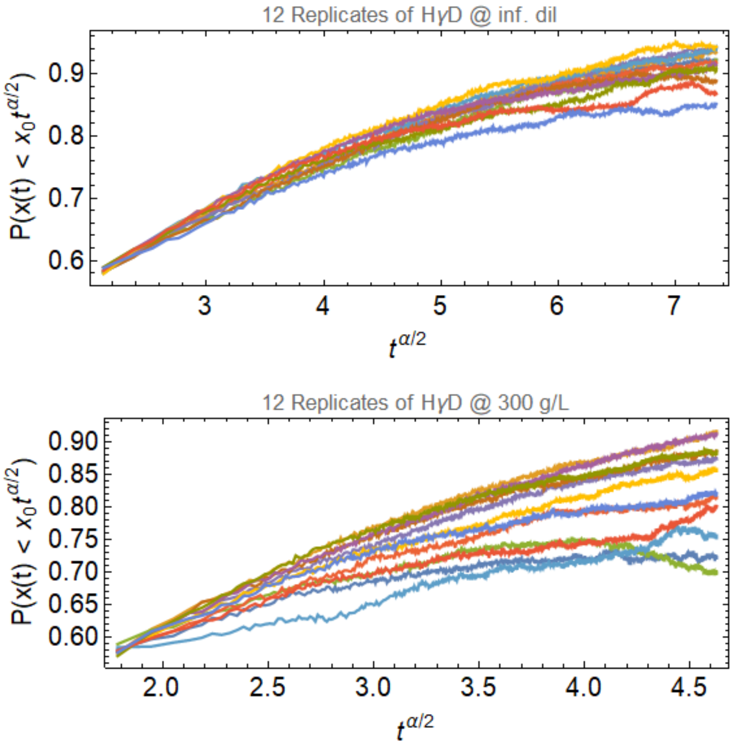

Figure 5.6: Growing sphere analysis of 12 replicates of 2 $\mu$s simulations of H$\gamma$D at (top) infinite dilution and (bottom) 300 g/L. An increasing probability indicates the presence of confinement, while a constant probability indicates Brownian diffusion[147].

The non-Gaussian PDF of dilute H$\gamma$D is a strong indication of a RWF-like diffusion. However, the introduction of macromolecular crowding recovers Gaussian-like behavior, suggesting a loss of the underlying geometric confinement due to macromolecular crowding. To determine whether the Gaussian PDF is simply a coincidence, a second test is applied to test for confined motion. The growing sphere analysis[147] tests the probability of a particle (in our case, the interdomain distance) being contained within a sphere that grows proportionally to the anomalous diffusion exponent:

$$P(x \leq x_0 t^{\alpha/2}) \approx \frac{1}{N(t)} \sum_{j=1}^{N(t)} \mathcal{H}(x_j(t) - x_0 t^{\alpha/2}) \tag{5.7}$$

where $\mathcal{H}(x)$ is the Heaviside function, x(t) is the absolute value of the displacement after lagtime t, $\alpha$ is the anomalous diffusion coefficient, and $x_0$ is a free parameter (in thic case, chosen as $\langle x(t=1)\rangle$[140]. If no underlying geometric confinement is present, the sphere grows at the same rate as the displacements and the probability remains constant. If confinement is present, the sphere grows faster than the displacements, and the probability increases. In Figure 5.6, both dilute and crowded simulations show increasing probabilities, confirming that geometric confinement is still present despite a Gaussian displacement PDF in the crowded simulations.

Characterization of the diffusion shows ergodic motion that follow a random walk on a fractal model for both dilute and crowded cases. However, crowding affects diffusion by restricting the displacements of the interdomain distance and the intermittent domain separations. This results in a loss in the heavy tails of the displacement distribution in the crowded case. No CTRW motions could be identified from the displacements, however, any ergodic fluctuations must be removed to definitively rule out CTRW diffusion. In the next section, the ACF will be fit to a noisy CTRW model in order to identify ergodicity breaking in the presence of ergodic fluctuations.

### 5.2.5 Noisy CTRW and future work

As previously discussed, Gaussian noise superimposed over a non-ergodic process, or a noisy CTRW, can make a dynamic process appear ergodic to many tests for ergodicity breaking[134]. Therefore, to effectively analyze an underlying non-ergodic process, one must be able to isolate such process from any ergodic noise. Since the noisy CTRW can serve as an analogy for protein motion on a rugged energy landscape with deep wells[149], Hu et al.[136] proposed that these wells can be identified by conformational clustering, allowing one to coarse-grain a trajectory into a CTRW between clusters. They continue to show that this coarse-grained trajectory shows non-ergodic behavior. To analyze the fluctuations, Hu et al. fit the aging ACF to a relaxation model:

$$C(t;T) = c_1 \exp[-(t/\tau)^{\beta}] + c_2 B(t/T, \alpha, 1-\alpha) + c_3 \tag{5.8}$$

where the first term is a stretched exponential term for the thermal fluctuations and the last two terms are the ACF relaxation model for a subdiffusive CTRW[150]. In this equation, t is the lagtime, T is the trajectory length, $\alpha$ is the anomalous diffusion coefficient, B(z, a, b) is the incomplete Beta function. The remaining variables are fit parameters. An improved fit of ACF to the noisy CTRW model should indicate the presence of an underlying CTRW process.

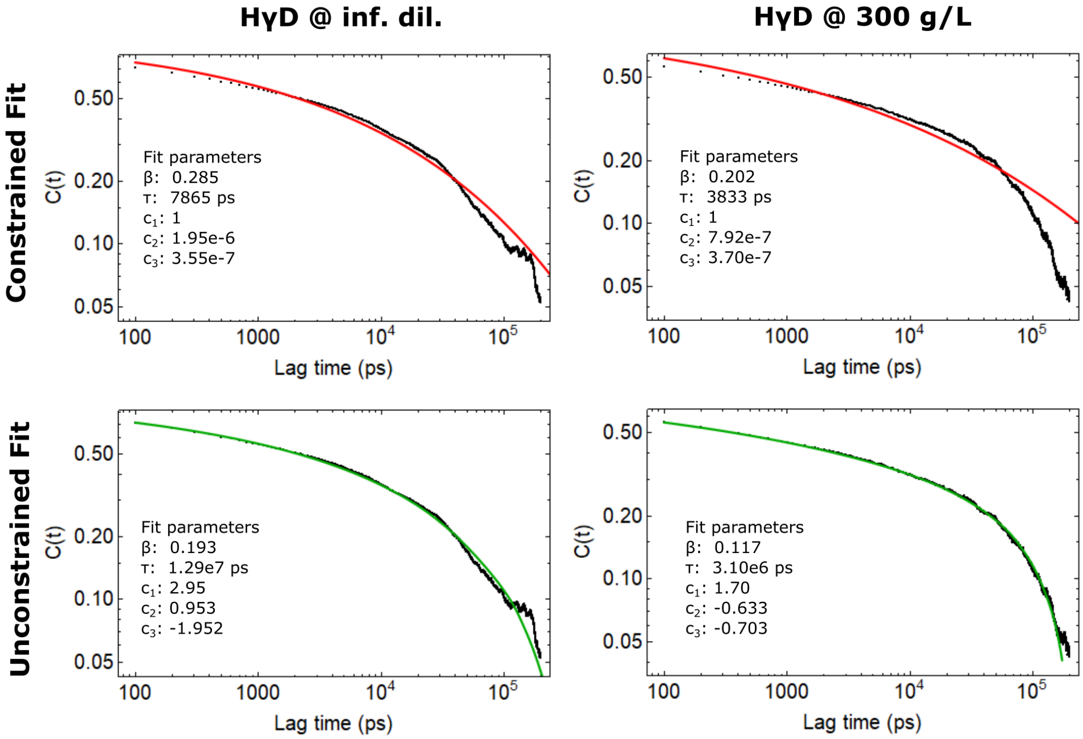

Figure 5.7: Plots of the fitted models of the noisy CTRW relaxation model[136, 134, 150] with the calculated ACF. Plots of the dilute and crowded simulations are displayed in the left and right columns, respectively. Fit parameters, $c_1, c_2$, and $c_3$, were constrained to be positive in the top row plots while the bottom plot shows the fit without constraints. The calculated ACF, constrained fit, and unconstrained fit are plotted in black, red, and green, respectively. The ACFs were calculated and averaged over 12 replicates of 2 $\mu$s trajectories.

The methodology described by Hu et al.[136] was used to determine if a similar underlying aging process is present in H$\gamma$D (work covering trajectory coarse-graining with the CCTN method is covered in Appendix E). When fitting the ACF to the noisy CTRW relaxation model, constraints were placed on the fit parameters. Since $c_2$, and $c_3$ represent the variance and the mean squared value of $\delta(t)$ (where $\delta(t) = x(t) - \overline{x}$ and $t$ is time)[150], these values must be constrained to positive values. With the constraints applied (top row of Figure 5.7), $c_2$ and $c_3$ are fit to values on the order of $10^{-6}$ and $c_1$ fit to nearly 1. This results in a fit that is lacking any CTRW fluctuations. Removal of the parameter constraints (bottom row of Figure 5.7), results in an improved fit at larger lagtimes ($t > 2 \times 10^4$ ps). However, since negative fit constants are not physically rational, such a fit should not be considered true. Though the fits suggest that no underlying subdiffusive CTRW can be identified in the ACF, the observation time dependence of the ACF (from Figure 5.1C) shows a clear sign of aging. Identifying an appropriate model by which the aging ACF follows will be important to identifying the underlying mechanism that drives aging in the interdomain fluctuations.

## 5.3 Conclusions

The observation time dependence of the ACF is considered a clear sign of ergodicity breaking. However, comparisons of the ensemble and temporal averaging and tests for mixing all seem to indicate that diffusion is ergodic. Furthermore, fits of the noisy CTRW model to the ACF fails to provide a fit with reasonable parameters. When the fit parameters are constrained to positive values, all terms related to CTRW vanish. This means that the tests for ergodicity breaking are unable to detect any underlying CTRW. Since ergodic noise is capable of masking aging from ergodicity tests, it is difficult to definitely rule out any ergodicity breaking. Nevertheless, non-stationarity of the ACF is still present, meaning that separating aging motions from the ergodic components still remains an important target for future work.

## 5.4 Methods

### 5.4.1 System preparation and equilibration

The initial protein coordinates of H$\gamma$D was built from the crystal structures deposited into the Protein Data Bank (PDB ID code 1HK0)[26]. Histidine protonation states were set to be the same as those published in the solution state NMR structure of the P23T variant of D-crystallin (PDB ID code 2KFB)[76]. Protein atoms were parameterized with the

CHARMM36 force field[31]. The crystal structure waters were kept and the proteins were solvated in a cubic TIP3P[77] water box measuring 80 Å on a side. The system was neutralized with chloride counterions. The single protein systems contained a total of 48,309 atoms. All system preparation was performed using the VMD 1.9.1 software package[78].

A 20 ns pre-production simulation equilibration was performed with NAMD 2.9[79]. The prepared systems were minimized for 10,000 steps in the NPT ensemble at 310 K and 1 atm. Protein heavy atoms were restrained with harmonic positional restraints and were gradually relaxed over 200 ps. NAMD was parameterized with the smooth particle mesh Ewald method[80, 81] for long-range electrostatic interactions, a real space interaction cutoff at 11 Å, and an integration time step of 2 fs/timestep. The RESPA algorithm[81] was used with a timestep of 4 fs for electrostatic forces, 2 fs for nonbonded forces, and 1 fs for bonded forces. Hydrogen covalent bonds were held fixed using the SHAKE[83] and SETTLE[84] algorithms. Constant temperature and pressure was maintained using a Langevin thermostat and a Nos-Hoover-Langevin piston[86, 85].

### 5.4.2 Microsecond time scale molecular dynamics simulations

Production simulations were performed on the Anton 2 supercomputer, a special-purpose computer for molecular dynamics simulations of biomolecules[34]. Protein and solvent atoms were parameterized with the CHARMM36[31] and TIP3P[77] forcefields, respectively. The multigrator scheme[87] was used to integrate Newtons equation of motion at 2.5 fs/timestep. Using the RESPA algorithm[82], long-range nonbonded, short-range nonbonded, and bonded forces were calculated at a timestep of 7.5 fs, 2.5 fs, and 2.5 fs, respectively. Long-range electrostatic forces were calculated using the k-Gaussian split Ewald method[88]. Hydrogen covalent bonds were held fixed using the SHAKE[83] algorithm. Constant temperature and pressure was maintained using Nose-Hoover chains[89] and the Martyna-Tobias-Klein barostat[86], respectively. Long time scale single protein simulations of H$\gamma$D was run for 50 $\mu$s, and the first 6 $\mu$s of simulation was discarded for equilibration of interdomain motions. 12 replicates of shorter 2 $\mu$s and 10ns simulations were prepared using conformations taken the 50 $\mu$s trajectory (a different conformation was extracted every 3 $\mu$s). Velocities were resampled and equilibrated (using the protocol mentioned before) before 2 $\mu$s of production simulation on Anton 2 or 10 ns of simulation on local resources. Frames were written every 100.8 ps and 100 fs for the 2 $\mu$s and 10 ns simulations, respectively.

### 5.4.3 Four protein simulation preparation

To prepare the four protein MD simulations (300 g/L concentration), a four protein oligomer conformation was extracted from a multi-conformation Monte-Carlo simulation[96] using conformational libraries taken from clustered conformations of the single protein 50 $\mu$s trajectory. To equilibrate the solvent, the four proteins were displaced such that each protein is separated by at least 12 Å from another protein. The conformations were then solvated and equilibrated using NAMD 2.7[79] for 20 ns as described before. On the Anton 2 supercomputer, the structures were equilibrated where centers of mass restraints were placed between the pairs of all four proteins, allowing rotation of the protein but restricting translational motion. The restraints were gradually relaxed over the course of 480 ns of equilibration. The system was then run for 50 $\mu$s of production simulation, and the first 6 $\mu$s of the trajectory was discarded before analysis. 3 replicates of 2 $\mu$s were prepared in the same manner as the single protein simulations.

### 5.4.4 Trajectory analysis

The interdomain motions were measured by the distance between centers of mass of the backbone alpha carbons in the N- and C-terminal domains. The interdomain linker and unstructured C-terminal strand were excluded from the center of mass calculation. Thus, the N- and C-terminal domains were defined by the alpha carbons of residues 1-80 and 87-169, respectively.

# Bibliography

[1] Allen Foster and S. Resnikoff. The impact of Vision 2020 on global blindness. *Eye*, 19(10):1133–1135, oct 2005.

[2] R J Truscott and R C Augusteyn. Changes in human lens proteins during nuclear cataract formation. *Experimental eye research*, 24(2):159–70, 1977.

[3] Wei Wang, William Yan, Kathy Fotis, Noela M. Prasad, Van Charles Lansingh, Hugh R. Taylor, Robert P. Finger, Damian Facciolo, and Mingguang He. Cataract Surgical Rate and Socioeconomics: A Global Study. *Investigative Opthalmology & Visual Science*, 57(14):5872, jan 2017.

[4] Rupert R.A. Bourne, Seth R. Flaxman, Tasanee Braithwaite, Maria V. Cicinelli, Aditi Das, Jost B. Jonas, Jill Keeffe, John Kempen, Janet Leasher, Hans Limburg, Kovin Naidoo, Konrad Pesudovs, Serge Resnikoff, Alex Silvester, Gretchen A. Stevens, Nina Tahhan, Tien Wong, Hugh R. Taylor, Peter Ackland, Aries Arditi, Yaniv Barkana, Banu Bozkurt, Richard Wormald, Alain Bron, Donald Budenz, Feng Cai, Robert Casson, Usha Chakravarthy, Nathan Congdon, Tunde Peto, Jaewan Choi, Reza Dana, Maria Palaiou, Rakhi Dandona, Lalit Dandona, Tueng Shen, Iva Dekaris, Monte Del Monte, Jenny Deva, Laura Dreer, Marcela Frazier, Leon Ellwein, James Hejtmancik, Kevin Frick, David Friedman, Jonathan Javitt, Beatriz Munoz, Harry Quigley, Pradeep Ramulu, Alan Robin, James Tielsch, Sheila West, Joao Furtado, Hua Gao, Gus Gazzard, Ronnie George, Stephen Gichuhi, Victor Gonzalez, Billy Hammond, Mary Elizabeth Hartnett, Minguang He, Flavio Hirai, John Huang, April Ingram, Charlotte Joslin, Rohit Khanna, Dwight Stambolian, Moncef Khairallah, Judy Kim, George Lambrou, Van Charles Lansingh, Paolo Lanzetta, Jennifer Lim, Kaweh Mansouri, Anu Mathew, Alan Morse, David Musch, Vinay Nangia, Maurizio Battaglia, Fernando Yaacov, Murugesan Raju, Luca Rossetti, Jinan Saaddine, Mya Sandar, Janet Serle, Rajesh Shetty, Pamela Sieving, Juan Carlos Silva, Rita S. Sitorus, Jaime Tejedor, Miltiadis Tsilimbaris, Jan van Meurs, Rohit Varma, Gianni Virgili, Jimmy Volmink, Ya Xing, Ning Li Wang, Peter Wiedemann, and Yingfeng Zheng. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. *The Lancet Global Health*, 5(9):e888–e897, 2017.

[5] M. A. Wride. Lens fibre cell differentiation and organelle loss: many paths lead

to clarity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1568):1219–1233, 2011.

[6] S. Bassnett, Y. Shi, and G. F. J. M. Vrensen. Biological glass: structural determinants of eye lens transparency. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1568):1250–1264, apr 2011.

[7] Joseph Horwitz, Michael P Bova, Lin-Lin Ding, Dana a Haley, and Phoebe L Stewart. Lens $\alpha$-crystallin: Function and structure. *Eye*, 13(3):403–408, may 1999.

[8] Michael A. Wride. Cellular and molecular features of lens differentiation: A review of recent advances. *Differentiation*, 61(2):77–93, 1996.

[9] J Horwitz. Alpha-crystallin can function as a molecular chaperone. *Proceedings of the National Academy of Sciences of the United States of America*, 89(21):10449–53, 1992.

[10] Carolyn N. Kingsley, William D. Brubaker, Stefan Markovic, Anne Diehl, Amanda J. Brindley, Hartmut Oschkinat, and Rachel W. Martin. Preferential and Specific Binding of Human $\alpha$B-Crystallin to a Cataract-Related Variant of $\gamma$S-Crystallin. *Structure*, 21(12):2221–2227, dec 2013.

[11] Mireille Delaye, John I. Clark, and George B. Benedek. Identification of the scattering elements responsible for lens opacification in cold cataracts. *Biophysical journal*, 37(3):647–56, mar 1982.

[12] Ronald H.H. Kröger, Melanie C.W. Campbell, Rejean Munger, and Russell D. Fernald. Refractive index distribution and spherical aberration in the crystalline lens of the African cichlid fish haplochromis burtoni. *Vision Research*, 34(14):1815–1822, jul 1994.

[13] Jennifer C Boatz, Matthew J Whitley, Mingyue Li, Angela M Gronenborn, and Patrick C. A. van der Wel. Cataract-associated P23T $\gamma$D-crystallin retains a native-like fold in amorphous-looking aggregates formed at physiological pH. *Nature Communications*, 8(May):15137, may 2017.

[14] Yongting Wang, Sarah Petty, Amy Trojanowski, Kelly Knee, Daniel Goulet, Ishita Mukerji, and Jonathan King. Formation of Amyloid Fibrils In Vitro from Partially Unfolded Intermediates of Human $\gamma$C-Crystallin. *Investigative Opthalmology & Visual Science*, 51(2):672, feb 2010.

[15] Katerina Papanikolopoulou, Ishara Mills-Henry, Shannon L Thol, Yongting Wang, Abby A R Gross, Daniel A Kirschner, Sean M Decatur, and Jonathan King. Formation of amyloid fibrils in vitro by human gammaD-crystallin and its isolated domains. *Molecular vision*, 14(December 2007):81–9, 2008.

[16] Nicholas J Ray, Damien Hall, and John A Carver. Deamidation of N76 in human $\gamma$S-crystallin promotes dimer formation. *Biochimica et biophysica acta*, 1860(1 Pt B):315–24, 2016.

[17] Peter G Hains and Roger J W Truscott. Post-translational modifications in the nuclear region of young, aged, and cataract human lenses. *Journal of proteome research*, 6(10):3935–43, oct 2007.

[18] Eugene Serebryany and Jonathan a. King. The $\beta\gamma$-crystallins: Native state stability and pathways to aggregation. *Progress in Biophysics and Molecular Biology*, 115(1):32–41, 2014.

[19] George B. Benedek. Cataract as a protein condensation disease: The Proctor lecture. *American Journal of Ophthalmology*, 125(2):279, feb 1998.

[20] Ajay Pande, Kalyan S. Ghosh, Priya R. Banerjee, and Jayanti Pande. Increase in surface hydrophobicity of the cataract-associated P23T mutant of human $\gamma$d-Crystallin is responsible for its dramatically lower, retrograde solubility. *Biochemistry*, 49(29):6122–6129, 2010.

[21] William D Brubaker and Rachel W Martin. H, C, and N assignments of wild-type human $\gamma$S-crystallin and its cataract-related variant $\gamma$S-G18V. *Biomolecular NMR assignments*, 6(1):63–7, apr 2012.

[22] Payel Das, Jonathan a King, and Ruhong Zhou. Aggregation of $\gamma$-crystallins associated with human cataracts via domain swapping at the C-terminal $\beta$-strands. *Proceedings of the National Academy of Sciences of the United States of America*, 108(26):10514–9, jun 2011.

[23] Sergi Garcia-Manyes, David Giganti, Carmen L. Badilla, Ainhoa Lezamiz, Judit Perales-Calvo, Amy E M Beedle, and Julio M. Fernández. Single-molecule Force Spectroscopy Predicts a Misfolded, Domain-swapped Conformation in human $\gamma$D-Crystallin Protein. *Journal of Biological Chemistry*, 291(8):4226–4235, feb 2016.

[24] Karuna Dixit, Ajay Pande, Jayanti Pande, and Siddhartha P. Sarma. Nuclear Magnetic Resonance Structure of a Major Lens Protein, Human $\gamma$c-Crystallin: Role of the Dipole Moment in Protein Solubility. *Biochemistry*, 55(22):3136–3149, 2016.

[25] Zhengrong Wu, Frank Delaglio, Keith Wyatt, Graeme Wistow, and Ad Bax. Solution structure of $\gamma$S-crystallin by molecular fragment replacement NMR. *Protein Science*, 14(12):3101–3114, dec 2005.

[26] Ajit Basak, Orval Bateman, Christine Slingsby, Ajay Pande, Neer Asherie, Olutayo Ogun, George B. Benedek, and Jayanti Pande. High-resolution X-ray Crystal Structures of Human $\gamma$D Crystallin (1.25Å) and the R58H Mutant (1.15Å) Associated with Aculeiform Cataract. *Journal of Molecular Biology*, 328(5):1137–1147, may 2003.

[27] Fangling Ji, Jinwon Jung, Leonardus M I Koharudin, and Angela M Gronenborn. The Human W42R D-Crystallin Mutant Structure Provides a Link between Congenital and Age-related Cataracts. *Journal of Biological Chemistry*, 288(1):99–109, jan 2013.

[28] Fangling Ji, Leonardus M I Koharudin, Jinwon Jung, and Angela M Gronenborn. Crystal structure of the cataract-causing P23T $\gamma$D-crystallin mutant. *Proteins*, 81(9):1493–8, sep 2013.

[29] Shannon L. Flaugh, Melissa S. Kosinski-Collins, and Jonathan King. Interdomain side-chain interactions in human $\gamma$D crystallin influencing folding and stability. *Protein Science*, 14(8):2030–2043, aug 2005.

[30] Eugene Serebryany and Jonathan a. King. Wild-type Human $\gamma$D-crystallin Promotes Aggregation of Its Oxidation-mimicking, Misfolding-prone W42Q Mutant. *Journal of Biological Chemistry*, 290(18):11491–11503, 2015.

[31] Robert B. Best, Xiao Zhu, Jihyun Shim, Pedro E.M. Lopes, Jeetain Mittal, Michael Feig, and Alexander D. MacKerell. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone $\phi$, $\psi$ and side-chain $\chi$ 1 and $\chi$ 2 Dihedral Angles. *Journal of Chemical Theory and Computation*, 8(9):3257–3273, 2012.

[32] Kyle a Beauchamp, Yu-Shan Lin, Rhiju Das, and Vijay S Pande. Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *Journal of chemical theory and computation*, 8(4):1409–1414, apr 2012.

[33] Thomas J. Lane, Diwakar Shukla, Kyle A. Beauchamp, and Vijay S. Pande. To milliseconds and beyond: Challenges in the simulation of protein folding. *Current Opinion in Structural Biology*, 23(1):58–65, 2013.

[34] David E. Shaw, J.P. Grossman, Joseph A. Bank, Brannon Batson, J. Adam Butts, Jack C. Chao, Martin M. Deneroff, Ron O. Dror, Amos Even, Christopher H. Fenton, Anthony Forte, Joseph Gagliardo, Gennette Gill, Brian Greskamp, C. Richard Ho, Douglas J. Ierardi, Lev Iserovich, Jeffrey S. Kuskin, Richard H. Larson, Timothy Layman, Li-Siang Lee, Adam K. Lerer, Chester Li, Daniel Killebrew, Kenneth M. Mackenzie, Shark Yeuk-Hai Mok, Mark A. Moraes, Rolf Mueller, Lawrence J. Nociolo, Jon L. Peticolas, Terry Quan, Daniel Ramot, John K. Salmon, Daniele P. Scarpazza, U. Ben Schafer, Naseer Siddique, Christopher W. Snyder, Jochen Spengler, Ping Tak Peter Tang, Michael Theobald, Horia Toma, Brian Towles, Benjamin Vitale, Stanley C. Wang, and Cliff Young. Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. In *SC14: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 41–53. IEEE, nov 2014.

[35] Binbin Wang, Changhong Yu, Yi-Bo Xi, Hong-Chen Cai, Jing Wang, Sirui Zhou, Shiyi Zhou, Yi Wu, Yong-Bin Yan, Xu Ma, and Lixin Xie. A novel CRYGD mutation (p.Trp43Arg) causing autosomal dominant congenital cataract in a Chinese family. *Human Mutation*, 32(1):E1939–E1947, jan 2011.

[36] H. J. Aarts, N. H. Lubsen, and J. G. Schoenmakers. Crystallin gene expression during rat lens development. *European Journal of Biochemistry*, 183:31–36, 1989.

[37] C. N. Nagineni and S. P. Bhat. Lens fiber cell differentiation and expression of crystallins in co-cultures of human fetal lens epithelial cells and fibroplasts. *Experimental Eye Research*, 54:193–200, 1992.

[38] Z. Ma, G. Piszczek, P.T. Wingfield, Y.V. Sergeev, and J.F. Hejtmancik. The G18V CRYGS mutation associated with human cataracts increases $\gamma$s-crystallin sensitivity to thermal and chemical stress. *Biochemistry*, 48:7334–7341, 2009.

[39] Srinivasu Karri, Ramesh Babu Kasetti, Venkata Pulla Rao Vendra, Sushil Chandani, and Dorairajan Balasubramanian. Structural analysis of the mutant protein d26g of human gammas-crystallin, associated with coppock caraeact. *Molecular Vision*, 19:1231—1237, 2013.

[40] Michael A. DiMaurro, Sandip K. Nandi, Cibin T. Raghavan, Rajiv Kuman Kar, Benlian Wang, Anirban Bhuania, Ram H. Nagaraj, and Ashis Biswas. Acetylation of gly1 and lys2 promotes aggregation of human gammad-crystallin. *Biochemistry*, 53:7269—7282, 2014.

[41] Yizhi Liu, Xinyu Zhang, Lixia Luo, Mingxing Wu, Ruiping Zeng, Gang Cheng, Bin Hu, Bingfen Liu, Jack J. Liang, and Fu Shang. A novel alphab-crystallin mutation associated with autosomal dominant congenital lamellar cataract. *Investigative Opthamology & Visual Science*, 47(3):1069—1075, 2006.

[42] Priya R. Banerjee, Shadakshara S. Puttamadappa, Ajay Pande, Alexander Shekhtman, and Jayanti Pande. Increased hydrophobicity and decreased backbone flexibility explain the lower solubility of a cataract-linked mutant of $\gamma$D-crystallin. *Journal of Molecular Biology*, 412(4):647–659, 2011.

[43] S. V. Bharat, A. Shekhtman, and J. Pande. The cataract-associated V41M mutant of human $\gamma$S-crystallin shows specific structural changes that directly enhance local surface hydrophobicity. *Biochenm. Biophys. Res. Commun.*, 443:110—114, 2010.

[44] Daumantas Matulis and Rex Lovrien. 1-anilino-8-naphthalene sulfonate anion-protein binding depends primarily on ion pair formation. *Biophysical Journal*, 74(1):422–429, 1998.

[45] J J Ory and L J Banaszak. Studies of the ligand binding reaction of adipocyte lipid binding protein using the fluorescent probe 1, 8-anilinonaphthalene-8-sulfonate. *Biophysical Journal*, 77(2):1107–1116, 1999.

[46] Dmitry Sheluho and Sharon H. Ackerman. An Accessible Hydrophobic Surface Is a Key Element of the Molecular Chaperone Action of Atp11p. *Journal of Biological Chemistry*, 276(43):39945–39949, 2001.

[47] Anne Bertolotti Christian Münch. Exposure of hydrophobic surfaces initiates aggregation of diverse als-causing superoxde dismutase-1 mutants. *Journal of Molecular Biology*, 399:512—525, 2010.

[48] Ashutosh Tiwari, Amir Liba, Se Hui Sohn, Sai V. Seetharanab, Osman Bilsel, C. Robert Matthews, P. John Hart, Joan Selverstone Valentine, and Lawrence J. Hayward. Metal deficiency increases aberrant hydrophobicity of mutant superoxide dismutases that cause amyotophic lateral sclerosis. *Journal of Biological Chemistry*, 284(27746—27758), 2009.

[49] M. P. Williamson. Using chemical shift perturbation to characterize ligand binding. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 73:1–16, 2013.

[50] S. B. Shuker, P. J. Hajduk, R. P. Meadows, and SW Fesik. Discovering high affinity ligands for proteins: Sar by nmr. *Science*, 274:1531—1534, 1996.

[51] T. ten Brink, C. Aguirre, T. E. Exner, and I. Krimm. Performance of protein-ligand docking with simulated chemical shift perturbations. *Journal of Chemical Information and Modeling*, 55:275—283, 2015.

[52] S. Huang and X. Zou. Efficient molecular docking of nmr structures: Application to hiv-1 protease. *Protein Science*, 16:43—51, 2007.

[53] Jaime L. Stark and Robert Powers. Application of nmr and molecular docking in structure-based drug discovery. *Topics in Current Chemistry*, 326:1—34, 2011.

[54] A. J. van Dijk, R. Boelens, and A. M. Bonvin. Data-driven docking for the study of biomolecular complexes. *FEBS Journal*, 272:293—312, 2005.

[55] Csaba Hetényi and David van der Spoel. Efficient docking of peptides to proteins without prior knowledge of the binding site. *Protein science : a publication of the Protein Society*, 11(7):1729–1737, 2002.

[56] Bogdan Iorga, Denyse Herlem, Elvina Barré, and Catherine Guillou. Acetylcholine nicotinic receptors: Finding the putative binding site of allosteric modulators using the "blind docking" approach. *Journal of Molecular Modeling*, 12(3):366–372, 2006.

[57] T. Konuma, Y.-H. Lee, Y. Goto, and K. Sakurai. Principal component analysis of chemical shift perturbation data of a multiple-ligand-binding system for elucidation of respective binding mechanism. *Proteins*, 81:107–118, 2013.

[58] Md Zahid Kamal, Jamshaid Ali, and Nalam Madhusudhana Rao. Binding of bis-ANS to Bacillus subtilis lipase: a combined computational and experimental investigation. *Biochimica et biophysica acta*, 1834(8):1501—1509, August 2013.

[59] G. Weber and L. B. Young. Fragmentation of bovine serum albumin by pepsin. *The Journal of Biological Chemistry*, 239(5):1415—1423, 1964.

[60] A. J. Shaka, P. B. Barker, and R. J. Freeman. Computer-optimized decoupling scheme for wideband applications and low-level operation. *Journal of Magnetic Resonance*, 64(547—552), 1985.

[61] F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfiefer, and A. Bax. Nmrpipe: A multidimensional spectral processing system based on unix pipes. *Journal of Biomolecular NMR*, 6:277—293, 1995.

[62] W. F. Vranken, W. Boucher, T. J. Stevens, R. H. Fogh, A. Pjon, M. Llinas, E. L. Ulrich, J. L. Markley, J. Ionides, and E. D. Laue. The ccpn data model for nmr spectroscopy: Development of a software pipeline. *Proteins*, 59(4):687—696, 2005.

[63] Garrett M. Morris, Ruth Huey, William Lindstrom, Michel F. Sanner, Richard K. Belew, David S. Goodsell, and Arthur J. Olson. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry*, 30(16):2785–2791, December 2009.

[64] Johann Gasteiger and Mario Marsili. A new model for calculating atomic charges in molecules. *Tetrahedron Letters*, 19(34):3181–3184, 1978.

[65] Oleg Trott and Arthur J. Olson. Software news and update AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 2010.

[66] I. A. Mills, S. L. Flaugh, M. S. Kosinski-Collins, and J. A. King. Folding and stability of the isolated Greek key domains of the long-lived human lens proteins $\gamma$D-crystallin and $\gamma$S-crystallin. *Protein Science*, 16(11):2427–2444, 2007.

[67] Benjamin G. Mohr, Cassidy M. Dobson, Scott C. Garman, and Murugappan Muthukumar. Electrostatic origin of in vitro aggregation of human $\gamma$-crystallin. *The Journal of Chemical Physics*, 139(12):121914, sep 2013.

[68] Usha P. Andley. Crystallins in the eye: Function and pathology. *Progress in Retinal and Eye Research*, 26(1):78–98, jan 2007.

[69] Hans Bloemendal, Wilfried de Jong, Rainer Jaenicke, Nicolette H Lubsen, Christine Slingsby, and Annette Tardieu. Ageing and vision: structure, stability and function of lens crystallins. *Progress in biophysics and molecular biology*, 86(3):407–85, nov 2004.

[70] J. Fielding Hejtmancik. Congenital cataracts and their molecular genetics. *Seminars in Cell and Developmental Biology*, 19(2):134–149, 2008.

[71] Kate L. Moreau and Jonathan A. King. Cataract-Causing Defect of a Mutant $\gamma$-Crystallin Proceeds through an Aggregation Pathway Which Bypasses Recognition by the $\alpha$-Crystallin Chaperone. *PLoS ONE*, 7(5):e37256, may 2012.

[72] Veniamin N Lapko, Andrew G Purkiss, David L. Smith, and Jean B Smith. Deamidation in Human $\gamma$S-Crystallin from Cataractous Lenses Is Influenced by Surface Exposure. *Biochemistry*, 41(27):8638–8648, jul 2002.

[73] Li Xiao and Barry Honig. Electrostatic contributions to the stability of hyperthermophilic proteins. *Journal of Molecular Biology*, 289(5):1435–1444, jun 1999.

[74] Eugene Serebryany, Jaie C. Woodard, Bharat V. Adkar, Mohammed Shabab, Jonathan A. King, and Eugene I. Shakhnovich. An Internal Disulfide Locks a Misfolded Aggregation-prone Intermediate in Cataract-linked Mutants of Human $\gamma$D-Crystallin. *Journal of Biological Chemistry*, 291(36):19172–19183, 2016.

[75] Kyle W. Roskamp, David M. Montelongo, Chelsea D. Anorma, Diana N. Bandak, Janine A. Chua, Kurtis T. Malecha, and Rachel W. Martin. Multiple Aggregation Pathways in Human $\gamma$S-Crystallin and Its Aggregation-Prone G18V Variant. *Investigative Opthalmology & Visual Science*, 58(4):2397, apr 2017.

[76] Jinwon Jung, In-Ja L. Byeon, Yongting Wang, Jonathan King, and Angela M. Gronenborn. The Structure of the Cataract-Causing P23T Mutant of Human $\gamma$D-Crystallin Exhibits Distinctive Local Conformational and Dynamic Changes. *Biochemistry*, 48(12):2597–2609, mar 2009.

[77] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, 1983.

[78] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, feb 1996.

[79] James C Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D Skeel, Laxmikant Kalé, and Klaus Schulten. Scalable molecular dynamics with NAMD. *Journal of computational chemistry*, 26(16):1781–802, dec 2005.

[80] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh Ewald: An N log( N ) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, jun 1993.

[81] Ulrich Essmann, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen. A smooth particle mesh Ewald method. *The Journal of Chemical Physics*, 103(19):8577–8593, 1995.

[82] H. Grubmüller, H. Heller, A. Windemuth, and K. Schulten. Generalized Verlet Algorithm for Efficient Molecular Dynamics Simulations with Long-range Interactions. *Molecular Simulation*, 6(1-3):121–142, mar 1991.

[83] Jean Paul Ryckaert, Giovanni Ciccotti, and Herman J.C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, 1977.

[84] Shuichi Miyamoto and Peter A. Kollman. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of Computational Chemistry*, 13(8):952–962, 1992.

[85] Scott E. Feller, Yuhong Zhang, Richard W. Pastor, and Bernard R. Brooks. Constant pressure molecular dynamics simulation: The Langevin piston method. *The Journal of Chemical Physics*, 103(11):4613–4621, 1995.

[86] Glenn J Martyna, Douglas J Tobias, and Michael L Klein. Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics*, 101(5):4177–4189, sep 1994.

[87] Ross A. Lippert, Cristian Predescu, Douglas J. Ierardi, Kenneth M. Mackenzie, Michael P. Eastwood, Ron O. Dror, and David E. Shaw. Accurate and efficient integration for molecular dynamics simulations at constant temperature and pressure. *Journal of Chemical Physics*, 139(16), 2013.

[88] Yibing Shan, John L. Klepeis, Michael P. Eastwood, Ron O. Dror, and David E. Shaw. Gaussian split Ewald: A fast Ewald mesh method for molecular simulation. *Journal of Chemical Physics*, 122(5), 2005.

[89] Glenn J Martyna, Michael L Klein, and Mark Tuckerman. NoséHoover chains: The canonical ensemble via continuous dynamics. *The Journal of Chemical Physics*, 97(4):2635–2643, aug 1992.

[90] Paolo Mereghetti, Razif R. Gabdoulline, and Rebecca C. Wade. Brownian Dynamics Simulation of Protein Solutions: Structural and Dynamical Properties. *Biophysical Journal*, 99(11):3782–3791, dec 2010.

[91] Michael Martinez, Neil J. Bruce, Julia Romanowska, Daria B. Kokh, Musa Ozboyaci, Xiaofeng Yu, Mehmet Ali Öztürk, Stefan Richter, and Rebecca C. Wade. SDA 7: A modular and parallel implementation of the simulation of diffusional association software. *Journal of Computational Chemistry*, 36(21):1631–1645, aug 2015.

[92] Razif R Gabdoulline and Rebecca C Wade. On the Contributions of Diffusion and Thermal Activation to Electron Transfer between Phormidium laminosum Plastocyanin and Cytochrome f : Brownian Dynamics Simulations with Explicit Modeling of Nonpolar Desolvation Interactions and Electron Transfer Even. *Journal of the American Chemical Society*, 131(26):9230–9238, jul 2009.

[93] Adrian H. Elcock, Razif R. Gabdoulline, Rebecca C. Wade, and J.Andrew McCammon. Computer simulation of protein-protein association kinetics: acetylcholinesterase-fasciculin. *Journal of Molecular Biology*, 291(1):149–162, sep 1999.

[94] William L. Jorgensen and Julian Tirado-Rives. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, 110(6):1657–1666, mar 1988.

[95] Jeffry D. Madura, James M. Briggs, Rebecca C. Wade, Malcolm E. Davis, Brock A. Luty, Andrew Ilin, Jan Antosiewicz, Michael K. Gilson, Babak Bagheri, L. Ridgway Scott, and J. Andrew McCammon. Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program. *Computer Physics Communications*, 91(1-3):57–95, 1995.

[96] Vera Prytkova, Matthias Heyden, Domarin Khago, J. Alfredo Freites, Carter T. Butts, Rachel W. Martin, and Douglas J. Tobias. Multi-Conformation Monte Carlo: A Method for Introducing Flexibility in Efficient Simulations of Many-Protein Systems. *The Journal of Physical Chemistry B*, 120(33):8115–8126, aug 2016.

[97] Nicholas Metropolis and S Ulam. The Monte Carlo Method. *Journal of the American Statistical Association*, 44(247):335, sep 1949.

[98] Elena Papaleo, Matteo Tiberti, Gaetano Invernizzi, Marco Pasi, and Valeria Ranzani. Molecular Determinants of Enzyme Cold Adaptation: Comparative Structural and Computational Studies of Cold- and Warm-Adapted Enzymes. *Current Protein & Peptide Science*, 12(7):657–683, nov 2011.

[99] Sandra Maguid, Sebastian Fernandez-Alberti, and Julian Echave. Evolutionary conservation of protein vibrational dynamics. *Gene*, 422(1-2):7–13, 2008.

[100] Sandra Maguid, Sebastián Fernández-Alberti, Gustavo Parisi, and Julián Echave. Evolutionary Conservation of Protein Backbone Flexibility. *Journal of Molecular Evolution*, 63(4):448–457, oct 2006.

[101] Elena Papaleo, Marco Pasi, Laura Riccardi, Ilaria Sambi, Piercarlo Fantucci, and Luca De Gioia. Protein flexibility in psychrophilic and mesophilic trypsins. Evidence of evolutionary conservation of protein dynamics in trypsin-like serine-proteases. *FEBS Letters*, 582(6):1008–1018, 2008.

[102] A. Hoyoux, I. Jennes, P. Dubois, S. Genicot, F. Dubail, J. M. Francois, E. Baise, G. Feller, and C. Gerday. Cold-Adapted -Galactosidase from the Antarctic Psychrophile Pseudoalteromonas haloplanktis. *Applied and Environmental Microbiology*, 67(4):1529–1535, apr 2001.

[103] S. Davail, G. Feller, E. Narinx, and C. Gerday. Cold adaptation of proteins. Purification, characterization, and sequence of the heat-labile subtilisin from the antarctic psychrophile Bacillus TA41. *The Journal of biological chemistry*, 269(26):17448–53, jul 1994.

[104] H Kobori, C W Sullivan, and H Shizuya. Heat-labile alkaline phosphatase from Antarctic bacteria: Rapid 5' end-labeling of nucleic acids. *Proceedings of the National Academy of Sciences of the United States of America*, 81(21):6691–6695, 1984.

[105] Glenn C. Johns and George N. Somero. Evolutionary Convergence in Adaptation of Proteins to Temperature: A 4-Lactate Dehydrogenases of Pacific Damselfishes (Chromis spp.). *Molecular Biology and Evolution*, 21(2):314–320, 2004.

[106] Kaare Teilum, Johan G. Olsen, and Birthe B. Kragelund. Protein stability, flexibility and function. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1814(8):969–976, 2011.

[107] Alexandre Cipolla, Salvino D'Amico, Roya Barumandzadehs, André Matagnes, and Georges Feller. Stepwise adaptations to low temperature as revealed by multiple mutants of psychrophilic $\alpha$-amylase from antarctic bacterium. *Journal of Biological Chemistry*, 286(44):38348–38355, 2011.

[108] Salvino D'Amico, Charles Gerday, and Georges Feller. Temperature Adaptation of Proteins: Engineering Mesophilic-like Activity and Stability in a Cold-adapted $\alpha$-Amylase. *Journal of Molecular Biology*, 332(5):981–988, oct 2003.

[109] Georges Feller. Psychrophilic Enzymes: From Folding to Function and Biotechnology. *Scientifica*, 2013:512840, jan 2013.

[110] Andrey Karshikoff, Lennart Nilsson, and Rudolf Ladenstein. Rigidity versus flexibility: The dilemma of understanding protein thermal stability. *FEBS Journal*, 282(20):3899–3917, 2015.

[111] Anni Linden and Matthias Wilmanns. Adaptation of Class-13 $\alpha$-Amylases to Diverse Living Conditions. *ChemBioChem*, 5(2):231–239, feb 2004.

[112] C Vetriani, D L Maeder, N Tolliday, K. S.-P. Yip, T J Stillman, K L Britton, D W Rice, H H Klump, and F T Robb. Protein thermostability above 100 C: A key role for ionic interactions. *Proceedings of the National Academy of Sciences*, 95(21):12300–12305, oct 1998.

[113] Andrey Karshikoff and Rudolf Ladenstein. Ion pairs and the thermotolerance of proteins from hyperthermophiles: a traffic rule' for hot roads. *Trends in Biochemical Sciences*, 26(9):550–557, sep 2001.

[114] Jie Chen, Huimin Yu, Changchun Liu, Jie Liu, and Zhongyao Shen. Improving stability of nitrile hydratase by bridging the salt-bridges in specific thermal-sensitive regions. *Journal of Biotechnology*, 164(2):354–362, 2012.

[115] Lilja B. Jónsdóttir, Brynjar Ö. Ellertsson, Gaetano Invernizzi, Manuela Magnúsdóttir, Sigríur H. Thorbjarnardóttir, Elena Papaleo, and Magnús M. Kristjánsson. The role of salt bridges on the temperature adaptation of aqualysin I, a thermostable subtilisin-like proteinase. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1844(12):2174–2181, dec 2014.

[116] C. Nick Pace, Hailong Fu, Katrina Lee Fryar, John Landua, Saul R. Trevino, Bret A. Shirley, Marsha McNutt Hendricks, Satoshi Iimura, Ketan Gajiwala, J. Martin Scholtz, and Gerald R. Grimsley. Contribution of Hydrophobic Interactions to Protein Stability. *Journal of Molecular Biology*, 408(3):514–528, may 2011.

[117] U. Deva Priyakumar. Role of Hydrophobic Core on the Thermal Stability of ProteinsMolecular Dynamics Simulations on a Single Point Mutant of Sso7d. *Journal of Biomolecular Structure and Dynamics*, 29(5):961–971, apr 2012.

[118] Mireille Delaye and Annette Tardieu. Short-range order of crystallin proteins accounts for eye lens transparency. *Nature*, 302(5907):415–417, mar 1983.

[119] R J Siezen, M R Fisch, C Slingsby, and G B Benedek. Opacification of gamma-crystallin solutions from calf lens in relation to cold cataract formation. *Proceedings of the National Academy of Sciences of the United States of America*, 82(6):1701–5, 1985.

[120] Andor J. Kiss, Amir Y Mirarefi, Subramanian Ramakrishnan, Charles F Zukoski, Arthur L Devries, and Chi-Hing C Cheng. Cold-stable eye lens crystallins of the Antarctic nototheniid toothfish Dissostichus mawsoni Norman. *The Journal of experimental biology*, 207(Pt 26):4633–49, dec 2004.

[121] Megha H. Unhelkar, Vy T. Duong, Kaosoluchi N. Enendu, John E. Kelly, Seemal Tahir, Carter T. Butts, and Rachel W. Martin. Structure prediction and network analysis of chitinases from the Cape sundew, Drosera capensis. *Biochimica et Biophysica Acta - General Subjects*, 1861(3):636–643, 2017.

[122] Juan Alfredo Freites, Xuhong Zhang, Eric K Wong, Rachel W Martin, Douglas J Tobias, and Carter T Butts. k-Cores in Dynamic Networks as a Tool for Analysis of Protein Structure Cohesiveness. In preparation.

[123] Carolyn N Kingsley, Jan C Bierma, Vyvy Pham, and Rachel W Martin. $\gamma$S-Crystallin Proteins from the Antarctic Nototheniid Toothfish: A Model System for Investigating Differential Resistance to Chemical and Thermal Denaturation. *The Journal of Physical Chemistry B*, 118(47):13544–13553, nov 2014.

[124] Noah C Benson and Valerie Daggett. A chemical group graph representation for efficient high-throughput analysis of atomistic protein simulations. *Journal of bioinformatics and computational biology*, 10(4):1250008, aug 2012.

[125] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[126] Julie D. Thompson, Desmond G. Higgins, and Toby J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680, 1994.

[127] A M Stadler, C J Garvey, A Bocahut, S Sacquin-Mora, I Digel, G J Schneider, F Natali, G M Artmann, and G Zaccai. Thermal fluctuations of haemoglobin from different species: adaptation to temperature via conformational dynamics. *Journal of The Royal Society Interface*, 9(76):2845–2855, nov 2012.

[128] Amber Goodchild, Neil F W Saunders, Haluk Ertan, Mark Raftery, Michael Guilhaus, Paul M G Curmi, and Ricardo Cavicchioli. A proteomic determination of cold adaptation in the Antarctic archaeon, Methanococcoides burtonii. *Molecular Microbiology*, 53(1):309–321, may 2004.

[129] Sepideh Parvizpour, Jafar Razmara, Mohd Shahir Shamsir, Rosli Md Illias, and Abdul Munir Abdul Murad. The role of alternative salt bridges in cold adaptation of a novel psychrophilic laminarinase. *Journal of Biomolecular Structure and Dynamics*, 1102(August):1–18, 2016.

[130] W.-T. Li, John W. Shriver, and John N. Reeve. Mutational Analysis of Differences in Thermostability between Histones from Mesophilic and Hyperthermophilic Archaea. *Journal of Bacteriology*, 182(3):812–817, feb 2000.

[131] S Parthasarathy and M R Murthy. Protein thermal stability: insights from atomic displacement parameters (B values). *Protein engineering*, 13(1):9–13, jan 2000.

[132] K.L Britton, K.S.P Yip, S.E Sedelnikova, T.J Stillman, M.W.W Adams, K. Ma, D.L Maeder, F.T Robb, N. Tolliday, C. Vetriani, D.W Rice, and P.J Baker. Structure determination of the glutamate dehydrogenase from the hyperthermophile Thermococcus litoralis and its comparison with that from Pyrococcus furiosus 1 1Edited by R. Huber. *Journal of Molecular Biology*, 293(5):1121–1132, nov 1999.

[133] A. V. Weigel, B. Simon, M. M. Tamkun, and D. Krapf. Ergodic and nonergodic processes coexist in the plasma membrane as observed by single-molecule tracking. *Proceedings of the National Academy of Sciences*, 108(16):6438–6443, 2011.

[134] Jae-Hyung Jeon, Eli Barkai, and Ralf Metzler. Noisy continuous time random walks. *The Journal of Chemical Physics*, 139(12):121916, sep 2013.

[135] Jean Philippe Bouchaud and Antoine Georges. Anomalous diffusion in disordered media: Statistical mechanisms, models and physical applications. *Physics Reports*, 195(4-5):127–293, 1990.

[136] Xiaohu Hu, Liang Hong, Micholas Dean Smith, Thomas Neusius, Xiaolin Cheng, and Jeremy C. Smith. The dynamics of single protein molecules is non-equilibrium and self-similar over thirteen decades in time. *Nature Physics*, 12(2):171–174, nov 2015.

[137] Haw Yang, Guobin Luo, Pallop Karnchanaphanurach, Tai-Man Louie, Ivan Rech, Sergio Cova, Luying Xun, and X Sunney Xie. Protein conformational dynamics probed by single-molecule electron transfer. *Science (New York, N.Y.)*, 302(5643):262–6, 2003.

[138] Wei Min, Guobin Luo, Binny J. Cherayil, S. C. Kou, and X. Sunney Xie. Observation of a power-law memory Kernel for fluctuations within a single protein molecule. *Physical Review Letters*, 94(19):1–4, 2005.

[139] Yasmine Meroz, Igor M. Sokolov, and Joseph Klafter. Test for determining a subdiffusive model in ergodic systems from single trajectories. *Physical Review Letters*, 110(9):1–4, 2013.

[140] Yasmine Meroz and Igor M. Sokolov. A toolbox for determining subdiffusive mechanisms. *Physics Reports*, 573:1–29, 2015.

[141] Jae-Hyung Jeon, Aleksei V Chechkin, and Ralf Metzler. Scaled Brownian motion: a paradoxical process with a time dependent diffusivity for the description of anomalous diffusion. *Phys. Chem. Chem. Phys.*, 16(30):15811–7, 2014.

[142] Jae Hyung Jeon, Natascha Leijnse, Lene B. Oddershede, and Ralf Metzler. Anomalous diffusion and power-law relaxation of the time averaged mean squared displacement in worm-like micellar solutions. *New Journal of Physics*, 15, 2013.

[143] Marcin Magdziarz and Aleksander Weron. Anomalous diffusion: Testing ergodicity breaking in experimental data. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 84(5):1–5, 2011.

[144] I. Y. Wong, M. L. Gardel, D. R. Reichman, Eric R. Weeks, M. T. Valentine, A. R. Bausch, and D. A. Weitz. Anomalous diffusion probes microstructure dynamics of entangled F-actin networks. *Physical Review Letters*, 92(17):30–33, 2004.

[145] Yasmine Meroz, Victor Ovchinnikov, and Martin Karplus. Coexisting origins of subdiffusion in internal dynamics of proteins. *Physical Review E*, 95(6):062403, jun 2017.

[146] Jae Hyung Jeon, Hector Martinez Seara Monne, Matti Javanainen, and Ralf Metzler. Anomalous diffusion of phospholipids and cholesterols in a lipid bilayer and its origins. *Physical Review Letters*, 109(18):1–5, 2012.

[147] Vincent Tejedor, Olivier Bénichou, Raphael Voituriez, Ralf Jungmann, Friedrich Simmel, Christine Selhuber-Unkel, Lene B. Oddershede, and Ralf Metzler. Quantitative Analysis of Single Particle Trajectories: Mean Maximal Excursion Method. *Biophysical Journal*, 98(7):1364–1372, apr 2010.

[148] A. Rahman. Correlations in the Motion of Atoms in Liquid Argon. *Physical Review*, 136(2A):A405–A411, oct 1964.

[149] H Frauenfelder, S. Sligar, and P. Wolynes. The energy landscapes and motions of proteins. *Science*, 254(5038):1598–1603, dec 1991.

[150] S Burov, R Metzler, and E. Barkai. Aging and nonergodicity beyond the Khinchin theorem. *Proceedings of the National Academy of Sciences*, 107(30):13228–13233, jul 2010.

[151] Gregory L. Warren, C. Webster Andrews, Anna-Maria Capelli, Brian Clarke, Judith LaLonde, Millard H. Lambert, Mika Lindvall, Neysa Nevins, Simon F. Semus, Stefan Senger, Giovanna Tedesco, Ian D. Wall, James M. Woolven, Catherine E. Peishoff, and Martha S. Head. A Critical Assessment of Docking Programs and Scoring Functions. *Journal of Medicinal Chemistry*, 49(20):5912–5931, oct 2006.

[152] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, mar 2008.

# Appendix A

# Supporting information for ANS Docking to the $\gamma$S-G18V

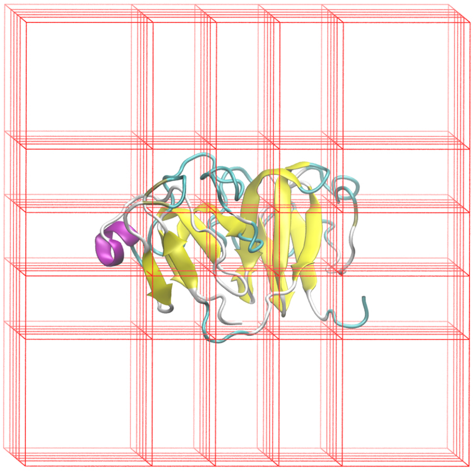

Figure A.1: Search space boundaries used in initial rigid receptor docking to NMR conformations of $\gamma$S-crystallin. A total of 27 docking runs were performed for each NMR structure. Search spaces were placed in a 3 x 3 x 3 grid with a 10 (Å) overlap between search spaces. Boxes were sized so enough space was available to sample ligand conformations on the entire surface of the protein.

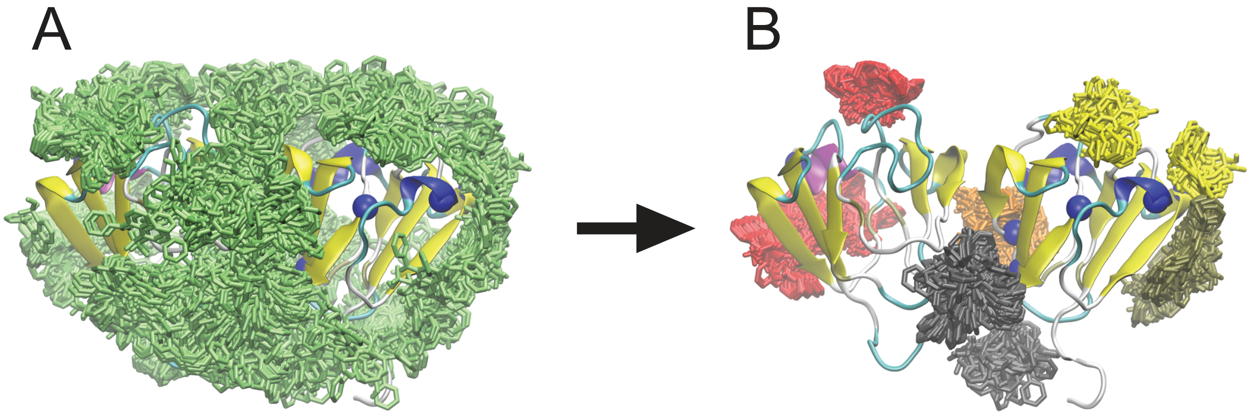

Figure A.2: Clustering of rigid docking results to define flexible binding sites ($\gamma$S-WT shown as example). (A) Set of poses containing both hydrophobic and electrostatic contacts necessary for fluorescence. The resulting set covers nearly the entire surface of the protein. (B) Clustering of the pose set resulted in 20 binding sites (several of the clusters were overlapping). Pose clusters near highly perturbed residues (shown with blue VDW spheres) were picked to define search spaces for flexible docking. Clusters are color coordinated for each search space.

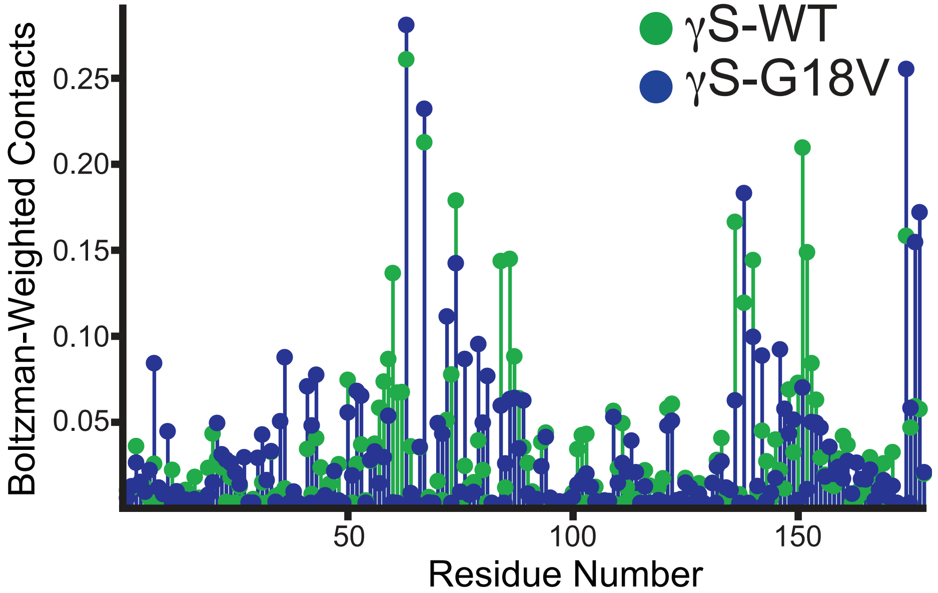

Figure A.3: ANS-residue contact frequencies for $\gamma$S-WT (green) and $\gamma$S-G18V (blue) from docking simulations. Although non-specific binding is observed for both proteins, the contact frequencies show more ANS binding for $\gamma$S-G18V than $\gamma$S-WT, with maximum binding localized near the interdomain interface

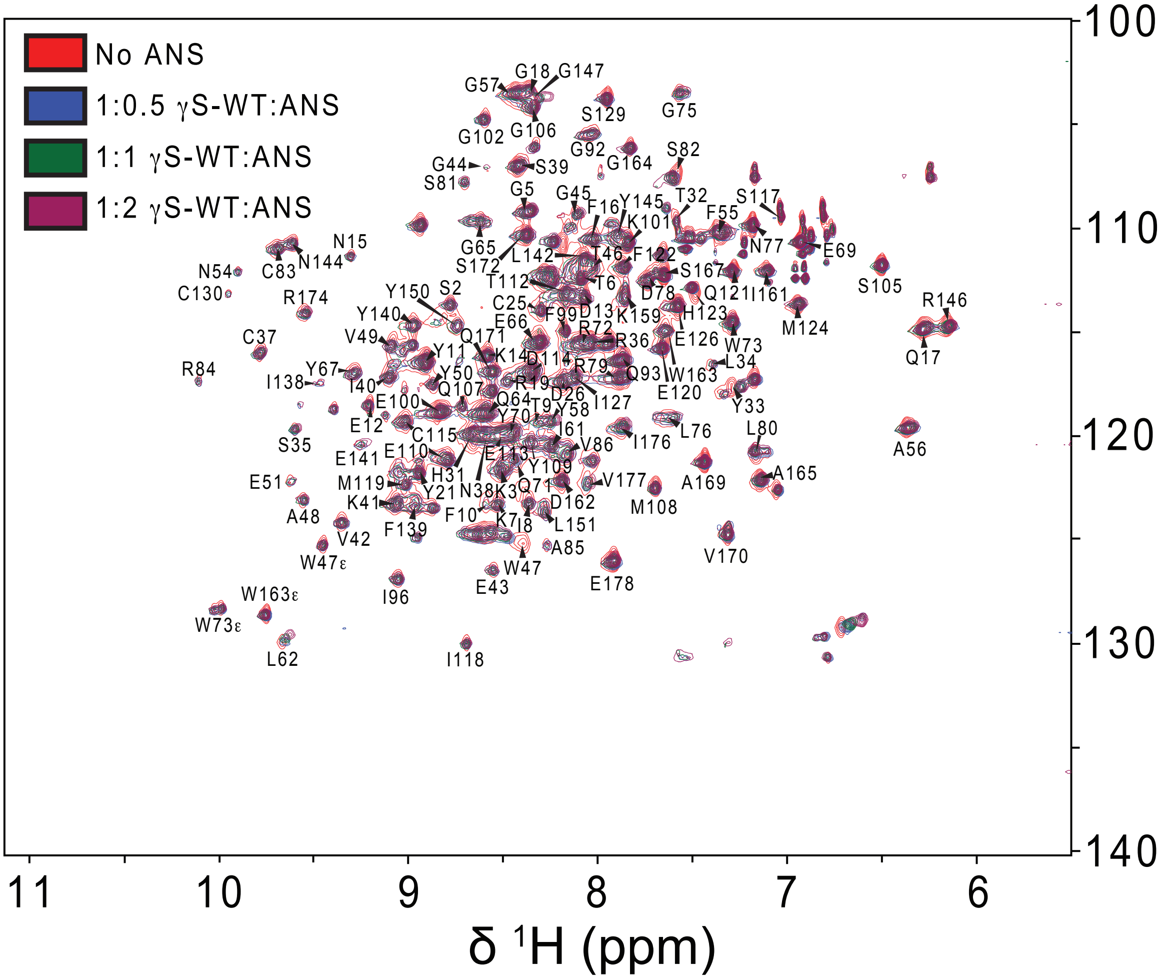

Figure A.4: $^1$H-$^{15}$N HSQC spectra of $^{15}$N labelled $\gamma$S-WT with increasing concentrations of ANS. Ratios of $\gamma$S:ANS were at 1:0, 1:0.5, 1:1, and 1:2 where the concentration of protein was approximately 0.3 mM. Spectra were acquired at 25 °C. Residues were assigned based on previous assignments of $\gamma$S-WT [21].

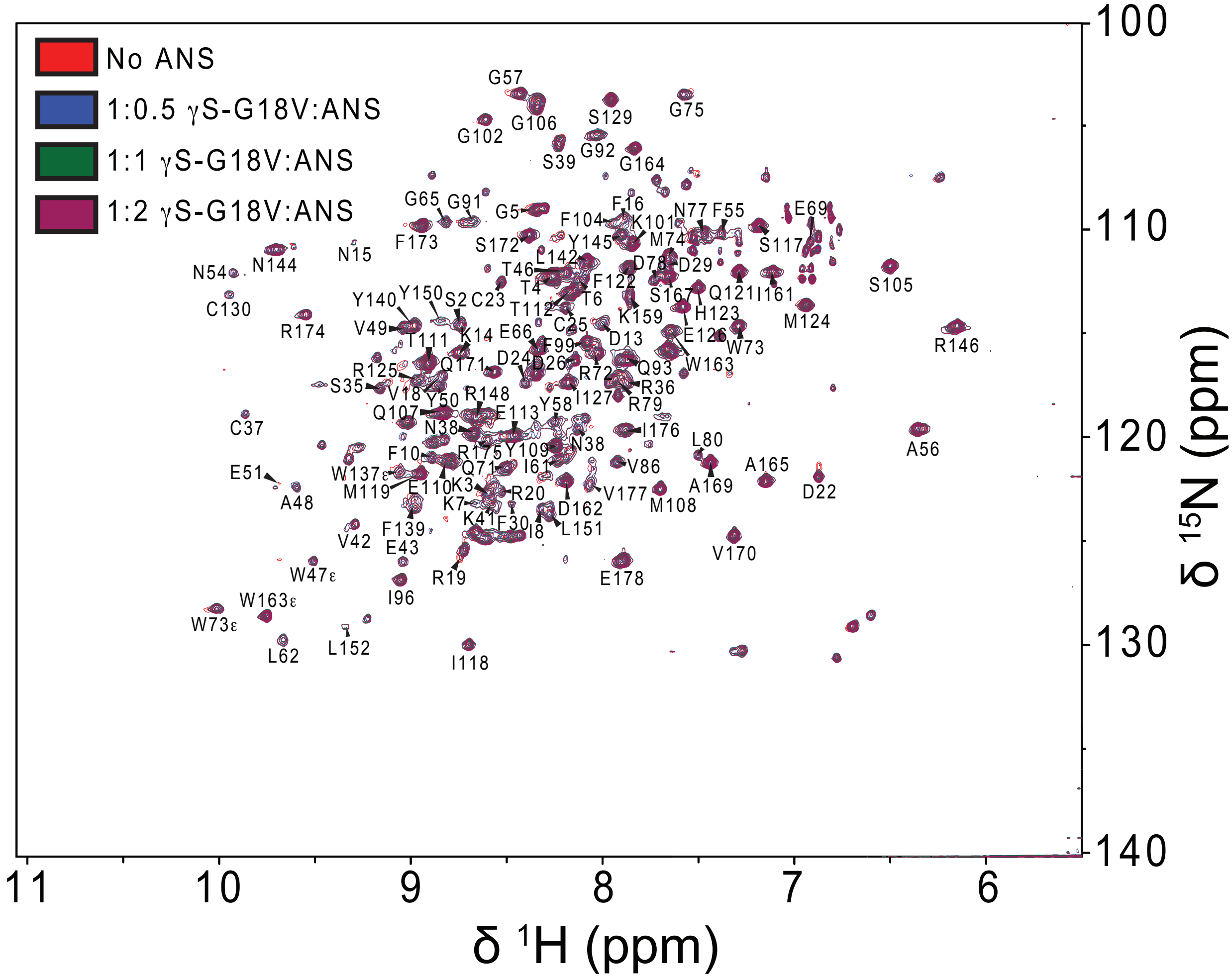

Figure A.5: $^1$H-$^{15}$N HSQC spectra of $^{15}$N labelled $\gamma$S-G18V with increasing concentrations of ANS. Ratios of $\gamma$S:ANS were at 1:0, 1:0.5, 1:1, and 1:2 where the concentration of protein was approximately 0.3 mM. Spectra were acquired at 25 °C. Residues were assigned based on previous assignments of $\gamma$S-G18V [21].

# Appendix B

# Overview of the Docking Workflow

## B.1 Background

Molecular docking can provide important structural information on bound ligands to known structures of receptors. This method often employs a simple energy function and search algorithm in order to quickly generate and rank bound conformations, often for the purpose of processing large libraries of compounds and/or receptors. Though some scoring functions provide binding energies for their poses, the standard error is too large to be predictive[65, 151]. Furthermore, benchmarks of several docking programs found that the scoring functions can reproduce co-crystallographic conformations within a set of top ranked poses, but are unable to identify the correct structure as the top scored pose[151]. Thus, the scoring functions serves best as a metric to rank the most plausible poses generated from the search algorithm, rather than pinpointing the exact binding mode. For this purpose, molecular docking serves as a fast, yet qualitative method to generate sets of top-ranked poses for further inquiry.

In this work, I seek the identify the protein-ligand binding sites. Since virtual screening efforts usually involve screening a particular active site, care should be taken when blindly docking to an entire protein. To address this, I incorporated a docking protocol involving two stages of docking with two stages of filtering to cluster and remove poses based on prior knowledge from experiments. An outline of the docking protocol is illustrated in Figure B.1.

## B.2 Blind docking

For this docking protocol, Autodock Vina[65] was chosen for its relatively fast operation. In the first step, docking simulation search spaces were distributed across the protein's entire surface. Since the search algorithm becomes much less effective with larger search spaces, a grid of 27 smaller 30 x 30 x 30 Å search spaces were used. The second step involves running the docking simulation for each search space, each of the 20 superimposed NMR conformations, and each protein variant, resulting in 1,080 runs for step 2. Since the goal of blind docking is to populate the binding sites for clustering, each run was configured to report the top 20 binding modes, rather than the default top 9 binding modes.

## B.3 Define Binding Sites by Clustering

In the third step, the collective pose set is clustered (using RMSD clustering in VMD[78]). The chosen RMSD cutoff of 5 Å was sufficiently large enough to cluster together poses with different orientations, but still small enough to distinguish between binding sites. The resulting 50+ clusters are then filtered for interactions with strongly binding CSP residues. Clusters with no heavy atoms within 8 Å (the interaction cutoff of the Vina scoring function[65]) of a strong binding residues are removed from the set. The remaining clusters form the new search spaces for a flexible refinement.

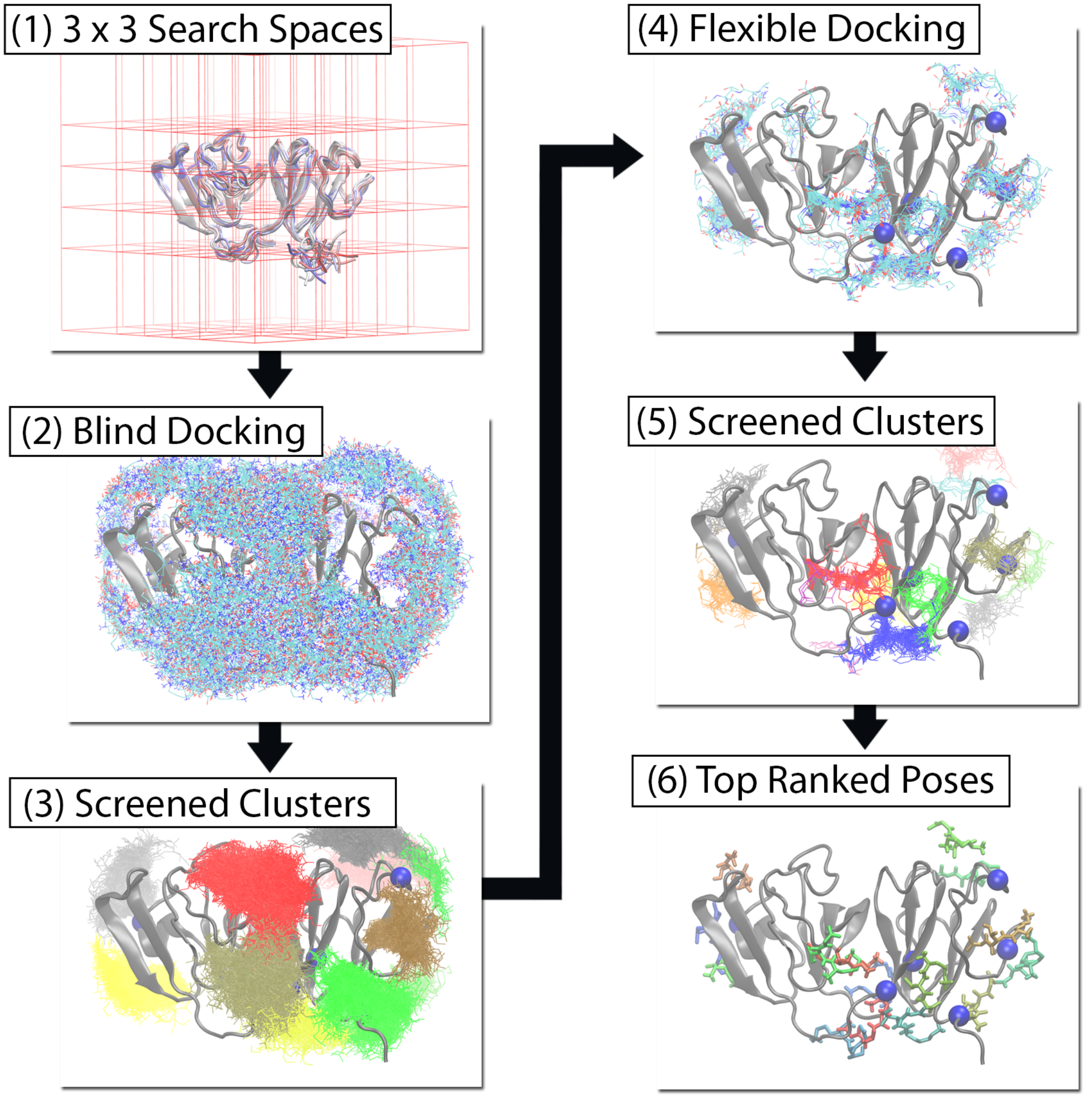

Figure B.1: Illustration of the docking workflow used to identify ANS binding sites. 1) The 27 search spaces shown in red outline around 21 superimposed NMR conformations. 2) Collective set of docked poses to a rigid protein. 3) Resulting pose clusters after filtering for contacts. Strong binding residues from CSP mapping are shown in blue spheres. 4) Docked pose set bound to a flexible protein. 5) Docked poses from flexible receptor docking. 6) The best scored poses for each binding site. These poses are then visually inspected for their contacts.

## B.4 Flexible Refinement and Post-Filtering

For flexible docking, residues with strong and weak binding CSP values were set as flexible, and the ligand is redocked to the flexible protein. The pose set is then clustered and refiltered for contacts with strong binding residues. The set is further filtered for the contacts necessary for fluorescence. This requires polar interactions with the ANS sulfonate and hydrophobic interactions with the conjugated rings[44]. The final set of poses include the top ranked poses for each binding site that satisfies all the criteria for a fluorescent pose that contributes to the observed CSP. Each pose is then individually inspected for differences in binding modes between $\gamma$S-WT and $\gamma$S-G18V. The three identified binding sites at the site of mutation, behind the cysteine loop, and at the interdomain interface have stronger binding with $\gamma$S-G18V directly resulting from changes in protein conformation linked to the G18V point mutation.

# Appendix C

# Supporting information for simulations of W42R H$\gamma$D

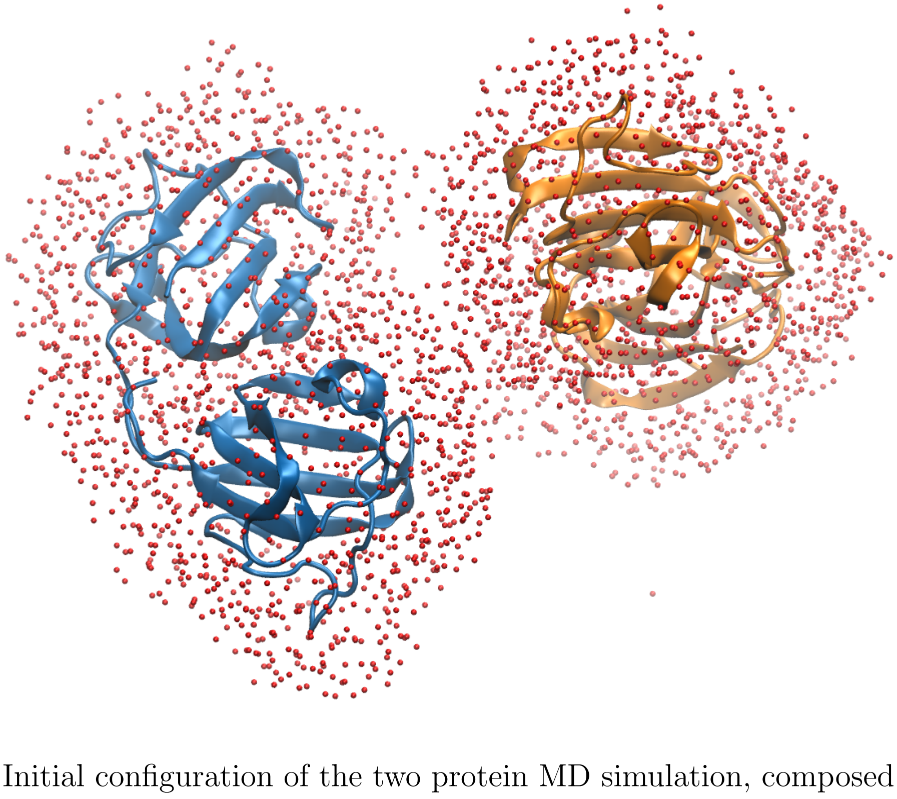

Figure C.1: Initial configuration of the two protein MD simulation, composed to two copies of W42R H$\gamma$D and their solvation shells (shown in cartoon and CPK representation, respectively) extracted from the single protein MD simulation. Proteins are placed farther than 9 Åfrom another such that the solvating waters are not overlapping.

Table C.1: List of residues with an increase in relative solvent accessibility ($\Delta$RSA) $> 0.13$. These residues are highlighted on snapshots of WT H$\gamma$D and W42R H$\gamma$D in figure 3.2. Hydrophobic residues that participate in strong interprotein interaction from MC simulations are shown in bold text.

| Residue No. | Residue Name | $\Delta$RSA |
|---|---|---|
| **53** | **LEU** | **0.282** |
| 54 | GLN | 0.305 |
| **56** | **PHE** | **0.294** |
| **69** | **MET** | **0.322** |
| **71** | **LEU** | **0.148** |
| 72 | SER | 0.203 |
| 81 | ILE | 0.186 |
| 140 | GLY | 0.254 |
| 141 | ARG | 0.230 |
| 142 | GLN | 0.497 |
| **144** | **LEU** | **0.432** |
| 167 | ARG | 0.170 |
| 169 | VAL | 0.241 |
| 170 | ILE | 0.339 |
| 171 | ASP | 0.345 |
| 172 | PHE | 0.134 |

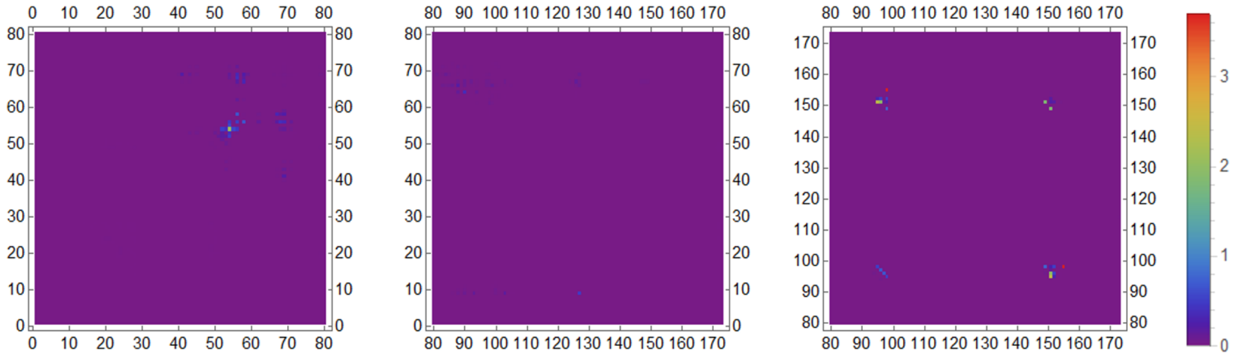

Figure C.2: Contact map of the two-protein MD simulation. Values are reported as # contacts/frame. From the left to right, N-N, N-C, and C-C interactions are plotted. Contacts were defined by a general distance cutoff of 3.5 Åbetween heavy atoms.

# Appendix D

# Supporting figures for thermal stability studies on H$\gamma$S, T$\gamma$S1 and $\gamma$S2

Table D.1: Residue frequency over 667 aligned $\gamma$-crystallin sequences. Residues appearing in either human or toothfish $\gamma$S-crystallins are highlighted in bold. Residues in parenthesis correspond to the residue at that position in H$\gamma$S, T$\gamma$S1, and T$\gamma$S2, respectively.

| | Residue position | | | |
|---|---|---|---|---|
| | 131 (KNV) | 103 (DND) | 159 (KKN) | 156 (EEE) |
| Ala (A) | - | 0.3% | 1.6% | - |
| Arg (R) | 2.7% | 0.1% | 34.9% | 0.1% |
| Asn (N) | **44.2%** | **40.6%** | **22.2%** | 2.4% |
| Asp (D) | 0.1% | **38.7%** | - | 15.1% |
| Cys (C) | 0.1% | - | 1.0% | 0.7% |
| Gln (Q) | 3.6% | - | 0.7% | 2.1% |
| Glu (E) | - | 5.1% | 0.1% | **66.9%** |
| Gly (G) | - | 0.3% | 0.9% | 0.1% |
| His (H) | 33.0% | - | 1.9% | 6.3% |
| Ile (I) | 0.3% | 0.1% | 0.1% | - |
| Leu (L) | - | 0.1% | 0.4% | 0.3% |
| Lys (K) | **9.7%** | 0.6% | **4.0%** | 0.1% |
| Met (M) | 0.9% | - | - | 0.3% |
| Phe (F) | - | 0.1% | 0.1% | - |
| Pro (P) | - | - | - | - |
| Ser (S) | 0.6% | 0.7% | 26.7% | 0.6% |
| Thr (T) | 0.6% | - | 1.0% | 0.1% |
| Trp (W) | - | - | - | 0.1% |
| Tyr (Y) | - | - | 0.1% | 0.4% |
| Val (V) | **2.2%** | - | 0.1% | - |

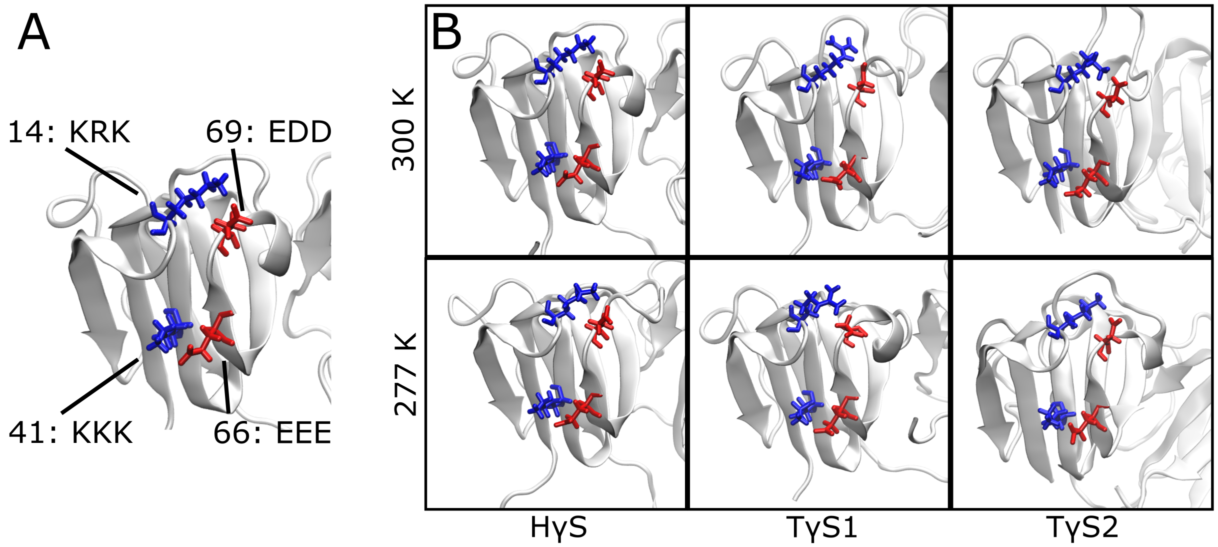

Figure D.1: A) Snapshot of the stabilizing salt bridges in the N-terminal domain of H$\gamma$S. Each residue is labeled with the H$\gamma$S residue number, and the residue name at that position for H$\gamma$S, T$\gamma$S1, and T$\gamma$S2, respectively. (B) Snapshots of the homologous residues to the stabilizing salt-bridges in H$\gamma$S. For both panels, residues are rendered in licorice representation and colored by residue type (Red: Acidic, Blue: Basic, Green: Polar, White: Hydrophobic). The protein backbone is rendered in cartoon representation and colored white

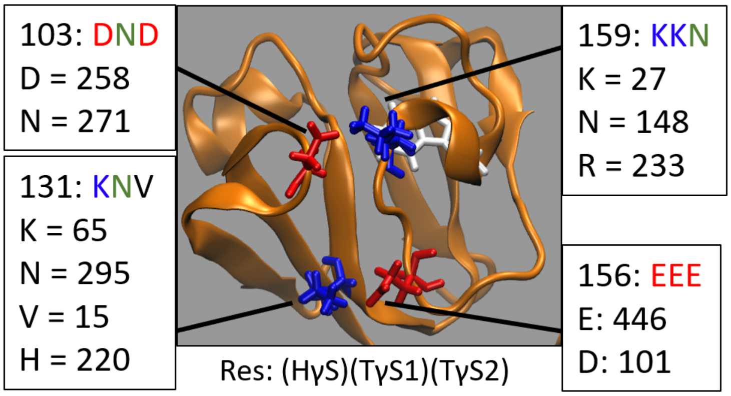

Figure D.2: Summary of residue frequencies over 667 $\gamma$-crystallin sequences. Frequencies are shown for four salt bridging residues in the C-terminal domain that bridge between regions of increased fluctuation in toothfish $\gamma$-crystallins at 300 K.

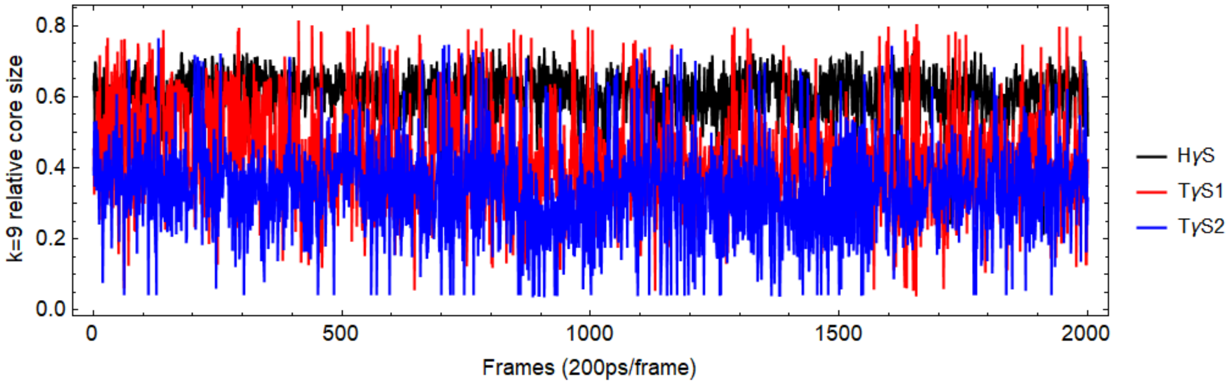

Figure D.3: Time evolution of the relative 9-core size for the last 400 ns of the MD simulation trajectory at 300 K.

# Appendix E

# Coarse-graining of interdomain motion using conformational clustering

In chapter 5, fluctuations in the interdomain distance of H$\gamma$D show a non-ergodic aging behavior as evidenced by the observation time dependence of the auto correlation function (ACF). One of the most well known models of non-ergodic motion is the continuous time random walk (CTRW) with divergent waiting times[140]. The noisy CTRW[134] has been analyzed for its analogy to non-ergodic motions in biological systems with thermal fluctuations. In the Conformational Cluster Transition Network (CCTN)[136] method, a molecular dynamics trajectory is coarse-grained in a CTRW between conformational clusters. This method assumes that structures contained within a cluster are close in the energy landscape, composing a local minimum. Therefore, the transition network describes a transition between wells in the energy landscape and conformations within a cluster describe fluctuations within a local minimum.

Conformational clustering was performed on 2 $\mu$s trajectories of H$\gamma$D using Gromacs 4.6[152]. Conformations were fitted and clustered using protein heavy atoms and an RMSD cutoff of 2.0 Å. After clustering, the medoid conformation of each cluster was calculated. With the cluster index known at each point in time, the time evolution of the cluster, the interdomain distance of the medoid conformation is reported vs. time and analyzed (Figure E.1).

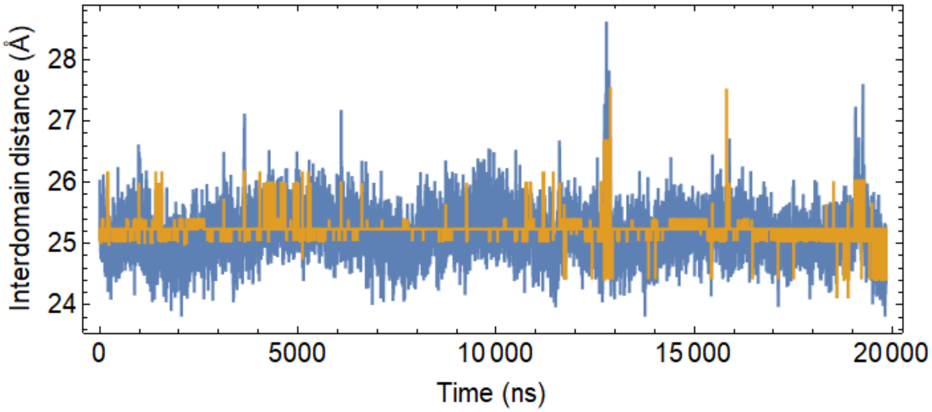

Figure E.1: Time evolution of the interdomain distance of a 2 $\mu$s single protein trajectory of H$\gamma$D. The distances from the coarse-grained and original trajectories are plotted in orange and blue, respectively.

Figure E.1 shows that, when the time trace is course grained, the protein motions are mainly short excursions from a central cluster. Similar analysis on phosphoglycerate kinase (PGK) (shown in the supporting figure S12a in the publication by Hu et al.[136]), shows transitions between distinct states that persist for microseconds (compared to H$\gamma$D, which shows short excursions from a single state). A possible explanation for this is that (PGK) is a three domain protein where the interdomain dynamics are measured between two separated, non-interacting domains, while H$\gamma$D measures the dynamics of two tightly bound domains with a hydrophobic interdomain interface. Exploring such differences may explain why the H$\gamma$D fluctuations begin to converge, while that of PGK continues to age.