Modelling pathological conformations in transthyretin amyloidosis

Matthew Carter Childers

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Valerie Daggett, Chair

Roland Strong

Wendy Thomas

Program Authorized to Offer Degree:

Bioengineering

University of Washington

Abstract

Modelling pathological conformations in transthyretin amyloidosis

Matthew Carter Childers

Chair of the Supervisory Committee:

Valerie Daggett, Professor

Bioengineering

The amyloidoses are a set of fatal disorders in which proteins aggregate and form fibrils that deposit in tissues throughout the body. Amyloid diseases are challenging to study as critical events in amyloid formation occur on timescales that span several orders of magnitude and involve heterogeneous, interconverting protein conformations. Consequently, there are few structural models of protein conformations associated with amyloid pathologies. The development of more effective technologies to diagnose and treat amyloid disease requires both a map of those conformations and an understanding of the molecular mechanisms that drive aggregation. Here, I present the results of simulations that benchmark the *il*mm molecular dynamics simulation package against experimental data; simulations that model the conformational changes to transthyretin monomers that are predicted to occur prior to aggregation; simulations of host-guest pentapeptides used to construct conformational libraries that have applications in protein design; and simulations of a redesigned transthyretin variant that is engineered to have reduced conformational stability.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am grateful to the members of my committee: Drs. Valerie Daggett, Lutz Maibaum, Roland Strong, and Wendy Thomas, whose time, commitment, and invaluable advice have propelled my work forward. I would also like to thank Clare Towse, Alissa Bleem, Dylan Shea, Tatum Prosswimmer, and my other peers in the Daggett Group for many inspiring conversations and enduring friendship. I am grateful for financial support provided by the Bioengineering Cardiovascular Training Grant, National Institutes of Health, and the UW Department of Bioengineering. This work was also supported by generous computational resources provided by the National Energy Research Scientific Computing Center, supported by the DOE Office of Biological Research, which is supported by the US Department of Energy under Contract DE-AC02-05CH11231 as well as the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation Grant ACI-1548562.

DEDICATION

To Mom and Dad for the sage advice and unconditional love that sparked an everlasting flame.

Chapter 1. BACKGROUND AND SIGNIFICANCE

1.1 Amyloid Disease

The hallmark of amyloid diseases is the deposition of insoluble fibrils, formed via the aggregation of proteins, in specific tissues throughout the body. During the fibril formation process, i.e. amyloidogenesis, proteins undergo conformational changes that enable them to aggregate and form cytotoxic intermediate species, i.e. soluble oligomers, large aggregates, and finally mature amyloid fibrils. Traditional biophysical techniques have been unable to resolve the structure(s) of the intermediate conformations formed during amyloidogenesis, which has impeded the development of therapeutic and diagnostic technologies to address amyloidosis. Therapeutically targeting this species is ideal because such a strategy would not disrupt native protein function, would target the cytotoxic conformations of amyloid proteins, and may yield insights into the development of therapeutics for diverse amyloid diseases.

1.2 Transthyretin

Human transthyretin (TTR, a.k.a. prealbumin or thyroxine-binding prealbumin) is a 55 kDa homotetrameric protein that contains four 127-residue monomers – each arranged in a β-sandwich topology (Blake *et al.*, 1971, 1978). Each monomer is comprised of eight β-strands arranged into two four-stranded β-sheets and an α-helix (Figure 1.1). The β-strands are sequentially named A-H and the corresponding sheets are referred to as the DAGH and CBEF sheets. A TTR dimer is formed when two dimers associate at strands H and H', leading to an 8-stranded β-sheet: DAGH-H'G'A'D'. Two dimers come together with the DAGH-H'G'A'D' sheets facing one another to form the thyroxine (T4) binding channel at the center of the four monomers. TTR primarily functions as a transporter of T4 in the serum and cerebrospinal fluid (CSF) and of retinol via an association with retinol binding protein (RBP) (Davis *et al.*, 1972; Jaarsveld *et al.*, 1973; Southwell *et al.*, 1992; Johnson *et al.*, 2012).

TTR is associated with several amyloid diseases (Sekijima, 2015), which are characterized by the misfolding and subsequent deposition of TTR into amyloid fibrils, and include senile systemic amyloidosis (SSA) (Yazaki and Higuchi, 2014), familial amyloid cardiomyopathy (FAC)

(Ruberg *et al.*, 2009), and familial amyloidotic polyneuropathy (FAP) (Planté-Bordeneuve *et al.*, 2013). These amyloidoses are distinguished based on the primary sequence of TTR as well as the specific tissues in the body where amyloid deposits occur (Sipe *et al.*, 2014). For example, SSA is caused by the misfolding and deposition of wild type (WT) TTR and affects ~25% of supra-octogenarians, although this figure has been disputed and the true prevalence may be higher due to misdiagnosis of TTR-associated amyloid (Tung-Chen and Arnau, 2018). In SSA, amyloid deposition can occur throughout the body, but primarily affects the heart, ligaments, and tenosynovium (Sipe *et al.*, 2016). In contrast, FAC and FAP are caused by misfolding and deposition of mutant forms of TTR and are distinguished based on the affected tissues in the body: in FAC, deposits are found in the heart; in FAP, deposits can be found in the nervous system and other tissues (Saraiva *et al.*, 1983; Saraiva *et al.*, 1984; Kelly, 1998).

Over 100 TTR single amino acid mutations associated with amyloidosis have been reported in the literature (Saraiva, 2001; Adams *et al.*, 2017); however, the clinical significance of many TTR mutants is not known as the WT variant forms amyloid and there have been few systematic studies of TTR mutations. In a study of the kinetic and thermodynamic stabilities of disease-associated TTR variants, Sekijima *et al.* show that mutations can modulate the kinetic and thermodynamic stabilities of TTR and that the tissue-specific secretion efficiency of TTR is correlated with a combination of both the kinetic and thermodynamic stabilities of TTR variants (Sekijima *et al.*, 2005). Based on these results, Sekijima *et al.* propose that competition between endoplasmic reticulum-associated degradation (ERAD) and endoplasmic reticulum-assisted folding (ERAF) controls the secretion efficiency of TTR in various tissues and thus underlies tissue-specific deposition of TTR amyloid. However, that study was limited to 23 mutations that are located in diverse regions of the TTR structure and the mutations had diverse effects on the dynamics of TTR aggregation; consequently, the extent to which mutations modulate the amyloid-forming propensity of TTR, the conformational states populated during amyloidogenesis, and the specific tissues in which TTR amyloid deposits occur have not been definitively established.

TTR is an ideal protein to model the molecular mechanisms of amyloidosis, as it has well-defined native structure and aggregation proceeds from a conformation that retains some characteristics of the native state (Figure 1.2). TTR amyloidogenesis has been primarily studied under an acid-mediated pathway, which is hypothesized to occur *in vivo* within the lysosomal compartment of the cell (Colon and Kelly, 1992; Lai *et al.*, 1996). The first, and rate-limiting, step

in the acid-mediated pathway of TTR amyloidogenesis is dissociation of the tetramer (Foss *et al.*, 2005). Under physiological conditions, tetramer dissociation is a slow process and the tetramer is the dominant species between pH 5 and 7 (Lai *et al.*, 1996). Aggregation of monomeric species occurs between pH 3.5-5; at lower pH, the tetramer dissociates but no fibrils form. Cross-linking and analytical centrifugation experiments establish that the monomer is the base aggregation unit and that the tetramer dissociates first at the dimer-dimer interface (yielding 2 dimers) and then at the dimer interface (yielding 4 monomers). Following tetramer dissociation, aggregation proceeds via downhill polymerization of TTR monomers (McCutchen and Kelly, 1993; Lashuel *et al.*, 1998; Nettleton *et al.*, 1998; Schneider *et al.*, 2001; Hurshman *et al.*, 2004). An engineered TTR variant (TTR F87M/L110M) that is monomeric at neutral pH (monomeric-TTR or M-TTR) but does not aggregate confirms that dissociation of the tetrameric structure alone does not initiate amyloidogenesis. That is, the monomeric species must undergo further conformational changes prior to the formation of aggregation-competent monomers (Jiang *et al.*, 2001). These aggregation-competent monomers then aggregate to form soluble oligomeric species, which are cytotoxic (Reixach *et al.*, 2004). and assemble into higher-order amyloid species and ultimately mature fibrils. Mutations to TTR can alter this aggregation pathway in diverse ways; for example, some mutations promote tetramer dissociation or allow for tetramer dissociation at higher pH, while others destabilize the monomeric subunits and promote the formation of aggregation competent monomers (McCutchen *et al.*, 1993, 1995; Hammarström *et al.*, 2002; Sekijima *et al.*, 2005).

1.3 Computational and Experimental Studies of Transthyretin Amyloidosis

The transient and conformationally heterogeneous nature of the aggregation-competent monomers has complicated their structural characterization at the atomic level; however, several groups have used a diverse array of techniques to probe the structural features of aggregation-competent TTR monomers. Early efforts analyzed hydrogen-deuterium exchange rates of backbone amides at different pHs. Liu *et al.* found that residues within the CBEF sheet are less protected than those in the DAGH sheet and that the interior strands (B, E, A, and G) are more protected than the edge strands (Liu *et al.*, 2000b). This information coupled with prior analysis of Trp fluorescence changes in TTR (Lai *et al.*, 1996) and the observation of large number of pathogenic mutations in the CBEF sheet (João and Saraiva, 1995) led to the hypothesis that the

aggregation-competent monomer is defined by conformational changes in the CBEF sheet with enhanced lability in strands C and D and that the core aggregation unit is comprised of strands BEF and AGH. Later, interpretation of this hydrogen-deuterium exchange data in context with native-state data indicated that the observed lability in strands C and D is also observed under native conditions and that the core aggregation unit includes strands A, B, E, and G (Liu *et al.*, 2000a). Analysis of pathogenic and amyloid-suppressing variants of TTR using hydrogen-deuterium exchange demonstrated that while sequence variants alter protection factor magnitudes, strands A, B, E, and G constitute the core aggregation unit.

Solution NMR experiments performed on M-TTR showed that strand H and the AB, GH, DE, and EF loops (the loops are named according to the β-strands that they connect) had the greatest amounts of conformational exchange. In contrast to the hydrogen-deuterium exchange experiments, this study indicates that minimal conformational exchange occurs in the CBEF sheet and concludes instead that destabilization of the DAGH sheet defines the aggregation-competent state (Lim *et al.*, 2013). Subsequent HSQC NMR results support these conclusions and magic-angle-spinning solid state NMR experiments using selective labeling schemes show that the AB loop becomes disordered in the aggregation-competent state (Lim *et al.*, 2016a-b). The discrepancies between the hydrogen-deuterium exchange experiments and other NMR studies can be partially explained by the different timescales monitored by the experiments. Additionally, the preparation protocols for the hydrogen-deuterium exchange experiments likely resulted in a population of oligomers in the sample that could confound the results.

Computational methods have also been employed to investigate the conformational changes that occur to TTR at the outset of amyloid formation. Prior results from the Daggett lab predict that the destabilization of the DAGH sheet and the formation of a non-standard type of secondary structure, dubbed α-sheet, define the aggregation-competent monomer (Armen *et al.*, 2004a). Subsequent simulations of TTR variants showed that certain pathological mutations also result in α-sheet formation in TTR and that the protective mutant T119M impedes α-sheet formation (Steward *et al.*, 2008). This α-sheet secondary structure was also observed in MD simulations employing a different force field and simulation package (Yang *et al.*, 2006a). Other classical MD simulations performed under physiological conditions have investigated the impact of single amino acid mutations on the flexibility, hydrogen bond stability, and edge-strand stability of TTR monomers (Yang *et al.*, 2003, 2006b; Lei *et al.*, 2004; Steward *et al.*, 2008; Banerjee *et*

al., 2010; Rodrigues *et al.*, 2010). Other computational studies have examined the conformational flexibility of TTR tetramers (Das *et al.*, 2014; Saldaño *et al.*, 2017; Zanotti *et al.*, 2017), dissociation of TTR dimers at low pH (Xue *et al.*, 2014), the dynamics of ligand binding to TTR (Cianci *et al.*, 2015; Cao *et al.*, 2017), and the conformational properties of model amyloid systems derived from TTR peptides (Lee and Na, 2016).

1.4 The α-sheet hypothesis

This work contributes to the α-sheet hypothesis of amyloid disease. In 1951, Pauling and Corey proposed several possible secondary structures that could be formed by polypeptides including the α-helix, β-sheet, and one referred to as a 'polar-pleated sheet' (Pauling and Corey, 1951). Like β-sheet secondary structure, polypeptide chains that adopt the 'polar-pleated sheet' structure form extended conformations; however, in polar-pleated sheets the carbonyl peptide groups are aligned on one side of the backbone and the amide peptide groups on the other. Pauling and Corey predicted that the 'polar-pleated sheet' would be less stable than other secondary structures. This prediction has been upheld as evidenced by the predominance of α-helices and β-sheets observed in protein X-ray crystal structures. Later, Daggett and coworkers observed the formation of the 'polar-pleated sheet' in MD simulations of multiple proteins associated with amyloid diseases, including transthyretin, D67H lysozyme, and the prion protein (Daggett, 2006). The observed 'polar-pleated sheet' conformation was redubbed 'α-sheet'. α-sheet is defined as an extended type of secondary structure in which sequential residues alternate between the right-handed (α_R) and left-handed (α_L) helical regions of Ramachandran space (respectively centered on φ, ψ = -87°, -49° and on φ, ψ = 45°, 92°) (Figure 2). The observation of this otherwise rare flavor of secondary structure in multiple amyloid-associated proteins with diverse native sequences and topologies led to the development of the α-sheet hypothesis of amyloid disease. The α-sheet hypothesis posits that α-sheet secondary structure is present during the early stages of amyloidogenesis, namely within aggregation-competent monomers and soluble oligomers, and that the unique biophysical properties of the secondary structure contribute to or drive protein aggregation.

The structural and dynamical heterogeneity of amyloid proteins has hindered attempts to resolve the conformations(s) populated during aggregation; consequently, direct structural

validation of the α-sheet hypothesis has not been obtained. However, several observations support the presence of a unique structural motif during the early stages of aggregation that is hypothesized to be α-sheet. First, Kayed *et al.* report the existence of an oligomer-specific antibody (henceforth the A11 antibody) that recognizes oligomeric and protofibrillar conformations of many amyloid species including α-synuclein, islet amyloid polypeptide (IAPP), polyglutamine, lysozyme, human insulin, and a prion peptide 106-126 (Kayed *et al.*, 2003; Glabe and Kayed, 2006; Kayed and Glabe, 2006). Importantly, the A11 antibody recognizes all of these proteins in a sequence- and native-topology-independent fashion and the A11 antibody does not recognize the native or fibrillar states. This suggests the existence of a backbone structure that is common to the on-pathway aggregation of many amyloid species. Second, Torii optimized the geometries of model α-sheet peptides at the B3LYP/6-31+G(2df,p) level of density functional theory (DFT) and calculated theoretical FTIR spectra of idealized α-sheet structures (Torii, 2008). The resulting spectra in the amide I region displayed a prominent signal at ~1670 cm^{-1} that is distinct from α-helical and β-sheet secondary structure. This signal has been observed during the aggregation of multiple amyloid species, including the aggregation of a variant of an 11-residue peptide from TTR that is frequently used as a model for amyloidogenesis (Hilaire *et al.*, 2018) and in peptides designed to adopt α-sheet secondary structure (Maris *et al.*, 2018). Third, Xu utilized AFM data and proposed that the initial stages of aggregation may be driven by a dipole moment intrinsic to the base aggregation units, a property satisfied by α-sheet (Xu, 2007; Maris *et al.*, 2018). Finally, compounds designed to interact with α-sheet secondary structure have inhibited aggregation in diverse amyloid systems including transthyretin (Hopping *et al.*, 2014; Kellock *et al.*, 2016), amyloid-$\beta_{1\text{-}42}$ (Hopping *et al.* 2014; Kellock *et al.*, 2016; Maris *et al.*, 2018), islet amyloid polypeptide (Kellock *et al.*, 2016), and PSMα-1 from *Staphylococcus aureus* (Bleem *et al.*, 2017).

Although rare, there are several instances of α-strand secondary structure observed in X-ray crystal structures of proteins, including synaptotagmin (PDB ID:1RSY residues 162-3, 205-7), hen egg white lysozyme (PDB ID:1HF4, residues 72-5), and a potassium channel from *Streptomyces lividans* (PDB ID: 1BL8, residues 74-80) (Armen *et al.*, 2004a). α-strand secondary structure has also been observed in crystal structures of a capped tripeptide Boc-Ala_L-$^{allo}Ile_D$-Ile_L-OMe (Di Blasio *et al.*, 1994) and capped diphenyl-glycine containing peptides (Pavone *et al.*, 1998; De Simone *et al.*, 2000). In addition to MD simulations of α-sheet conversion performed by

the Daggett group using the *in lucem* molecular mechanics (Beck *et al.*, 2000-2019) simulation package and the Levitt *et al.* force field (Levitt *et al.*, 1995), β- to α-sheet conversion has also been observed in TTR using the AMBER parm94 force field (Yang *et al.*, 2006a) and modeled in simple poly-alanine peptides using the GROMOS96 force field (Hayward and Milner-White, 2011). Thus, several lines of evidence emanating from structural, computational, and phenomenological perspectives provide indirect support for the α-sheet hypothesis and indicate that the development of peptides to interact with α-sheet secondary structure is a promising strategy to inhibit amyloid aggregation.

1.5 This Work

This work begins with a validation of contemporary molecular dynamics force fields and molecular modeling packages. To accomplish this, I performed simulations of two model protein systems (the engrailed homeodomain and ribonuclease H) using four contemporary molecular dynamics force fields and molecular modelling packages: *il*mm (the in-house MD software authored and maintained by the Daggett Group), AMBER, GROMACs, and NAMD. The results of these simulations were compared against over 3,100 experimentally derived data points obtained using biophysical and spectroscopic techniques. This study was published in the *Journal of Physical Chemistry B* and is included as Chapter 2 (Childers and Daggett, 2018). The agreement between the results of simulations performed with *il*mm and the experimental data for these model protein systems supports the use of MD simulations to model more complex systems – such as transthyretin during amyloidogenesis – where substantially less experimental data are available for comparison.

Next, I performed 21 simulations of transthyretin under amyloidogenic conditions with a net sampling time in excess of 10 microseconds, to my knowledge this is the most extensive set of all-atom simulations of transthyretin performed to date. These simulations include wild-type transthyretin along with six variants harboring pathological mutations and were performed under amyloidogenic conditions (low pH). In-depth analysis of these simulations was used to identify conformational changes that occur to transthyretin monomers prior to the formation of amyloid species. These include changes to the secondary structure (included as Chapter 3) and tertiary structure (included as Chapter 4) of TTR. My analysis was used to propose a mechanism by which α-sheet secondary structure forms in transthyretin (detailed in Chapter 3). I then tested the

proposed mechanism by generating a transthyretin variant with altered propensity to form α-sheet, included as Chapter 5.

In Chapters 6 and 7 I present the results of an extensive set of MD simulations of pentapeptides that were used to identify the intrinsic conformational preferences (i.e. the preferential main chain and side chain conformations) of L- and D- amino acids in both homochiral and heterochiral polypeptide chains. These studies resulted in a set of conformational libraries that can be used in protein and peptide design applications. The first study, a comparison of main chain conformations sampled in Glycine and Alanine based peptides was published in *Protein Engineering Design and Selection* (Childers *et al.*, 2016) and is included as Chapter 6. The second study, a comparison of side chain conformations in Glycine and Alanine based peptides was published in *Protein Engineering Design and Selection* (Childers *et al.*, 2018) and is included as Chapter 7. Supplemental tables and figures are included as Appendix A.

1.6 Figures

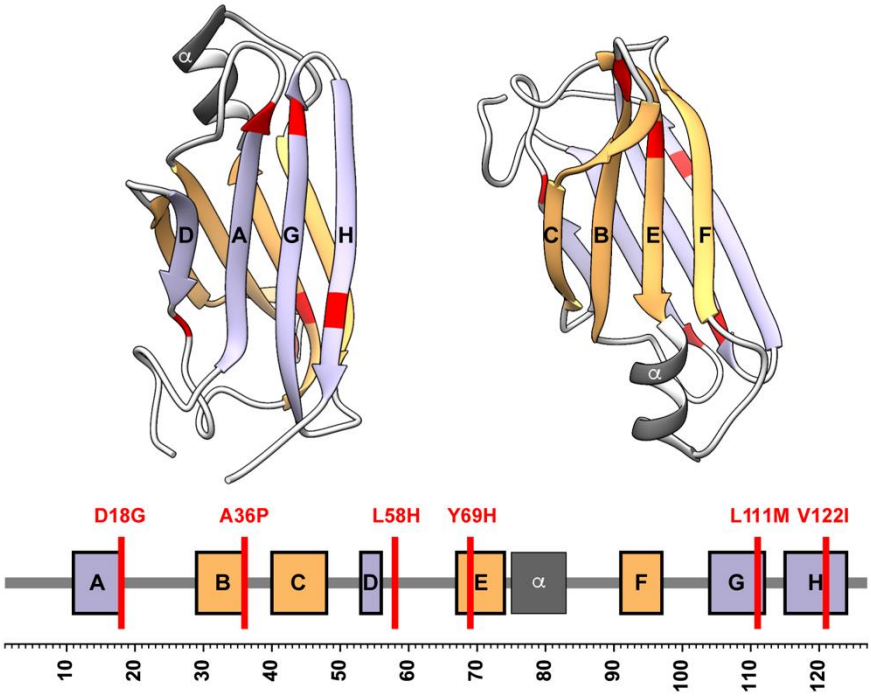


Figure 1.1. The X-ray crystal structure of transthyretin.

A transthyretin monomer in the crystallographic conformation is shown with the strands (A-H) individually labeled A-H and colored based on their organization into sheets. The DAGH sheet is colored purple and the CBEF sheet is colored orange. Red ribbons indicate the positions of pathological mutations that have been simulated in this work.

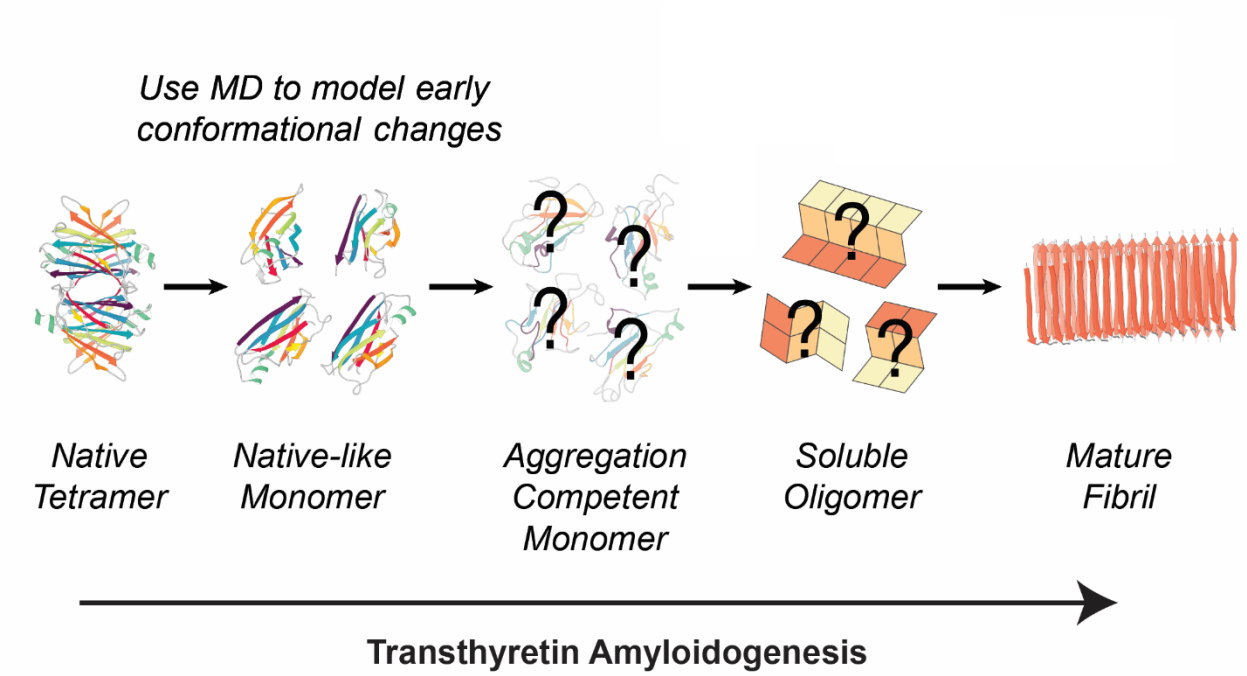


Figure 1.2. Schematic of transthyretin aggregation.

Transthyretin aggregation can be initiated in acidic environments, which triggers dissociation of the native tetramer into monomers. The monomeric species then undergo conformational changes that result in aggregation-competent conformations which assemble into toxic soluble oligomers and eventually mature fibrils.

Chapter 2. VALIDATING MOLECULAR DYNAMICS SIMULATIONS AGAINST EXPERIMENTAL OBSERVABLES IN LIGHT OF UNDERLYING CONFORMATIONAL ENSEMBLES

2.1 Summary

Far from the static, idealized conformations deposited into structural databases, proteins are highly dynamic molecules that undergo conformational changes on temporal and spatial scales that may span several orders of magnitude. These conformational changes, often intimately connected to the functional roles that proteins play, may be obscured by traditional biophysical techniques. Over the past 40 years, molecular dynamics (MD) simulations have complemented these techniques by providing the "hidden" atomistic details that underlie protein dynamics. However, there are limitations of the degree to which molecular simulations accurately and quantitatively describe protein motions. Here we show that although four molecular dynamics simulation packages (AMBER, GROMACS, NAMD, and ilmm) reproduced a variety of experimental observables for two different proteins (engrailed homeodomain and RNase H) equally well overall at room temperature, there were subtle differences in the underlying conformational distributions and the extent of conformational sampling obtained. This leads to ambiguity about which results are correct, as experiment cannot always provide the necessary detailed information to distinguish between the underlying conformational ensembles. However, the results with different packages diverged more when considering larger amplitude motion, for example, the thermal unfolding process and conformational states sampled, with some packages failing to allow the protein to unfold at high temperature or providing results at odds with experiment. While most differences between MD simulations performed with different packages are attributed to the force fields themselves, there are many other factors that influence the outcome, including the water model, algorithms that constrain motion, how atomic interactions are handled, and the simulation ensemble employed. Here four different MD packages were tested each using best practices as established by the developers, utilizing three different protein force fields and three different water models. Differences between the simulated protein behavior using two different packages but the same force field, as well as two different packages with different force fields but the same water models and approaches to restraining motion, show how other factors can influence the behavior,

and it is incorrect to place all the blame for deviations and errors on force fields or to expect improvements in force fields alone to solve such problems.

2.2 Introduction

Molecular dynamics (MD) simulations, "virtual molecular microscopes", employ computational methods to probe the dynamical properties of atomistic systems and proffer insights into molecular behavior. Beginning with the report of a 9.2 ps simulation of bovine pancreatic trypsin inhibitor (BPTI) in 1977 (McCammon, 1977), MD simulations have provided the means to visualize proteins in action and to investigate that paradigmatic relationship between form and function (Dodson *et al.*, 2008; Vlachakis *et al.,* 2014). When taken in context with experimental results, MD simulations can drive discoveries in protein design (Childers *et al.* 2017; Kiss *et al.* 2013), protein folding (Daggett *et al.*, 1996; Bond *et al.,* 1997; Duan *et al.*, 1998), and other spheres of protein science (Lee *et al.*, 2009). However, two factors limit the predictive capabilities of MD: First, lengthy simulations may be required to correctly describe certain dynamical properties (i.e. the *sampling problem*) (Grossfield *et al.,* 2009). Second, insufficient mathematical descriptions of the physical and chemical forces that govern protein dynamics may yield biologically meaningless results (i.e., the *accuracy problem*) (Lopes *et al.,* 2015; van Gunsteren *et al.,* 2018). To increase our confidence in the ability of MD simulations to provide meaningful results for arbitrary proteins and peptide systems, it is necessary to benchmark computational results against experimental data.

Improved computational infrastructure (Shaw *et al.,* 2014), software (Beberg *et al.,* 2009; Larson *et al.,* 2009), and parallelization schemes (Voter *et al.,* 1998; Bowers *et al.,* 2007) allow contemporary simulations to probe increasingly larger systems at time scales approaching those of experiment (Freddolino *et al.,* 2006; Shaw *et al.,* 2010; Lindorff-Larsen *et al.,* 2011; Prinz *et al.,* 2011). However, the requisite simulation times to accurately measure dynamical properties are rarely known *a priori*; instead, simulations are deemed "sufficiently long" when some observable quantity has "converged". In the context of molecular simulation, Sawle and Ghosh argue that convergence is a misnomer and show that the time scales required to satisfy the most stringent tests of "convergence" or "self-consistency" vary from system to system and are dependent on the method used to assess convergence (Sawle *et al.,* 2016). This behavior is mirrored in the analysis of MD simulations: we have shown that the overall level of insight into the dynamics of a system

can be modulated by the type of analysis performed and the level of detail described (Benson *et al.,* 2012). Thus, how long is "sufficiently long" remains unanswered. A wealth of information can be obtained from simulations that probe native state dynamics as well as conformational changes that result in excursions from the native state. However, for slow dynamical processes like the folding of typical globular proteins and the "nonfolding" of intrinsically disordered proteins (Henriques *et al.,* 2015; Stanley *et al.,* 2015), the requisite time scales remain out of reach for the time being, at least for conventional MD simulations.

Approximations built into the mathematical forms of MD force fields and their associated parametrizations give rise to the accuracy problem. These force fields are empirical and begin with parameters obtained from high-resolution experimental data and quantum mechanical calculations for small molecules, and then they are modified to reproduce different experimental properties or desired behaviors (Halgren, 1995; Halgren, 1996a-d; Halgren and Nachbar, 1996; Waldher *et al.,* 2010; Weiner *et al.,* 1984; Chen *et al.,* 2002; Ercolessi *et al.,* 1994). Over time, modification of these parameters has yielded improved force fields with similar functional forms (Lindorff-Larsen *et al.,* 2012). In addition, it is important to note that while usually the focus, or blame, is on the force field, it is not just the potential energy function and associated parameters that determine the results of MD simulations. Protein dynamics are often more sensitive to the protocols used for integration of the equations of motion, treatment of nonbonded interactions, and various unphysical approximations.

The most compelling measure of the accuracy of a force field is its ability to recapitulate and predict experimental observables. However, there are challenges associated with this method of validation (van Gunsteren *et al.,* 2018). Namely, the experimental data used for validation are averages over space and time; the underlying distributions and time scales associated with these averages are often obscured. Consequently, correspondence between simulation and experiment does not necessarily constitute a validation of the conformational ensemble(s) produced by MD, i.e. multiple, and possibly diverse, ensembles may produce averages consistent with experiment. This is underscored by simulations that demonstrate how force fields can produce distinct pathways of the lid-opening mechanism of adenylate kinase that nevertheless sample the crystallographically identified "open" and "closed" conformers (Unan *et al.,* 2015). Furthermore, extensive simulations of the villin headpiece demonstrated that while MD-derived folding rates and native state structures had good agreement with experiment, the folding pathways and

denatured state properties were force-field-dependent (Piana *et al.,* 2011). In addition, experimental observables may be derived using relationships that are functions of molecular conformation and are themselves associated with some degree of error. For example, most chemical shift predictors produce chemical shifts from molecular structures via training against high-resolution structural databases, not solely via calculations from first principles.

Here, we address the extent to which multiple simulations performed for 200 ns each agree with experimental data. Multiple short simulations yield better sampling of protein conformational space than a single simulation with total sampling time equal to the aggregate sampling time of multiple small simulations (Kazmirski *et al.,* 1998; Caves *et al.,* 1998). As simulations see increased usage, particularly by those not trained in the method, it is important to place quantitative bounds on the extent to which these simulations agree with experimental data and to understand the limits of their ability to explain experimental findings. Consequently, we have compared how three force fields (AMBER ff99SB-ILDN (Lindorff-Larsen *et al.*, 2010), Levitt *et al.* (Levitt *et al.*, 1995), and CHARMM36 (Huang and MacKerell, 2013)) used within four MD packages (AMBER (Pearlman *et al.,* 1995; Case *et al.,* 2005; Salomon-Ferrer *et al.,* 2013), *in lucem* molecular mechanics (*il*mm) (Beck *et al.,* 2000-2019), GROMACS (Abraham *et al.,* 2015), and NAMD (Phillips *et al.,* 2005)) agree with a diverse set of experimental data for two globular proteins with distinct topologies: the Engrailed homeodomain (EnHD) and Ribonuclease H (RNase H) (Figure 2.1). The *Drosophila* engrailed homeodomain has 54 residues arranged into three α-helices (denoted HI–HIII) and constitutes the DNA-binding domain of the larger transcription factor in which it is found (Clarke *et al.*, 1994). RNase H is an endonuclease α/β protein composed of 155 residues organized into five α-helices (denoted α_A–α_E) and a single, five-stranded β-sheet (denoted β_1–β_5) that hydrolyzes the RNA strand in double-stranded RNA–DNA hybrid (Katayanagi *et al.*, 1992).

2.3 Methods

2.3.1 *Molecular Dynamics Simulations*

The initial coordinates for simulations of EnHD were obtained from the 2.1 Å resolution X-ray crystal structure solved by Clarke *et al.* (PDB ID: 1ENH) (Clarke *et al.*, 1994). The initial coordinates for simulations of RNase H were obtained from the 1.48 Å resolution crystal structure

solved by Katayanagi *et al.* (PDB ID: 2RN2) (Katayanagi *et al.,* 1992). Crystallographic solvent atoms were removed from these structures, and then, conventional molecular dynamics simulations were performed using four software package–force field combinations: *in lucem* molecular mechanics (*il*mm) (Beck *et al.,* 2000-2019) with the Levitt et al. force field (Levitt *et al.,* 1995), AMBER with the Amber ff99SB-ILDN force field (Lindorff-Larsen *et al.,* 2010), GROMACS (Abraham *et al.,* 2015) with the Amber ff99SB-ILDN force field (Lindorff-Larsen *et al.,* 2010), and NAMD (Phillips *et al.,* 2005) with the CHARMM36 force field (Huang *et al.,* 2013; MacKerell *et al.,* 1998; Mackerell *et al.,* 2004). The simulations were performed under conditions consistent with those under which the experimental data were obtained. Simulations of EnHD were performed at neutral pH (7.0) at 298 K, and simulations of Rnase H were performed at acidic pH (5.5, histidine residues protonated) at 298 K. All simulations were performed in triplicate for 200 ns using periodic boundary conditions, explicit water molecules, and "best practice parameters", as determined by recent papers in the literature by authors of the software packages [AMBER (Janowski *et al.,* 2016; Chaudhuri *et al.,* 2011; Okal *et al.,* 2014), *il*mm (Beck *et al.,* 2004), GROMACS (Kimanius *et al.,* 2015), and NAMD (Bernardi *et al.,* 2014)], and their associated force fields typically contain many adjustable parameters. Here, we aimed to strike a balance between keeping these parameters consistent and adjusting them only when necessary for specific force fields/MD package combinations. Within each package/force field combination, simulation methods were kept constant for the two proteins. Simulations of the native state were performed at 298 K, and thermal unfolding was simulated at 498 K. Details for the initial preparation of the systems and the 298 K simulations for each force field follow. We maintain that the algorithms used and associated input control parameters are as important as the force field *per se* in determining simulated behavior. This information may also be of use in meta-analyses that evaluate force field/MD software. Also, as there is some overlap in the protein (AMBER and GROMACS) and water (GROMACS and NAMD) force fields used by different programs, the simulations will be referred to throughout by the name of the simulation package used.

2.3.1.1 AMBER

Simulations were performed with the AMBER14 package and ff99SB-ILDN force field (Lindorff-Larsen *et al.,* 2010). Explicit hydrogen atoms were modeled onto the X-ray structure using the *leap* module, and each protein was solvated with explicit TIP4P-EW (Horn *et al.,* 2004)

waters in a periodic, truncated octahedral box that extended 10 Å beyond any protein atom. Each system was then minimized in three stages. First, solvent atoms were minimized for 500 steps of steepest descent minimization followed by 500 steps of conjugate gradient minimization in the presence of 100 kcal mol^{-1} restraints on protein atoms. Second, solvent atoms and protein hydrogens were minimized for 500 steps of steepest descent minimization followed by 500 steps of conjugate gradient minimization in the presence of 100 kcal mol^{-1} restraints on protein heavy atoms. Third, all atoms were minimized for 500 steps of steepest descent minimization followed by 500 steps of conjugate gradient minimization in the presence of 25 kcal mol^{-1} restraints on protein C_α atoms. After minimization, systems were heated to 298 K during 6 successive stages. In each stage, the system temperature was increased by 50 K over 200 ps (25 000 steps) using the canonical NVT (constant **n**umber of particles, **v**olume, and **t**emperature) ensemble (25 kcal mol^{-1} restraints on protein C_α atoms were present during each stage). After the system temperature reached 298 K, the systems were equilibrated over 7 successive stages. During the first 5 stages, the systems were minimized for 1000 steps (500 steps of steepest descent followed by 500 steps of conjugate gradient minimization), and restraints on protein C_α atoms were decreased from 5 to 1 kcal mol^{-1}. Next, the systems were equilibrated using the NVT ensemble for 500 000 steps (1 ns), and then the NPT ensemble for an additional 500 000 steps (1 ns) in the presence of 0.5 kcal mol^{-1} restraints was present on protein C_α atoms. Production dynamics were then performed using the isobaric–isothermal NPT (constant **n**umber of particles, **p**ressure, and **t**emperature) ensemble using a 2 fs time step, and coordinates were saved every picosecond for analysis. The SHAKE algorithm was used to constrain the motion of hydrogen-containing bonds (Ryckaert *et al.,* 1977; Miyamoto *et al.,* 1992). Long-range electrostatic interactions were calculated using the particle mesh Ewald (PME) method.

2.3.1.2 GROMACS

Simulations were performed with GROMACS version 5.0.6 (Abraham *et al.,* 2015) and the AMBER ff99SB-ILDN (Lindorff-Larsen *et al.,* 2010) force field. Hydrogen atoms were modeled onto the X-ray structure using *pdb2gmx* prior to solvation with TIP3P (Jorgensen *et al.,* 1983) waters in periodic, cubic boxes that extended 10 Å beyond any protein atom. Solvent molecules were replaced with counterions until the system was neutralized. Throughout the following stages, a Verlet cutoff scheme (Pall *et al.,* 2013) was employed with a 10 Å cutoff for

both electrostatic and van der Waals interactions, and LINCS was employed to constrain bonds (Hess *et al.,* 1977). Electrostatic interactions were calculated using PME. The solvated systems were minimized for 50 000 steps using steepest descent minimization. Systems were then equilibrated over two stages in the presence of positional restraints on protein atoms. First, systems were equilibrated in the NVT ensemble for 50 000 steps followed by equilibration in the NPT ensemble for an additional 50 000 steps. Finally, production dynamics were performed in the NPT ensemble with a 2 fs time step, and coordinates were saved every picosecond for analysis.

2.3.1.3 *il*mm

Simulations were performed with the *in lucem* molecular mechanics (*il*mm) package and Levitt *et al.* force field (Levitt *et al.,* 1995) using our standard protocols. Explicit hydrogen atoms were modeled onto the X-ray crystal structures prior to steepest descent minimization for 1000 steps. Each protein was solvated with explicit flexible 3-center (F3C) (Levitt *et al.,* 1997) water molecules in a periodic, cubic box that extended 10 Å beyond any protein atom, with the solvent density set to the experimental value at 298 K (0.997 g mL^{-1}) (Kell *et al.,* 1967). Solvent atoms were then minimized for 1000 steps and equilibrated for 500 steps (1 ps) prior to additional, separate minimization of the solvent (500 steps) and protein (500 steps) atoms. Conventional molecular dynamics simulations were then performed using the microcanonical NVE (constant **n**umber of particles, **v**olume, and **e**nergy) with a target temperature of 298 K. The equations of motion were propagated using a 2 fs time step with a 10 Å force-shifted nonbonded cutoff (Levitt *et al.,* 1995; Beck *et al.,* 2005) and coordinates were saved every picosecond for analysis. In contrast to the other software packages described here, *il*mm with the Levitt *et al.* force field (Levitt *et al.,* 1995) subscribes to a molecular-level representation and natural Boltzmann sampling through use of the NVE ensemble rather than trying to control macroscopic variables, such as temperature and pressure, in these microscopic systems. Such temperature and pressure coupling lead to very frequent scaling of the velocities, which in turn provides discontinuous trajectories. This may not be an issue if conformational sampling is desired as opposed to pathways for those conformational changes, but *il*mm was developed with the objective of characterizing both "kinetic" pathways and "equilibrium" states. In addition, *il*mm does not restrain atomic motion via algorithms such as LINCS and SHAKE nor introduce artificial periodicity into the molecular system via algorithms such as PME (Beck *et al.,* 2005).

2.3.1.4 NAMD

Simulations were performed with NAMD version 2.10 (Phillips *et al.,* 2005) and the CHARMM36 force field (Huang *et al.,* 2013). Hydrogen atoms were modeled onto the X-ray crystal structures using *psfgen* prior to solvation with TIP3P waters (Jorgensen *et al.,* 1983) in a periodic box that extended 10 Å beyond any protein atom. Next, minimization was performed in two phases. In the first stage, minimization was performed for 20 000 steps with all hydrogen-containing bonds constrained and protein atoms fixed. In the second stage, minimization was performed for 1000 steps with all protein backbone atoms fixed and an additional 1000 steps with no fixed atoms. After minimization, systems were heated to 298 K over 10 000 steps with harmonic restraints on backbone atoms that were gradually decreased from 5.0 to 0 kcal mol^{-1}. After heating, systems were equilibrated for 100 000 steps (200 ps) in the NPT ensemble. Finally, production dynamics were performed in the NVT ensemble using a 2 fs time step. van der Waals interactions were truncated with a switching potential, and coordinates were saved every picosecond for analysis. Electrostatic interactions were calculated via PME summation, and SHAKE was used to constrain bonds.

2.3.2 *High-Temperature Unfolding Simulations*

In addition to the native state simulations described above, we also performed high-temperature unfolding simulations of EnHD with the same force fields as for the native simulations. The high-temperature unfolding protocols were similar to those for the native state simulations, with the following changes. The simulations were performed at 498 K in triplicate for 10 ns. Nonbonded cutoffs were reduced by 2 Å, and structures were saved every 0.2 ps for analysis. To keep solvent molecules in the liquid state, the pressure was set to ∼26 atm (Haar *et al.,* 1984) for simulations with GROMACS, AMBER, and NAMD; for *il*mm, the solvent density was set to the experimental value at 498 K and ∼26 atm (0.829 g mL^{-1}) (Kell *et al.,* 1967; Haar *et al.,* 1983). For AMBER, the two final equilibration phases were reversed, with NPT equilibration (500 000 steps) followed by NVT equilibration (500 000 steps), and the NVT ensemble was employed for production dynamics. For NAMD, the NPT ensemble was employed for production dynamics.

2.3.3 *Analysis of MD Data*

After production dynamics were completed, all trajectories were converted into an *il*mm-compatible format and analyses were performed identically for all trajectories. Unless otherwise specified, 298 K analyses were performed on an ensemble created by pooling all three replicate simulations, and the first 20 ns of each simulation were excluded from analysis, yielding 5.4×10^5 structures in each ensemble.

2.3.4 *Calculation of Gross Structural Changes and Dynamic Fluctuations*

The root mean squared deviations (RMSDs) were calculated by aligning each frame to the crystal structure, and fluctuations (RMSFs) were calculated by aligning each frame to the average structure calculated after the 20 ns equilibration period. The C_α atoms for all "core" residues were included in the alignment and subsequent calculations. "Core" residues refer to all residues except for flexible N- and C-terminal residues (EnHD, residues 8–53; RNase H, residues 5–142). Experimental *B*-factors were compared with RMSFs via Equation 2.1, where *B* = experimental *B*-factors (Willis *et al.*, 1975).

$$\mathrm{RMSF} = (3B / 8\pi^2)^{1/2}$$

(Equation 2.1)

2.3.5 *Calculation of NMR Chemical Shifts*

Simulated chemical shifts were calculated with the SHIFTX2 program. (Han *et al.*, 2011) Experimental data for EnHD and RNase H were obtained from the Biological Magnetic Resonance Bank (BMRB) (Ulrich *et al.*, 2008) entries 15536 (Religa *et al.*, 2008) and 1657 (Yamazaki *et al.*, 1991), respectively. Chemical shifts were calculated for 1% of the ensemble (one frame every 100 ps). For RNase H, the re-referenced chemical shifts provided by the RefDB were used (Zhang *et al.*, 2003).

2.3.6 *Calculation of NMR Scalar Coupling Constants*

We calculated $^3J_{H\alpha,HN}$ scalar coupling constants from MD simulations using the Karplus relation (Karplus, 1963) (Equation 2.2) by taking the average of the coupling constants calculated for each frame in the simulation. The coefficients for the Karplus equation (C_0, C_1, and C_2) were

obtained from the literature, and we used 7 different parametrizations (Habeck *et al.*, 2005; Ludvigsen *et al.*, 1991; Schmidt *et al.*, 1999; Smith *et al.*, 1991; Vuister *et al.*, 1993; Hu *et al.*, 1997; Pardi *et al.*, 1984). Experimental data for EnHD and RNase H were obtained from the BMRB entries 15536 (Religa *et al.*, 2008) and 1657 (Yamazaki *et al.*, 1991), respectively.

$$^{3}J(\theta) = C_0 + C_1 \cos\theta + C_2 \cos^2\theta$$

(Equation 2.2)

2.3.7 *Calculation of NOE Satisfaction*

Experimental nuclear Overhauser effect (NOE) values for EnHD and RNase H were obtained from the BMRB, entries 15536 (Religa *et al.*, 2008) and 1657 (Yamazaki *et al.*, 1991), with 654 and 1428 NOEs, respectively. NOEs were classified as satisfied in simulations based on the inequality in Equation 2.3, where r is the distance between a pair of protons, and rUB is either the experimental upper bound restraint distance or 5 Å, whichever is greater.

$$\langle r^{-6} \rangle \leq r_{UB}$$

(Equation 2.3)

2.3.8 *Calculation of Generalized Amide S^2 Order Parameters*

Experimental order parameters for EnHD were obtained from BMRB entry 15336 (Religa *et al.*, 2008). Experimental order parameters for RNase H were obtained from the Supporting Information of Stafford *et al.* (Stafford *et al.*, 2015). MD-derived NH bond order parameters were calculated using the method described by Wong and Daggett using a 250 ps window for EnHD and a 10 000 ps window for RNase H given its ~9.7 ns tumbling time (Wong *et al.*, 1998). Final order parameters are reported by averaging the results from the three replicate simulations.

2.3.9 *Dihedral Angle Principal Component Analysis*

Principal component analysis (PCA) is frequently employed to investigate protein dynamics by systematically reducing the dimensionality of complex motions into simpler components (Kazmirski *et al.*, 1999). Here, we used PCA to investigate the dynamics of the Gly-rich loop in RNase H. Although Cartesian coordinates are normally employed in PCA calculations, we chose to use dihedral angle PCA of selected residues to investigate dynamics in the Gly-rich

loop region of RNase H using the sine and cosine components of the ϕ and ψ dihedral angles of residues 11–22 as input.

2.3.10 *Transition State Identification*

The protein folding/unfolding transition states were identified from the high-temperature unfolding simulations by using a previously established conformational clustering method (Li and Daggett 1994; Li and Daggett, 1996; Li and Daggett., 1998). Using this method, a pairwise C_α RMSD matrix describes the conformational similarity of all structures in $n \times n$-dimensional space (where n = the number of frames in each simulation). This matrix is then projected into three dimensions such that frames with similar conformations are clustered in the reduced dimensional space. We then choose an ensemble of conformers prior to the onset of a significant conformational change to model the transition state structure.

2.3.11 *Transition State Evaluation*

To assess how well the putative transition state ensemble chosen from MD simulations agrees with experimental observations, we performed a comparative analysis between MD-derived *S*-values and ϕ-values (Daggett *et al.*, 1996). *S*-values, or structure index values, are semiquantitative equivalents of experimental ϕ-values (Fersht *et al.*, 1991; Fersht *et al.*, 1993) and incorporate secondary and tertiary structure components. *S*-values are defined as $S = (S_{2^\circ})(S_{3^\circ})$, where S_{2° describes the secondary structure component and S_{3° describes the tertiary structure component, and provide the fraction of native structure present in the transition state on a per-residue basis. The fraction of native secondary structure, S_{2°, for residue i is defined as the fraction of time that the dihedral angles of residues $i - 1$, i, and $i + 1$ spend within ± 35° of the values in the X-ray structure. The fraction of native tertiary structure, S_{3°, for residue i is defined as the number of tertiary contacts present in the transition state divided by the number of tertiary contacts present in the native state. Tertiary contacts were defined for interactions that heavy atoms in the residue of interest form with others separated by two or more residues (≤ 5.4 Å for C–C contacts and ≤ 4.6 Å for all others).

2.4 Results & Discussion

2.4.1 *Native State Sampling*

The X-ray and NMR-derived structures deposited in the Protein Data Bank (PDB, www.rcsb.org) (Berman *et al.*, 2000) are averages over space and time, washing over the subtle and varied excursions from the native state frequently taken by globular proteins in solution (Petsko *et al.*, 1996). One of the chief goals of MD simulations is to explore such excursions beyond the native state. Before comparing the simulations against NMR observables, we first examined their conformational sampling relative to the starting native structures. We calculated the C_α RMSDs and RMSFs for EnHD and RNase H by aligning the core residues to either the starting structure (RMSD) or the average MD-ensemble structure (RMSF). For EnHD, the core residues were 8–53, and for RNase H, the core residues were 5–142 (Figure 2.1). For both proteins in all force fields, the C_α RMSDs reached stable values by 100 ns. Averaging over all three replicate simulations per MD package, we found that EnHD had average C_α RMSDs ranging from 0.6 Å (GROMACS and NAMD) to 0.8 Å (AMBER) to 1.0 Å (*il*mm) and that RNase H had average Cα RMSDs ranging from 1.3 Å (GROMACS) to 1.4 Å (NAMD) to 1.5 Å (AMBER) to 2.4 Å (*il*mm) (Figure 2.2 and Table 2.3). Independent of the protein system, GROMACS and NAMD produced narrow C_α RMSD distributions with little variation between replicate simulations, whereas AMBER and *il*mm produced broader distributions with more variation between simulation replicates) (Figure 2.2 and Table 2.3).

We analyzed the per-residue contributions to the C_α RMSDs and found that, for EnHD, all force fields consistently produced some deviation from the native conformation in two regions. The first region was HIII, the C-terminal helix that contains many DNA-binding residues. In *il*mm and AMBER, HIII frayed at the C-terminus; however, the nature of this local unfolding was force-field-dependent. In one *il*mm simulation, HIII rotated away from the hydrophobic core, exposing W48 to solvent. After sampling this near-native conformation, HIII began to refold. In one AMBER simulation, the C-terminal residues lost helical structure (Figure A.2.1, A.2.2). Partial loss of helical structure was also present for GROMACS and NAMD, but to a lesser extent (Figure A.2.2). The second region with significant deviation from the native conformation was the loop connecting HI and HII. In *il*mm, residues 24 and 25 underwent a dihedral transition, reducing the chain length of the HI–HII loop and producing a 1.9 Å C_α RMSD for residue 29. A similar event

was observed with AMBER; however, instead of a single flip, multiple residues in the HI–HII loop experienced small dihedral angle shifts, resulting in a 2.7 Å C_α RMSD for residue 29 (Figure A.2.3). The heightened degree of motion is supported by experiment. Stollar *et al.* found that the HI–HII loop undergoes conformational fluctuations on the μs–ms time scale. (Stollar *et al.*, 2003) These initial displacements seen in MD may be precursors to more significant conformational changes (Stollar *et al.*, 2003). For RNase H, the major contributions to the RMSD came from 5 different regions of the protein: the Gly-rich loop (residues 11–22), the β_2–β_3 loop (residues 28–30), the α_A–β_4 loop (residue 59–63), the handle region (residues 81–101), and the β_5–α_E loop (residue 121–127) (Figure 2.1).

We then compared the simulated C_α RMSFs to crystallographic *B*-factors and found that AMBER and *il*mm had substantially lower root mean squared error (RMSE) between the simulated and experimental values than GROMACS or NAMD (Table 2.1). The correlation coefficients were high for all force fields, ranging between 0.82 and 0.87. The lower RMSEs for AMBER and *il*mm can be traced to the HII–HIII loop, N-terminal residues, and C-terminal residues (Figure A.2.4). The dynamics in the AMBER and *il*mm simulations lead to the lower errors relative to GROMACS and NAMD. Lower correspondence between the *B*-factors and simulation was observed for RNase H for all MD packages (correlation coefficients all below 0.7), but the associated RMSEs were low (less than 0.3 Å) (Table 2.2).

2.4.2 *Comparison with NMR Observables*

Next, we assessed the ability of MD simulations to reproduce four types of NMR observables: chemical shifts, nuclear Overhauser effect crosspeaks (NOEs), backbone NH order parameters, and scalar coupling constants.

2.4.2.1 Chemical Shifts

Chemical shifts report on the local electronic environments of distinct nuclei within proteins. We calculated the chemical shifts from our simulations using SHIFTX2 (Han *et al.*, 2011). For the calculation, the first 20 ns of each trajectory was excluded, and all three replicate simulations were combined into a single ensemble. To determine the level of sampling required to accurately report chemical shifts, we calculated the chemical shifts for run 1 of RNase H performed with AMBER using 1, 10, and 100 ps granularity and confirmed that subsampling the simulation

at 100 ps resulted in predicted shifts that were nearly identical for the same calculation run at full granularity (Figure A.2.5). All the MD packages in this study reproduced chemical shifts with errors comparable to those associated with SHIFTX2 (Figure 2.3 and Tables 2.1 and 2.2). For comparison, we also calculated the agreement for the X-ray crystal structures for EnHD and RNase H. Prior to chemical shift calculations, hydrogens were modeled onto the crystal structure using *il*mm, and the hydrogen atoms were minimized for 1000 steps of steepest descent minimization. For some nuclei, the MD-generated ensembles produced better agreement than the X-ray conformations. For example, the X-ray conformation of K17 in EnHD did not agree well with the experimental data, especially for the N, C_α, H_α, and H_N nuclei (Figure A.2.8). In general, MD ensembles produced chemical shifts that were more consistent with the experimental data than the X-ray structure alone, particularly for the N and H_N chemical shifts; however, little improvement was observed for C_α shifts (Figure A.2.8). We investigated the origins of this discrepancy and found that K17 had distinct hydrogen bonding patterns within different force fields (Table A.2.1). The rate of formation of a main chain hydrogen bond between K17 (donor) and A14 (acceptor) may modulate the predicted chemical shift for the amide hydrogen of K17 (Table A.2.1). Wagner *et al.* and others have demonstrated that hydrogen bond geometry can influence the chemical shifts for H_N nuclei and a subtle difference in hydrogen bonding patterns can contribute to enhanced correspondence between MD and experiment (Wagner *et al.,* 1983). Additionally, there were a few instances where MD simulations produced worse agreement with the experimental data than the reference X-ray structure. Overall, however, all the MD packages produced strong agreement with experimental chemical shifts, exhibited by the high correlation coefficients and low RMSEs (Figure 2.3 and Tables 2.1 and 2.2).

2.4.2.2 Nuclear Overhauser Effect Crosspeaks

Nuclear Overhauser effect crosspeaks (NOEs) relate to interproton distances and provide significant conformational information. We calculated NOEs from our simulations and stratified the results in two ways. First, we analyzed the percentage satisfaction for NOEs as a function of sequence separation: short-range NOEs refer to NOEs arising from residues $i \rightarrow i + 1$, $i + 2$; medium-range NOEs, $i \rightarrow i + 3$, $i + 4$, $i + 5$; and long-range NOEs, $i \rightarrow i > i + 5$. Second, we analyzed the number of NOE violations, stratified by violation distance. For comparison, we also calculated the NOE satisfaction for the EnHD and RNase H crystal structures. For crystal structure

analyses, we used the crystal structures containing hydrogens, as described above. For comparison, we also calculated the NOE satisfaction for the EnHD NMR ensemble (PDB ID: 2JWT). For EnHD (654 total NOEs), GROMACS, *il*mm, and NAMD ensembles had marginally better agreement with the NOE data (96%, 97%, and 96%, respectively) than the crystal structure alone (95%), while AMBER had marginally worse agreement (94%) (Table A.2.2). Deviations from the level of agreement with the crystal structure were small (< 2%), and there was little variation among replicate simulations. In addition, the total number of NOE violations was small. Across all force fields, the mean number of violations was 28 with an average violation distance of 0.625 Å. The number of severe NOE violations (i.e., those with a violation distance > 2 Å) was also small (AMBER, 2, mean distance = 2.6 Å; GROMACS, 2, mean distance = 2.6 Å; *il*mm, 1, mean distance = 2.7 Å; NAMD, 4, mean distance = 3.0 Å); however, there were 12 severe violations in the first NAMD replicate (Table A.2.2). For RNase H (1428 total NOEs), the X-ray structure had marginally better agreement with the experimental NOE data than any MD-generated ensemble (X-ray, 98%; AMBER, 97%; GROMACS, 97%; *il*mm, 95%; NAMD, 97%) (Table A.2.3). Again, deviations from the level of agreement with the crystal structure were small (<3.5%), and there was little variation in the agreement among replicate simulations. Across all force fields, the mean number of violations was 47 with an average violation distance of 0.775 Å. There were a number of severe NOE violations (i.e., those with a violation distance >2 Å): AMBER, 3, mean distance = 3.5 Å; GROMACS, 2, mean distance = 3.5 Å; *il*mm, 5, mean distance = 2.9 Å; NAMD, 4, mean distance = 4.3 Å) (Table A.2.3).

Next, we grouped NOEs by the residues with which they were associated and found that several residues had force-field-dependent NOE satisfaction. For example, Leu 26 of EnHD, located in the HI–HII loop, is associated with 50 NOEs: 34 were satisfied by all MD simulations, the X-ray structure, and the NMR ensemble; 1 was never satisfied; and 15 had model-dependent satisfaction (Table A.2.4). Of the 15 NOEs with model-dependent satisfaction, there were 4 NOEs satisfied by the X-ray structure and not the NMR ensemble and 10 satisfied by the NMR ensemble and not the X-ray structure (Table A.2.4). The satisfaction of these NOEs was dependent on the rotameric state of Leu 26, which was in the *t, g+* conformation in the X-ray structure and the *g–, t* conformation in the NMR ensemble (Figure 2.4). In AMBER, GROMACS, and *il*mm, Leu 26 alternated between these two conformations; however, simulations performed with NAMD largely retained the X-ray conformation for Leu 26, populating the *t, g+* conformation for 97% of the

simulation (Figure 2.4, Table A.2.5). While no one model satisfied all the NOEs associated with Leu 26, AMBER, GROMACS, and *il*mm had better agreement with the NOEs (and on par with the NMR ensemble) than the crystal structure or NAMD. In this instance, rotameric exchange was necessary to satisfy the NOEs that were not present in the crystal structure. The χ_1 torsional potentials for Ile, Leu, Asn, and Asp were modified in the ff99SB-ILDN force field, which we used in the AMBER and GROMACS simulations. The improvements made in this force field likely contributed to the modeling of L26 in the AMBER and GROMACS simulations; however, control simulations employing the ff99SB force field were not performed, so the degree of improvement cannot be quantified here. Although we found that 3 of the 4 MD packages had better agreement than the crystal structure with respect to the solution behavior of L26, we cannot say which MD-generated ensemble best agreed with the "true" behavior of L26 in solution. We found that the force fields/packages had variable populations and lifetimes of the two primary rotamer conformations (Tables S2.5 and S2.6). Furthermore, the HI–HII loop undergoes conformational exchange on the μs–ms time scale, indicating that the MD simulations may not have captured the full extent of dynamics associated with the side chain of this residue (Stollar *et al.*, 2003).

2.4.2.3 Scalar Coupling Constants

Scalar coupling constants can be related to various dihedral angles via the Karplus relation (Equation 2.2). Here we calculated the $^3J_{HN,H\alpha}$ coupling constants, which are related to the ϕ dihedral angle, using seven different parameter sets obtained from the literature for the Karplus equation (Habeck *et al.*, 2005; Ludvigsen *et al.*, 1991; Schmidt *et al.*, 1999; Smith *et al.*, 1991; Vuister *et al.*, 1993; Hu *et al.*, 1997; Pardi *et al.*, 1984). Table A.2.7 shows that the choice of Karplus parameters affects the level of agreement between simulation and experiment, with the Schmidt *et al.* and Smith *et al.* parameter sets consistently producing higher RMSEs. Overall, the Habeck parameter set, which was derived by applying Bayesian regression models to high-resolution data from ubiquitin, produced the best agreement, with correlation coefficients ranging from 0.80 to 0.89 and RMSDs ranging from 0.8 to 0.98 Hz (Figure 2.5, Table A.2.7). While most residues had excellent agreement with the experimental data, some residues, such as E22, had poor agreement independent of force field or parameter set, and some residues, such as N41, had force-field-dependent agreement (Figure A.2.9). The N41 coupling constant was not described well in the X-ray structure, the AMBER ensemble, or the NAMD ensemble; however, both *il*mm and

GROMACS-derived ensembles had excellent agreement (*il*mm error, 0.04 Hz; GROMACS error, 0.5 Hz). In all MD-generated ensembles, N41 sampled two regions of Ramachandran space: one located in the P_{IIL}, basin and one located on the boundary between the β and P_{IIL} basins (Figure A.2.9). The ratio of sampling in these two regions dictated the ensemble-averaged value for ϕ and, in turn, the coupling constant. Upon further analysis, we found that, structurally, N41 functions as a dynamic helix cap, with both the backbone carbonyl and side chain carboxamide group forming multiple hydrogen bonds with the N-terminal residues of HIII (Figure A.2.9 and Table A.2.8). These data suggest that force-field-specific hydrogen bonding patterns for N41 may have contributed to the level of agreement with experimental data; however, it is also possible that intrinsic ϕ/ψ preferences for individual amino acids, which are known to be force-field-dependent (Vymetal *et al.*, 2013), may have also influenced the agreement for N41. Ultimately, however, it is not possible to assess which MD-generated ensemble produced the best prediction for in-solution behavior, as numerous ϕ/ψ distributions can yield ensembles that satisfy the experimental data. The difficulties in evaluating simulated and experimental coupling constants are exacerbated by the fact that 2–4 ϕ-values map to a single coupling constant and use of the Karplus relation itself can introduce error in the form of the three coefficients.

2.4.2.4 S^2 Generalized Order Parameters

NMR-derived generalized S^2 order parameters of NH groups report on the local extent of motion of the polypeptide chains. We calculated the backbone order parameters for EnHD and RNase H separately for each simulation, and the reported values were averaged over the three replicates. For EnHD, there was good correspondence between the simulated and experimental values, with correlation coefficients ranging from 0.71 to 0.94 and RMSEs ranging from 0.7 to 1.3 (Figure 2.6 and Table 2.1). Across all force fields, the N-terminal residues and turns had the greatest error. Furthermore, although *il*mm had above-average errors for N-terminal residues, it produced better agreement for helical residues, particularly residues 10–14 and 48–51 (Figure 2.6 and Figure A.2.10). There was also good correspondence between simulation and experiment for RNase H (Figure 2.7 and Table 2.2). The level of agreement for Gly 15 was force-field-dependent, with *il*mm producing the best agreement (Figure 2.7 and Figure A.2.11). Gly 15 is within the Gly-rich loop, which, along with the $\beta_5\alpha_E$ loop, coordinates the DNA/RNA hybrid prior to catalysis. We performed dPCA to explore the conformational distributions of the Gly-rich loop. Our analysis

was aided by multiple X-ray and NMR structures of RNase H and several homologues. The following structures were included in our analysis: RNase H from *E. coli* (X-ray, PDB ID: 2RN2), RNase H from *E. coli* (NMR, PDB ID: 1RCH), RNase H from *Thermus thermophilus* (X-ray, PDB ID: 1RIL), a stabilized RNase H variant from *E. coli* (X-ray, PDB ID: 1GOA), and two structures of RNase H D210N from *Homo sapiens* in a complex with DNA/RNA hybrids (X-ray, PDB ID: 2QKB, in a complex with a 20-mer DNA/RNA hybrid; X-ray, PDB ID: 2QKK, in a complex with a 14-mer DNA/RNA hybrid). The first two principal components described 53% and 12% of the variance within the data set, and Gly15 had strong weights. There were several highly populated regions in PC space (Figure 2.8). Of these, one corresponded to the unbound conformation of the Gly-rich loop observed in solution (denoted by a blue arrow in Figure 2.8), and another corresponded to the bound conformation of the loop observed in solution (denoted by the red arrow in Figure 2.8). Figure 2.8c breaks down the sampling of PC space by force field and simulation number. Visualization of the PC maps shows that *il*mm was the only force field/MD package that sampled both the bound and unbound conformations; furthermore, the unbound solution conformations were the dominant conformers sampled by *il*mm. In contrast, the remaining force fields primarily sampled the X-ray conformation and another region that was not observed in the experimental data (far left of Figure 2.8). Moreover, even with 600 ns of aggregate simulation time, AMBER, GROMACS, and NAMD were unable to achieve the level of sampling seen in *il*mm with 300 ns of aggregate simulation time (i.e., *il*mm reached this degree of sampling in <100 ns). While there are dramatic differences in the sampling, there is a limit to the extent to which we can determine the level of agreement between simulation and experiment, as the number, distribution, and interconversion rates of Gly-rich loop conformations cannot be derived from static structures alone.

2.4.2.5 Global Comparison with NMR Observables

To assess the overall agreement of the modeled dynamics with experimental observables, we calculated the χ^2 statistic using the method of Panelopulos *et al.* (Pantelopulos *et al.*, 2015). χ^2 was calculated using (Equation 2.4), where N is the total number of types of experimental observables, $RMSD^{MD,experiment}$ is the RMSD between MD and experiment for a set of observables, and σ is the error associated with each individual measurement. Prior to calculation of the χ^2 statistic, we excluded data for residues that were poorly modeled in at least 3 of the 4 force

field/software package sets. Data were excluded on a per-protein, per-data-type basis. Residues were considered poorly modeled if the absolute value of the difference between the experimental and MD-derived values was greater than a data-type-specific cutoff. The cutoffs were set as twice the mean of the absolute value of the difference between the experimental and MD-derived values for all data points (Figure A.2.12). The data associated with EnHD were organized into 8 types of observables: N chemical shifts, C_α chemical shifts, C_β chemical shifts, C' chemical shifts, H_α chemical shifts, H_N chemical shifts, $^3J_{HN,H\alpha}$ coupling constants, and backbone NH order parameters. The data associated with RNase H were organized into 5 types of observables: N chemical shifts, C_α chemical shifts, H_α chemical shifts, H_N chemical shifts, and backbone NH order parameters. A value of 0.78 was used for the error associated with the coupling constants, as determined by Beauchamp *et al.* (Beauchamp *et al.*, 2012). A value of 0.1 was used for the error associated with the backbone NH order parameters. Nucleus-specific errors were used for the chemical shift data, based on the rms errors calculated for SHIFTX2: 0.44 ppm (C_α), 0.52 ppm (C_β), 0.53 ppm (C), 0.12 ppm (H_α), 0.17 ppm (H_N), and 1.12 ppm (N). In the case of RNase H, the expected rms error for SHIFTX2 versus the chemical shifts in the RefDB (Zhang *et al.*, 2003) entry for 2RN2 were used for N, C_α, H_α, and H_N chemical shifts (1.48, 0.79, 0.14, and 0.30 ppm, respectively). For EnHD and RNase H, the χ^2-values fell within the expected distribution (Figure 2.9) (Table A.2.9).

2.4.3 *High-Temperature Unfolding*

High temperature simulations have been used to probe protein folding/unfolding pathways (Daggett *et al.*, 1996; Bond *et al.*, 1997) and to aid in the design of thermostable protein variants (Childers *et al.*, 2017). In prior studies, putative protein (un)folding transition states have been identified from high-temperature simulations for multiple proteins including EnHD (Gianni *et al.*, 2003), c-Myb (Gianni *et al.*, 2003), chymotrypsin inhibitor 2 (Daggett *et al.*, 1996), barnase (Bond *et al.*, 1997; Li and Daggett, 1998), and others. We evaluated the ability of the different MD packages and force fields to model known components of the (un)folding pathway of EnHD. First, we evaluated the sampling of the transition state of protein (un)folding by comparing MD-derived *S*-values to experimentally derived ϕ-values (Gianni *et al.*, 2003). The ϕ- and *S*-values reflect the degree of structure present in the transition state along the sequence. The *S*-values were calculated

for the putative transition state ensembles identified via conformational clustering (Figure A.2.13), and the averages over the 3 replicates showed decent agreement with the experimentally derived ϕ-values for *il*mm ($R = 0.70$) and moderate correspondence for AMBER ($R = 0.48$), GROMACS ($R = 0.50$), and NAMD ($R = 0.35$) (Table A.2.10). Next, we examined whether any of the unfolding simulations sampled the known EnHD folding intermediate, the structure of which was first predicted computationally in 2000 (Mayor *et al.,* 2000; Mayor *et al.,* 2003) and later confirmed by NMR in 2005 (Religa *et al.,* 2005). This intermediate structure was observed in simulations performed with *il*mm, but not AMBER, GROMACS or NAMD. Finally, we examined whether the unfolding simulations sampled the denatured state after 10 ns at high temperature. Simulations performed with GROMACS and *il*mm sampled highly denatured states, followed by AMBER, whereas simulations performed with NAMD retained significant native-like structure over the course of the simulation (Figure 2.10). The unfolding pathway of EnHD was strongly dependent on the choice of force field and simulation software. Only one force field/software combination, *il*mm with the Levitt *et al*. force field, sampled transition state, intermediate state, denatured state structures, and kinetics of unfolding consistent with the experimental results for this system (Mayor *et al.,* 2000; Mayor *et al.,* 2003; Religa *et al.,* 2005). While GROMACS sampled somewhat appropriate transition-state-like conformations, significant helical structure was lost after passing through the transition state, resulting in increased sampling of denatured state conformers but without sampling the obligatory intermediate structures (Mayor *et al.,* 2003; Religa *et al.,* 2005). NAMD also sampled transition-state-like conformations, but in contrast to GROMACS, the protein never unfolded, preventing sampling of the intermediate and denatured states over the course of the simulations (Figure 2.10).

2.4.4 *Summary and Outlook*

The ultimate objective of MD simulations is to visualize dynamic behaviors and structural conformations that cannot be described by experimentally derived structures alone. Usually, the inability of MD simulations to produce conformational ensembles that are consistent with experiment is blamed on the force field. However, the force field is not the sole determinant of MD-simulated behavior or accuracy; if it were, the results from the simulations performed with GROMACS and AMBER would be more similar, as the same force field was used with both

packages. To this end, we encourage authors of MD studies to include the standard input files used to model and perform any published simulations. This will facilitate meta-analyses and discussions of the impact of MD software-specific parameters on modeled behavior. Given the differences that we observed, ongoing and future validation efforts must account for both the force field parametrizations as well as the methods and approximations used to propagate systems in time without conflating the two. This includes approximations that may introduce artificial periodicity, such as PME (Beck *et al.*, 2005 and references therein), or overly constrain the simulated systems, such as LINCS and SHAKE. In addition, the choice of simulation ensemble can play a role in the dynamics, as suggested by the inability of AMBER, GROMACS, and NAMD to reproduce the known unfolding behavior of EnHD. We believe that the improved ability of *il*mm to sample conformational intermediates lies in the flexibility and conformational changes that are made possible when such approximations are not allowed to constrain or impede molecular motions. In addition, the force fields use different force constants on the dihedral angles; the Levitt et al. force field uses a value of 0 for the barrier to readily allow conformational transitions subject to steric and electrostatic interactions. AMBER and CHARMM force fields use values of 1.13/1.88 kcal/mol and 1.36/1.46 kcal/mol for Φ/Ψ, respectively. These barriers, while relatively small, may also aid in retention of the starting structure such that native states are well-maintained and unfolding, even at high temperature, is discouraged. Another possibility lies in the use of microcanonical ensemble in *il*mm, which allows for Boltzmann sampling of an isoenergetic surface of the conformational landscape over continuous reaction pathways without frequent scaling of the velocities, which dampen motion.

In cases where dynamic behavior is required, restricted sampling results in worse agreement with experimental data that reflect protein dynamics in solution. We observed several instances of this in our simulations, particularly with NAMD. For example, simulations of EnHD performed with NAMD had the lowest C_α RMSDs relative to the crystal structure, but they were unable to recover NOEs associated with Leu 26 that were satisfied by the other MD packages. In addition, the coupling constant for N41 as predicted by NAMD was nearly identical to the value present in the X-ray structure. However, the X-ray structure is not in agreement with the solution behavior of N41, and *il*mm, GROMACS, and AMBER (both using the same protein force field but different TIP4P-EW and TIP3P water models, respectively) showed improved agreement with experiment. Finally, after 10 ns of simulation at 498 K, NAMD (using the CHARMM36 force

field and TIP3P water model) was unable to produce significantly denatured structures comparable to those produced by the other MD packages, and instead, all three helices remained intact and packed. This highlights a more general issue. Many force field/MD packages have been developed to maintain the conformation of the starting crystal structure by attenuating or impeding dynamics in a variety of ways, and this is generally viewed as desirable. However, this becomes problematic if one is interested in characterizing larger scale native protein dynamics or protein unfolding.

Another crucial component to improving MD software is to establish more complex means of comparing experimental and computational results in a systematic and quantitative fashion. As we have shown here, when observables were evaluated at a coarse level of detail, these MD packages showed similar agreement with experiment with respect to the native state. However, when individual residues were examined in detail, significant differences were observed in native dynamics. A prominent difference was the behavior of L26 in EnHD. While three of the four MD packages recovered many NOEs missing from the crystal structure, the underlying dynamics of L26 (i.e., the populations and lifetimes of different rotameric states) varied significantly. This observation highlights not the shortcomings of MD, but the limitations of the data used to assess the computational results. That is, the available experimental data for L26, for example, are unable to identify which MD package best models the L26 dynamics. Recent studies examining different force fields (AMBER, CHARMM, and OPLS) and their ability to correctly model the strength of salt bridges (Debiec *et al.,* 2014) and main chain propensities (Vymetal *et al.*, 2013) in simple model systems highlight how consideration of detailed interactions and behavior reveals dramatic differences between these force fields/MD packages. More broadly, however, the experimental data presented here are unable to distinguish, with high confidence, which MD-generated ensemble best approximates native protein behavior in solution. Starker differences were seen for protein unfolding, in which case only one of the four MD packages produced unfolding trajectories consistent with experiment. Moving forward, access to the underlying distributions that give rise to the experimental observables will be necessary to improve the quality of MD and to better detect when a simulation does or does not appropriately model protein dynamics.

The results presented here, along with other recent force field validation efforts, show that contemporary force fields produce models that are in similar agreement with experimental results. Additionally, these results show that certain force fields agree better with certain observables than others. Considering these results, it is not possible to prescribe a “best” model; instead, models

should be selected on the basis of the information sought. For example, simulations of intrinsically disordered proteins typically require significant alterations to model parameters to obtain accurate results; here is a case where it is undesirable to constrain conformational sampling. Along those lines, additional conformational sampling, obtained through longer or more numerous conventional MD simulations, may aid in the identification of an "optimal" model, but it is not guaranteed to do so. Pantelopulos *et al.* found that longer simulation times yielded better agreement with experiment, independent of the force field chosen (Gianni *et al.*, 2003). In an evaluation of order parameter agreement, Bowman found that a larger aggregate sampling time yielded better agreement for side chain methyl group order parameters, but essentially no change in the level of agreement for backbone order parameters (Bowman *et al.*, 2016). That study also concluded that the aggregate simulation time and method used to calculate observables affected the level of agreement more than the force field (Bowman *et al.*, 2016). However, these comparisons were all between MD packages that constrain motion, and here we found that flexible molecular representations and simulation protocols that do not artificially restrain motion provide greater sampling of conformational space in shorter periods of time. Nonetheless, increased conformational sampling, whether obtained through longer or more numerous simulations or choice of MD/FF package, should facilitate model selection in cases where the data clearly indicate that observation of the dynamics across longer time scales is necessary.

2.5 Conclusions

Our results show that the MD programs and force fields studied here show comparable agreement overall with experimental data for the native state. However, we observed instances where the MD packages generated distinct conformational ensembles that agreed equally well with the experimental data. This underscores the fact that agreement with experimental data is necessary, but not sufficient, to validate atomistic simulations. The four MD package/force field combinations unquestionably produced distinct ensembles. For example, hydrogen bond networks, including both the residues engaged in the networks as well as the frequency of different interactions, were variable across the MD packages. While these differences in dynamics may be small in magnitude, such dynamic modes form the background over which more extensive conformational changes occur. Ultimately, quantitative comparisons between such rapid, small

amplitude motions and experimental data should enhance our ability to isolate the “True” ensembles present in solution.

2.6 Tables

Table 2.1. Correspondence between simulation and experiment for EnHD

	AMBER		GROMACS		*il*mm		NAMD	
	R[b]	**RMSE[c]**	**R[b]**	**RMSE[c]**	**R[b]**	**RMSE[c]**	**R[b]**	**RMSE[c]**
B–factor	0.82	0.65	0.87	1.24	0.83	0.4	0.83	1.1
Cα CS[a]	0.98	0.83	0.98	0.83	0.96	0.95	0.98	0.75
Cβ CS	0.99	0.66	0.99	0.7	0.99	0.82	0.99	0.71
C CS	0.90	1.69	0.88	1.63	0.91	1.84	0.92	1.57
N CS	0.84	1.02	0.85	1.1	0.82	0.92	0.86	0.86
Hα CS	0.92	0.13	0.92	0.13	0.87	0.17	0.92	0.14
H CS	0.75	0.38	0.77	0.37	0.79	0.35	0.82	0.34
$^3J_{HN,H\alpha}$	0.68	1.10	0.82	0.70	0.70	0.94	0.83	0.87
S^2	0.71	0.13	0.90	0.08	0.78	0.11	0.94	0.07

a. CS = Chemical shift
b. R = Pearson's correlation coefficient
c. RMSE = root mean squared error

Table 2.2. Correspondence between simulation and experiment for RNase H

	AMBER		GROMACS		*il*mm		NAMD	
	R[b]	**RMSE[c]**	**R[b]**	**RMSE[c]**	**R[b]**	**RMSE[c]**	**R[b]**	**RMSE[c]**
B–factor	0.67	0.32	0.62	0.38	0.49	0.37	0.68	0.33
Cα CS[a]	0.98	0.82	0.98	0.78	0.97	1.19	0.99	0.73
N CS	0.92	2.48	0.94	2.32	0.86	3.16	0.93	2.34
Hα CS	0.93	0.26	0.94	0.23	0.86	0.36	0.93	0.25
H CS	0.83	0.42	0.85	0.39	0.77	0.47	0.86	0.37
S^2	0.85	0.10	0.90	0.09	0.83	0.11	0.88	0.09

a. CS = chemical shift
b. R = Pearson's correlation coefficient
c. RMSE = root mean squared error

Table 2.3. Average (± standard deviation) C_α RMSDs for EnHD and RNase H native state simulations. The first 20 ns of each simulation was excluded in this calculation.

	AMBER		GROMACS		*il*mm		NAMD	
EnHD	**Avg. (Å)**	**St. Dev. (Å)**	**Avg. (Å)**	**St. Dev. (Å)**	**Avg. (Å)**	**St. Dev. (Å)**	**Avg. (Å)**	**St. Dev. (Å)**
Run 1	0.67	0.10	0.59	0.11	1.07	0.26	0.60	0.09
Run 2	0.95	0.22	0.64	0.14	0.85	0.16	0.56	0.08
Run 3	0.75	0.13	0.58	0.09	1.01	0.16	0.57	0.10
Ensemble	0.79	0.20	0.61	0.12	0.98	0.22	0.57	0.09
RNase H	**Avg. (Å)**	**St. Dev. (Å)**	**Avg. (Å)**	**St. Dev. (Å)**	**Avg. (Å)**	**St. Dev. (Å)**	**Avg. (Å)**	**St. Dev. (Å)**
Run 1	1.57	0.22	1.35	0.14	2.50	0.12	1.44	0.23
Run 2	1.45	0.16	1.37	0.12	2.28	0.26	1.50	0.21
Run 3	1.58	0.24	1.30	0.13	2.35	0.33	1.38	0.12
Ensemble	1.53	0.22	1.34	0.14	2.38	0.27	1.44	0.20

2.7 Figures

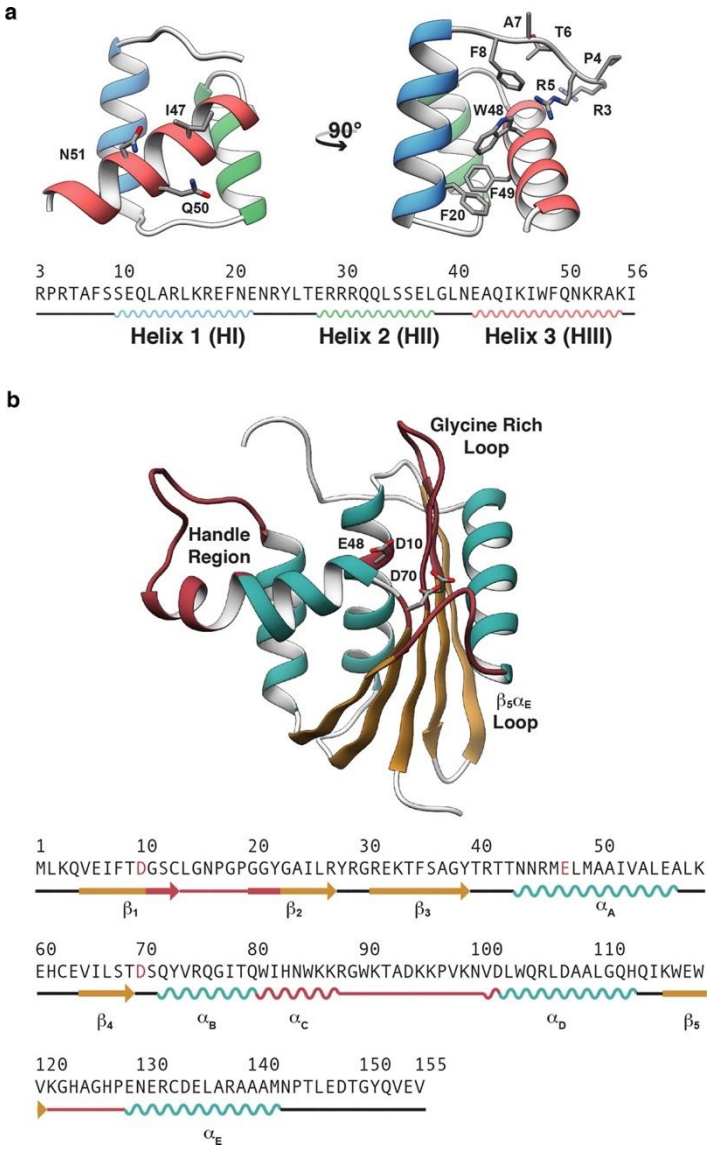


Figure 2.1. X-**ray crystal structures of the engrailed homeodomain and ribonuclease H**.
(a) The crystal structure and sequence of the engrailed homeodomain are shown. In the front view (left), the DNA-binding residues in HIII are represented as balls and sticks; on the side view (right), the DNA-binding residues on the N-terminus as well as four aromatic residues within the hydrophobic core are shown as balls and sticks. (b) Here the crystal structure and sequence of ribonuclease H are shown with the β-sheet colored tan, α-helices cyan, and functional regions burgundy

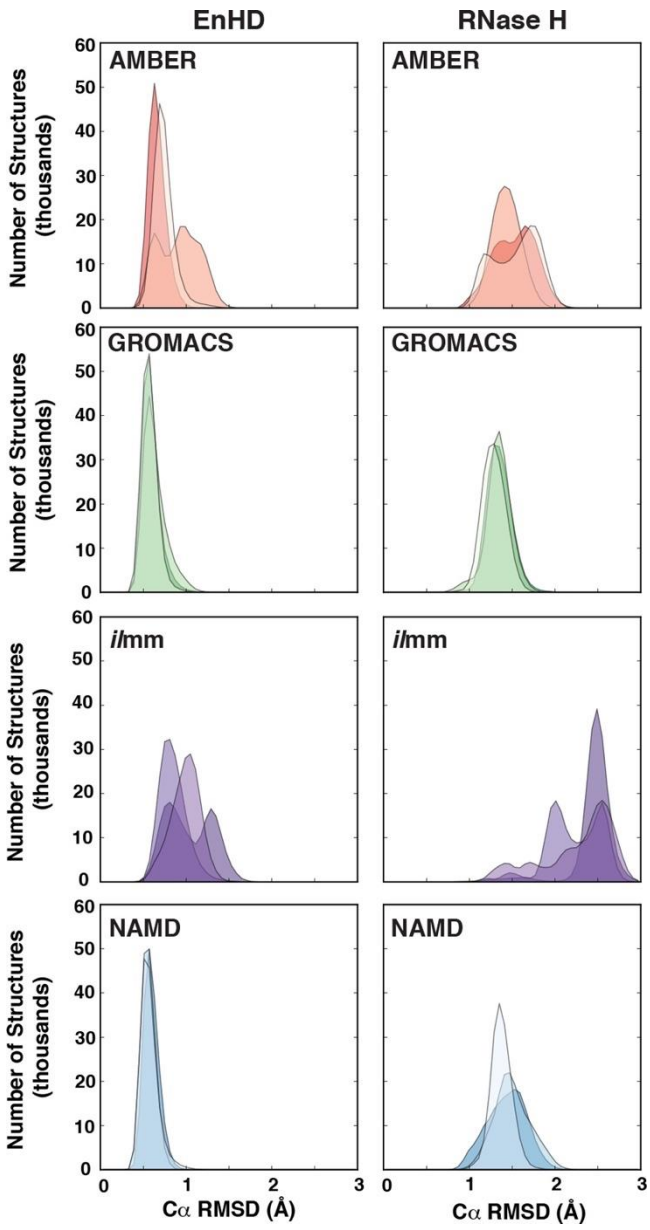


Figure 2.2. Distribution of Cα root mean squared deviations for EnHD and RNase H. Overlaid histograms of Cα RMSDs were constructed for each of the three replicate simulations of EnHD (left) and RNase H (right) for simulations performed with AMBER (orange), GROMACS (green), *il*mm (purple), and NAMD (blue) at 298 K.

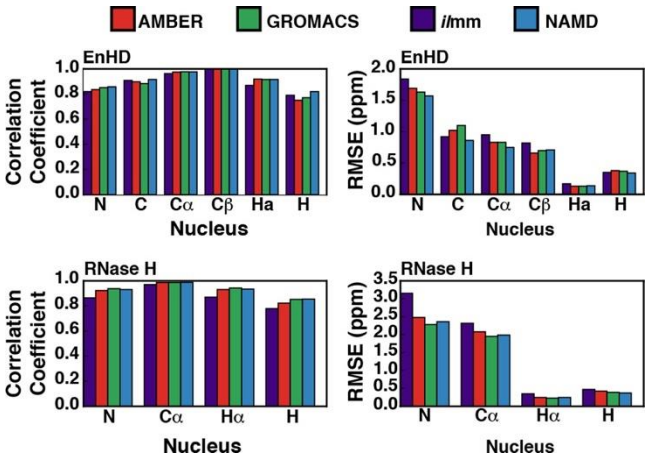


Figure 2.3. Correspondence between MD-derived and experimental chemical shifts.

The correlation coefficients (left column) and RMSEs (right column) for the chemical shift correspondence for EnHD (top row) and RNase H (bottom row) are shown, stratified by nucleus type.

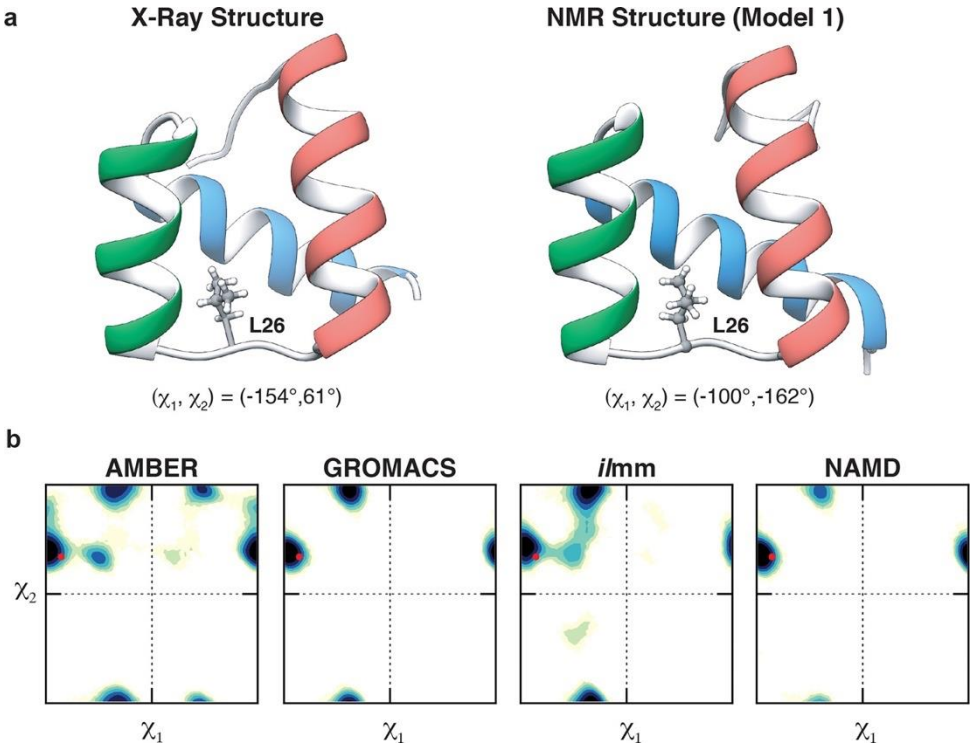


Figure 2.4. Conformational heterogeneity in Leu 26.

(a) Leu 26 occupies distinct rotameric conformations in the X-ray structure and NMR ensemble of EnHD. In the X-ray crystal structure (left), L26 occupies the *t*, *g*+ conformation while it occupies the g–, *t* conformation in the NMR ensemble (right). (b) Side chain χ_1/χ_2 dihedral angle maps for the different MD packages. The red point denotes the conformation of L26 in the X-ray structure.

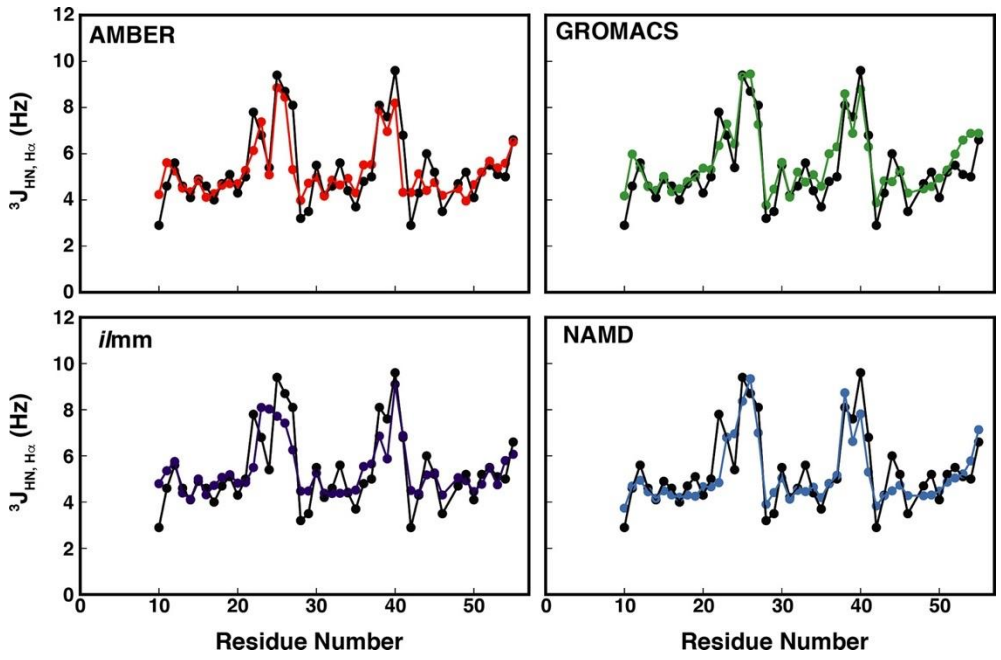


Figure 2.5. MD simulations reproduce $^3J_{HN,H\alpha}$ coupling constants for EnHD.

Simulated vs experimental coupling constants are plotted as a function of residue number. Here, the Habeck *et al.* parameters have been used.

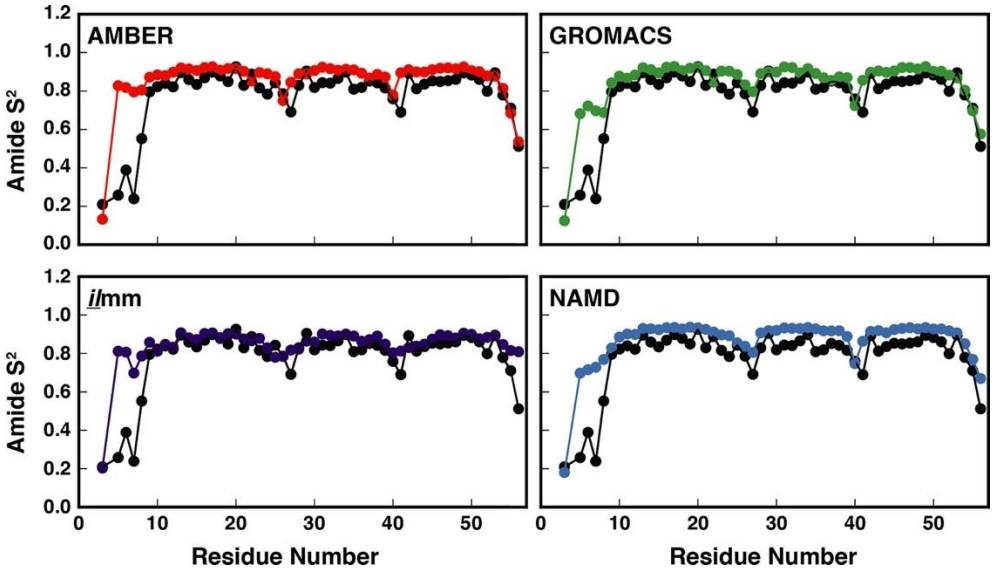


Figure 2.6. Correspondence between experimental and MD-derived order parameters for EnHD.

The experimental (black points) and MD-derived order parameters (red points, AMBER; purple points, *il*mm; green points, GROMACS; and blue points, NAMD) are plotted as a function of residue number for EnHD. Excellent correspondence was observed for all force fields except at the N-terminus. A larger number of simulations or significantly longer simulations are required for MD to reproduce the order parameters for highly flexible terminal residues that may become trapped in local minima.

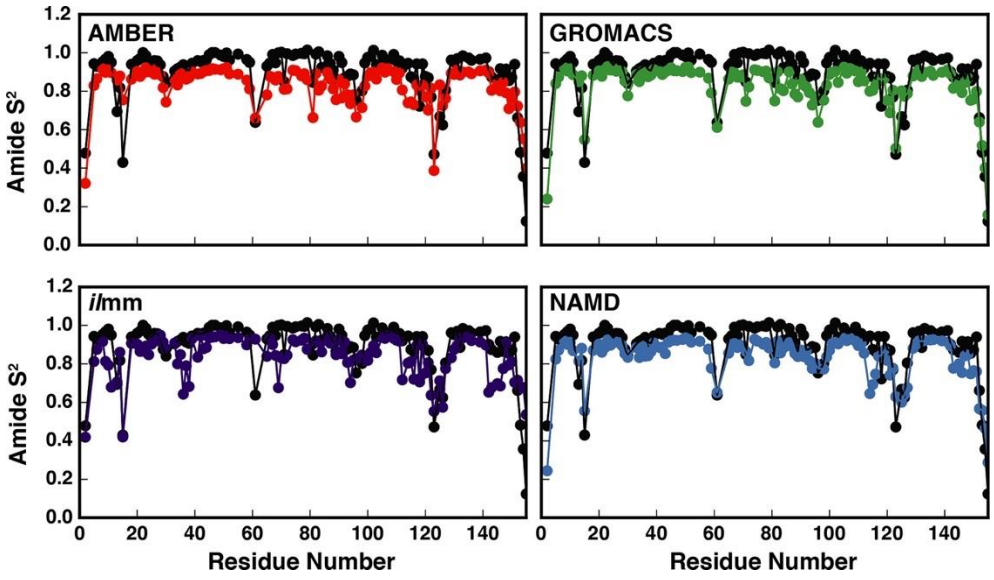


Figure 2.7. Correspondence between experimental and MD-derived order parameters for RNase H.

The experimental (black points) and MD-derived order parameters (red points, AMBER; purple points, *il*mm; green points, GROMACS; blue points, NAMD) are plotted as a function of residue number for RNase H.

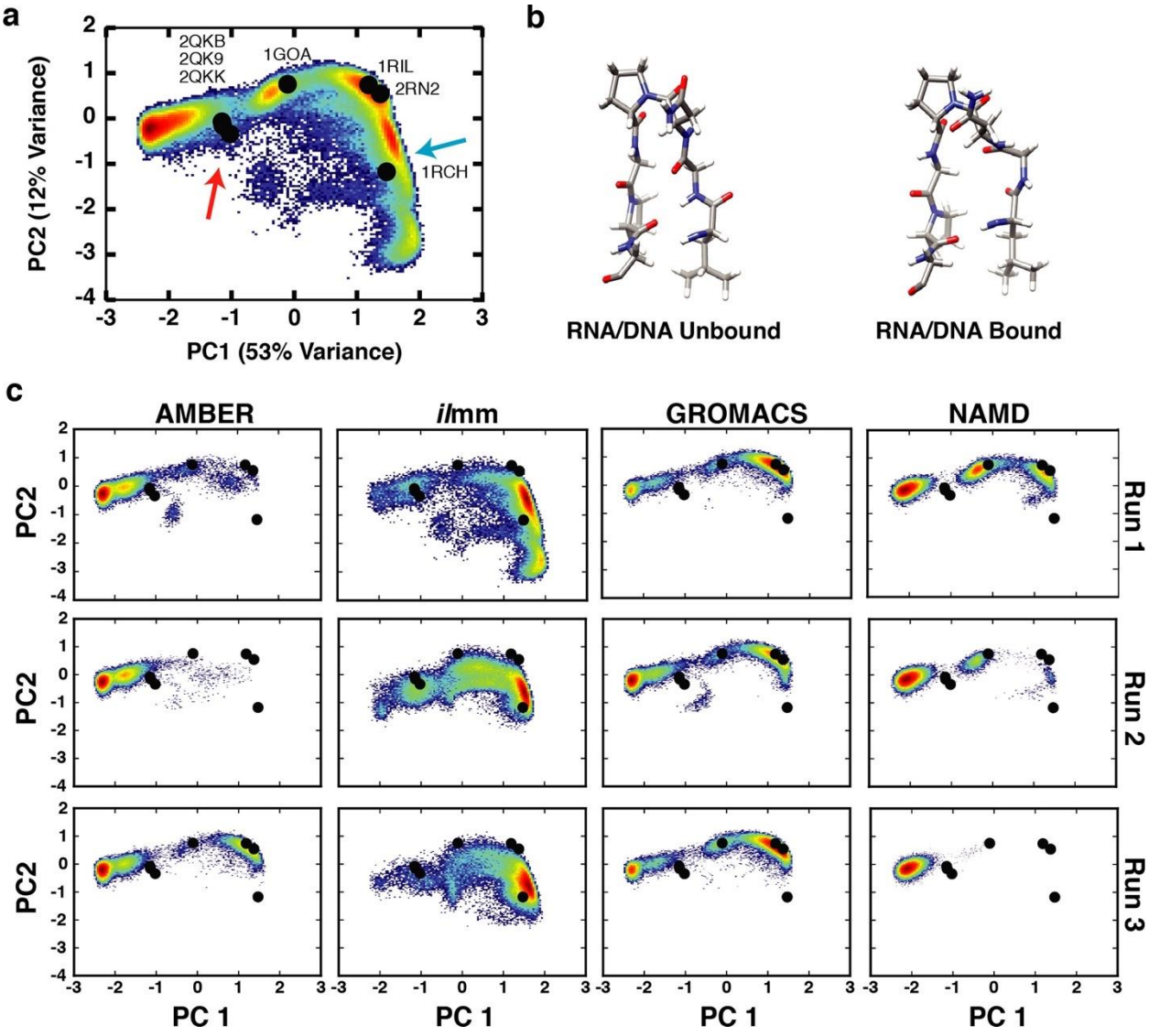


Figure 2.8. Conformational heterogeneity in the Gly-rich loop.

(a) The dPCA landscape for the residues in the Gly-rich loop, constructed using conformations aggregated from the experimental reference structures and MD simulations, maps the conformational heterogeneity in the Gly-rich loop. Black points denote the location of RNase H reference structures within the dPCA landscape. The blue arrow denotes the region corresponding to the unbound conformation of the Gly-rich loop in solution (b, left). The red arrow denotes the region corresponding to the DNA/RNA bound conformation of the Gly-rich loop in solution (b, right). (c) Conformations sampled by the Gly-rich loop in MD simulations.

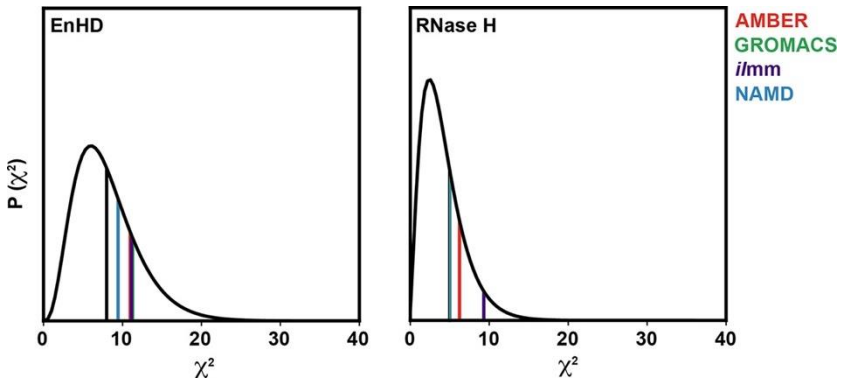


Figure 2.9. Global correspondence between simulation and experiment as assessed by the χ^2 statistic.

Each plot shows the χ^2 distribution (black curves) for the degrees of freedom associated with each comparison (8 degrees of freedom for EnHD, left; 5 degrees of freedom for RNase H, right). The vertical lines denote the χ^2-value calculated for each force field/software package combination (AMBER, red; GROMACS, green; *il*mm, purple; NAMD, blue) as well as the expectation value for that distribution (black).

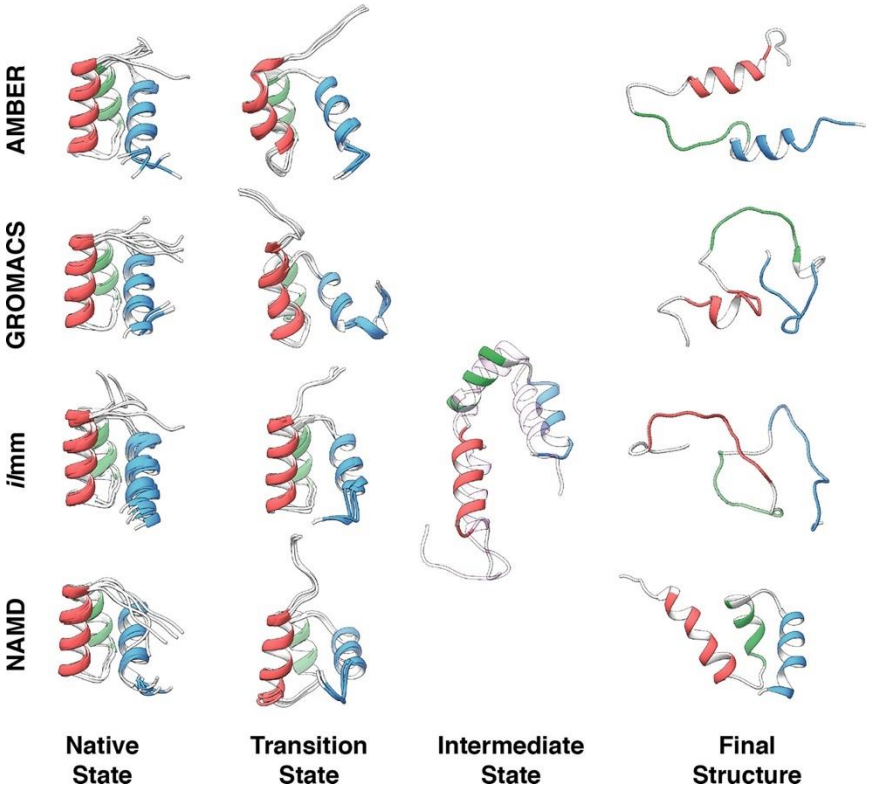


Figure 2.10. Survey of conformations populated during unfolding of EnHD.

Native state: Five snapshots, extracted from simulation one of the 298 K simulations at 0, 50, 100, 150, and 200 ns, serve as a visual reference of native state sampling. Transition state: For each of the four MD packages, the transition state is represented by three overlaid structures extracted from the simulation that best modeled the experimentally determined transition state. Intermediate state: One intermediate state structure was extracted from the *il*mm simulation that best modeled the experimentally determined intermediate state. Model 1 of the experimentally determined intermediate state structure is shown as a transparent ribbon and aligned against the MD-derived intermediate structure. Final structure: To represent the extent of unfolding that occurred during high-temperature MD, the most-disrupted final structures from the 498 K replicates are shown.

Chapter 3. DRIVERS OF SECONDARY STRUCTURE CONVERSION IN TRANSTHYRETIN AT THE OUTSET OF AMYLOID FORMATION

3.1 Summary

Amyloid diseases are a set of fatal disorders in which proteins aggregate to form fibrils that deposit in tissues throughout the body. Amyloid diseases are challenging to study because amyloid formation occurs on timescales that span several orders of magnitude and involve heterogeneous, interconverting protein conformations. The development of more effective technologies to diagnose and treat amyloid disease requires both a map of the conformations sampled during amyloidogenesis and an understanding of the molecular mechanisms that drive this process. In prior molecular dynamics simulations of amyloid proteins, we observed the formation of a nonstandard type of secondary structure, called α-sheet, that is associated with the pathogenic conformers in amyloid disease – the soluble oligomers. However, the detailed molecular interactions that drive the conversion from β-sheet to α-sheet remain elusive. Here we use molecular dynamics simulations to interrogate a critical event in transthyretin aggregation: the formation of aggregation-competent, monomeric species. We show that conformational changes in one of the two β-sheets in transthyretin enable solvent molecules and polar side chains to form electrostatic interactions with main chain peptide groups that facilitate and modulate conversion to α-sheet secondary structure. Our results shed light on the early conformational changes that drive transthyretin towards the α-sheet structure associated with toxicity. Delineation of the molecular events that lead to aggregation at atomic resolution can aid strategies to target the early, critical toxic soluble oligomers.

3.2 Introduction

Human transthyretin (TTR) is a homotetrameric protein comprised of four 127-residue monomers, each with a β-sandwich topology (Blake *et al.*, 1971, 1978). The monomeric subunits contain eight β-strands arranged into two four-stranded β-sheets and an α-helix. The β-strands are labelled A-H and the corresponding sheets are referred to as the DAGH and CBEF sheets (Figure

3.1). A TTR dimer is formed when two monomers associate at strands H and H', leading to two 8-stranded β-sheets: DAGH-H'G'A'D' and CBEF-F'E'B'C'. In the tetrameric conformation two dimers associate with the DAGH-H'G'A'D' sheets facing one another. TTR primarily functions as a transporter of T4 in the serum and cerebrospinal fluid and of retinol via an association with retinol binding protein (Davis *et al.*, 1972; Jaarsveld *et al.*, 1973; Southwell *et al.*, 1992; Johnson *et al.*, 2012); Vieira and Saraiva (Vieira and Saraiva, 2014) and Alsheheri *et al.* (Alshehri *et al.*, 2015) review other recently discovered roles of TTR.

Although the broad features of TTR amyloidogenesis have been determined, the conformational changes to TTR monomers that result in aggregation-competent species, as well as the ways that mutations modulate these conformational changes, remain to be fully fleshed out. Knowledge of such processes is necessary for the development of atomistic models of protein aggregation and for the engineering of therapeutic strategies to mitigate amyloidosis by targeting aggregation-competent monomers and toxic, soluble oligomers. Two factors obscure atomic-level characterization of the misfolding and subsequent oligomerization of amyloid species by both computational and experimental methods. First, the molecular processes involved in amyloid diseases occur on timescales that span several orders of magnitude. Second, the molecular details of the misfolding and oligomerization are challenging to obtain because of the numerous and heterogeneous species formed during amyloidogenesis.

In prior computational studies, we predicted that the early stages of TTR misfolding are characterized by the destabilization of the DAGH sheet via the formation of a nonstandard secondary structure, α-sheet (Armen *et al.*, 2004a). α-sheet is defined as an extended chain secondary structure in which sequential residues alternate between the right-handed (α_R) and left-handed (α_L) helical regions of Ramachandran space (respectively centered on $\varphi, \psi = -87°, -49°$ and on $\varphi, \psi = 45°, 92°$) (Armen *et al.*, 2004b). Like β-sheet secondary structure, polypeptide chains that adopt α-sheet structure form extended conformations; however, unlike β-sheet structure, the carbonyl groups are aligned on one side of the main chain and the amide groups on the other. Subsequent simulations of TTR variants showed that certain pathological mutations also result in α-sheet formation in TTR (Steward *et al.*, 2008). The observation of this secondary structure in multiple amyloid-associated proteins with diverse native sequences and topologies led to the development of the α-sheet hypothesis of amyloid disease (Daggett, 2006; Bi and Daggett, 2018). The α-sheet hypothesis posits that α-sheet secondary structure is present during the early stages

of amyloidogenesis, namely within aggregation-competent monomers and soluble oligomers, and that the unique biophysical properties of the secondary structure contribute to or drive protein aggregation. Follow-up experimental studies demonstrated that peptides designed to interact with α-sheet secondary structure impede aggregation of TTR(Hopping *et al.*, 2014; Kellock *et al.*, 2016) and other (Hopping *et al.*, 2014; Kellock *et al.*, 2016; Bleem *et al.*, 2017; Maris *et al.*, 2018; Paranjapye and Daggett, 2018; Shea *et al.*, 2019) amyloid proteins. However, the detailed molecular mechanisms that drive conversion of β-sheet to α-sheet structure as well as the factors (such as native topology and amino acid composition) that modulate conversion warrant further investigation.

Here, we present the results of molecular dynamics (MD) simulations of WT TTR as well as 6 pathogenic variants (D18G (Hammarström *et al.*, 2003), A36P (Salvi *et al.*, 2012), L58H (Nichols *et al.*, 1989), Y69H (Ziskin *et al.*, 2015), L111M (Ranløv *et al.*, 1992), and V122I (Jacobson *et al.*, 1990)). These specific mutations were selected to sample potential effects of mutations in various regions of the native TTR structure. We use our simulations to probe the mechanism by which secondary structure conversion occurs in TTR. Specifically, we found that exposure of the DAGH sheet to solvent in the monomeric state triggers a set of conformational changes in the DAGH sheet that allows solvent molecules and polar side chains to interact with main chain carbonyl groups and drive secondary structure conversion. Based on these results, we speculate that conformational changes occur more readily in the DAGH sheet because Nature has 'designed' the CBEF sheet to be solvent-exposed while the DAGH sheet is in a buried channel in the native tetramer.

3.3 Methods

3.3.1 *Model building*

The coordinates of the 1.7 Å crystal structure of wild type (WT) transthyretin (PDB ID: 1TTA) were obtained from the Protein Data Bank (PDB) (Berman *et al.*, 2000). The 1TTA crystal structure was chosen since it contains coordinates for the full-length protein and did not require modeling of the unstructured N- and C- termini. When duplicate side chains were present (residues C10, M13, L17, K48, E63, D74, K80, L82, R104, and T119), the first rotameric state (conformer

A) was chosen. Models of the six mutant forms of TTR (D18G, A36P, L58H, Y69H, L111M, and V122I) were generated via *in silico* mutations of the WT structure (chain A of the 1TTA X-ray structure) using the Dynameomics rotamer library (Towse *et al.*, 2016a). As previously noted, mutations to TTR rarely result in a significant deviation from the crystal structure present in 1TTA (Palaninathan, 2012). Each of the seven TTR variants was modeled under the most common *in vitro* model of amyloidogenic conditions: acidic pH (~pH 4.4) via protonation of Asp (called Ash, net charge of 0), Glu (called Glh, net charge of 0), and His (called Hip, net charge of +1), residue.

3.3.2 *Molecular dynamics simulations*

The starting structures described above were prepared for molecular dynamics (MD) simulations using the *in lucem* molecular mechanics (*il*mm) package (Beck *et al.*, 2000), which was recently validated against a set of >3,100 experimental observables (Childers and Daggett, 2018). First, missing hydrogen atoms were modeled on the crystal structure and then minimized for 500 steps. Next, all atoms were minimized via steepest descent minimization for 1000 steps. Next, the proteins were solvated in a water box that extended at least 10 Å beyond any protein atom and the box volume was adjusted to reproduce the experimental density at 310 K (0.992 g/ml) (Kell, 1967). Solvent atoms were minimized for 1000 steps, equilibrated for 500 steps, and then minimized again for 500 steps. Finally, all protein atoms were minimized for 500 additional steps. Production MD simulations were performed using *il*mm (Beck *et al.*, 2000) with the Levitt *et al.* force field (Levitt *et al.*, 1995) and the flexible three-center (F3C) water model (Levitt *et al.*, 1997). Simulations were performed using the microcanonical NVE (constant **n**umber of particles, **v**olume, and **e**nergy) ensemble with periodic boundary conditions, a 10 Å force-shifted non-bonded cutoff (Beck *et al.*, 2005), and a 2 fs timestep. Coordinates were saved every picosecond for analysis. All (n = 21) production simulations were performed at 310 K, acidic pH, and in triplicate for 500 nanoseconds (ns), for an aggregate simulation time in excess of 10 μs.

3.3.3 *Simulation analysis*

Unless specified otherwise, MD simulations were analyzed using *il*mm and the first 25 ns of the production portions of the simulations were excluded from analysis. All protein images were prepared using UCSF Chimera (Pettersen *et al.*, 2004).

Root-mean-squared deviations The C_α root mean squared deviation (RMSD) was calculated for TTR monomers as a function of time and residue number was calculated after alignment to the 'core residues' (residues 12-123) of the minimized starting structure. RMSDs were reported for 'core residues' as well as for residues within individual strands.

Secondary structure detection Secondary structure types were classified based on repeated patterns of main chain dihedral angles. Residues were classified as having α-sheet secondary structure if at least three sequential residues adopted alternating α_R (ϕ,ψ = -87° ± 35°, -49° ± 35°) and α_L (ϕ,ψ = 45° ± 35°, 92° ± 35°) main chain dihedral angle motifs. Residues were classified as having β-sheet secondary structure if at least three sequential residues adopted β (ϕ,ψ = -115° ± 32.5°, 130° ± 50°) main chain dihedral angle motifs. Residues were classified as having α-helix secondary structure if at least three sequential residues adopted α_R (ϕ,ψ = -87° ± 35°, -49° ± 35°) main chain dihedral angle motifs.

Contact analysis Protein-protein and protein-solvent hydrogen bonds were identified when the distance between the donor-acceptor pair was less than 2.6 Å and no greater than 45° from linearity. Two side chains were considered in contact with one another if at least one pair of atoms from the two side chains were within 5.4 Å (carbon-carbon interactions) or 4.6 Å (other interactions) of one another. Side chain interaction networks were built using Cytoscape (Shannon *et al.*, 2003).

Statistical methods We evaluated whether there were statistically significant differences in the dynamics between the DAGH and CBEF sheets in the set of 21 MD simulations. To do so, we calculated the average α-sheet secondary structure content, average hydrogen bond population, average sampling of main chain geometries, and several network-level properties of solvent-exposed side chain interactions for both sheets in all (n = 21) simulations. Then, unpaired, two-sided Student's T-tests (as implemented in *scipy*) were performed for the null hypothesis that the DAGH and CBEF sheets have identical dynamic behaviors. Properties of the side chain interaction networks were calculated using the *NetworkAnalyzer* (Doncheva *et al.*, 2012) tool within *Cytoscape* (Shannon *et al.*, 2003).

3.4 Results

A total of 21 MD simulations of monomeric TTR were performed under amyloidogenic conditions. 7 unique TTR sequences were simulated: wild-type human TTR and six pathogenic mutations: D18G, A36P, L58H, Y69H, L111M, and V122I. Independent replicate ($n = 3$) simulations were performed for each unique structure. The sequences and simulations covered in this study are summarized in Table 3.1. During analysis, we found that the simulated structures shared dynamic characteristics and we focused our final analysis and discussion using an ensemble averaging approach. In the ensemble average, we combined the results from all ($n = 21$) simulations to highlight conformational changes that were independent of sequence. Results for specific simulations are provided in the supplementary information and are discussed separately to illustrate shared behavior and points of divergence.

3.4.1 *Nonnative conformations were sampled during MD*

We identified conformational changes to the tertiary structure of TTR monomers by calculating the C_α root-mean-squared deviation (RMSD) of structures sampled by MD relative to the minimized reference structure. C_α RMSD values were obtained after alignment to a set of 'core' residues, which included residues in strands A, B, E, and G (residue numbers 11-18, 28-36, 65-74, and 103-112, respectively). This set of 'core' residues excluded the primary sources of conformational heterogeneity within the simulations: the edge strands and loops connecting the secondary structure elements. C_α RMSD values were calculated for residues 11-123, which excludes flexible N- and C-terminal residues (Figure 3.2, Table 3.2). The TTR monomers had an average 3.7 Å C_α RMSD relative to the minimized reference structure, although there was variation: run 1 of the L58H simulations had the lowest RMSD (2.5 Å) and run 2 of the L58H simulations had the largest RMSD (7.2 Å) (Table 3.2). We averaged the C_α RMSD for all 21 simulations as a function of time and residue number (Figure 3.2). The average results as a function of time identified a period of rapid changes to TTR monomers during the first 50 nanoseconds of the trajectories followed by a period of slower movement away from the native structure (Figure 3.2A). Large C_α RMSD values for strands D and H correspond to partial dissociation of these strands from the DAGH sheet. As expected, results obtained for individual simulations showed

variation in these trends: not all simulations experienced large-scale conformational changes (Figure A.3.1, A.3.2).

3.4.2 *Secondary structure conversion in transthyretin*

In addition to changes in the tertiary structure, we monitored changes in the secondary structure of TTR. As in previous studies (Armen *et al.*, 2004a-b), we observed changes in the secondary structure of the DAGH sheet, specifically, the loss of β-sheet and formation of α-sheet secondary structure (Figure 3.3). We quantified the secondary structure content in TTR simulations by calculating the average percentage of simulation time that each residue adopted β-sheet, α-sheet, or α-helical main chain conformations. Secondary structure conformations were analyzed over two portions of the simulation: 25 – 500 ns and 475 – 500 ns; the latter represents the conformations populated at the end of the simulations (Figure 3.4). Overall, the α-helix and the CBEF sheet had the highest fraction of native secondary structure content averaged over all simulations (Figure 3.4A). On the timescale studied, the CBEF sheet was more stable than the DAGH sheet and had a higher β-sheet content while the DAGH sheet had higher α-sheet content. Within the CBEF sheet, residues V32, F33, K70, V71, Y69, R34, and V93 retained their native secondary structure for more than 90% of the aggregate simulation time (25-500 ns). Strands B and E retained more native-like secondary structure than any other strands. Residues in the CBEF sheet largely evaded conversion to α-sheet, the most frequent exception was for the residues that were close to the α-helix and in the least-structured region of the CBEF sheet (residues 28-29, 48-49, and 73-74). Strands D (residues 53-56) and H (residues 115-123) in the DAGH sheet had the least amount of native β-sheet structure, and strand A (residues 13-18) had the highest α-sheet content (>23 %, 25-500 ns) (Figure 3.4A). While the extent of α-sheet formation was variable from simulation to simulation, the specific residues that converted were consistent (Figure A.3.3).

Visualization of the secondary structure content as a function of time showed that β-sheet structure within the CBEF sheet was more stable than within the DAGH sheet. At the end of the simulations (475-500 ns), on average, the CBEF sheet had 3% α-sheet content (Figure 3.4B). In the DAGH sheet, β-sheet structure was rapidly lost with a rapid increase in the coil content, followed by a slower increase in α-sheet content. At the end of the simulations, on average, the DAGH sheet had 24% α-sheet content (Figure 3.4B). The differential rates of α-sheet conversion

in the DAGH and CBEF sheets suggest that properties of the β-sheets themselves, such as topology and amino acid composition, may alter the propensity to undergo conversion to α-sheet secondary structure.

The average behavior demonstrated a contrast between the secondary structures adopted by the CBEF and DAGH sheets, but there was also variability in the total amount and rate of α-sheet formation within the DAGH sheet between replicate simulations (Figure A.3.3, A.3.4). The extent of α-sheet formed on the timescale studied varied from simulation to simulation. In nine simulations (WT runs 2 and 3; D18G run 3; L58H runs 2 and 3; Y69H run 1; L111M runs 1 and 2; and V122I run 3) or 43% of all simulations the DAGH sheet had > 30% α-sheet content. In twelve simulations (WT run 1; D18G runs 1 and 2; A36P runs 1, 2, and 3; L58H run 1; Y69H runs 2 and 3; L111M run 3; V122I runs 1 and 2) or 57% of all simulations, the DAGH sheet had < 30% α-sheet content. Across the set of 21 TTR simulations, residues the DAGH sheet sampled α-sheet secondary structure for 23.9 ± 22.4% of the final 25ns of the simulations while the CBEF sheet sampled α-sheet structure for 2.8 ± 3.9% of the final 25ns of the simulations ($p << 0.01$) (Figure A.3.5).

The orientation of the α-sheet was consistent in stable α-sheets: the carbonyl groups were always oriented toward strand D in the DAGH sheet and the amide groups toward strand H. Within the DAGH sheet, there was also a consistent pattern as to which residues converted to α_R (M13, K15, L17, L55, T106, A108, L110, Y116, T118, and A120) and which converted to α_L (L12, V14, V16, D18, E54, H56, I107, A109, L111, S117, T119, and V121).

3.4.3 *Conformational changes precede secondary structure conversion*

These simulations provided enough conformational sampling to analyze the mechanism of conversion to α-sheet secondary structure. Conversion to α-sheet structure in the DAGH sheet occurred via a peptide-plane flipping mechanism that preserved the positions of the C_α and side chain atoms (Figure 3.3) (Armen *et al.*, 2004b; Hayward, 2008). During the peptide plane flip, two residues (residues *i* and *i+1*) converted from β-sheet to α-sheet simultaneously (residue *i* to the α_R region of Ramachandran space and residue *i+1* to the α_L region). We consistently observed three conformational changes that preceded peptide plane flipping: (1) a population of distorted main chain geometries, (2) a loss of main chain – main chain hydrogen bonding patterns, and (3) a

reorganization of solvent-exposed side chain – side chain interaction networks. Each of these processes is described in more detail below.

(1) Prior to conversion to α-sheet, residues in the DAGH sheet heavily sampled distorted main chain geometries in which the main chain groups that form the peptide bond were skewed out of the hydrogen bonding plane of the sheet, which we call 'pleated main chain geometry.' To analyze the conformational sampling of the main chain dihedral angles during conversion from β-sheet to α-sheet, we defined two dihedral angles that monitor the degree of pleating: θ_1 (defined by atoms H_α, C_α, C, and O) and θ_2 (defined by atoms H, N, C_α, and H_α). We classified the main chain geometry of residues in the DAGH and CBEF sheets based on the values of θ_1 for residue *i* and θ_2 of residue *i+1* (Figure 3.5). We found that pleated peptide plane geometries were sampled *en route* to α-sheet conversion and that pleated geometries allow for interactions between solvent molecules and main chain carbonyl groups (Figure 3.6).

Formation of pleated peptide planes was observed for residues in the DAGH sheet and to a lesser extent in the CBEF sheet (Figure 3.7). Peptide bonds in the DAGH sheet adopted a mix of pleated and α-sheet like conformations, whereas peptide bonds in the CBEF sheet tended to adopt a mix of pleated and β-sheet like conformations (Figure 3.7). Across the set of 21 TTR simulations, native β-sheet like main chain geometries among residues within the CBEF sheet were retained to a greater extent than among residues within the DAGH sheet ($37 \pm 4\%$ vs $24 \pm 5\%$, $p << 0.001$). In contrast, non-native α-sheet like main chain geometries were more prevalent in the DAGH sheet relative to the CBEF sheet ($13 \pm 9\%$ vs $3 \pm 3\%$, $p << 0.001$). The average sampling of pleated main chain geometries was more similar, with the DAGH sheet sampling pleated geometries slightly more frequently than the CBEF sheet ($63 \pm 6\%$ vs $59 \pm 3\%$, $p = 0.009$) (Figure A.3.6).

Strand H was most likely to form pleated conformations, and strand B was least likely to form pleated conformations. Within the DAGH sheet, the peptide bonds between residues 11 and 12, 53 and 54, 54 and 55, 107 and 108, 115 and 116, and 117 and 118 were the most susceptible to pleating (Figure 7). Within the CBEF sheet, peptide bonds between residues 34 and 35, 42 and 43, 44 and 45, 67 and 68, and 95 and 96 were the most susceptible to pleating (Figure 3.7). Residues that retained native-like β-sheet conformations, particularly F33, were the least susceptible. Peptide bonds involving a Gly or Pro residue (P11, P43, G47, G53, and G67) had above-average sampling of pleated peptide bond conformations due to the altered steric constraints

for these residue types. In pleated conformations, the peptide main chain carbonyl (of residue *i*) and amide (of residue *i+1*) groups were approximately perpendicular to both the main chain and the plane of main chain hydrogen bonds in the sheet. The carbonyl groups were oriented away from the hydrophobic core of the β-sandwich and towards solvent while the amide groups were oriented towards the hydrophobic core and away from solvent (Figure 3.7). Pleating of residues within the DAGH sheet was associated with a reduction in the formation of regular, β-sheet-like hydrogen bonding patterns, a reduction in the end-to-end distance of the β-strands, and the intermittent exposure of carbonyl groups to solvent.

(2) Formation of pleated main chain geometries and loss of native β-sheet secondary structure was accompanied by a concomitant loss of native main chain hydrogen bonding patterns in the DAGH sheet (Figure 3.8). Within the DAGH sheet, the bond between residues 14 (donor) and 55 (acceptor) was the least stable (populated for 4% of the net simulation time) and the bond between residues 15 and 107 was the most stable (77%). In the DAGH sheet, the most persistent hydrogen bonds were those that were present in both β-sheet and α-sheet conformations. The CBEF sheet retained native main chain hydrogen bonding patterns to a greater extent and the residues with the highest retention of native-like hydrogen bonding patterns were least likely to form pleated main chain conformations or convert to α-sheet secondary structure (Figure A.3.7). Across the set of 21 TTR simulations, native main chain to main chain hydrogen bonds among residues within the CBEF sheet were sampled to a greater extent than among residues within the DAGH sheet (63 ± 8% vs. 34 ± 7%, $p << 0.001$). The difference in hydrogen bond stability was reduced when only the interior strands were considered (strands BE: 65 ± 12%, vs. strands AG: 53 ± 15%, $p = 0.007$). And the difference in hydrogen bond stability was magnified when only the edge strands were considered (strands CB, EF: 61 ± 8%, vs. strands DA, GH: 24 ± 13%, $p << 0.001$) (Figure A.3.8).

Within the CBEF sheet, the bond between residues 97 (donor) and 67 (acceptor) was the least stable (populated for 35% of the net simulation time) and the bond between residues 33 and 70 was the most stable (91%). Overall, main chain native hydrogen bonds in the CBEF sheet were almost twice as stable (present for 63% of the net simulation time) as those in the DAGH sheet (34%). In the CBEF sheet, the average hydrogen bond stability did not vary across the three strand interfaces BC: 61%, CE: 65%, EF: 62%. In the DAGH sheet, the average hydrogen bond stability

varied across the three strand interfaces DA: 25%, AG: 53%, GH: 24% (Figure 3.8). The most stable interface was between strands A and G, which have a parallel topology. The least stable interfaces involved edge strands (D and H) that run antiparallel to their partners and strands that dissociated from the protein to some degree over the course of the simulations.

(3) To determine whether side chain dynamics occurred prior to conversion to α-sheet, we calculated the fraction of the net simulation time that side chains spent in contact with one another and compared the results of our simulations to the minimized WT TTR structure. We found that structural changes along the main chain were accompanied by altered side chain dynamics. Specifically, there were changes in the side chain–side chain interaction network among solvent-exposed residues (M13, K15, L17, E54, H56, R104, T106, A108, L110, S112, S115, S117, T119, V121, and T123) within the DAGH sheet (Figure 3.9, Figure A.3.9). For example, in the X-ray conformation (i.e. the tetrameric state) residue L17 forms inter-chain contacts between residues T119 and V121 across the dimer-dimer interface and intra-chain contacts with K15 and L110. In the reference structure (a minimized WT monomer), L17 also formed contacts with K15 and L110. In MD under amyloidogenic conditions, L17 largely retained its native side chain contacts and formed contacts with K15 for 88% of the aggregate simulation time and contacts with L110 for 97% of the aggregate simulation time. However, the loss of interactions between L17 and T119/V121 that were present in the X-ray structure provided sufficient conformational flexibility to allow L17 to form new interactions with A108 for 89% of the aggregate simulation time (Figure 3.9). Some residues also lost contacts that were present in the X-ray and reference conformations. For example, in the reference conformation T106 formed contacts with M13, K15, R104, A108, T119, and V121. In MD under amyloidogenic conditions, these contacts were present for 99%, 57%, 60%, 17%, 3%, and 2% of the time, respectively (Figure 3.9). T106 also formed new contacts with E54 (3%) and T123 (38%). These heterogeneous interactions worsened the packing of side chains on the DAGH surface and intermittently exposed main chain peptide groups to solvent. Dynamic side chain-side chain interactions were also observed for some residues on the CBEF surface, but a subset retained reference-state-like interactions: H31, F33, W41, P43, K70, and E72 (Figure A.3.10). These residues maintained their native side chain-side chain interactions and were resistant to conversion to α-sheet secondary structure. F33 was the most stable residue and maintained contacts observed in the reference state with H31, W41, P43, K70, and E72 for 94%, 97%, 99%, and 99% of the aggregate simulation time (Figure A.3.10). Only one reference-state

interaction was infrequently observed: F33 interacted with K35 for 9% of the aggregate simulation time. Additionally, F33 formed few new contacts during MD and interacted with residues D39, S46, and E92 for 5%, 4%, and 11%, respectively, of the aggregate simulation time (Figure A.3.10).

Across the set of 21 TTR simulations, there were significant differences in the properties of the solvent exposed side chain interaction networks in the DAGH and CBEF sheets. On average, residues in the CBEF sheet had a higher edge per node ratio than residues in the DAGH sheet (5.1 ± 0.3 vs. 4.1 ± 0.5, $p << 0.001$), thus residues in the CBEF formed contacts with more neighbors than did residues in the DAGH sheet. The CBEF sheet also had higher degrees of network centralization and heterogeneity than the DAGH sheet (0.27 ± 0.05 vs. 0.21 ± 0.06, $p = 0.001$ and 0.40 ± 0.03 vs. 0.33 ± 0.05, $p << 0.001$, respectively). Finally, the clustering coefficient of the CBEF sheet interaction network was slightly greater than that of the DAGH sheet (0.65 ± 0.03 vs. 0.62 ± 0.03, $p = 0.04$) (Figure A.3.11). To measure the average 'strength' of the solvent exposed interaction networks, we calculated the average sum of all edge weights in the DAGH (purple) and CBEF (orange) sheets using data from the final 475ns of each (n = 21) simulation. The edge weights correspond to the fraction of simulation time that a given residue-residue contact was observed. We found that contacts formed among residues the CBEF sheet were stronger (i.e. there were more and/or more frequent contacts observed) than in the DAGH sheet (23.4 ± 1.4 vs. 18.0 ± 3.1, $p << 0.001$). The difference in the 'strength' of 'native' contacts was more pronounced (19.9 ± 1.5 vs. 14.0 ± 2.3, $p << 0.001$), but there was no significant difference in the 'strength' (3.5 ± 0.9 vs. 4.0 ± 1.2, $p = 0.13$) of 'non-native' contacts (Figure A.3.12).

The formation of pleated peptide bond geometries, loss of native main chain hydrogen bonding patterns, and dynamic rearrangements of side chain interaction networks observed in MD simulations preceded conversion from β-sheet to α-sheet secondary structure. These dynamic features were observed to a greater extent in the DAGH sheet (where the majority of α-sheet secondary structure was observed) than in the CBEF sheet; however, they were observed in the DAGH sheet for simulations that did and did not form α-sheet. Based on these observations, we propose that these dynamic fluctuations promote secondary structure conversion in TTR.

3.4.4 *Electrostatic interactions drive peptide plane flipping*

As described above, changes in both the secondary and tertiary structure occurred prior to the transition to α-sheet. To determine whether a consistent set of interatomic interactions participated in or promoted conversion, we tracked atomic contacts formed by the main chain peptide groups during peptide plane flips. Peptide plane flips were detected via analysis of the main chain ϕ and ψ dihedral angles. First, Ramachandran maps were constructed for individual residues in the DAGH and CBEF sheets of TTR. These maps were analyzed with the program *Minimum Energy Path Surface Analysis* (*MEPSA,* Marcos-Alcalde *et al.*, 2015) which enables analysis of 3D energy landscapes using transition state theory, to identify the locations (ϕ and ψ) of energetic barriers separating the β, α_R, and α_L regions of Ramachandran space. The precise barrier locations varied with amino acid identity and position. Based on the positions of the transition barriers identified by MEPSA, Ramachandran space was partitioned into 4 states: (β-sheet: β, right-handed α-helix: α_R, and left-handed α-helix α_L, and other: o). In Ramachandran space, there are two pathways for a residue to convert from the β region to the α_R region and two pathways for conversion from the β region to the α_L region. For conversion from β to α_R, the favorable pathway passes through ψ = 0° and the less favorable pathway through ψ = +/- 180°. For conversion from β to α_L, the favorable pathway passes through ϕ = 0° and the less favorable pathway through ϕ = +/- 180°. Using the MEPSA-identified transition barriers, we algorithmically identified β-sheet to α-sheet transitions passing through the more favorably pathways among residues in the DAGH sheet. Each single transition was analyzed on a per-residue, per-simulation basis and the timepoint assigned to each transition was the final timepoint before the residue crossed the MEPSA-identified barrier. In total, 116 transitions were included in the analysis.

We analyzed the atomic interactions formed by the main chain carbonyl and amide groups over a 2 ns window centered on the transition timepoints and found that β-sheet to α-sheet peptide plane flips were associated with the formation of hydrogen bonds with main chain carbonyl groups (Figure 3.10). Prior to the peptide plane flip, the probability of forming a hydrogen bond between a water molecule and the main chain carbonyl group increased steadily. And among all 116 transitions, 75% of the peptide plane flips occurred in the presence of a carbonyl-solvent hydrogen bond at the moment of the transition (Figure 3.10). After the transition, the probability of the formation of a carbonyl-solvent hydrogen bond remained stable at ~30%. These persistent post-

transition hydrogen bonds were primarily due to residues at the edges of the DAGH sheet with solvent-exposed peptide groups. 16% of peptide plane flips occurred in the presence of a carbonyl to side chain hydrogen bond at the moment of the transition (Figure 3.10). A side chain to main chain hydrogen bond participated in transitions for residues, such as L110, that are surrounded by polar side chains. After the transition, the probability of forming a side chain to carbonyl hydrogen bond increased to ~65%, which was primarily due to interactions between the carbonyl groups of residues in strand H, which are surrounded by several polar side chains. Overall, 91% of the peptide plane flips occurred in the presence of a hydrogen bond with the carbonyl group of the residue that converted into the α_R region. The remaining 9% of transitions occurred in the absence of a hydrogen bond; however, transitions occurring in the absence of a hydrogen bond were more likely to occur if residues in neighboring strands had previously converted to α-sheet. This suggests that later transitions do not require water or side chain interactions to catalyze the transition and instead they are forced to convert because of the buildup of a dipole across the main chain of a neighboring strand. That is, when several carbonyl groups align in one strand, it is unfavorable to have a carbonyl in a neighboring strand directed at them.

Analysis of a single transition (for residues T106 and I107 of run 3 of the WT simulation) illustrates the observed electrostatic-mediated transition mechanism in detail (Figure 3.11). Initially, T106 and I107 sampled main chain dihedral angle conformations near their initial β-sheet positions (0 - ~10 ns, Figure 3.11). During this time, the carbonyl group of T106 primarily formed native main chain to main chain hydrogen bonds. Next, T106 and I107 sampled a broader range of pleated main chain conformations within the β-region, during which time the carbonyl group of T106 formed hydrogen bonds with solvent molecules (~10 - ~70 ns, Figure 3.11). During the peptide plane flip (at 71.075 ns), a solvent molecule formed a hydrogen bond with the carbonyl group of T106 (Figure 3.11). After the transition, the T106 and I107 remained in the α-sheet conformation for the duration of the simulation, and only main chain to main chain hydrogen bonds were formed with the carbonyl group of T106 (Figure 3.11)

3.5 Discussion

3.5.1 *Aggregation safeguards are subverted in the monomeric state of TTR*

TTR is a highly amyloidogenic protein and formation of amyloid precedes from a native-like monomer. This high degree of amyloidogenicity likely exerts evolutionary pressure on the transthyretin-like protein (TLP) fold, which has been observed in bacteria, plants, and animals (Hennebry, 2009). That is, Nature has 'designed' mechanisms to evade amyloid formation by members of the TLP fold. First, the tetrameric structure of the TLP is highly favored and stable under physiological conditions. For TTR from *H. sapiens*, WT homotetramers exchange slowly under physiological conditions (Schneider *et al.*, 2001), with an extrapolated *in vitro* half-life of 293 years (Lai *et al.*, 1997). Furthermore, mutations that kinetically stabilize TTR (such as T119M, Harrison *et al.*, 1991) prolong human lifespans by 5-10 years (Hornstrup *et al.*, 2013). Second, the native tetrameric structure of TTR 'protects' the least stable regions of the structure. For example, it is known that parallel and mixed β-sheets are typically less stable than anti-parallel β-sheets (Richardson, 1977; Sheridan *et al.*, 1979). The two parallel strands in TTR (strands A and G) are buried in the dimer:dimer interface, but become solvent-exposed upon tetramer dissociation. Third, TTR monomers incorporate several structural features to avoid edge-strand to edge-strand aggregation (Richardson and Richardson, 2002), such as the presence of irregular dihedral angles that distort the geometry of edge strands (e.g. P43 distorts strand C geometry and H56 forms a β-bulge that disrupts strand D geometry). Finally, networks of chaperone proteins, such as HSP90 (Oroz *et al.*, 2017) and BiP (Susuki *et al.*, 2009) in *H. sapiens*, can recognize and sequester aggregation-competent TTR conformers. Many of these protective mechanisms are undermined during aggregation of TTR. For example, protonation of ionizable residues at low pH promotes tetramer dissociation and initiates aggregation. Tetramer dissociation also results in the exposure of strand H to solvent, which is significant as this strand does not contain any known structural features that protect against aggregation (Richardson and Richardson, 2002). Furthermore, strand H dissociated in multiple TTR simulations, which also resulted in the exposure of strand G to solvent, which similarly lacks anti-aggregation safeguards.

3.5.2 *The DAGH sheet is susceptible to destabilization*

The transient and conformationally heterogeneous nature of the aggregation-competent monomers has complicated their structural characterization at the atomic level; however, several groups have used a diverse array of techniques to probe their structure. Early efforts analyzed hydrogen-deuterium exchange (HDX) rates of main chain amides at different pHs. Liu *et al.* found that residues within the CBEF sheet are less protected than those in the DAGH sheet and that the interior strands (B,E,A, and G) are more protected than the edge strands (Liu *et al.*, 2000b). This information coupled with prior analysis of Trp fluorescence changes in TTR (Lai *et al.*, 1996) and statistics demonstrating that a large number of pathogenic mutations occurred in the CBEF sheet (João and Saraiva, 1995) led to the hypothesis that the aggregation-competent monomer is defined by conformational changes partially in the CBEF sheet with enhanced lability in the C and D strands and that the core aggregation unit was comprised of strands BEF and AGH. Later, interpretation of hydrogen-deuterium exchange data of the aggregation-competent monomers in context with native-state data revealed that the observed lability in strands C and D is also observed under native conditions and that the core components of the aggregation-competent monomer includes strands A, B, E, and G (Liu *et al.*, 2000a). Analysis of pathogenic and amyloid-suppressing variants of TTR using hydrogen-deuterium exchange demonstrate that while sequence variants altered the magnitude of the protection factors, strands A, B, E, and G constitute the core component of aggregation-competent monomers.

In contrast to results obtained from HDX, relaxation dispersion NMR experiments performed on M-TTR show that most conformational exchange occurs in the DAGH sheet: strand H and several loops (the AB, GH, DE, and EF; the loops are named according to the β-strands that they connect) have the greatest extent of exchange. These relaxation dispersion NMR experiments indicate minimal conformational exchange in the CBEF sheet and conclude that destabilization of the DAGH sheet defines the aggregation-competent state (Lim *et al.*, 2013). Subsequent HSQC NMR results support these conclusions and magic-angle-spinning solid state NMR experiments using selective labeling schemes show that the AB loop becomes disordered in the aggregation-competent state (Lim *et al.*, 2016a-b). The discrepancy between insights from the earlier hydrogen-deuterium exchange experiments and more recent NMR studies may be partially explained by the

different timescales monitored by the experiments and the potential for a variety of conformational states to contribute to the HDX results.

Our MD-derived results mirrored the more recent NMR-derived observations. As was observed in solution NMR experiments, we found that conformational changes in the DAGH sheet occurred on faster timescales than in the CBEF sheet. This highlights a fundamental difference in the properties of the CBEF and DAGH sheets. The CBEF sheet appears 'designed' to be solvent-exposed whereas the DAGH sheet appears 'designed' to be buried in the tetramer. Similarly, the residues in the DAGH sheet form one of the functional sites of TTR (the T4 binding site) and are thus subject to evolutionary pressure to optimize function, possibly at the cost of structural stability. In contrast, the more stable main chain hydrogen bonding patterns and side chain interaction networks in the CBEF sheet suggest that these residues have been optimized for structural stability. Additionally, tetramer dissociation results in the loss of numerous interactions made by residues in the DAGH sheet that form contacts with other monomers. For example, residues Y116, T118, A120, and V122 (all located in strand H) form main chain to main chain hydrogen bonds with strand H' at the monomer:monomer interface, but all 8 of these hydrogen bonds are lost in the monomeric species. We observed that strand H was the most susceptible to conformational exchange, and we observed dissociation of strand H in multiple simulations. We also observed increased dynamics in several loops, such as the AB loop, which is structured under physiological conditions but disordered under amyloidogenic conditions (Das *et al.*, 2014). Based on our results, the loss of anti-aggregation safeguards facilitates conformational changes and allows solvent molecules to participate in and drive secondary structure changes in TTR. This suggests that water, which is already known to aid in protein folding (Collet, 2011) and unfolding (Bennion and Daggett, 2003) can also actively contribute to protein mis-folding.

3.5.3 *Insights into factors that drive α-sheet conversion*

We have proposed that the formation of α-sheet secondary structure is the underlying mechanism of aggregation common to amyloid proteins and that the presence of α-sheet defines the soluble oligomeric intermediates that are responsible for amyloid-associated cytotoxicity (Daggett, 2006). The simulations discussed here refine this hypothesis and indicate that an electrostatic-mediated mechanism drives the formation of α-sheet secondary structure in TTR. The conversion begins as residues adopt pleated main chain conformations, resulting in the loss of

native inter-strand hydrogen bonds, the formation of non-native hydrogen bonds between main chain carbonyl and solvent or polar side chain atoms, and a reorganization of solvent-exposed side chain interaction networks. When the DAGH sheet was sufficiently destabilized, electrostatic interactions formed with carbonyl groups mediated peptide plane flips from β-sheet to α-sheet secondary structure. These electrostatic interactions were observed in the form of interactions with solvent molecules (75% of transitions), side chain polar atoms (16% of transitions), and main chain carbonyl groups from neighboring strands (9% of transitions). We propose that hydrogen bonds facilitated transitions by stabilizing unfavorable pleated main chain geometries, by stabilizing the transition state conformation of the peptide plane flip, and by 'guiding' residues across the free energy barriers separating the β and $\alpha_{R/L}$ regions of Ramachandran space. The transitions not mediated by a hydrogen bond were instead mediated by electrostatic repulsion among the carbonyl groups in neighboring strands.

Taken in context with experimental observations, our results provide further evidence that during TTR amyloidogenesis, conformational changes within the DAGH sheet follow tetramer dissociation. Our simulations, which provide full atomic resolution, indicate that the initial conformational changes to the DAGH sheet on the sub-μs timescale results in the conversion to α-sheet secondary structure. By systematically evaluating molecular interactions at a fine level of detail, only accessible through MD, we have identified a mechanism behind the formation of a structure associated with protein aggregation and toxicity. In these simulations, once α-sheet formed, it remained stable and we predict that it will remain present over longer timescales, particularly during the formation of toxic, soluble oligomers, as has been experimentally observed directly for Aβ (Shea *et al.*, 2019) and indirectly for TTR (Hopping *et al.*, 2014; Kellock *et al.*, 2016).

It is known that the structures formed by β-sheets are the results of competition between optimal main chain hydrogen bonding patterns and side chain packing (Richardson, 1981). Here, we have observed that changes to native β-sheet structures can tip this balance. In the case of TTR, an acidic environment alters the distribution of charges in TTR (via modification of Asp, Glu, and His residues), disrupts quaternary interactions, and causes a reorganization of side chain packing and other tertiary interactions. Changes in the main chain structure and secondary structure elements are then a response to disruptions in the tertiary structure, providing another example of how tertiary level interactions can affect secondary structure as has been shown previously to occur

in protein folding, unfolding, and misfolding (Bond *et al.*, 1997; Wong *et al.*, 2000; Alonso and Daggett, 2000; Scott *et al.*, 2006). In future studies, incorporation of these findings along with a dissection of other factors that modulate β-sheet structure should lead to new 'rules' for the design of stable β-sheets, α-sheets, or of molecules created to interact with sheet-like main chain structures.

3.6 Conclusions

The simulations described here reveal a set of conformational changes that define the formation of aggregation-competent, monomeric TTR species shared by multiple pathological variants, leading to three broader implications. First, as we have observed α-sheet formation in MD simulations of multiple amyloid systems, the electrostatic-mediated conversion described here may be a mechanism common to other amyloid proteins. If this is the case, synergistic computational and experimental studies may be used to refine models of the early stage of amyloidogenesis and open up avenues for 'one-size-fits-all' therapeutic and diagnostic strategies to identify and treat amyloid disease. Second, that the mechanism was largely independent of amino acid sequence is significant and suggests that a single strategy may be employed to address the diverse pathological outcomes associated with TTR amyloid. Finally, if our computational predictions are borne out in experimental studies, then iterative computational and experimental results can be used to better map – and target – pathogenic conformers associated with amyloid pathology.

3.7 Tables

Table 3.1. Summary of simulated transthyretin variants

Transthyretin Variant	Number of Replicates	Simulation Length (ns)
Wild Type	3	500
D18G	3	500
A36P	3	500
L58H	3	500
Y69H	3	500
L111M	3	500
V122I	3	500

Table 3.2. Time averaged C_α RMSD of transthyretin variants

Transthyretin Variant	Run Number	C_α RMSD (Å) Res. 11-123	C_α RMSD (Å) Res. 11-114
WT	**1**	4.27	3.70
	2	3.41	3.38
	3	2.64	2.64
D18G	**1**	3.74	3.62
	2	3.49	3.13
	3	4.25	4.34
A36P	**1**	2.76	2.69
	2	3.57	3.52
	3	3.72	3.71
L58H	**1**	2.52	2.56
	2	7.24	7.52
	3	3.01	2.97
Y69H	**1**	4.00	4.02
	2	4.10	4.07
	3	4.58	4.52
L111M	**1**	3.72	3.29
	2	2.91	2.81
	3	3.41	3.38
V122I	**1**	3.03	2.86
	2	2.99	2.89
	3	3.67	3.68

3.8 Figures

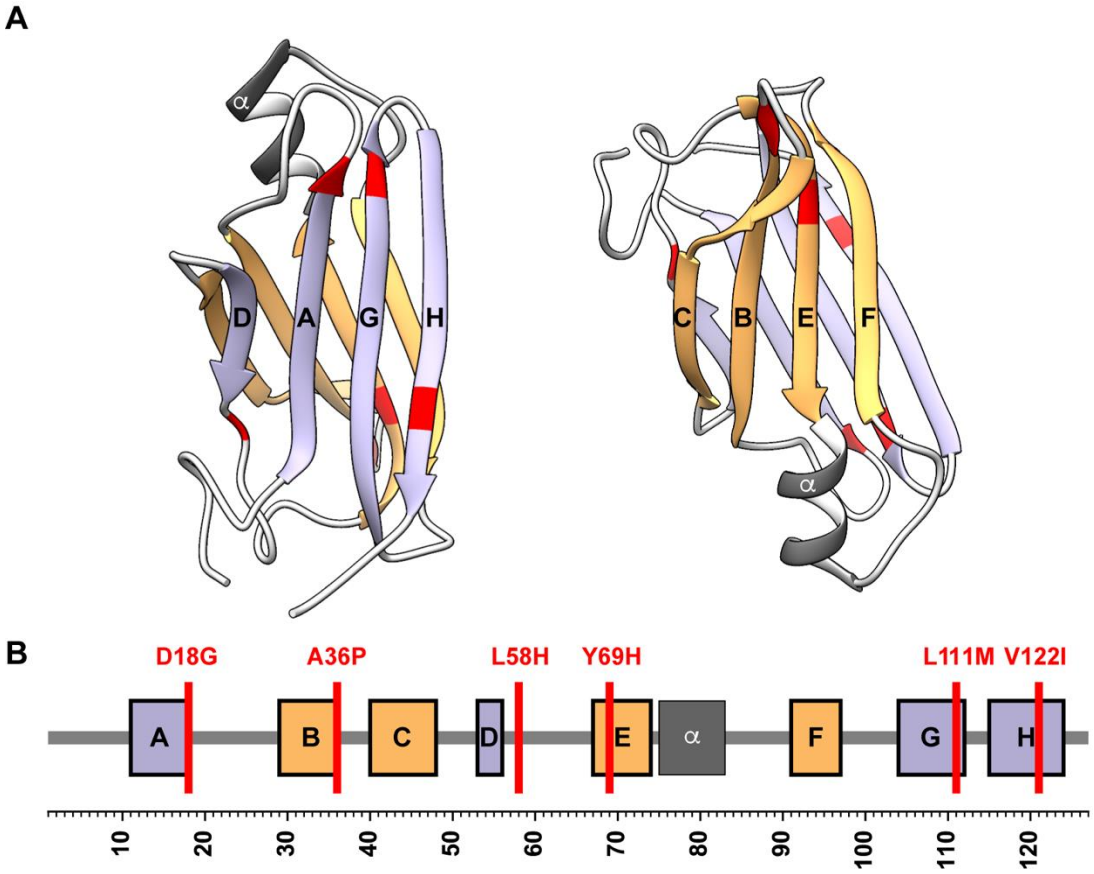


Figure 3.1. The X-ray structure of a transthyretin monomer

(A) Monomeric transthyretin has a β-sandwich topology and is composed of two four-stranded β-sheets: the DAGH sheet (purple) and the CBEF sheet (orange) and a single α-helix (α). Red ribbons denote the position of pathogenic TTR mutations simulated in this study. (B) The primary structure of human transthyretin. The positions of mutant variants that have been simulated in this study (D18G, A36P, L58H, Y69H, L111M, and V122I) are colored red.

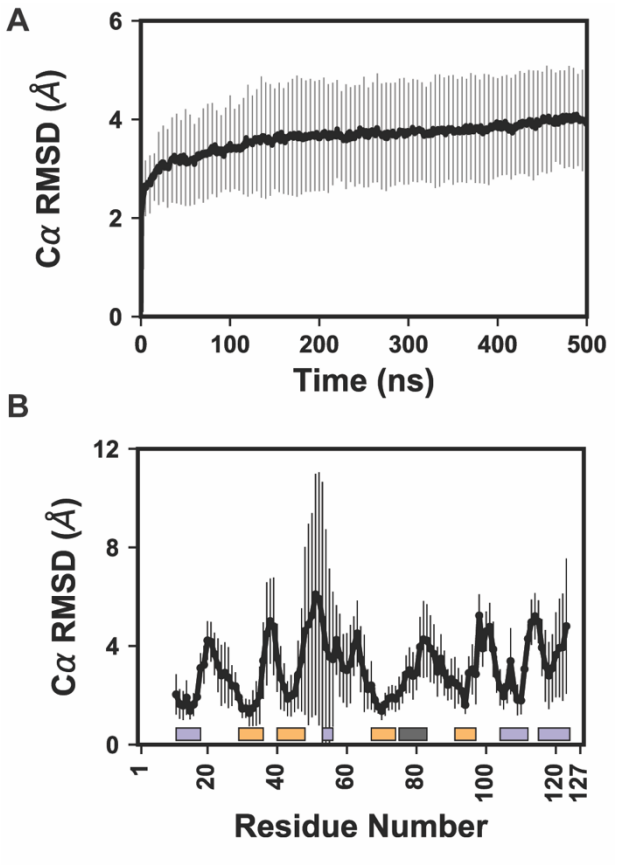


Figure 3.2. Cα RMSD as a function of time and residue number.

(A) The average Cα RMSD of all low pH TTR simulations (21 total simulations) as a function of time shows that on average TTR monomers sampled conformations approximately 2-4 Å from the minimized crystal structure (error bars denote standard deviation) and that the average Cα RMSD increased continuously over the 500 ns time scale. (B) The average Cα RMSD of all low pH TTR simulations (21 total simulations) as a function of residue number shows that the most flexible residues were the AB loop, BC hairpin turn, CD loop, strand D, DE loop, α-helix-EF loop, FG loop, GH loop, and strand H. (Figure 2B). The colored bars denote the positions of the strands that compose the DAGH (purple) and CBEF (orange) sheets. Data are shown for the 'core' TTR residues and include all timepoints from all simulations

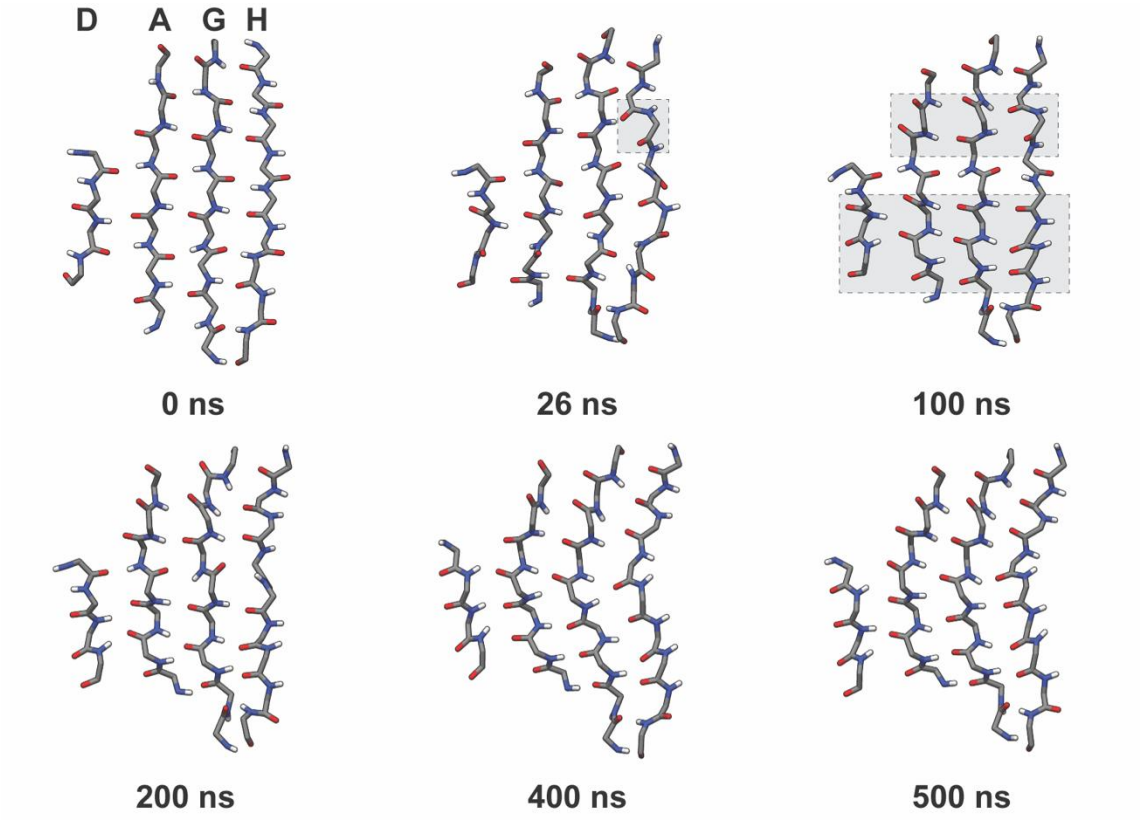


Figure 3.3. Conversion from β-sheet to α-sheet secondary structure in the DAGH sheet. Formation of α-sheet structure in the DAGH sheet is shown for a single simulation (WT TTR, run 3). At 0 ns, native β-sheet secondary structure was present in the DAGH sheet. At 26 ns, the native β-sheet secondary structure was disrupted for several residues (highlighted in the grey box). At 100 ns, α-sheet secondary structure formed for several residues in two regions of the sheet (highlighted in the grey boxes). This disruption in the β-sheet remained until the final residues transitioned into β-sheet between 200 and 400 ns. The resulting α-sheet secondary structure remained for the duration of the simulation and strands D and H remained attached to the sheet.

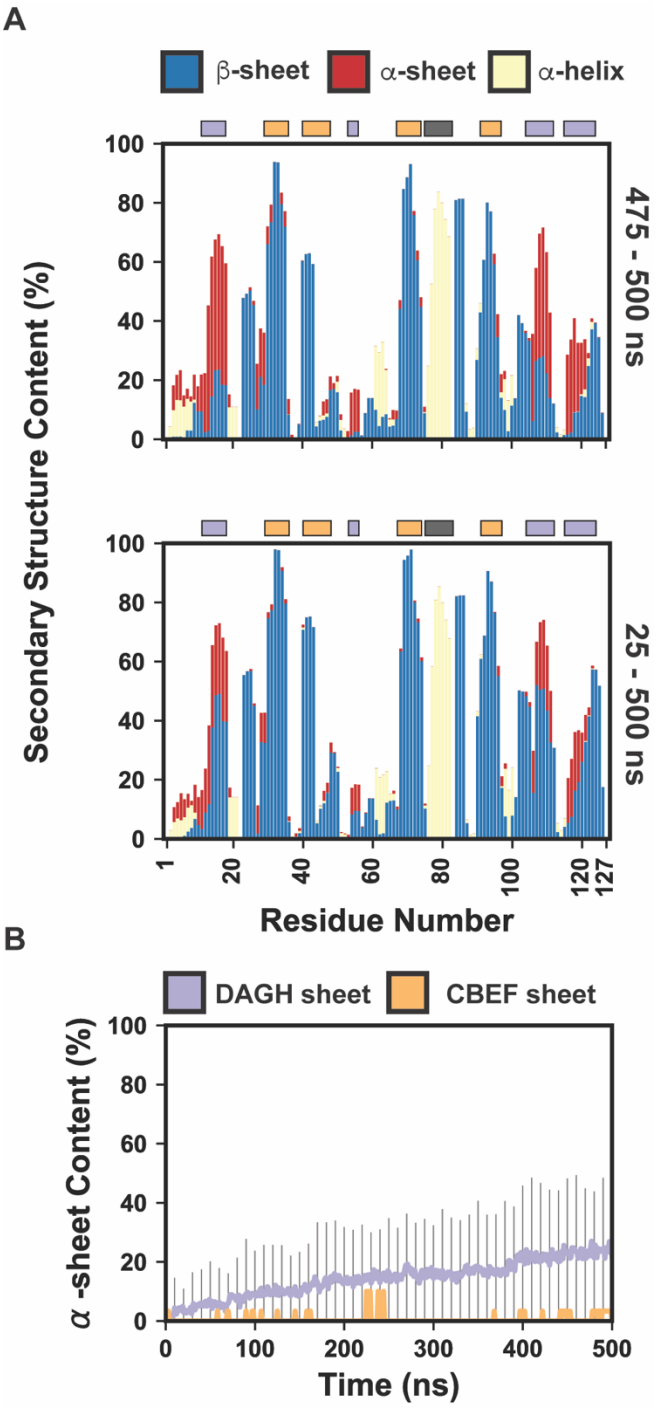


Figure 3.4. Average secondary structure content as a function of residue number and time. (A) The average β-sheet (blue), α-sheet (red), and α-helix (yellow) secondary structure content averaged overall all simulations plotted as a function of residue number over two timescales: 475 – 500 ns and 25 – 500 ns. (B) The average α-sheet secondary structure content of the DAGH (purple) and CBEF (orange) sheets plotted as a function of time suggests that complete conversion to α-sheet secondary structure in the DAGH for all simulations would occur on the 1 – 5 μs

timescale. α-sheet secondary structure rarely formed in the CBEF sheet. When it did form, the residues closest to the α-helix (the least structured region of the sheet) were most likely to convert.

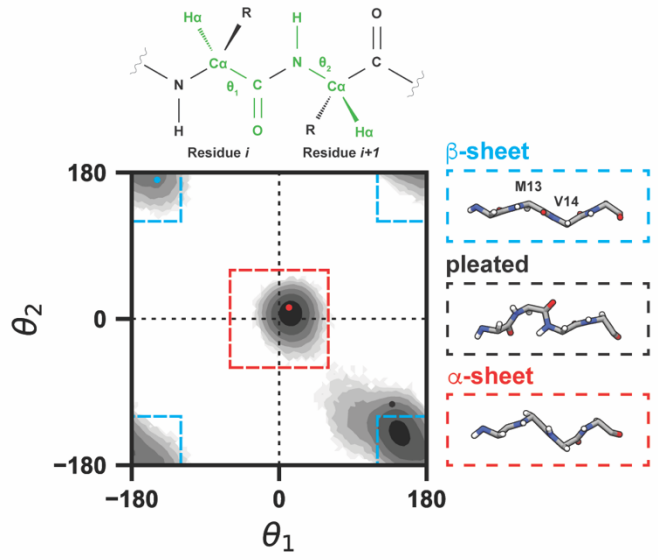


Figure 3.5. Defining pleated peptide plane geometries in the DAGH and CBEF sheets.

Two dihedral angles were defined to monitor the extent of peptide plane pleating: θ_1 (defined by atoms H_α, C_α, C, and O) and θ_2 (defined by atoms H, N, C_α, and H_α). The map shown here plots θ_1 and θ_2 for residues M13 and V14 from run 2 of the WT TTR simulations and readily identifies pleated conformations. In the map, darker colors indicate greater sampling. In native β-sheet like conformations, θ_1^i and θ_2^{i+1} have values near (180°, 180°); in α-helical conformations, θ_1^i and θ_2^{i+1} have values near (15°, -120°); and in α-sheet like conformations, θ_1^i and θ_2^{i+1} have values near (0°, 0°). The θ_1^i vs θ_2^{i+1} maps were divided into three continuous regions to classify peptide plane geometry. The first region, outlined in blue, corresponds to β-sheet like secondary structure and is defined by: ($150° < \theta_1^i < 180°$, $150° < \theta_2^{i+1} < 180°$); ($-180° < \theta_1^i < -150°$, $150° < \theta_2^{i+1} < 180°$); ($-180° < \theta_1^i < -150°$, $-180° < \theta_2^{i+1} < -150°$); and ($150° < \theta_1^i < 180°$, $-180° < \theta_2^{i+1} < -150°$) (outlined in blue). The second region, outlined in red, corresponds to α-sheet like secondary structure and is defined by: ($-30° < \theta_1^i < 30°$, $-30° < \theta_2^{i+1} < 30°$). The final region includes all remaining areas of the θ_1^i vs θ_2^{i+1} map and is indicative of pleated main chain geometry. The colored points on the map correspond to structures extracted from the simulation that are shown to the right.

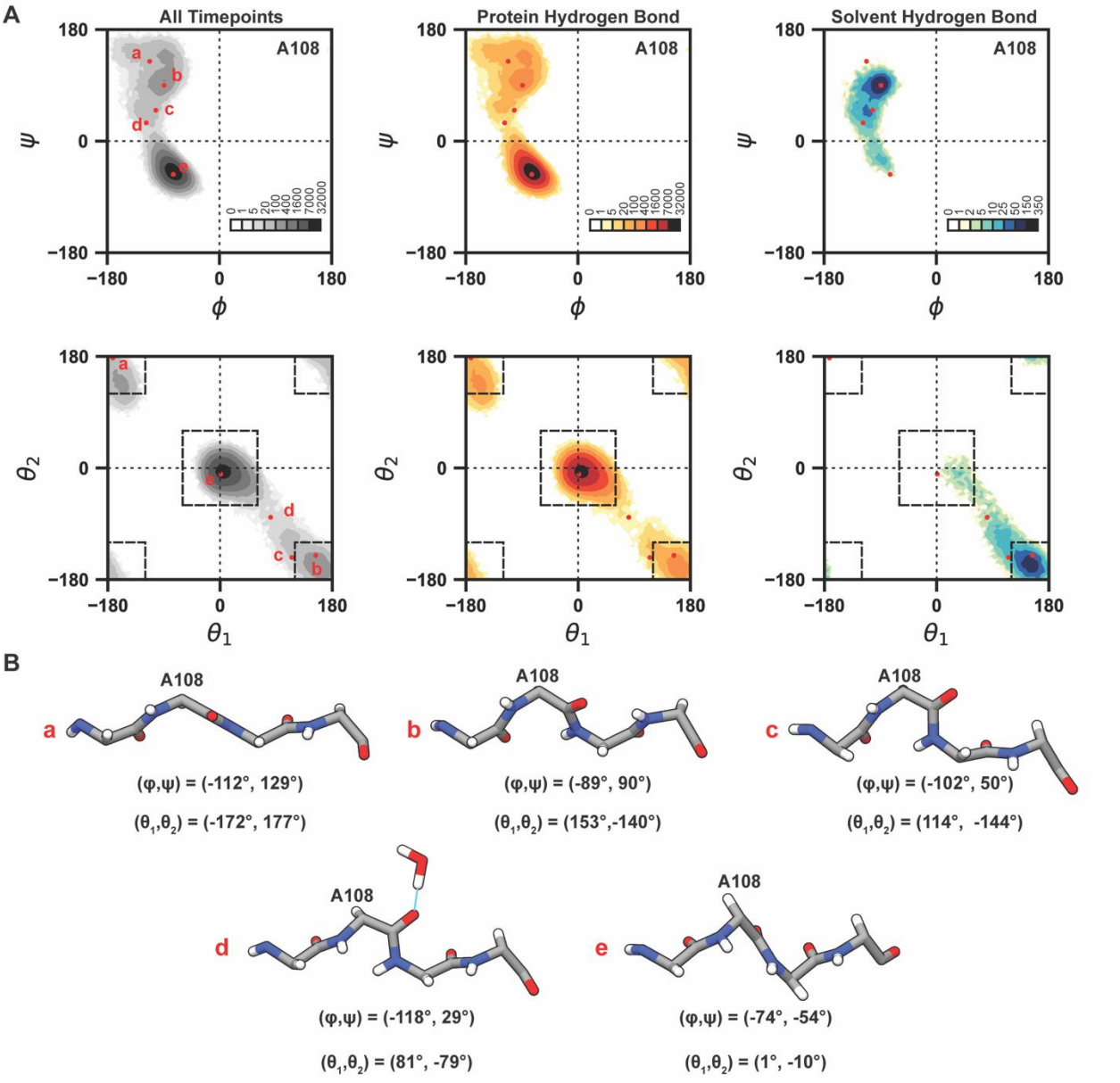


Figure 3.6. Visualizing peptide plane pleating during secondary structure conversion.

(A) ϕ vs ψ *(top row)* and θ_1^i vs θ_2^{i+1} *(bottom row)* maps taken from A108 and A109 from run 2 of the WT TTR simulations show conformations sampled during secondary structure conversion. For each set of dihedral angles, three maps are presented: one contains all time points *(left)*, one contains only those timepoints for which the carbonyl group of A108 formed a hydrogen bond with another protein atom *(center)*, and one contains only those timepoints for which the carbonyl group of A108 formed a hydrogen bond with a solvent molecule *(right)*. The annotated points on

the maps correspond to structures extracted from the simulation that are shown in panel B. (B) Before the transition to α-sheet secondary structure, the main chain atoms adopt pleated conformations that allow solvent molecules to form hydrogen bonds with carbonyl groups. In the native conformation (structure a), the main chain adopts β-sheet structure. As the ϕ vs ψ move along the diagonal $\psi = -\phi$ towards the origin (structure b), the main chain pleats: it shortens in length and 'folds' like an accordion. In pleated conformations for which either ϕ or ψ deviates from the diagonal, one of the main chain peptide groups shifts out of the β-sheet hydrogen bonding plane towards either the hydrophobic core of the protein or solvent (structure c). Solvent molecules were then able to form hydrogen bonds with the backbone in 'pleated' conformations (structure d), facilitating the conversion to α-sheet secondary structure (structure e).

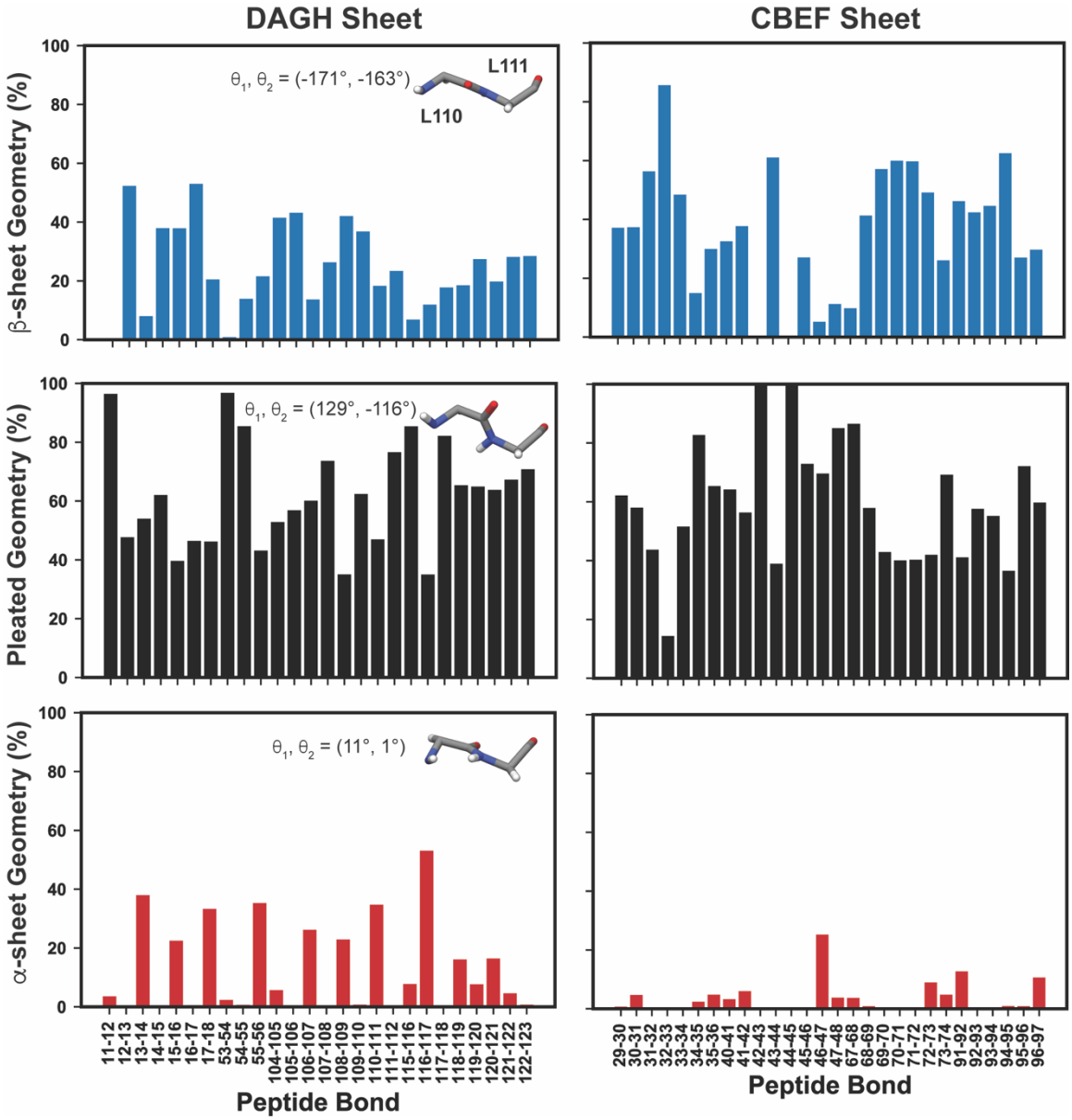


Figure 3.7. Sampling of pleated peptide plane geometries in the DAGH and CBEF sheets. (A) The fraction of the simulation time during which β-sheet-like (top, blue) pleated (middle, black), and α-sheet-like (bottom, red) peptide plane geometries were populated during the simulations in the DAGH (top) and CBEF (bottom) sheets. Peptide bonds involving Pro or Gly residues had high populations of pleated conformations due to altered steric constraints for these residues. Example inset conformations of the peptide main chain for residues L110 and L111 obtained from run 3 of the WT TTR simulation show representative structure of β-sheet-like, pleated, and α-sheet-like main chain geometry.

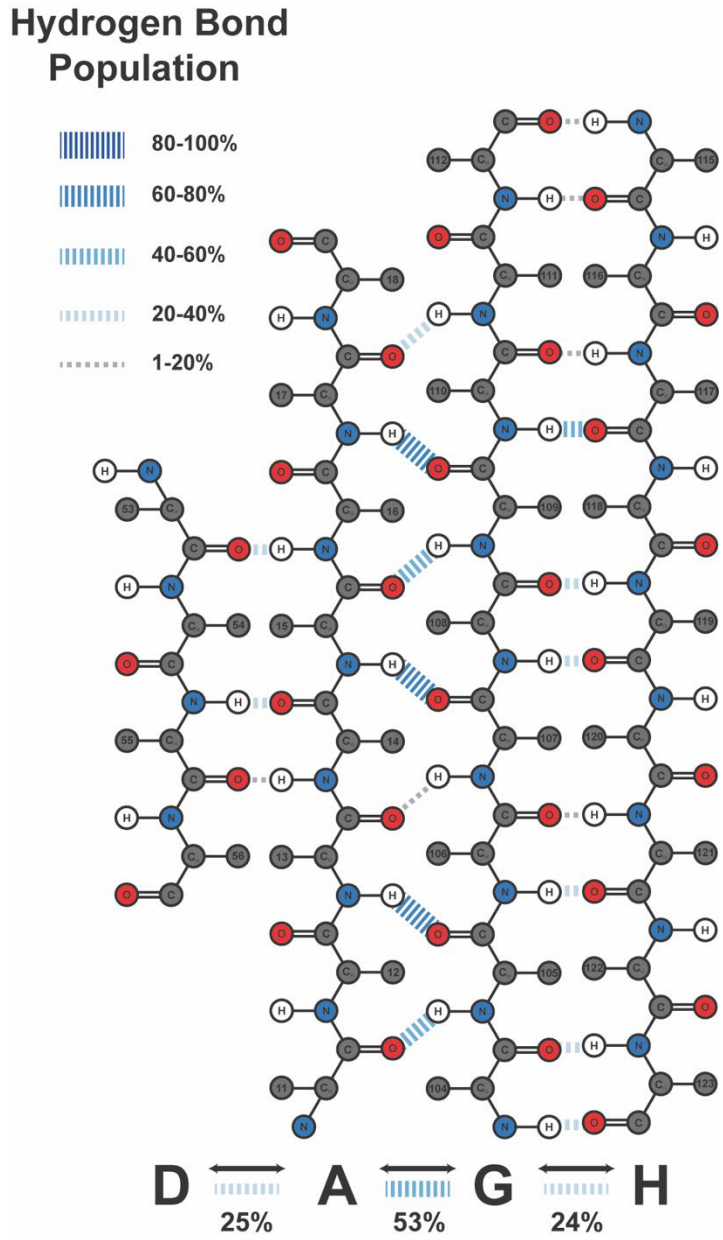


Figure 3.8. Native hydrogen bonding patterns are lost in the DAGH sheet.

The main chain atoms of the peptide main chain for residues in the DAGH sheet are represented as circles in the conformation present in the X-ray structure. The hydrogen bonds are represented as dashed blue lines and are only shown for bonds present after hydrogen atoms were modeled onto the X-ray structure (i.e. reference state hydrogen bonds). The color and width of the hydrogen bonds are proportional to the average percentage of time during which these hydrogen bonds were present. The bonds connecting the strand names at the bottom of the schematic correspond to an average over all reference-state hydrogen bonds between the respective strands.

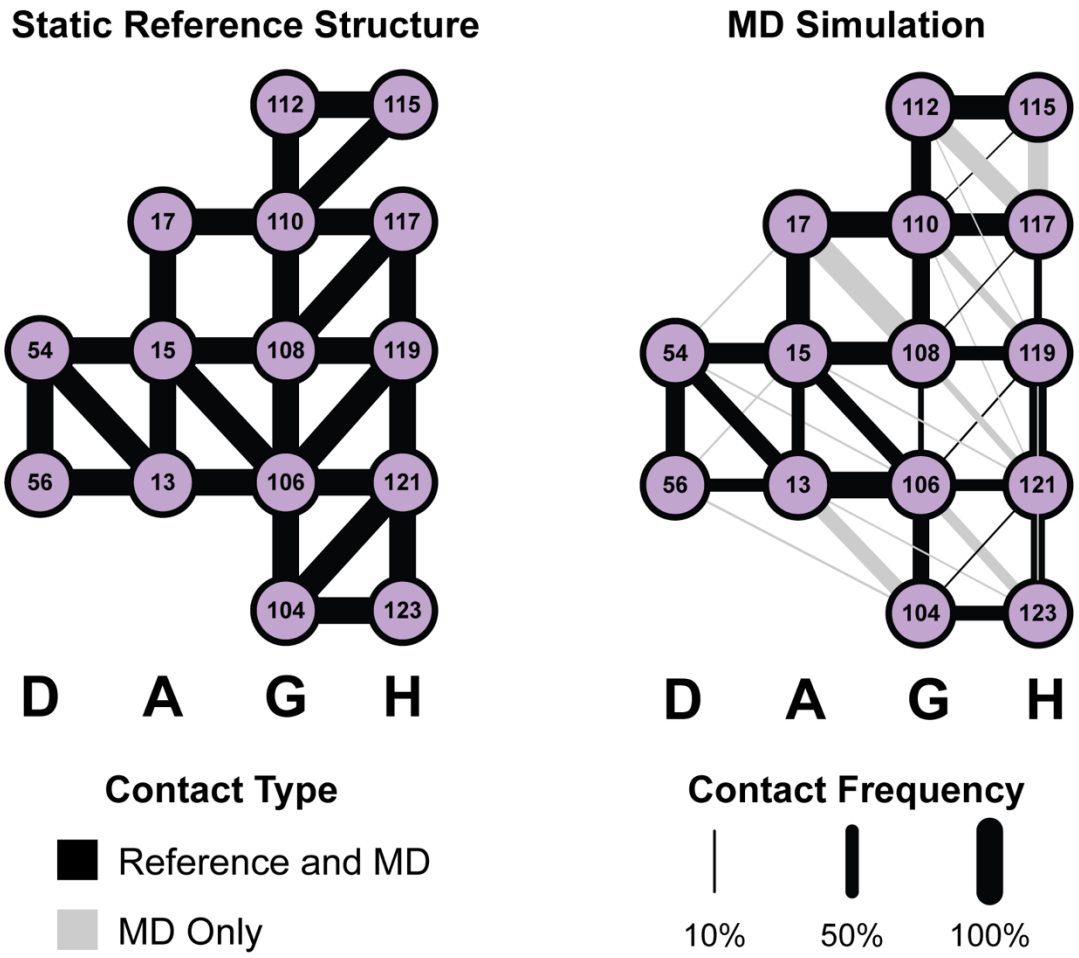


Figure 3.9. Dynamic reorganization of the solvent-exposed side chain interaction network in the DAGH sheet.

The solvent-exposed side chain interaction network composed of residues in the DAGH sheet is represented as a graph. Each residue is a node and edges indicate that the side chains of the connected residues were in contact during the simulation. The width of the edges is proportional to the percentage of simulation time that the contact was present for. Reference state contacts (i.e. those present in the minimized starting structure) are colored black edges and contacts only observed during MD are colored grey. The network on the left-hand side shows the interaction network present in the reference structure (since there is only one conformation for this model, all interactions have been set to their max value – 100%). The network on the right-hand side shows the same interaction network using data averaged from all MD simulations.

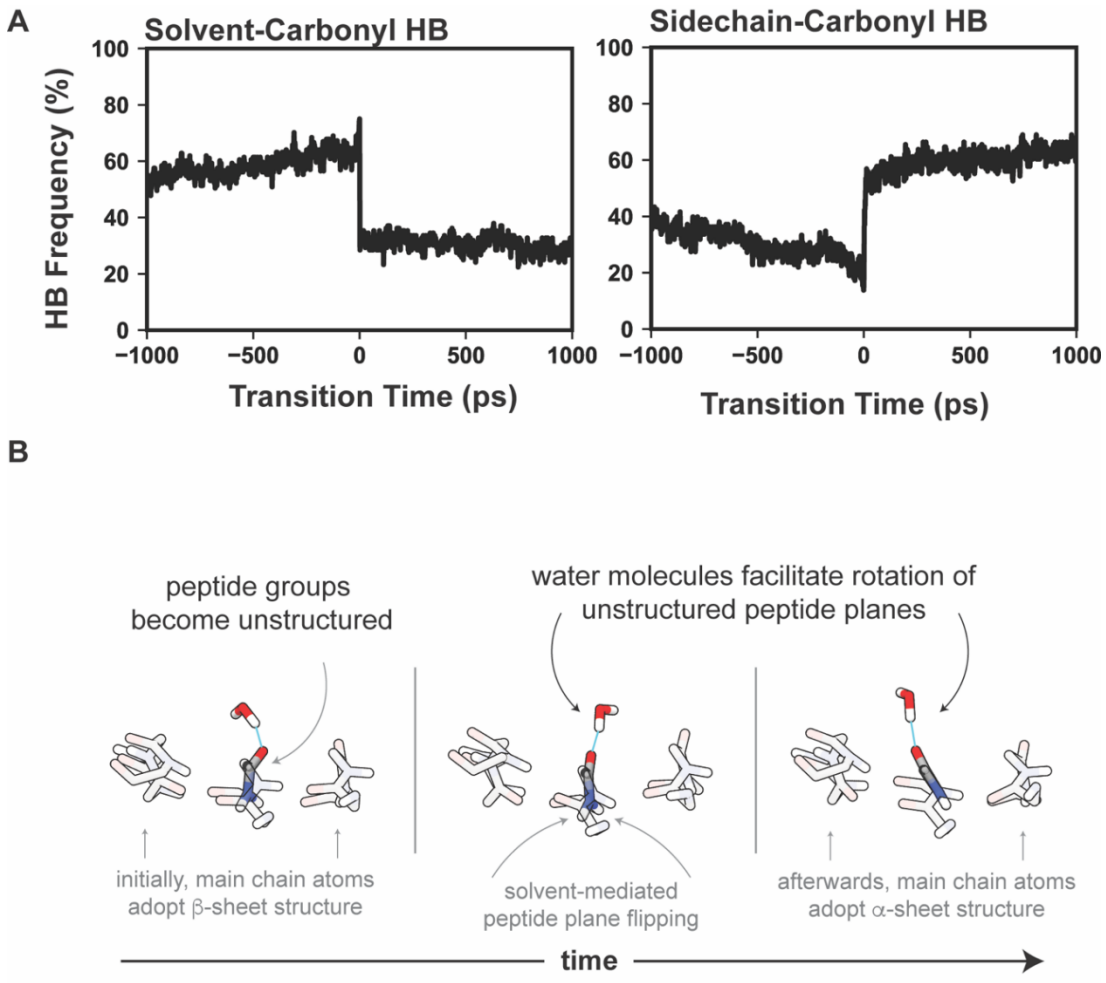


Figure 3.10. Carbonyl to solvent hydrogen bond frequency during transitions to α-sheet secondary structure.

(A) The transition timepoints identified from our MEPSA analysis were used to construct a dataset of 116 stable transitions to α-sheet secondary structure. For each transition, we measured whether a main chain carbonyl to solvent hydrogen bond was present for 1ns before and after the transition. Next, we aligned the transitions and calculated the probability of the formation of a carbonyl to solvent hydrogen bond prior to and after the transition. *Solvent to carbonyl hydrogen bonds (left)* In the nanosecond leading up to the transition, the hydrogen bond probability increases up to 75%, then rapidly drops immediately after the transition and remains stable for a further nanosecond. *Side chain to carbonyl hydrogen bonds (right)* In the nanosecond leading up to the transition, the hydrogen bond probability decreases to 16%, then increases after the transition. Panel (B) illustrates the participation of a water molecule during a peptide plane flip.

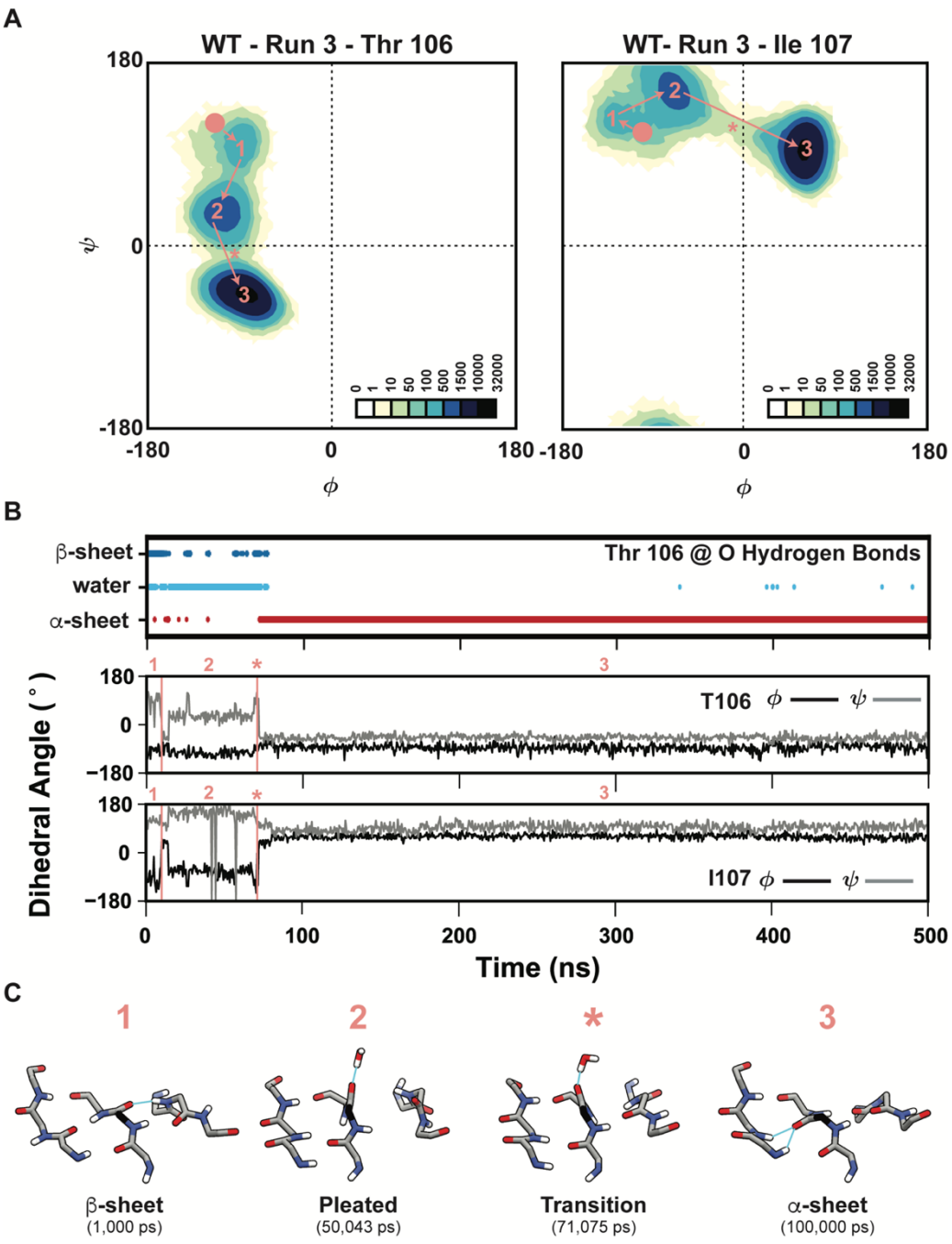


Figure 3.11. Water-mediated transition from β-sheet to α-sheet secondary structure.

The dynamics of T106 and I107 in run 3 of the WT simulation illustrate the dynamics of the β-sheet to α-sheet transition. (A) Ramachandran plots of T106 and I107 show the distribution of conformations for these residues in dihedral angle space darker colors (e.g. black and blue) indicate higher sampling. The pink points indicate the initial ϕ/ψ values for T106 and I107, numerals indicate the major conformational states occupied by the residues, and the asterisks indicate the

ϕ/ψ values sampled during the β- to α-sheet transition. T106 and I107 initially move from their starting conformation to pleated conformations, then approach the barriers between the β/α_R and β/α_L regions before transitioning to the α_R and α_L regions. (B) The upper plot shows hydrogen bonds formed by the carbonyl oxygen atom of Thr 106 with V121 (β-sheet conformation, blue), solvent (pleated and transition conformations, cyan) or M13/V14 (α-sheet conformation, red). The lower plots show the main chain dihedral angles ϕ and ψ as a function of time for T106 and I107. (C) Snapshots of the β-sheet to α-sheet transition. The Cα atom of T106 is colored black. At 1.000 ns, the carbonyl of T106 formed a hydrogen bond with V121. By 50.043 ns, T106 and I107 had adopted pleated structures that were stabilized by intermittent hydrogen bonds with solvent molecules. At 71.075 ns, the transition timepoint, a solvent molecule engaged the peptide main chain and facilitated the transition to an α-sheet conformation. After the transition, the carbonyl group of T106 formed a stable, bifurcated hydrogen bonds with M13 and V14.

Chapter 4. MODELLING AGGREGATION-COMPETENT CONFORMATIONS OF TRANSTHYRETIN MONOMERS

4.1 Summary

During pathological amyloid formation (amyloidogenesis) proteins undergo conformational changes that allow them to self-aggregate and assemble into insoluble, fibrillar structures. Intermediate structures, known as soluble oligomers, that form during this process typically contain 2-24 monomeric subunits and are cytotoxic. Prior to the formation of these soluble oligomers, monomeric species must first adopt aggregation-competent conformations. Knowledge of the structures of these intermediate states is invaluable to the development of molecular strategies to arrest pathological amyloid aggregation. However, the highly dynamic and interconverting nature of amyloid species limits biophysical insight into structures formed during amyloidogenesis. Here, we use molecular dynamics simulations to probe conformational changes to one amyloid protein, transthyretin, and place our findings in context with experimental results. Our results refine the current model of transthyretin amyloidogenesis by showing that one region of the protein (the DAGH sheet) is susceptible to conformational changes in the monomeric state in an acidic environment. Furthermore, we show that changes in the tertiary structure of transthyretin can be associated with disruptions to the secondary structure. Finally, we leverage our computational results to produce experimentally testable hypotheses that can drive forward the current state-of-the-art modelling of the transthyretin amyloidogenesis.

4.2 Introduction

During amyloidogenesis proteins (both folded and natively disordered) self-aggregate into insoluble fibrils (Knowles *et al.*, 2014; Bleem and Daggett, 2017). During this 'reaction', aggregating proteins sample a multitude of conformational states including native-like monomers, aggregation-competent monomers, soluble oligomers, protofibrils, and mature fibrils. Each of these states harbors some degree of conformational heterogeneity. Thirty-six proteins from *Homo sapiens* and ten from other animals have been associated with the formation of pathological amyloid fibrils (Benson *et al.*, 2018). Functional, non-pathological amyloids have also been

identified in humans, fungi, and bacteria (Benson *et al.*, 2018). Despite the ubiquity of amyloid, a clear understanding of the structural 'landmarks' that are sampled *en route* to fibril formation as well as the dynamical processes that drive aggregation remain elusive. Increasing evidence frames the intermediate conformers – namely the soluble oligomers – as the agents responsible for amyloid pathologies (Shea *et al.*, 2019). Improved models of these intermediate conformers as well as a delineation of the pathways that connect native and fibrillar species are necessary to engineer therapeutic and diagnostic strategies to diagnose, treat, or prevent amyloid disease.

Accurate biophysical characterization of amyloid is challenging. From an experimental perspective, disordered, heterogeneous, and interconverting conformations are challenging to isolate and stabilize under conditions suitable for X-ray crystallography or NMR spectroscopy. From a computational perspective, critical events during amyloid reactions can occur across a wide range of timescales – complete aggregation from native species can take up to weeks *in vitro*, but important conformational changes occur on the sub-μs timescale. In light of this, iteration between computational and experimental techniques is an efficient strategy to map the amyloid landscape piece by piece, beginning with the conversion of native conformations into aggregation competent ones. Transthyretin (TTR, a *trans*porter of *thy*roxine and *retin*ol) is an ideal candidate for such an investigation, as it has several properties that distinguish it among amyloids. First, the native structure is known and well defined. TTR is a homotetrameric, 55 kDa protein comprised of 127-residue monomers, each arranged in a β-sandwich fold. Second, aggregation proceeds from a native-like monomeric species, and a high degree of native-like structure is thought to be present in oligomeric and fibrillar forms of TTR. However, TTR amyloidosis has also been studied using numerous techniques, and the literature provides lots of 'clues' into amyloid formation by TTR.

Several points must be considered before an accurate model of TTR oligomers can be produced. Although wild type (WT) TTR forms amyloid, mutations to the TTR sequence, proteolytic cleavage of TTR, and the conditions under which aggregation occurs all affect the distribution of oligomeric and fibrillar species formed *in vitro* and *in vivo*. At least two (annular and linear) (Dasari *et al.*, 2019) oligomeric architectures have been isolated; and at least two fibrillar architectures (type A and type B) have been isolated. Importantly, how factors, such as mutation, alter the dynamics of TTR to yield distinct oligomeric and fibrillar species is poorly understood.

The current model of oligomeric conformations of TTR is shaped by results from X-ray crystallography, NMR, computation, and protein design. Early clinical reports of TTR amyloid identified genetic variants associated with amyloid pathologies and showed that a large number of mutations occurred in the CBEF sheet of TTR, particularly in the edge strands. It was thought that structural changes to this sheet, which could be promoted by certain mutations, initiated aggregation. Currently, over 100 mutations to TTR have been reported and are distributed throughout the TTR sequence; however, the clinical and structural significance of a large share of these mutations is not known. Though numerous crystal structures of WT and variant TTR in various environments have been obtained, the native tetrameric structure of TTR is highly favored, and in nearly all cases, TTR crystallized as a tetramer in the native state. A subsequent analysis of 23 crystal structures showed that the small amplitude changes to the tertiary structure of TTR were not able to yield significant insights into amyloidogenesis (Hörnberg *et al.*, 2000). Currently, there are nearly 300 crystal structures of TTR from *H. sapiens* deposited into to the PDB, the majority of which have near-identical backbone conformations. One structure with a large Cα RMSD relative to the 1TTA structure (PDB ID 1G1O) is a designed triple mutant (G53S, E54D, L55S) that is highly amyloidogenic; however, the only apparent difference in the crystal structure is a shift in the registry of strand D and a rearrangement of the DE loop.

Various biophysical techniques have been employed to study amyloid formation by transthyretin. Hydrogen-deuterium exchange (HDX) was used to probe the solvent accessibility of TTR under amyloidogenic conditions and identify significant conformational changes during amyloidogenesis. The HDX results indicated unfolding of strands C and D; it was argued that destabilization of the CBEF sheet initiates aggregation (Olofsson *et al.*, 2004). However, more recent NMR studies employing solid state NMR and relaxation-dispersion experiments of WT and monomeric TTR (M-TTR, F87M/L110M) have predicted that the initial conformational changes to TTR monomers occur in the DAGH sheet (Lim *et al.*, 2013). Various NMR studies have predicted that dynamics in either the CBEF sheet or the DAGH sheet initiate aggregation, while others predict that native like DAGH and CBEF structures are present during aggregation (Olofsson *et al.*, 2004; Lim *et al.*, 2016b; d; Leach *et al.*, 2018). These disparate results make clear that the 'true' dynamics that initiate aggregation have yet to be definitively established and may be subject to modification by many variables.

Computational techniques have also been employed to investigate amyloid formation by TTR. In prior molecular dynamics studies, we have shown that the DAGH sheet of both WT and mutant TTR monomers undergoes a conversion from β-sheet to α-sheet secondary structure (Armen *et al.*, 2004a-b; Steward *et al.*, 2008). MD studies from other groups have predicted α-sheet conversion in the DAGH sheet (Yang *et al.*, 2006), a destabilization of hydrogen bonding patterns in the CBEF sheet (Rodrigues *et al.*, 2010), and a disruption of hydrogen bonding in the DAGH sheet (Yang *et al.*, 2005).

Protein redesign has also been used to map amyloid formation by transthyretin. Three sets of TTR designs have been developed to investigate its amyloid formation. The first contains three mutations within strand D (G53S, E54D, and L55S); these mutations result in a highly amyloidogenic variant with altered native structure in strand D (Eneqvist *et al.*, 2000). The mutant was designed based off the presence of a mutational 'hot spot' in strand D and showed that destabilization of strand D can promote aggregation. The second contains two mutations (F87M and L110M, a.k.a. monomeric-TTR or M-TTR) that result in a TTR variant that is natively monomeric and does not form amyloid unless partially denatured (Jiang *et al.*, 2001). This design was generated to destabilize the dimer:dimer interface and shows that conformational changes to the monomeric conformation of TTR are required for amyloid formation. The last set of designs contain various mutations to H88 (H88A, H88F, H88S, and H88Y) (Yokoyama *et al.*, 2017). These mutations were introduced to probe the role that H88 plays in TTR amyloid formation. Neutron diffraction studies showed that in the native state H88 mediates a hydrogen bond network that contributes to the stability of the TTR tetramer. The four designs showed that mutation at this position modulates the degree of amyloidogenicity of TTR. H88R and H88S increased the extent of amyloid formed; H88F and H88Y decreased the extent of amyloid formed, and H88A had no effect on the extent of amyloid formed (Yokoyama *et al.*, 2017).

Insights into TTR amyloidogenesis often contain contradictory predictions regarding the structure of TTR during the early stages of aggregation. The most apparent contradiction is whether the CBEF sheet or DAGH sheet is destabilized first. Several confounding factors likely contribute to this uncertainty. To address this issue, we used conventional molecular dynamics (MD) simulations to probe early conformational changes that occur to TTR monomers prior to the formation of aggregation competent species. Here, we characterize tertiary changes to the structures in our simulations, compare the conformational landscape observed in MD with

experimental data, propose possible aggregation-competent conformations, and generate testable hypotheses that we speculate may be used to correctly identify the conformation(s) of aggregation-competent TTR monomers.

4.3 Methods

The simulations presented in this chapter are the same set of simulations described in Chapter 3, the Model building and Molecular dynamics simulation methods sections have been reproduced for clarity.

4.3.1 *Model building*

The coordinates of the 1.7 Å crystal structure of wild type (WT) transthyretin (TTR, PDB ID: 1TTA) were obtained from the Protein Data Bank (PDB, Berman et al., 2000). The 1TTA crystal structure was chosen since it contains coordinates for the full-length protein and did not require modeling of the unstructured N- and C- termini. When duplicate side chains were present, the first rotameric state (conformer A, this affected residues C10, M13, L17, K48, E63, D74, K80, L82, R104, and T119) was chosen. Models of the six mutant forms of TTR (D18G, A36P, L58H, Y69H, L111M, and V122I) were generated via in silico mutations of the WT structure (chain A of the 1TTA X-ray structure) using the Dynameomics rotamer library. As previously noted, mutations to TTR rarely result in a significant deviation from the crystal structure present in 1TTA (Palaninathan, 2012). Each of the seven TTR variants was modeled under the most common in vitro model of amyloidogenic conditions: acidic pH via protonation of Asp (named Ash, net charge of 0), Glu (named Glh, net charge of 0), and His (named Hip, net charge of +1), residues.

4.3.2 *Molecular dynamics simulations*

The starting structures described previously were prepared for molecular dynamics (MD) simulations using the *in lucem* molecular mechanics (*il*mm) package, which was recently validated against a set of >3,100 experimental observables (Childers and Daggett, 2018). First, missing hydrogen atoms were modeled on the crystal structure and then minimized for 500 steps. This structure was used as the crystallographic ‘reference state’ during analysis. Next, all atoms were minimized via steepest descent minimization for 1000 steps. Next, the proteins were solvated in a

water box that extended at least 10 Å beyond any protein atom and the box volume was adjusted to reproduce the experimental density at 310 K (0.992 g/ml) (Kell, 1967). Solvent atoms were minimized for 1000 steps, equilibrated for 500 steps, and then minimized again for 500 steps. Finally, all protein atoms were minimized for 500 additional steps. Production MD simulations were performed using *il*mm (Beck *et al.*, 2000) with the Levitt *et al.* force field (Levitt *et al.*, 1995) and the flexible three-center (F3C; Levitt *et al.*, 1997) water model. Simulations were performed using the microcanonical NVE (constant **n**umber of particles, **v**olume, and **e**nergy) ensemble with periodic boundary conditions, a 10 Å force-shifted non-bonded cutoff (Beck *et al.*, 2005), and a 2 fs timestep. Coordinates were saved every picosecond for analysis. All (n = 21) production simulations were performed at 310 K, acidic pH, and in triplicate for 500 nanoseconds (ns), for an aggregate simulation time of 10.5 μs.

4.3.3 *Molecular dynamics analysis*

Unless specified otherwise, analyses were performed using *il*mm, the first 25 ns of each simulation was excluded when reporting average values, and simulations were sampled at 1 picosecond granularity.

RMSD calculations The Cα root mean squared deviation (RMSD) for each simulation as a function of time was calculated using *il*mm. Structurally stable residues in strands A (11-18), B (28-36), E (65-74), and G (103-112) were used for trajectory alignment, and the Cα RMSD was calculated for residues 11-123 (flexible N- and C- terminal residues were excluded). The Cα RMSD of individual edge strands was calculated after alignment to the structurally stable strands in the same sheet. The Cα RMSDs for strands D (residues 53-56) and H (115-123) were calculated after alignment to strands A (residues 11-18) and G (104-112). The Cα RMSDs for strands C (residues 40-48) and F (residues 90-97) were calculated after alignment to strands B (residues 28-36) and E (residues 65-74).

Pairwise RMSD matrix generation The Cα RMSD matrix was constructed via alignment of all simulations to residues within the strands (A:11-18, B:28-36, C:40-48, D:53-56, E:65-74, F:90-97, G:104-112, H:115-123) and the Cα RMSD was calculated for those same residues using *prody*. Simulations were sampled at 1 nanosecond granularity for construction of the RMSD matrix.

Multi-dimensional scaling Metric multi-dimensional scaling (MDS) was performed using the SMACOF algorithm as implemented in *sklearn.*

Contact analysis Intramolecular protein hydrogen bonds were identified when the distance between the donor-acceptor pair was less than 2.6 Å and no greater than 45° from linearity. Hydrophobic contacts were detected if two C atoms from separate residues were within 5.4 Å of one another. For the side chain to side chain network analysis, two side chains were considered in contact with one another if they formed a hydrogen bond or at least one pair of heavy atoms from the two side-chains were within 5.4 Å (carbon-carbon atom pairs) or 4.6 Å (polar – nonpolar atom pairs) of one another.

Solid State NMR Restraints Interatomic distances between the C and Ca atoms of residue pairs were measured as a function of time. Average distances between atoms *i* and *j* ($d^{avg}_{i,j}$) were calculated on a per-simulation basis and as an ensemble average using data from all (*n*=21) production simulations (Equation 1).

$$d_{i,j}^{avg} = \langle d_{i,j}^{-6} \rangle^{-1/6}$$

Equation 4.1

Figure Preparation Protein images were prepared with *UCSF Chimera*. Side-chain interaction networks were constructed with *Cytoscape* (Shannon *et al.*, 2003). The remaining figures were prepared with *matplotlib*.

Ensemble Averaging Significant changes to the tertiary structure of TTR were not observed as a function of mutation. To this end, some results described here are the result of an ensemble average across all TTR systems that have been simulated. Specific results for individual simulations are provided in the supplementary material and are discussed separately in the main text if necessary.

4.4 Results

4.4.1 *Overview of tertiary conformations sampled during MD*

We performed a total of 21 simulations of WT and variant (D18G, A36P, L58H, Y69H, L111M, V122I) TTR under amyloidogenic conditions (Figure 4.1). Each system was simulated in

triplicate for 500 ns, yielding a net sampling time of 10.5 μs. In a prior study, we showed that this degree of sampling was sufficient to observe changes to the secondary structure of TTR that are associated with amyloid formation. To quantify changes to the tertiary structure of TTR, we constructed a pairwise C_α RMSD matrix that included all simulations subsampled every 1 ns (Figure 4.2A). Most simulations sampled conformations within 6 Å of the crystallographic conformer and the average pairwise C_α RMSD among structures in this dataset was 3.2 Å. Extensive conformational changes occurred in two simulations: TTR^{WT} run 1 and TTR^{L58H} run 2, which had average C_α RMSDs of 4.3 Å and 7.2 Å relative to the crystallographic conformer, respectively. While the extent of sampling achieved was not exhaustive, several conformational changes that occurred in multiple simulations provide clues into the early events in TTR conversion that are 'hidden' to experimental techniques.

To identify the largest-amplitude conformational changes that occurred in our simulations, a multi-dimensional scaling (MDS) algorithm (Kruskal, 1964) was used to reduce the dimensionality of the pairwise C_α RMSD matrix from 10,500 to 3. The MDS algorithm embeds conformational dis-similarity data from the 10,500-dimensional C_α RMSD matrix into 3 abstract dimensions while preserving the relative distances between samples to the greatest extent possible (Figure 4.2B). MDS has been previously used to identify transition states in protein folding and to map conformational changes in other systems (Daggett *et al.*, 1996; Li and Daggett, 1996). Here, MDS was leveraged to identify early events in the conversion of TTR from the crystallographic conformation to aggregation-competent conformations. The embedded pairwise C_α RMSD data showed that most simulations sampled structures similar to the crystallographic conformer and also identified more pronounced excursions from the native state. Visual analysis of these excursions showed that the most significant conformational changes involved rearrangements and (partial) dissociation of the edge strands in TTR: strands C, D, F, and H (Figure 4.3).

4.4.2 *Amyloidogenic conditions promote edge strand dissociation in transthyretin*

The relative stabilities of the edge strands (C, D, F, and H) were determined by calculating the strand $C\alpha$ RMSD relative to the crystallographic conformation as well as the number of hydrogen bonds and hydrophobic interactions that were retained relative to the crystallographic strand interfaces: *i.e.* the BC (residues 28-36, 40-49), AD (residues 11-18, 53-56), EF (residues

66-74, 90-97), and GH (residues 104-112, 115-123) interfaces. Strands H and D were the least stable, strand C had intermediate stability, and strand F was the most stable (Figure 4.4). The dynamics of strand dissociation was analyzed in detail for 4 simulations that showed dissociation or partial dissociation of strands H (TTR^{WT} run 1), D (TTR^{D18G} run 3), C (TTR^{A36P} run 3), and F (TTR^{Y69H} run 2).

4.4.3 *Dissociation of strand H*

Dissociation of strand H was the most common topological change that occurred among TTR monomers in this dataset, and the propensity to dissociate was not affected by mutations to the sequence. Run 1 of the TTR^{WT} simulations served as an example of strand H dissociation. Dissociation initiated at both the N- and C- termini of strand H. At the N-terminus, dissociation began with the displacement of the GH loop from its native structure. In the crystallographic conformation, P113 in the GH loop forms a main chain to main chain hydrogen bond with F87 in the EF loop and four aromatic residues (F87, H88, Y114, and Y116) are well packed around the GH loop. After ~10 ns of sampling in MD, this main chain to main chain hydrogen bond was lost, packing interactions among the four aromatic residues were lost, and the GH loop projected towards solvent. Displacement of the GH loop occurred in all MD simulations in this study. At the C-terminus, residues in strand H progressively lost main chain to main chain hydrogen bonds with strand G, beginning with T123 and moving backwards through the sequence. Loss of these hydrogen bonds resulted in near complete solvation of these residues, which began to adopt disordered conformations. At 63 ns, P113 within the GH loop transitioned from the α-helical region of Ramachandran space and began sampling conformations in the β-strand region of Ramachandran space. This resulted in loss of hairpin structure in the GH loop and facilitated non-native interactions between residues in strand H and residues in the EF loop and strand F. Beginning at 135 ns, interactions between strands H and F strengthened for the remainder of the simulation: residues 117 and 119 formed hydrogen bonds with residues 92 and 94. While strand H dissociation was the most common unfolding event in these simulations, stand H rarely dissociated to this extent during the simulations as the hairpin structure of the GH turn remained stable in most simulations. Aside from the WT run 1 simulation, other instances of strand H dissociation were not coupled with refolding of strand H in native or nonnative conformations.

Instead, the dissociated strand H preferentially sampled disordered conformations in solvent (Figure 4.4, 4.5).

4.4.4 *Dissociation of strand D*

Conformational changes to strand D were the next most common topological change to TTR monomers in this dataset. In the crystallographic conformation, strand D associates with strand A via three main chain to main chain hydrogen bonds, a salt bridge formed by K15 and E52, and several hydrophobic interactions within the protein core. Strand D is much shorter in sequence than strand A (4 residues vs 8), but ordered conformations at the N- and C- termini of strand D cover the remaining surface area of strand A. At the N-terminus, residues 50-53 form a tight hairpin turn, the structure of which is further stabilized by the sidechains of S50 and S53. At the C-terminus, H56 forms a β-bulge and residues in the DE loop form a quasi-helix that shield the N-terminus of strand A from solvent. These may be 'designed' structural features to protect against aggregation (Richardson and Richardson, 2002). In some MD simulations, conformational changes to strand D and its neighboring loops resulted in the formation of non-native or disordered structures around strand D and the occasional exposure of strand A to solvent. As an example of strand D dissociation, we analyzed the dynamics of this region of run 3 of the TTR^{D18G} simulations. Dissociation of strand D began at 48 ns with the loss of structure and subsequent expansion of the hairpin turn formed by residues 50 – 53. This resulted in a loss of native main chain to main chain hydrogen bonds, the formation of non-native side chain to side chain contacts between residues 11 & 55 and 10 & 53, and a shift of strand D toward the N-terminus of strand A. The resulting conformation allowed for the formation of non-native interactions between the N-terminus and strand D as well as the formation of main chain to main chain hydrogen bonds between S46 and G47 (which are located in strand C) with H56 and L55. At 160 ns the CD loop hairpin turn and the quasi-helical DE loop unfolded, which allowed for more complete formation of a non-native CD hairpin as residues F44 and L58 came into contact with one another. Hairpin formation induced α-sheet structure in strands C and D in this simulation, but only after the CD and DE loops unfolded. This also induced α-sheet structure in the AB loop and to some extent in strand B; however, native like interactions in the CBEF sheet likely prevented complete formation of α-sheet structure in strand B, as previously discussed (Figure 4.4, 4.6).

4.4.5 *Dissociation of strand C*

In terms of both secondary and tertiary structure, the CBEF sheet of TTR was more stable than the DAGH sheet during MD simulations. Of the four edge strand interfaces in TTR, the BC loop has the least regular structure in the crystallographic conformation. This unique structure is partially induced by W41, the side chain of which extends across the sheet and forms contacts with residues in strands B, E, and F. These regions produce a highly twisted hairpin structure in the BC loop. As discussed above, dissociation of the C-terminal end of strand C may be coupled to unfolding of strand D and result in the formation of a non-native CD hairpin; however, there were also instances of dissociation of strand C alone. As an example, we analyzed the dynamics of this region in run 3 of the TTR^{A36P} simulations. Dissociation was initiated at 5.8 ns, at which time a main chain to main chain hydrogen bond between R34 and E42 was lost. In the crystallographic conformation, the W41 side chain is sandwiched between the aliphatic regions of the K35 and K70 side chains; however, expansion of the BC loop following the loss of the R34-E42 hydrogen bond promoted release of W41 from this pocket. For the remainder of the simulation, W41 was dynamic and sampled multiple conformations without returning to a native like conformation. At 29 ns, the C-terminal end of strand C began dissociating from the N-terminal end of strand B with the loss of a main chain to main chain hydrogen bond between residues 30 and 47. At 91 ns, the second hydrogen bond between R34 and E42 was broken. Shortly afterward, the K35 side chain formed hydrogen bonds with the main chain carbonyl groups of residues 38 and 39 in the BC loop, stabilizing the unfolded conformation. Some α-sheet formed in the C-terminal end of strand C (residues 45 – 48) but not in strand B, preventing re-association of the strands. By 240 ns, most of the transient interactions between strand C and D disappeared, but regular secondary structure was absent in the terminal ends of strand C, which instead populated disordered conformations. The bulge formed by residues 43 and 44 was only native-like main chain to main chain interaction left and may be responsible for maintaining strand C structure (Figure 4.4, 4.7).

4.4.6 *Dissociation of strand F*

Strands E and F constituted the most stable interface observed in these simulations. The origins of this stability are apparent in the crystallographic conformation. The EF interface has a large number of uninterrupted main chain to main chain hydrogen bonds, a strong network of side

chain to side chain interactions, and the terminal ends of strand F are anchored to other regions of TTR. For example, in the crystallographic conformation, the side chain of N98 forms stabilizing hydrogen bonds with main chain atoms in residues 64 and 103. These hydrogen bonds help maintain a compact FG loop geometry. Nevertheless, some conformational changes did occur to strand F during MD. As an example, we analyzed the dynamics of run 2 of the TTR^{Y69H} simulations. In this case, dissociation was initiated by dynamics in another region of the structure, which then propagated to strand F. After ~8 ns of sampling, native hydrogen bonds formed by the N98 side chain were lost and the side chain projected into solvent, resulting in an expansion of the FG loop and a loss of stabilizing interactions that maintained FG loop geometry. As in the A36P run 3 example, W41 unfolded from the K35/K70 sandwich, which allowed the BC hairpin turn to become oriented perpendicular to the rest of the sheet. These two conformational changes disrupted native main chain structure in strand F. Along the same timescale, the typical events associated with H-strand dissociation occurred. This caused a rearrangement of side chain packing within the hydrophobic core, as residues in strand F form hydrophobic contacts with residues in strand H in the crystallographic conformation. The 'inward' facing residues in strand F swapped from interacting with residues in strand H to interacting with residues in strand G. Then at 47.2 ns, T96, which is solvent exposed in the crystallographic conformation, flipped inward toward the hydrophobic core. The hydroxyl group on the T96 side chain formed transient hydrogen bonds with the main chain atoms of residues in strand E, further separating strands E and F. This also introduced a kink in strand F that initiated additional disruptions to the main chain structure and to hydrophobic core packing. The resulting conformation was semi-stable for the remainder of the simulation, although there were continual changes to hydrophobic packing and the FG loop structure (Figure 4.4, 4.8).

4.4.7 *Hydrophobic core packing*

In the MD simulations described here, strand dissociation and other topological changes were frequently associated with changes or disruptions to hydrophobic packing in the core of TTR. In a prior study, we found that the solvent-exposed side chain to side chain interaction networks in the DAGH and CBEF sheets had differential stabilities, which affected the propensities of these sheets to undergo secondary structure changes from β-sheet to α-sheet. Here, we investigated coupling between the side chain interaction network in the hydrophobic core and topological

changes to TTR. As expected, the side chain to side chain interaction network formed among residues in the hydrophobic core was more stable than that formed by surface-facing residues. Overall, contacts formed between residues in the interior strands (A, C, E, and G) were the most likely to be maintained. For example, contacts formed between V32-V71 and V14-I107 were maintained for 100% and 99% of the time during MD. Of the 95 residue contact pairs present in the reference structure, 37% were maintained for > 80% of the aggregate simulation time; 52% were maintained for 20 – 80% of the aggregate simulation time, and 12% were maintained for < 20% of the aggregate simulation time. The least stable of these contacts corresponded to topological changes in TTR. For example, the Y78-L111 contact indicated the presence of native-like geometry within the GH and EF loops, but was lost upon strand dissociation and subsequent projection of the GH loop into solvent. Of the 45 residue contact pairs not observed in the reference structure, 4% were present for > 80% of the aggregate simulation time, 67% were observed for 20 – 80 % of the aggregate simulation time, and 29% were present for < 20% of the aggregate simulation time (Figure 4.9).

Overall, the least stable regions of the hydrophobic core were the edge strands as well as the end of the sandwich farthest away from the α-helix. This region of TTR contains three inward facing Phe (F44 in strand C, F64 in the DE loop, and F95 in strand F) residues that are protected from solvent by the DE and FG loops. In the crystallographic structure, these aromatic residues form contacts with many neighbors within the hydrophobic core, but those contacts were disrupted by rearrangements to the hydrophobic core packing that are associated with topological changes in TTR. In the reference structure F44 contacts L12, V32, R34, T59, F64, and Y69. During MD, these interactions were present for 26%, 97%, 98%, 71%, 27%, and 77% of the aggregate simulation time, respectively. During MD, F44 formed transient contacts with E42 and L58 for 20% and 40% of the aggregate simulation time, respectively. This indicates that during TTR misfolding, F44 makes fewer contributions to the stability of the hydrophobic core in favor of interactions with residues in the edge strands C and F. In the reference structure, F64 contacts L12, F44, T59, Y69, F95, A97, N98, and Y105; these contacts were maintained during MD for 60%, 27%, 41%, 64%, 26%, 51%, 10%, and 54% of the aggregate simulation time, respectively. During MD, F64 formed transient contacts with L58 and V32 for 13% and 14% of the aggregate simulation time, respectively. This reflects an overall loss in the contribution of F64 to the topological stability of TTR under amyloidogenic conditions: the native interactions that were lost

were not replaced by new contacts. In the reference structure, F95 formed contacts with L12, F64, Y69, V71, V93, A97, Y105, and I107, which were maintained for 70%, 26%, 88%, 88%, 86%,40%, 86%, and 84% of the simulation time, respectively. During MD, F95 formed new contacts with V14, A109, T118, and L120 for 50%, 22%, 46%, and 34% of the aggregate simulation time. This reflects a tradeoff in the interaction preferences for F95: loss of interactions with residues in strands G and E were replaced by interactions in strand H (Figure 4.9, Table A.4.1).

4.4.8 *Comparison with experimental data*

In recent years, solid state NMR with selective labeling schemes has proved to be a reliable method of obtaining detailed structural restraints for mis- or un-folded conformations of TTR. Through selective labeling of specific C and C_α residue pairs, 14 interatomic distance restraints have been obtained from solid state NMR spectra for WT TTR (Lim *et al.*, 2016a-b). NMR signals were present if the specified atom pairs were within 6 Å of one another. Of these, 11 signals were present in both the native and amyloid states of TTR: L12–Y105, H31–S46, F33–Y69, G47–V30, L55–M13, I73–V30, I73–A91, F95–Y69, L110–Y116, L111–Y116, and V121–Y105; and 3 were present in the native state but absent in the amyloid state of TTR: P24–L17, A36-D39, and F87–Y114. In the amyloid state, the presence of the A36-D39 signal was dependent on the mixing time. The presence of the majority of these restraints in both the native and amyloid states suggests a preservation of native-like structure in the fibrillar state with some loss of structure in the AB and GH loops.

We calculated whether these 14 interatomic distance restraints were satisfied in MD simulations. The results for the native state of TTR were obtained using a single conformer – the reference state crystal structure. The results for the amyloid state of TTR were obtained by averaging over structures from the MD ensemble. Two calculations were obtained from MD data: the first used only structures from the WT simulations (n = 3) and the second included structures from all (n = 21) simulations. It is important to point out that the experimental results were obtained from TTR in the fibrillar state, whereas the MD results have been obtained for putative aggregation-competent conformations of TTR monomers. Thus, this comparison tests the degree to which aggregation competent species obtained computationally align with experimentally obtained restraints of fibrillar species. The computational results were in agreement with the

restraints present in both the native and amyloid states of TTR (Figure 4.10). All 11 of the signals present in both the native and amyloid states of TTR were satisfied by the MD data, both for the WT simulations (Figure 4.10A) and when all WT + mutant simulations were considered (Figure 4.10B). This suggests that the MD did not produce conformations that are 'too unfolded' or 'incorrectly' unfolded and that an appropriate level of native like structure has been preserved. The computational results were also in good agreement with the restraints present in the native state, but absent in the amyloid state of TTR (Figure 4.10). In both the WT and WT+mutant calculations, 2/3 restraints that were present in the native state were absent in the aggregation-competent state of TTR: F87-Y114 and A36-D39. Loss of these restraints was indicative of the conformational changes to the GH and BC loops which occurred during MD. The remaining restraint (P24–L17) indicates loss of structure in the AB loop and was satisfied in both the WT and WT+mutant calculations. This restraint was absent in 3/21 simulations (one simulation each from the L58H, Y69H, and V122I replicates) when the results were calculated for individual simulations. This suggests that unfolding of some regions, such as the AB loop, occur over longer timescales than have been probed by MD.

While these selective labeling schemes have provided detailed structural information, there are too few restraints to allow derivation of a structure. Additionally, there were no restraints found to be absent in the native state but present in the amyloid state of TTR. Instead, insights from the dataset have been used to refine models of TTR during aggregation. To aid in the refinement of models of TTR during aggregation and to provide testable hypotheses that can be used to assess computational results, we have calculated several additional restraints from our simulations that we believe could be informative. We calculated all 16,129 possible interatomic C - C_α distances from the MD ensemble. Of those, the majority (15,454 or 95.8%) were not satisfied in either the reference crystal structure state or the MD ensemble. The remaining restraints have the potential to provide clues into the structure of TTR during amyloidogenesis. 486 (3%) were satisfied in both the native and aggregation-competent states; 101 (0.6%) were satisfied in the aggregation-competent state but not the native state; and 88 (0.4%) were satisfied in the native state but not the aggregation-competent state. These final three categories, respectively, include restraints that indicate a preservation of native like structure during aggregation, formation of non-native structure during aggregation, and loss of native-like structure during aggregation. Table 4.1 lists 10 interatomic distances culled from this dataset that could be tested experimentally in order to

confirm or deny computational results and provide new insights into structural changes that precede the formation of aggregation-competent conformations of TTR monomers.

Our MD results have also uncovered a caveat to consider during the interpretation of NMR data: it is possible for conformational changes to occur that result in non-native like structure that nevertheless yield interatomic distances that produce NMR signals. For example, in run 3 of the V122I simulations, strand C, stand D and the CD loop formed distinctly non-native conformations; however, this simulation did agree with all experimentally derived results. Importantly, the L55-M13 interatomic distance was < 6 Å. Experimentally, this finding was interpreted to indicate that native like structure in the strand D: strand A interface is preserved in the amyloid state. Our simulations show that while the overall topological connections in this interface are native-like there are several structural differences relative to the crystal structure including the exposure of the C-terminal end of strand A and a change in the registry of the D:A interface (Figure 4.10C).

4.5 Discussion

4.5.1 *Conformational changes the early stages of TTR amyloidogenesis*

Available experimental data have suggested that destabilization of either the CBEF sheet or the DAGH initiates TTR un-folding during amyloidogenesis. These results have been obtained from distinct TTR sequences under subtly different environmental conditions and are likely confounded by the presence of multiple conformational states. Previous structural studies of TTR have indicated varying degrees of native-like structure within TTR aggregates, yet the results have been contradictory. Recent findings have lain methodological concerns on most of these studies, which have been performed under different environmental conditions that are known to affect the ultimate topological arrangement of mature TTR amyloid fibrils. Additionally, sample preparation methods may have resulted in a range of conformational states (e.g. tetramer, monomer, aggregation-competent monomer, oligomers, and fibrils) that contributed to the reported signals.

A recent series of solution and solid-state NMR experiments have improved sample preparation methods and performed experiments in a consistent way that facilitates comparisons with the MD results presented here. Relaxation-dispersion experiments using solution NMR of intermediate TTR species during aggregation have predicted stable, native-like structures for residues in the CBEF sheet, whereas the DAGH sheet was shown to undergo conformational

fluctuations at and below millisecond timescales. We speculate that these fluctuations may correspond to conversion to α-sheet structure which was observed in our simulations. In our MD study, the CBEF sheet was also more stable: the edge strands in the CBEF sheet had a reduced propensity to dissociate relative to the DAGH sheet and the side chain solvent exposed contact network was more consistent. Solid state NMR experiments using selective C and C_α labeling schemes mapped the degree of native like structure present in mature TTR fibrils. The resulting cross peaks showed native like CBEF and DAGH structures within fibrils, with the potential for dissociation of strand D and exposure of strand A. While there was good correspondence between computational and simulated results, the analysis has pointed to a few counterexamples that limit the degree to which these data demonstrate true 'native-like' structure. Furthermore, our analysis has revealed several potentially interesting residue pairs that merit investigation in future studies.

Overall, interpretation of the experimental and computational results has produced a clearer picture of transthyretin during amyloidogenesis. In a solution NMR HSQC spectrum of TTR amyloid intermediates, the observable resonances were mapped to residues in the C, B, E, and F strands. This further indicated stable, native like structure in the CBEF sheet and fluctuations/disorder in the DAGH sheet. Follow-up solid-state NMR experiments showed that the AB loop (located within the DAGH sheet) becomes disordered in the mature amyloid state. The MD results, which track the behavior of single molecules, suggest that conformational changes within the DAGH sheet are more likely as this region had high propensity to form α-sheet as well as the greatest propensity for strand dissociation (both strand D and strand H). MD also showed that conformational changes to the edge strands of the CBEF sheet also occur but on a slower timescale than in the DAGH sheet.

4.5.2 *Coupling of changes in the tertiary and secondary structures of TTR*

The discretization of secondary and tertiary interactions in the protein structure hierarchy kindles the subtle idea that the two levels of structural organization are unrelated, when in fact they can be very correlated. Prior MD studies have shown that tertiary contacts can influence the formation and dissolution of secondary structure elements and *vice versa*. In studies of barnase and protein A, tertiary contacts influenced the formation of secondary structure elements during protein folding (Bond *et al.*, 1997; Wong *et al.*, 2000; Alonso and Daggett, 2000; Scott *et al.*, 2006). For barnase, a hydrophobic interaction network centered on a Trp residue aided in the

nucleation of a neighboring α-helix. In this study, interactions at the tertiary structure level were shown to influence pathological conformational changes as well as protein folding. In D18G run 3, unfolding of strand D led to α-sheet conversion in strands C and B. In WT run 2, the rearrangement of strand D and the partial dissociation of strand C led to a separation of strands B and E near the α-helix in TTR and conversion to α-sheet for residues in this region followed. In run 2 of the L111M simulations, the BC turn and surrounding residues moved almost perpendicular to the CBEF sheet. In V122I run 3, the AB loop and α-helix separated from one another. This allowed for α-sheet conversion in the C-terminal end of the AB loop and N-terminal end of strand B, which later propagated to strands E and F. Taken together, these results contribute to the current model of aggregation-competent TTR monomers: destabilization and unfolding of the DAGH sheet occurs early in unfolding. We propose that loss of tertiary structure in the DAGH sheet can propagate to the CBEF sheet and induce α-sheet conversion.

4.5.3 *Future directions & testable hypotheses*

One limitation of the NMR dataset is that there were no restraints found to be absent in the native state but present in the amyloid state of TTR. Another limitation of the NMR data is that it is possible for conformational changes to occur that result in nonnative like structure that nevertheless yield signals that indicate a preservation of native like structure. To this end, we have predicted several residues pairs from our simulations for which a signal should be absent in the native state of TTR but present in the amyloid state of TTR. These hypotheses can be directly tested experimentally, which has the potential to (a) validate or invalidate our computational results and (b) provide additional insight into the conformation(s) of aggregation-competent TTR monomers. These simulations have also identified structural regions prone to misfolding under amyloidogenic conditions in TTR. We speculate that knowledge of these regions may allow for the design of TTR variants with altered stabilities. For example, A120L and A45L should stabilize the H and C strands, respectively, and inhibit edge strand dissociation.

4.6 Conclusions

The heterogeneous and dynamical nature of amyloid proteins has been a perennial frustration. The available experimental data that probe the structures of amyloid states are often

low-resolution or incomplete. Here, we have employed computational methods that track the dynamics of single molecules over time in order to better assess the early dynamics that result in aggregation-competent TTR conformations. Our computational results are in good agreement with available experimental data. Our results show that structural rearrangements and dissociation of the edge strands in TTR were the most common conformational change to TTR monomers. Furthermore, our results further point to the destabilization and rearrangement of the DAGH sheet as an initial event in the formation of aggregation-competent monomers. Finally, we have shown that results from our simulations can in turn be used to suggest experiments or mutations to TTR that could further refine models of toxic amyloid species, which should aid in the development of improved strategies to diagnose and treat amyloidosis.

4.7 Tables

Table 4.1. Predicted interatomic distance restraints

Atom 1	Atom 2	Reference Distance	MD Ensemble Distance	Description
A25 @ C	16 @ C_α	5.8	6.6	AB Loop Disorder
A25 @ C	T49 @ C_α	5.9	6.8	AB / CD Loop Separation
G53 @ C	P24 @ C_α	5.7	6.6	Strand D disorder / registry
L58 @ C	A45 @ C_α	5.6	7.2	Strand D disorder / registry
V65 @ C	N98 @ C_α	5.6	6.4	Strand F disorder
T75 @ C	H90 @ C_α	5.5	6.5	Strand F disorder
N98 @ C	P102 @ C_α	5.6	7.3	FG Loop disorder
P11 @ C	G57 @ C_α	8.7	5.6	Strand D disorder / registry
G47 @ C	L55 @ C_α	6.7	5.9	Association of strands C and D
L55 @ C	T49 @ C_α	6.4	5.7	Association of strands C and D

4.8 Figures

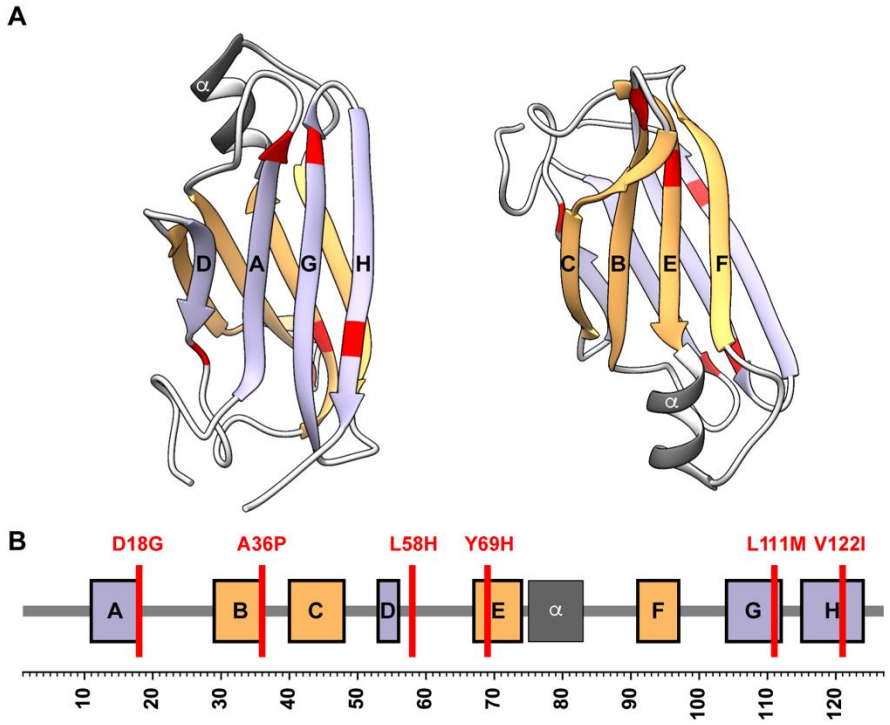


Figure 4.1. The sequence and crystallographic conformation of transthyretin.

(A) The crystallographic structure (PDB ID: 1TTA) of a wild type TTR monomer is shown. The DAGH sheet is colored purple, the CBEF sheet is colored orange, the α-helix is dark grey, and loops are light grey. The ribbons of residue positions where mutations introduced in this study are colored red. (B) The x-axis denotes residue positions in the TTR monomer. The colored rectangles correspond to secondary structure elements and are colored as in panel A. Vertical red bars indicate the individual mutations that were introduced in this study.

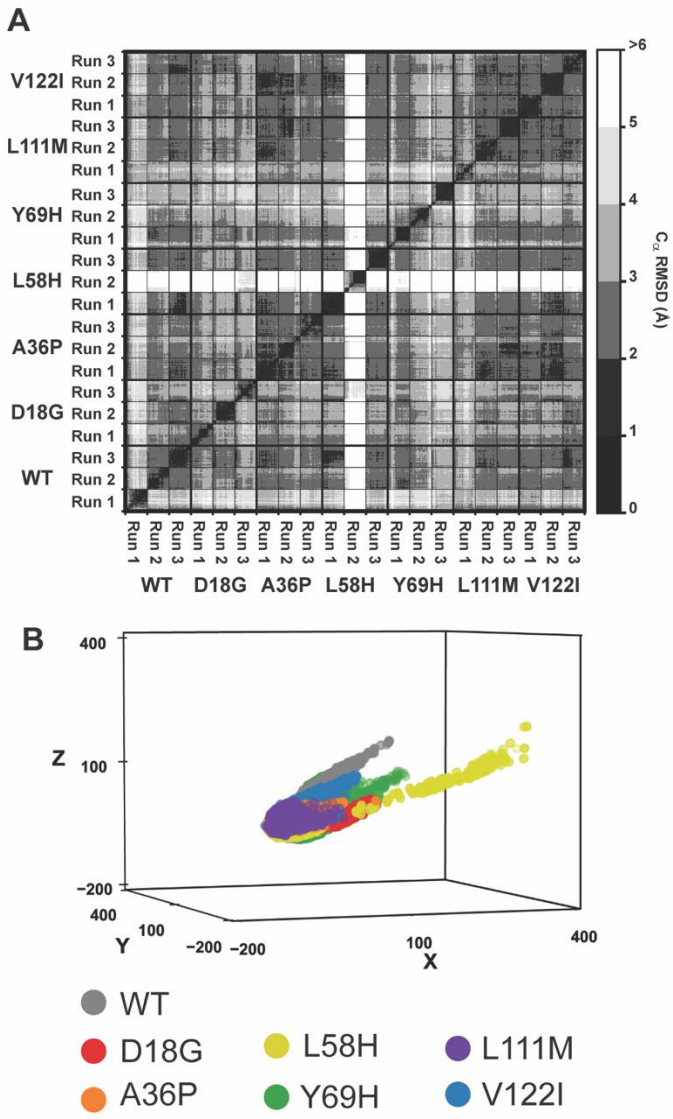


Figure 4.2. Tertiary conformations sampled during MD.

(A) The Cα RMSD matrix highlights conformational diversity within the topological structure of TTR monomers during MD simulations. The central diagonal line is a comparison of each structure in the dataset to itself and is always 0 Å. Submatrices along the diagonal that are surrounded by thin black lines are a comparison of each structure in a simulation to all other structures in that same simulation. Submatrices along the diagonal that are surrounded by thick black lines are a comparison of each structure in a system (WT or mutant TTR) to all other structures in that same system. Off diagonal elements are comparisons between different systems. Large off-diagonal RMSD values (white) indicate that unique conformations not observed in other simulations, e.g. L58H. Small off-diagonal RMSD values (black) indicate that a conformation is sampled in

multiple systems or simulations. (B) A 3D projection of the pairwise C_α RMSD dataset after application of the MDS algorithm. Points are plotted in a dimensionless, embedded-coordinate space and colored on a per-system basis (WT: grey, D18G: red, A36P: orange, L58H: yellow, Y69H: green, L111M: purple, V122I: blue).

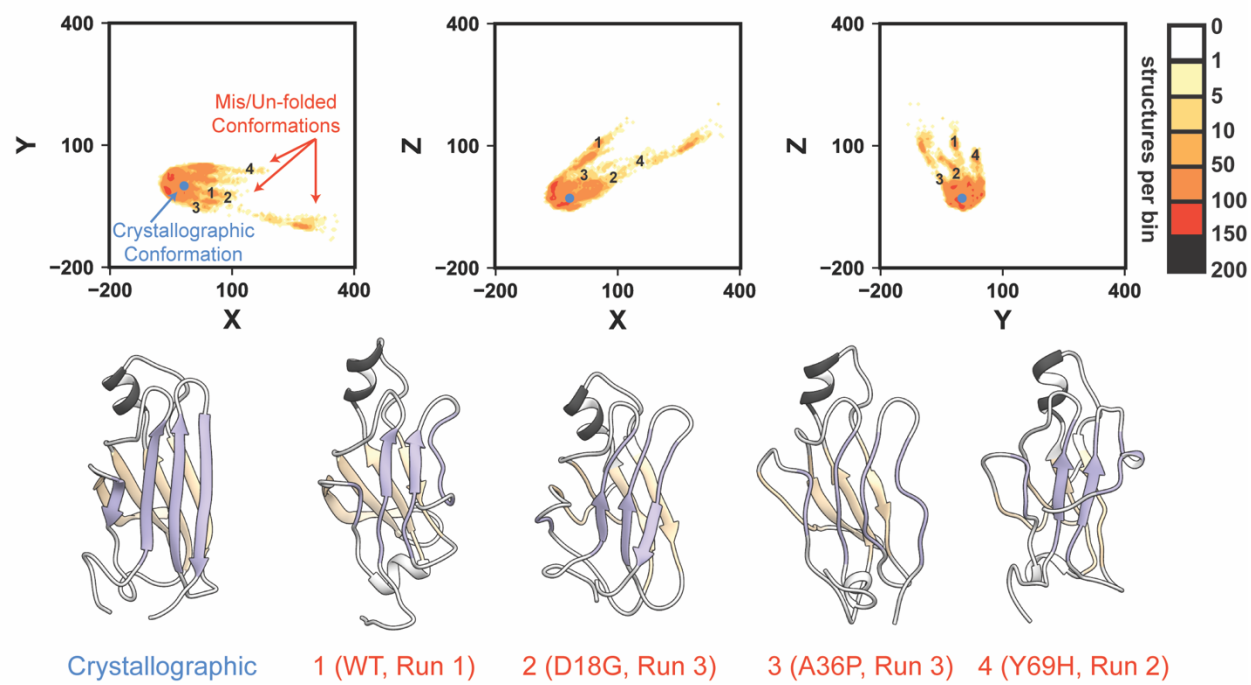


Figure 4.3. Multi-dimensional scaling highlights excursions from the native TTR conformation.

Each plot shows a two-dimensional histogram of the pairwise C_α RMSD dataset after application of the MDS algorithm. Values are plotted in a dimensionless, embedded-coordinate space and colored based on the extent of sampling observed. The blue circle represents the position of the crystallographic reference structure. The annotated points (A, B, C, and D) in the Y-Z plane correspond to the mis/unfolded conformations shown at the bottom of the figure.

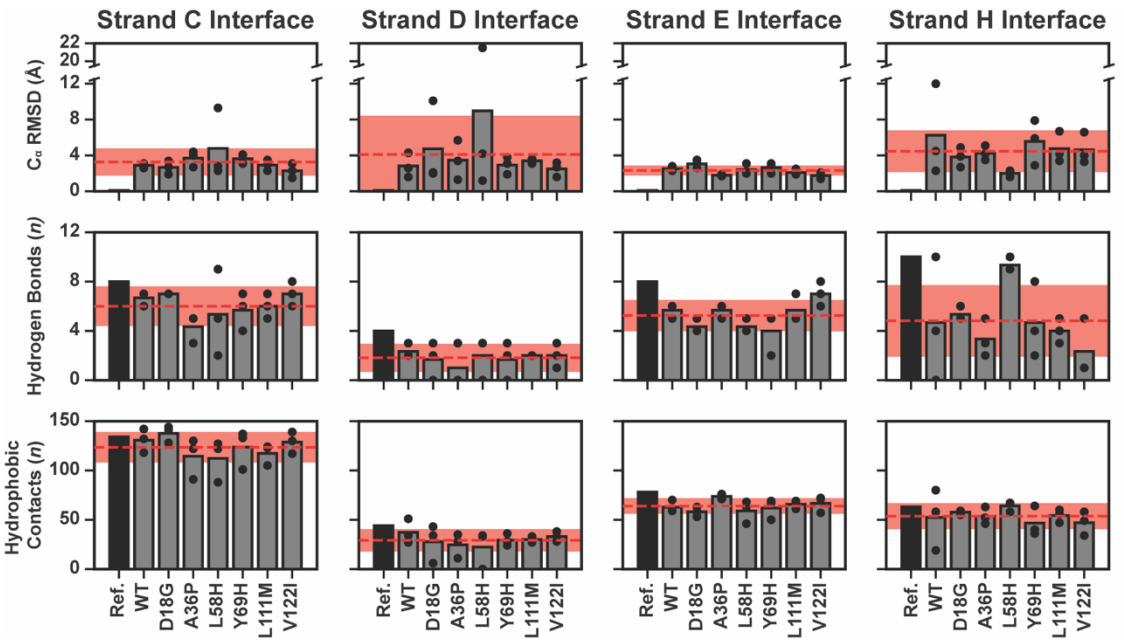


Figure 4.4. Structural stability of the edge strands

We evaluated the structural stability of the edge strands (C, D, F, and H) by calculating the C_α RMSD of the strand after alignment to other strands in the same sheet (i.e. strands D and H were aligned to strands A and G; C and F to E and G, top row), the number of hydrogen bonds formed at the strand interfaces (middle row), and the number of hydrophobic contacts formed at the strand interfaces(bottom row). The reported values were averaged over the final 25 ns of each simulation. Black bars correspond to the value present in the crystallographic reference structure. Each grey bar corresponds to the average value among the three replicate simulations for a given system. Each circle corresponds to the value obtained from a single simulation. The dashed red lines and shaded red regions correspond to the mean and standard deviation across all MD simulations.

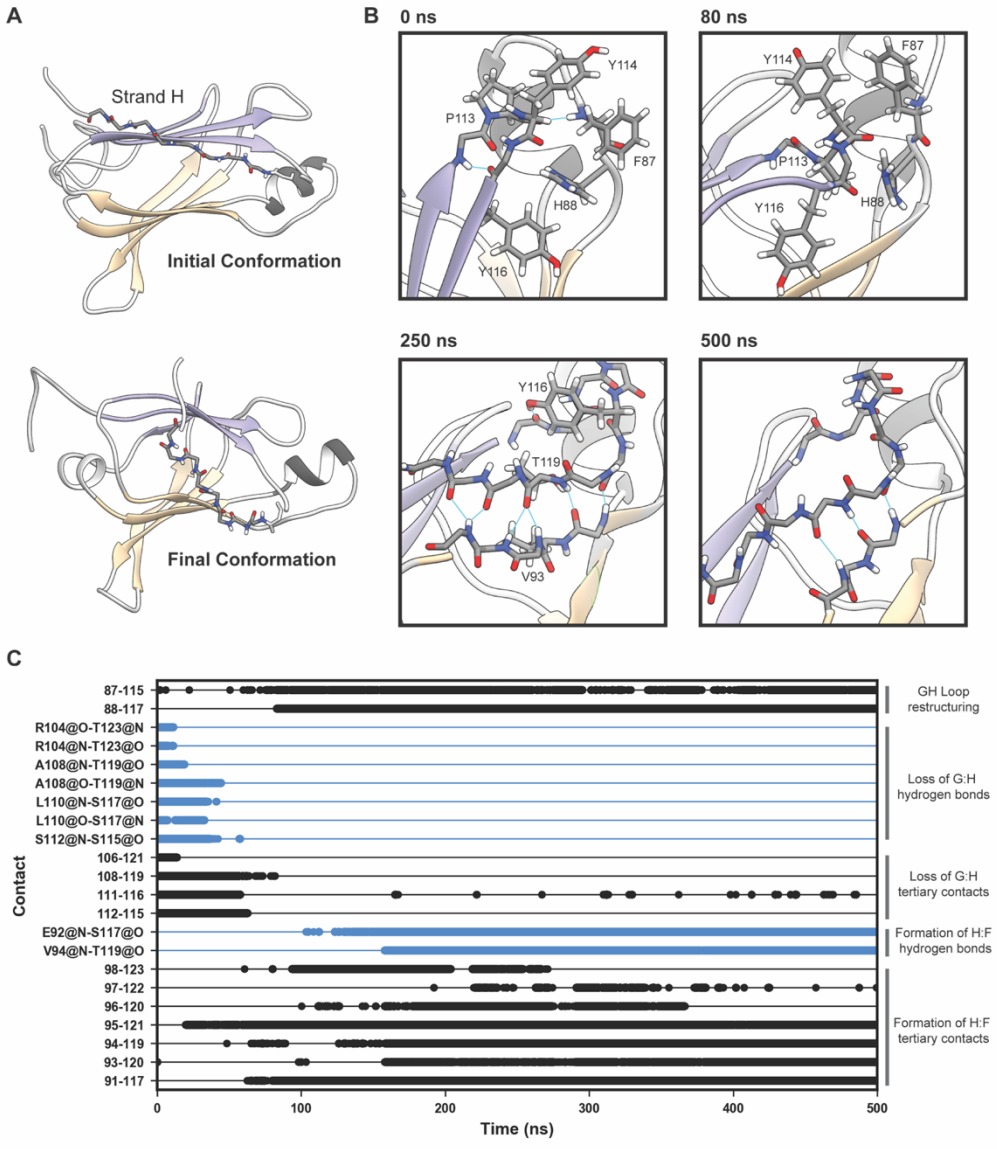


Figure 4.5. Dynamics of strand dissociation at the G:H interface

(A) The structures of the starting and final conformations of WT TTR, run 1 are shown with strand H in a stick representation. (B) Snapshots from the simulation illustrate the sequence of events associated with the dissociation of strand H. At 0 ns, the GH loop and helix-EF loop are packed together well, stabilized by interactions among several aromatic residues. Between 80 and 250 ns, native like structure in this region was lost: the GH loop lost hairpin structure and began to form interactions with strand F. By the end of the simulation, strand H formed main chain hydrogen bonds with strand F. (C) Interatomic hydrogen bonds (blue) and inter-residue hydrophobic interactions (black) that are diagnostic of strand H dissociation are shown as a function of time and organized according to the specific structural changes they are associated with.

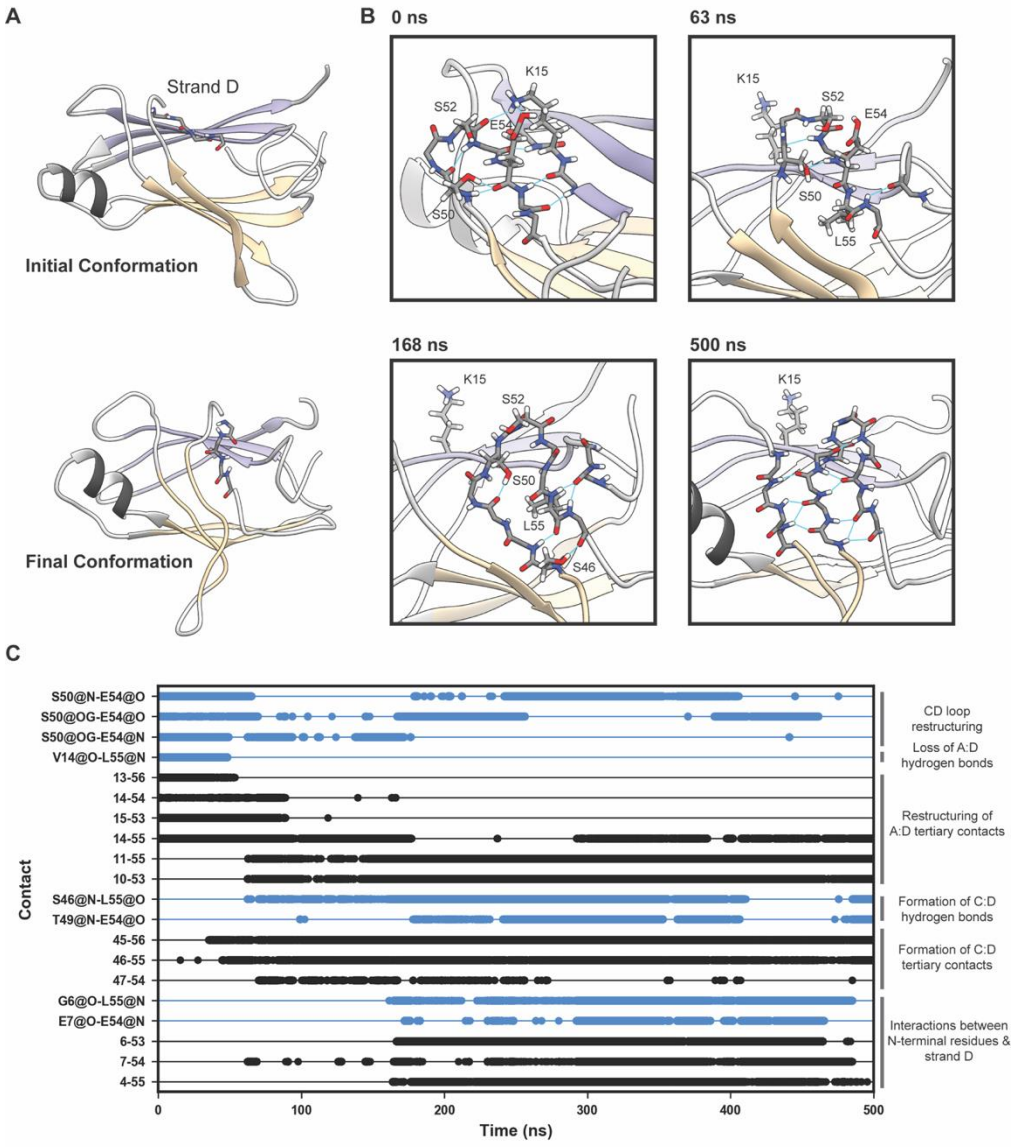


Figure 4.6. Dynamics of strand dissociation at the D:A interface

(A) In run 3 of the D18G TTR simulations, strand D dissociated from the DAGH sheet and formed a non-native α-sheet hairpin with strand C. The structures of the starting and final conformations are shown with strand D in a stick representation. (B) Snapshots from the simulation illustrate the sequence of events associated with the dissociation of strand D. (C) Interatomic hydrogen bonds (blue) and inter-residue hydrophobic interactions (black) that are diagnostic of strand D dissociation are shown as a function of time and organized according to the specific structural changes they are associated with.

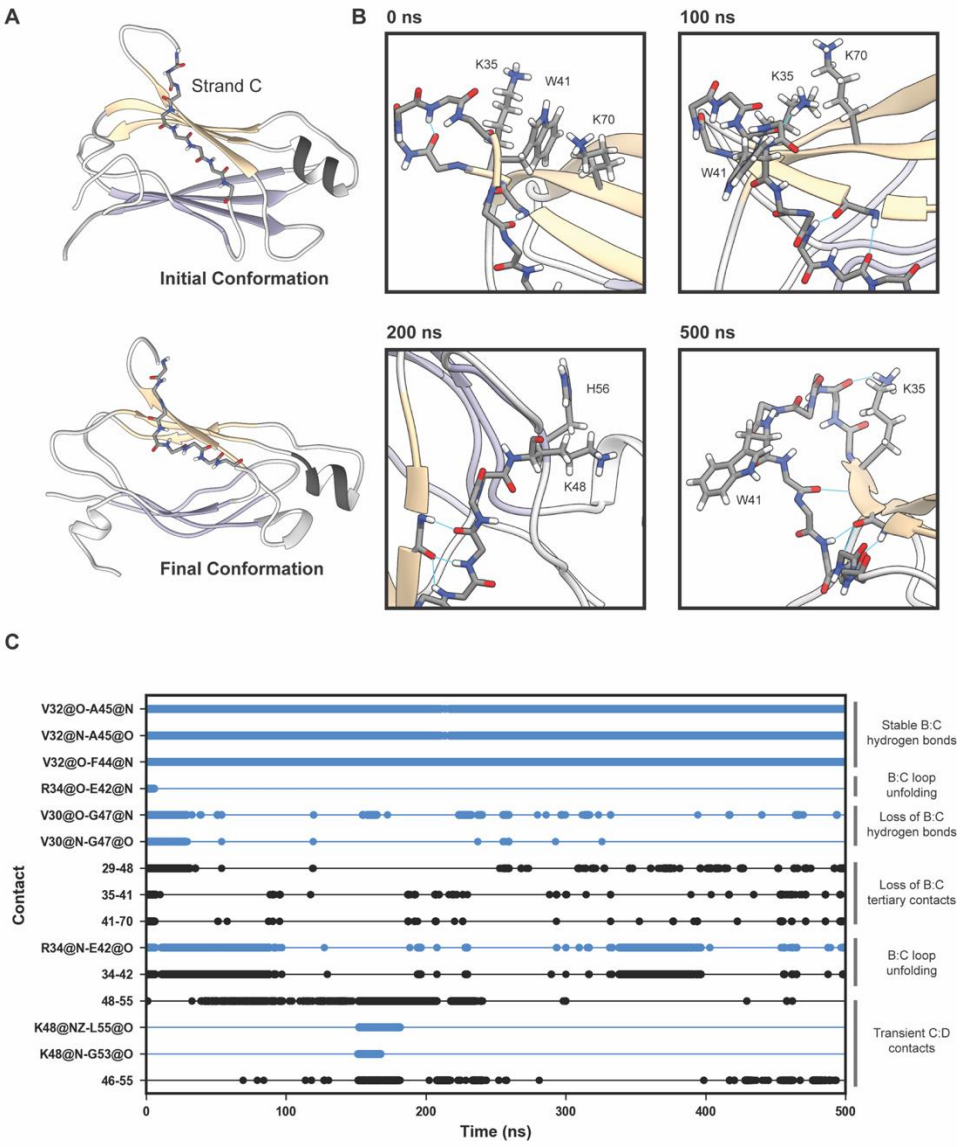


Figure 4.7. Dynamics of strand dissociation at the B:C interface

(A) In run 3 of the A36P TTR simulations, strand C dissociated from the CBEF sheet and formed transient, non-native interactions with strand D. The structures of the starting and final conformations are shown with strand C in a stick representation (B) Snapshots from the simulation illustrate the sequence of events associated with the dissociation of strand C. (C) Interatomic hydrogen bonds (blue) and inter-residue hydrophobic interactions (black) that are diagnostic of strand C dissociation are shown as a function of time and organized according to the specific structural changes they are associated with.

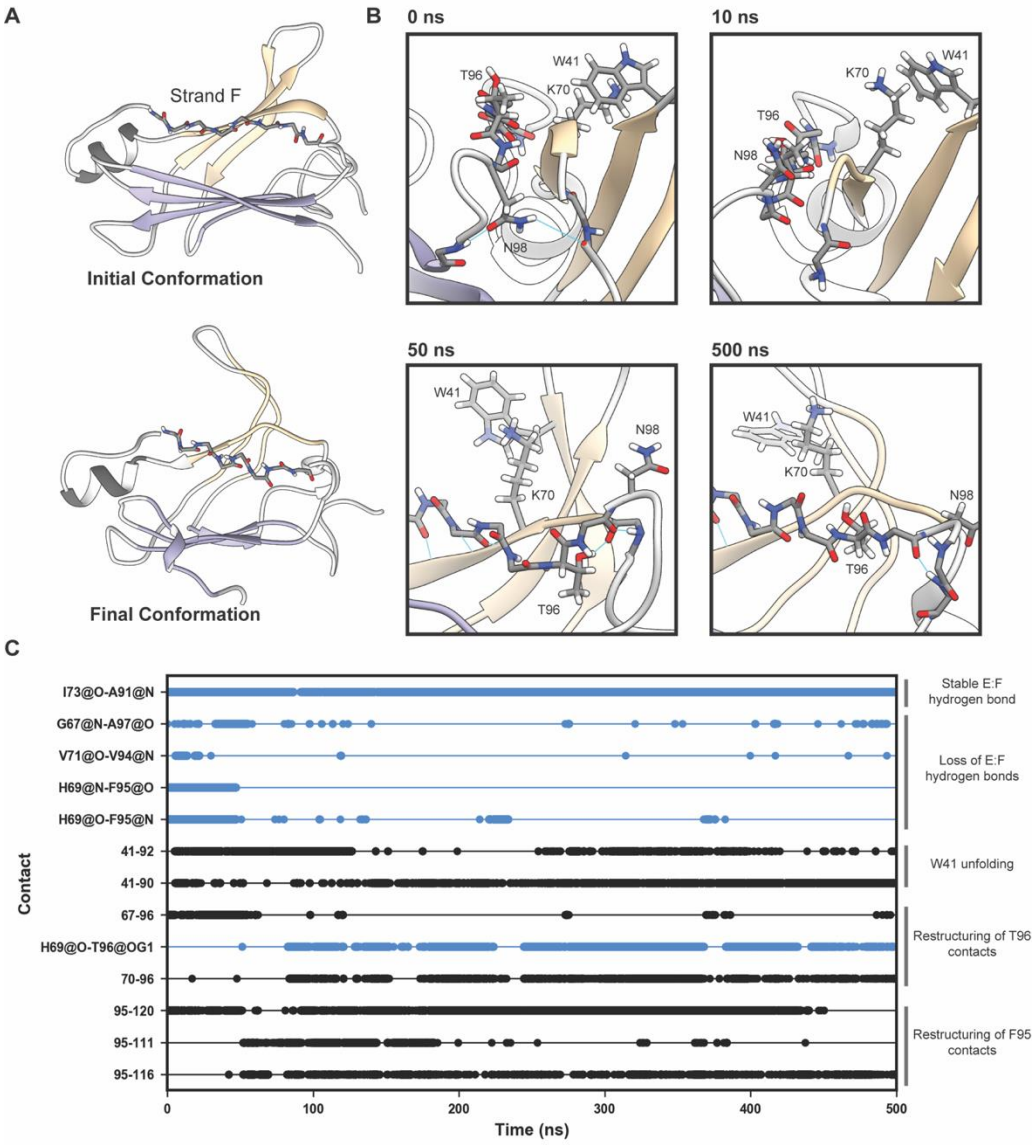


Figure 4.8. Dynamics of strand dissociation at the E:F interface

(A) In run 2 of the Y69H TTR simulations, strand F partially dissociated from the CBEF sheet. The structures of the starting and final conformations are shown with strand F in a stick representation. (B) Snapshots from the simulation illustrate the sequence of events associated with the dissociation of strand F. (C) Interatomic hydrogen bonds (blue) and inter-residue hydrophobic interactions (black) that are diagnostic of strand F dissociation are shown as a function of time and organized according to the specific structural changes they are associated

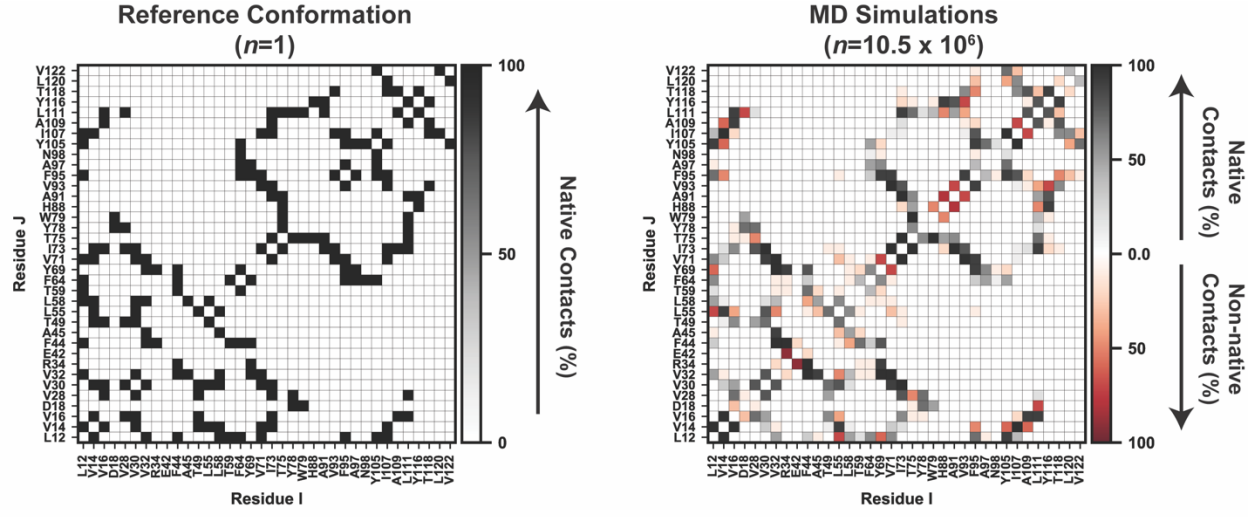


Figure 4.9. Stability of the side chain to side chain interaction network in the hydrophobic core of TTR monomers under amyloidogenic conditions.

The contact maps show the frequency of side chain – side chain interactions in the crystallographic reference structure (left) and MD ensemble (right). 'Native' interactions, or those present in the crystallographic reference structure, are colored on a scale from white (never in contact) to black (always in contact). 'Non-native' interactions, or those present only in the MD ensemble, are colored on a scale from white (never in contact) to red (always in contact).

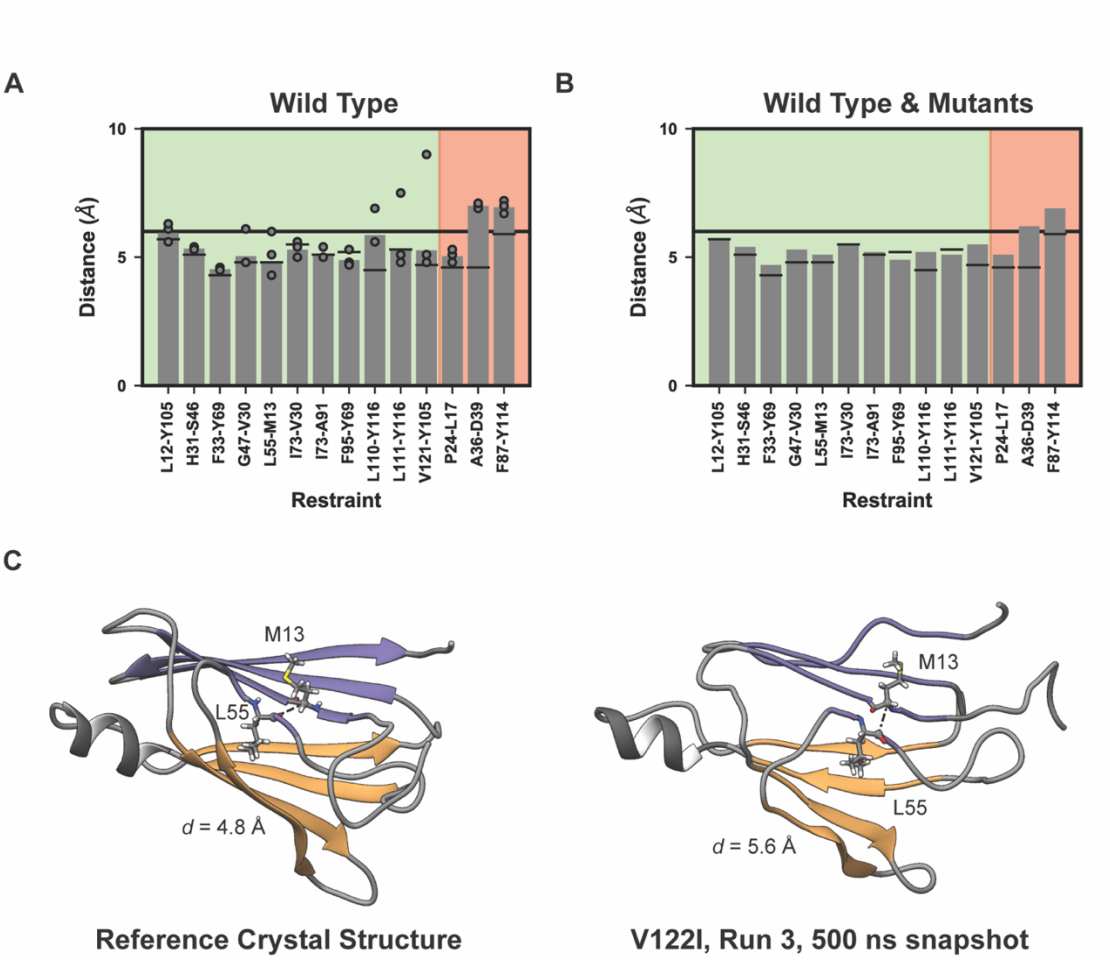


Figure 4.10. Molecular dynamics simulations reproduce experimentally observed interatomic distance restraints.

Solid state NMR proton-driven-spin diffusion (PDSD) experiments were performed on TTR in the native and amyloid states to uncover the degree to which native like structure is preserved in fibrillar states. 14 residue pairs were selectively labeled at the C and C_α atoms. Of these, 11 atom pairs resulted in an NMR signal (i.e. the atoms were < 6Å from one another) in *both* the native and fibrillar states (green shading) and 3 pairs resulted in a signal in the native state but not the fibrillar state (red shading). The r^{-6} (Equation 4.1) averaged interatomic distances for these restraints were calculated from the WT TTR MD simulations (A) and as an ensemble average over WT and all mutant TTR simulations (B). In A and B the grey bars give the average interatomic distance over all (A, n = 3; B, n = 21) simulations; short black lines indicate the distance in the reference structure; long black lines indicate the distance threshold for yielding a signal; and circles indicate

the distances obtained from individual simulations. The MD results showed good correspondence with experiment: all 11/11 restraints present in both the native and fibrillar states were satisfied by the MD ensemble and 2/3 restraints present in the native but absent in fibrillar states were satisfied by the MD ensemble. The restraints are labelled to that the first listed residue was selectively labeled on C ad the second residue was labeled on C_α. (C) The presence of a signal in both the native and fibrillar states has been interpreted as a preservation of native-like structure in the amyloid states of TTR. However, this is not necessarily the case. Here, a snapshot of the V122I run 3 simulation is compared against the reference crystal structure to highlight that a restraint (L55@C – M13@C_α) can be satisfied even though the structure is not native like. In this example, strand D has adopted a non-native structure: the CD loop is disordered and the registry of strand D vs strand A has been altered relative to the reference crystal structure, yet the interatomic distance is preserved.

Chapter 5. MODULATION OF SECONDARY STRUCTURE IN TRANSTHYRETIN THROUGH PROTEIN REDESIGN

5.1 Summary

A secondary structure conversion from β-sheet to α-sheet occurred in molecular dynamics simulations of transthyretin monomers. The conversion occurred via peptide plane flips of main chain peptide groups. Based on these results, a mechanism was proposed in which the transition from β-sheet to α-sheet is mediated by three types of electrostatic interactions: attractive interactions between main chain carbonyl groups and either water molecules or polar side chain atoms and repulsive interactions between main chain carbonyl groups from neighboring strands. To computationally test this proposed conversion mechanism, a mutant version of transthyretin was designed to favor β-sheet to α-sheet conversion. Here, molecular dynamics simulations of the design show that protein design can be used to modulate the extent of α-sheet formation observed in transthyretin under amyloidogenic conditions.

5.2 introduction

Transthyretin (TTR) from *Homo sapiens* is a homotetrameric 55 kDa protein comprised of four 127 residue monomers that adopt a β-sandwich topology. Each monomer is comprised of two four-stranded β-sheets (called the DAGH and CBEF sheets) and a single α-helix. Canonically, TTR in *H. sapiens* functions as a *trans*porter of *thy*roxine and *retin*ol in the plasma and cerebrospinal fluid and is one of the most abundant plasma proteins (Vieira and Saraiva, 2014). TTR is also a highly amyloidogenic protein associated with several amyloid diseases and can form amyloid deposits throughout the body including the heart, PNS, tenosynovium, meninges, and eyes, among others (Saraiva, 2001). Deposition of TTR amyloid can lead to secondary pathological consequences, such as osteoarthritis (Matsuzaki *et al.*, 2017). Aggregation of TTR occurs via a low pH-mediated pathway: below a pH of 5, the TTR tetramer dissociates into monomeric species, which undergo conformational changes prior to aggregating into toxic oligomeric species and, ultimately, mature amyloid fibrils (Lai *et al.*, 1997; Foss *et al.*, 2005).

Dissociation of the TTR tetramer is an obligate and rate-limiting step for TTR amyloidogenesis, a property that has been capitalized on therapeutically (Bulawa *et al.*, 2012).

Due to the amyloidogenic potential of transthyretin, the body employs a network of safeguards to regulate the pool of circulating TTR and protect against aggregation. These safeguards are present at multiple levels of biological organization. First, TTR has a relatively fast turnover in the plasma. The average half-life for the most abundant plasma protein (albumin) is ~19 days (Prinsen and De Sain-Van Der Velden, 2004) and in contrast the half-life of TTR is ~2 days (Tsegaye *et al.*, 2017). At least two chaperone proteins, Hsp90 and BiP, can recognize and bind to mis-folded and aggregation-competent conformations of TTR monomers (Susuki *et al.*, 2009; Oroz *et al.*, 2017). These chaperones can permit proper folding of TTR to a native conformation or mark it for degradation. The structure of TTR is optimized – via evolutionary pressure – for structural stability. The native conformation, the tetramer, is very stable under physiological conditions and is the dominant TTR conformation between pH 5 and 7. The tertiary structures of TTR monomers include anti-aggregation safeguards present in β-sandwich folds (Richardson and Richardson, 2002), including the presence of β-bulges and charged residues located on edge strands. In multiple molecular dynamics (MD) simulations of TTR monomers, we observed that the solvent-exposed side chain – side chain interaction network in the CBEF sheet was more stable than the solvent-exposed interaction network in the DAGH sheet. In the native tetrameric state, the CBEF sheet is solvent exposed, but the DAGH sheet is buried and forms the T4 binding site. Based on these observations, we hypothesized that the CBEF sheet is 'optimized' for structural stability whereas the DAGH sheet is 'optimized' for ligand binding and stabilization of the tetrameric conformation. Several factors can modulate the effectiveness of these safeguards. Disruptions in metabolic and proteostasis networks lead to dysregulation of the circulating pool of TTR and increasingly permit TTR amyloidogenesis during aging (Cohen, 2012). Mutations to TTR that decrease the thermodynamic or kinetic stabilities of the tetrameric species result in tetramer dissociation at higher pH values and an increase in the population of aggregation-competent monomeric species (Sekijima *et al.*, 2005). Pathologically, such mutations typically correspond to aggressive, early-onset amyloid diseases.

The Daggett lab has proposed that the α-sheet hypothesis underlies the aggregation of TTR and other amyloids. This hypothesis posits that α-sheet secondary structure is formed during the initial stages of aggregation and that the unique biophysical properties of α-sheet contribute to or

drive aggregation (Daggett, 2006; Bi and Daggett, 2018). α-sheet is a rare type of secondary structure first predicted by Pauling and Corey and later 're-discovered' by the Daggett lab in multiple MD simulations. α-sheet secondary structure is defined as an extended type of secondary structure in which sequential residues alternate between the right-handed and left-handed helical regions of Ramachandran space (abbreviated α_R and α_L, respectively). This results in a main chain conformation that is locally helical, but forms slightly curved, extended structures. Unlike extended β-sheet conformations in which peptide backbone groups alternate on two sides of the main chain, in α-sheet the backbone peptide carbonyl groups are aligned on one side of the main chain and the amide groups on the other. In multi-stranded α-sheets, this leads to the build-up of a molecular dipole across the main chain, and we propose that this structure participates in or drives protein aggregation. In an MD study of TTR monomers, we found that electrostatic interactions between solvent molecules and main chain carbonyl groups drive conversion from β-sheet to α-sheet secondary structure. In that study, we also predicted that three conformational changes occur prior to electrostatic-driven conversion: the loss of main chain – main chain hydrogen bonding patterns, the formation of distorted main chain geometries, and disruptions in side chain – side chain interaction networks. We also found that one sheet in TTR, the DAGH sheet, was susceptible to α-sheet conversion on the sub-μs timescale and that the CBEF sheet was resistant to α-sheet conversion on the sub-μs timescale. Here, our primary objective is to computationally test the proposed mechanism for β-sheet to α-sheet conversion in TTR using protein re-design to modulate the α-sheet propensity in TTR. We engineered a TTR variant with altered α-sheet content. We introduced a single mutation that destabilized the CBEF sheet and promoted α-sheet conversion.

5.3 Methods

The simulations presented in this chapter are the same set of simulations described in Chapter 3, the Model building and Molecular dynamics simulation methods sections have been reproduced for clarity.

5.3.1 *Multiple Sequence Alignment*

Homologous TTR sequences were identified using the *BLAST* (Altschul *et al.*, 1990) webserver provided by the National Center for Biotechnology Information (NCBI) (https://blast.ncbi.nlm.nih.gov/Blast.cgi). The amino acid sequence from the transthyretin precursor protein from *Homo sapiens* (NCBI Reference Sequence NP_000362.1) was queried against the non-redundant protein sequence database (nr) using the protein-protein *BLAST* (*blastp*) algorithm in July 2018. The sequence of the transthyretin precursor protein includes the mature transthyretin protein (127 residues) plus a 20-residue long signal peptide (for a total sequence length of 147 residues). A maximum of 5000 target sequences were requested using an expect threshold of 10, word size of 6, and a maximum of 0 matches in a query range. Alignments were scored using the BLOSUM62 substitution matrix with gap creation and extension costs of 11 and 1, respectively. The *BLAST* query returned 1,483 results, which were subsequently used in the construction of a multiple sequence alignment using the *Clustal Omega* webserver (Sievers *et al.*, 2011). The set of aligned sequences were then filtered to remove any 'predicted', 'hypothetical', 'partial', or 'synthetic' TTR sequences as well as any sequences corresponding to known pathologic mutations of *H. sapiens* TTR. After filtering, 860 sequences remained for analysis. The sequences were numbered according to two numbering schemes: the position in the MSA as well as the position corresponding to the *H. sapiens* TTR sequence. That is, the positions within target sequences with insertions, deletions, or gaps that were not present in the *H. sapiens* TTR sequence were ignored. In this numbering scheme, we assume that target sequences also adopt similar structures to the *H. sapiens* TTR and that insertions, deletions, and gaps do not significantly affect the structure, but this cannot be confirmed for all sequences.

5.3.2 *Sequence Analysis*

Sequences in the filtered MSA were analyzed to infer structural restrictions on mutations to the TTR, the positions of conserved residues, and the positions of covarying residue pairs. First, we calculated the frequency of each of the standard 20 amino acids, plus gaps, at each position in both of our numbering schemes. Second, we calculated the sequence entropy based on the Shannon entropy formula (Equation 5.1), where H is the Shannon entropy for a given position in

the MSA, *i* is an index over all possible amino acid types, N is the number of possible amino acid types, and P_i is the fraction of residues in a given MSA position of amino acid type *i*.

$$H = -\sum_{i=1}^{N} P_i log_e(P_i)$$

Equation 5.1

Results are presented graphically for the *H. sapiens* numbering scheme in the main text.

5.3.3 *Model building*

The coordinates of the 1.7 Å crystal structure of wild type (WT) transthyretin (TTR, PDB ID: 1TTA) were obtained from the Protein Data Bank (PDB; Berman *et al.*, 2000). The 1TTA crystal structure was chosen since it contains coordinates for the full-length protein and did not require modeling of the unstructured N- and C- termini. When duplicate side chains were present, the first rotameric state (conformer A, this affected residues C10, M13, L17, K48, E63, D74, K80, L82, R104, and T119) was chosen. A TTR variant with altered amyloidogenic potential was designed manually using *UCSF Chimera* and our pentapeptide-derived continuous rotamer library (Childers *et al.*, 2018). The design is called TTR-unstable monomer (TTR-um) and contains a single mutation relative to wild type: F33A, which was introduced to destabilize the CBEF sheet and promote conversion from β-sheet to α-sheet secondary structure. As previously noted, mutations to TTR rarely result in a significant deviation from the crystal structure present in 1TTA (Palaninathan, 2012). The TTR structures were modeled under the most common *in vitro* model of amyloidogenic conditions: acidic pH via protonation of Asp (named Ash, net charge of 0), Glu (named Glh, net charge of 0), and His (named Hip, net charge of +1), residues.

5.3.4 *TTR molecular dynamics simulations*

The starting structures were prepared for molecular dynamics (MD) simulations using the *in lucem* molecular mechanics (*il*mm) package, which was recently validated against a set of >3,100 experimental observables (Childers and Daggett, 2018). First, missing hydrogen atoms

were modeled on the crystal structure and then minimized for 500 steps. Next, all atoms were minimized via steepest descent minimization for 1000 steps. Next, the proteins were solvated in a water box that extended at least 10 Å beyond any protein atom and the box volume was adjusted to reproduce the experimental density at 310 K (0.992 g/ml) (Kell, 1967). Solvent atoms were minimized for 1000 steps, equilibrated for 500 steps, and then minimized again for 500 steps. Finally, all protein atoms were minimized for 500 additional steps. Production MD simulations were performed using *il*mm (Beck *et al.*, 2000) with the Levitt *et al.* force field (Levitt *et al.*, 1995), and the flexible three-center (F3C; Levitt *et al.*, 1997) water model. Simulations were performed using the microcanonical NVE (constant **n**umber of particles, **v**olume, and **e**nergy) ensemble with periodic boundary conditions, a 10 Å force-shifted non-bonded cutoff (Beck *et al.*, 2005), and a 2 fs timestep. Coordinates were saved every picosecond for analysis. The production simulations (n = 6) were performed at 310 K, acidic pH, and in triplicate for 500 nanoseconds (ns). We have previously reported on α-sheet formation in the WT simulations described here.

5.3.5 *Secondary structure propensity calculations*

As a simple way to calculate the intrinsic α-sheet propensity of amino acids, we analyzed the fraction of time that the three central residues in our AAXAA simulations adopted main-chain conformations consistent with three types of secondary structure: α-helix, β-sheet, and α-sheet. α-helix propensities were defined as the fraction of simulation time that the three central residues (AXA) adopted main-chain ϕ and ψ values within the α-helical region of Ramachandran space, α_R. β-sheet propensities were defined as the fraction of simulation time that the three central residues (AXA) adopted main-chain ϕ and ψ values within the β-sheet region of Ramachandran space, β. α-sheet propensities were defined as the fraction of simulation time that the three central residues (AXA) adopted main-chain ϕ and ψ values that alternated between the α_R and α_L regions of Ramachandran space, α_R and α_L.

5.4 Results

5.4.1 *Computationally derived insights into protein redesign*

α-sheet conversion in MD simulations In prior simulations of TTR, we found that three types of motion preceded α-sheet conversion: the formation of pleated main chain geometries, the loss of native hydrogen bonding patters, and a reorganization of side chain interaction networks. These changes were associated with the interaction of main chain carbonyl groups with water. To destabilize the CBEF sheet, a single mutation was introduced: F33A. This mutation disrupts the side chain interaction network in the CBEF sheet but does not result in a mutation at a highly conserved position in the TTR sequence (Figure 5.1). We hypothesized that a small disruption in the side chain interaction network at this site, which is highly stable in WT simulations, could sufficiently disrupt CBEF sheet dynamics and allow for α-sheet conversion.

Sequence conservation. Transthyretin-like proteins (TLPs) are found in diverse organisms from multiple kingdoms of life. Thus, sequence records contain a significant amount of information that reflects various evolutionary pressure(s) on the TLP fold. These evolutionary pressures could include a conservation of positions required for monomer folding, a conservation of positions required for tetramer assembly, and a conservation of positions that protect against aggregation. Based on the sequence entropy and sequence conservation (Figure 5.1), we identified several positions that were conserved and avoided making mutations at these positions. We assume that the high degree of conservation at these positions in our diverse set of TTR sequences indicates that these positions serve roles critical to the folding pathway of native structure (Figure 5.1).

Intrinsic α-sheet propensities We hypothesized that the amino acid identities at various positions in the TTR structure affect its ability to form α-sheet. To make wise design choices, we calculated the intrinsic α-sheet propensities of the 20 standard amino acids (Figure 5.1C).

5.4.2 *Design of TTR-um*

Using insights from MD simulations, the degree of conservation in homologous TTR sequences, intrinsic α-sheet propensities, and predictors of aggregation-prone regions, we engineered a TTR variant that is designed to form α-sheet in the CBEF sheet to a greater extent and/or on faster timescales than WT TTR. The TTR-um design contains a single mutation relative

to WT: F33A. Position 33 is located in the middle of strand B and the side chain faces the exterior of the monomer. In WT TTR, F33 stabilizes the solvent-exposed side chain – side chain interaction network in the CBEF sheet. We hypothesize that this residue significantly contributes to the stabilization of β-sheet structure in the CBEF sheet under amyloidogenic conditions. Ala was introduced at this position to eliminate the stabilizing contributions of F33 and to promote conversion to α-sheet secondary structure in the CBEF sheet (Figure 5.2).

5.4.3 *Dynamics of TTR-um*

The C_α RMSDs of the TTR variants were calculated after alignment to a set of 'core residues' (those in strands A, B, E, and G) to quantify the extent of deviation from the native monomeric conformation. C_α RMSDs were reported for residues 11-123 (the flexible N- and C-termini were excluded) and the first 20 ns of each simulation was excluded from the calculations. Overall, the TTR-um design resulted in a slight increase to the C_α RMSD relative to WT (3.8 Å vs 3.4 Å). Examination of the C_α RMSD as a function of residue number revealed the local conformational impact of the designed mutation. For TTR-WT (Figure 5.3), the regions with the highest C_α RMSD were the GH loop, strand H, the FG loop, and the CD loop. In TTR-um, the mutations led to an increase in the RMSD for residues in the BC loop and helix-EF loop (Figure 5.3).

5.4.4 *Altered conversion to α-sheet in a designed TTR variant*

In simulations of TTR (both of the WT sequence and structures harboring pathological mutations) we observed conversion from β-sheet to α-sheet secondary structure that was mediated by nonnative electrostatic interactions with carbonyl groups of the protein backbone. The primary objective of this study was to computationally test the proposed conversion mechanism through protein design. Mutations that destabilize the CBEF sheet and expose main chain peptide groups to solvent should promote α-sheet conversion in that sheet. To that end, the design objectives were supported by the conformations sampled during our simulations.

In WT TTR, residues in the DAGH sheet converted to α-sheet secondary structure in all three replicate simulations. The extent of conversion was lower in run 1 relative to runs 2 and 3, possibly due to the fact that strand H dissociated from the DAGH sheet prior to conversion in that

simulation (Figure 5.4). In runs 2 and 3 of the WT simulations, residues in strands D, A, G, and H all converted to α-sheet secondary structure. In simulations of TTR-um, the DAGH sheet was unaffected by the F33A mutation; in all three runs, residues in the DAGH sheet converted to α-sheet secondary structure (Figure 5.4).

In TTR-WT, residues in the CBEF sheet evaded conversion to α-sheet secondary structure. In prior studies, this was attributed to strong tertiary contacts formed between the side chain atoms of residues in neighboring strands that shielded main chain atoms. In contrast, the single mutation in the TTR-um design (F33A) altered the dynamics of the CBEF sheet and allowed for α-sheet to form. In runs 1 and 2, α-sheet formed in N-terminal residues of strand E and the C-terminal residues of strand F. In run 3, residues closest to the α-helix formed α-sheet like secondary structure in strands C, B, E, and F. While these residues did display a conformational hallmark of α-sheet – the alignment of carbonyl groups, the main chain geometry did not meet our strict geometric criteria for α-sheet as in each strand at least one of the residues in the α-sheet like conformation resided in the β-sheet region of Ramachandran space (Figure 5.5). Overall, TTR-um exhibited distinct conformational preferences relative to that of TTR-WT: a single mutation was able to modestly increase α-sheet conversion in the CBEF sheet (Figure 5.6).

In prior studies, we observed that fluctuating and transient side chain to side chain contacts were associated with the formation of pleated peptide plane geometries and exposed main chain peptide groups to solvent. We have proposed that these dynamics promote conversion to α-sheet structure. Thus, the mutations designed to interfere with α-sheet conversion should similarly alter these dynamics. In TTR-WT, changes to the side chain interaction network were observed among residues in the DAGH sheet. The side chain interaction network of DAGH sheet residues was unaffected relative to WT in the TTR-um design (Figure 5.7). In contrast, alterations to the side chain interaction network were observed in the CBEF sheet. In TTR-WT, several residues central to the CBEF sheet formed a stable network of contacts (H31, F33, K35, W41, K70, and E72). Mutation of one of these residues, F33A, was associated with a reorganization of this network. For example, in TTR-um, position 33 formed fewer interactions with positions 31, 41, 43, 70, and 72. As a response, there were increased contacts between positions 35 & 70 and 41 & 72 (Figure 5.8).

In TTR-WT, prior to conversion to α-sheet, main chain peptide groups formed pleated geometries and the extent of pleating was greater in the DAGH sheet (Figure 5.9) than in the CBEF

sheet (Figure 5.10). We have proposed that the formation of pleated geometries is critical to α-sheet conversion as pleated geometries correspond to reduced main chain hydrogen bonds and increased interactions between main chain carbonyl groups and either water molecules or polar side chain atoms. In TTR-um, the extent of pleating of the DAGH sheet residues was unaffected relative to WT (Figure 5.9). In contrast to the DAGH, the CBEF sheet of TTR has been 'designed' (by Nature) to form stable β-sheet conformations even when exposed to solvent. This was apparent in the TTR-WT simulations where less pleating of the main chain was observed (Figure 5.10). The F33A mutation in the TTR-um design was associated with an increase in the formation of pleated peptide groups in the CBEF sheet (Figure 5.10).

5.5 Discussion

5.5.1 *Probing mechanisms of amyloid formation through protein design*

One challenge in studying amyloid proteins from both computational and experimental perspectives is the degree of conformational heterogeneity associated with conformations sampled during amyloid formation. Protein design can circumvent some of these issues: altering the molecular properties of a system can facilitate efforts to isolate and probe specific conformations formed during amyloidogenesis. For example, Ricagno, Belloti, and Bolognesi have collaboratively designed variants of the protein involved in dialysis-related amyloidosis: β–2-microglobulin (β2M). Esposito *et al.* engineered stable (i.e. less aggregation-prone) β2M variations by substitution of Gly for Trp at either position 60 or 95 (Esposito *et al.*, 2008). Azinas *et al.* engineered a more stable (i.e. less aggregation-prone) β2M variant by introducing a Pro residue at position 53, which altered the conformational properties of an edge strand in β2M (Azinas *et al.*, 2011). Kihara *et al.* engineered Trp to Phe mutations at positions 60 and 95 in β2M, then introduced Trp residues at various positions to spectroscopically investigate β2M structural changes during fibril formation (Kihara *et al.*, 2006).

Amyloid formation of TTR has also been studied through protein design. Jiang *et al.* engineered a TTR variant (F87M/L110M) that is monomeric (a.k.a. monomeric TTR or M-TTR) and nonamyloidogenic at neutral pH, but forms amyloid under denaturing conditions (Jiang *et al.*, 2001). In a neutron diffraction study, the Mizuguchi group found evidence for a hydrogen bond network centered on His 88 that includes residues T75, W79, H88, S112, and P113 (Yokoyama *et*

al., 2012). Later, the same group introduced several mutations at position 88 (H88R, H88A, H88F, H88Y, and H88S) to alter the hydrogen bonding network (Wei *et al.*, 2017). H88R is a clinically reported mutation that leads to deposition of TTR amyloid in cardiac tissue. In that study, the mutations to H88 had diverse effects on the stability of TTR monomers and the extent of amyloid formation by TTR. H88R and H88S led to less stable monomers and 300-400% increase in the extent of amyloid formed; the H88F and H88Y mutations led to a 50% decrease in the extent of amyloid formed; and the H88A mutation did not significantly alter the extent of amyloid formed relative to wild type (WT). Analysis of the corresponding experimental structures showed that mutations that affect the hydrogen bond network around H88, which is located in the EF loop, resulted in altered aggregation properties. The Foguel group engineered an amyloid-prone variant of TTR via a Lys to Leu mutation at position 35 (Sant'Anna *et al.*, 2014). The authors of that study found evidence that K35 acts as a gatekeeper residue that protects against aggregation, possibly via electrostatic repulsion between Lys side chains between two monomeric units during aggregation. Thus, engineered variants of amyloid proteins can be used to map conformational changes during amyloidogenesis, probe the effects of specific residues and structural features on aggregation, and lead to insights for the development of therapeutic and diagnostic technologies for the treatment of amyloidosis.

Here, we have further demonstrated that computational protein design can be used to test proposed mechanisms of amyloid formation. We have proposed that α-sheet secondary structure is a critical conformation for the aggregation of TTR and other amyloids and a defining feature of soluble oligomeric states. An MD investigation of conversion to α-sheet resolved the molecular details of this conversion at the atomic level. Specifically, formation of α-sheet can be catalyzed by electrostatic interactions with main chain carbonyl groups, and this process becomes favorable when native β-sheet conformations are destabilized. The revelation of this mechanism, made possible through the unique strengths of MD, subsequently enabled a computational test of that mechanism through protein design. This computational test was successful: a TTR variant was engineered to have an increased ability to convert to α-sheet secondary structure. In the design, a slight disruption to the side chain interaction network of the CBEF sheet was met with an increase in the amount of α-sheet formed. The success of this design supports our proposed mechanism of α-sheet formation in TTR, which may also extend to other related amyloid proteins with β-sheet sandwich topologies (e.g. immunoglobin light chain, β2-microglobulin). The success of this design

also provides a hypothesis that can be directly tested experimentally. Since the TTR-um design promoted the formation of α-sheet, we hypothesize that it would aggregate at a faster rate relative to WT. The extent of fibril formation by TTR-um can be directly tested through aggregation assays that monitor the extent of fibrils formed through optical density or thioflavin T fluorescence.

5.5.2 *Assessing the impact of mutations on amyloid formation by transthyretin*

Numerous mutations to TTR that are associated with amyloid formation have been reported in the literature. However, the pathological significance of the majority of these mutations has not been established as the WT TTR sequence also forms amyloid. This problem becomes more complex in light of the diverse ways in which mutations could affect amyloid formation by TTR. First, mutations can interfere with the equilibrium distribution of TTR tetramers, dimers, and monomers in a pH-dependent fashion. Second, mutations can interfere with the mechanism or rate of monomer unfolding that occurs before amyloid assembly. Third, mutations can either stabilize or destabilize oligomeric and fibrillar conformations sampled during amyloid formation. We hypothesize that the mutation described here (F33A) would most likely interfere with the mechanism or rate of monomer unfolding and misfolding. In future studies, experimental investigation of amyloid formation by designed variants could represent one step forward to a delineation of the effects of arbitrary mutations on amyloid formation by TTR, an important goal to meet for optimal therapeutic interference with transthyretin amyloidosis.

Four pathological mutations at position 33 have been reported in the literature: F33I, F33L, F33C, and F33V. In these clinical reports, the various mutations at position 33 resulted in diverse symptoms. The first reported mutation at position 33 (F33I) arrived in 1984 with independent reports in two patients. Later, the F33L mutation was reported in a patient with carpal tunnel syndrome, peripheral neuropathy, mild cardiomyopathy, and mild GI issues. This was the first instance in which TTR-amyloidosis presented with symptomatic ascites, which occur in 10-20% of patients with AL amyloidosis (Myers *et al.*, 1998). Patients with the F33V and F33C variants presented with cardiac and renal involvement. Thus, seemingly any mutation at this position may be associated with amyloid formation, but the specific symptoms involved vary on either a patient or mutation-specific basis. This suggests that, as predicted, position 33 is critical for the structural stability of TTR monomers.

In contrast to the abundance of pathological mutations to TTR that have been reported in the literature, only two have been found that stabilize TTR, and both of these mutations exert their effects via stabilization of the TTR tetramer. The protective T119M mutation was first identified in a patient with a TTR sequence containing both the T119M mutation as well as the highly amyloidogenic V30M mutation (Hammarström *et al.*, 2001). Position 119 is located in the middle of strand H and participates in the formation of the monomer:monomer and dimer:dimer interfaces. The T119M mutation suppresses amyloid formation via kinetic stabilization of the TTR tetramer. This mechanism of stabilization can occur even in the presence of other, destabilizing mutations. T119M protects against amyloidogenesis by stabilizing the tetrameric state, the dissociation of which is the initial and rate limiting step in TTR amyloid formation (Kim *et al.*, 2016). The T119M mutation causes tetramers to dissociate 40 times slower and reassemble between 90 and 200 times faster relative to the WT protein (Hammarström *et al.*, 2002). The protective R104H mutation was first identified in family of Japanese patients that with TTR sequences containing both the R104H mutation as well as the V30M mutation (Terazaki *et al.*, 1999). Position 104 is located at the N-terminal end of strand G. The R104H mechanism is proposed to suppress amyloid formation via thermodynamic stabilization of tetrameric TTR (Almeida *et al.*, 2000; Sekijima *et al.*, 2006).

5.6 Conclusions

Transthyretin is one of the most extensively studied amyloid proteins. Age-related cardiac amyloidosis was first reported by Soyka in 1876; since that time, much progress has been made in understanding the incidence and prevalence of TTR amyloidosis, the biochemical conditions that promote amyloid formation, and the sequence of molecular events that result in amyloid fibrils. However, many questions remain. Notably, the structures of TTR soluble oligomers and amyloid fibrils have not been solved and the pathological significance of most TTR mutations has not been established. Here, we have demonstrated the utility of MD simulations to uncover and test mechanisms that drive amyloid formation at the atomic level. Future studies that capitalize on these insights and experimentally test our proposed mechanisms should make even greater strides towards a road map of amyloid formation by transthyretin.

5.7 Figures

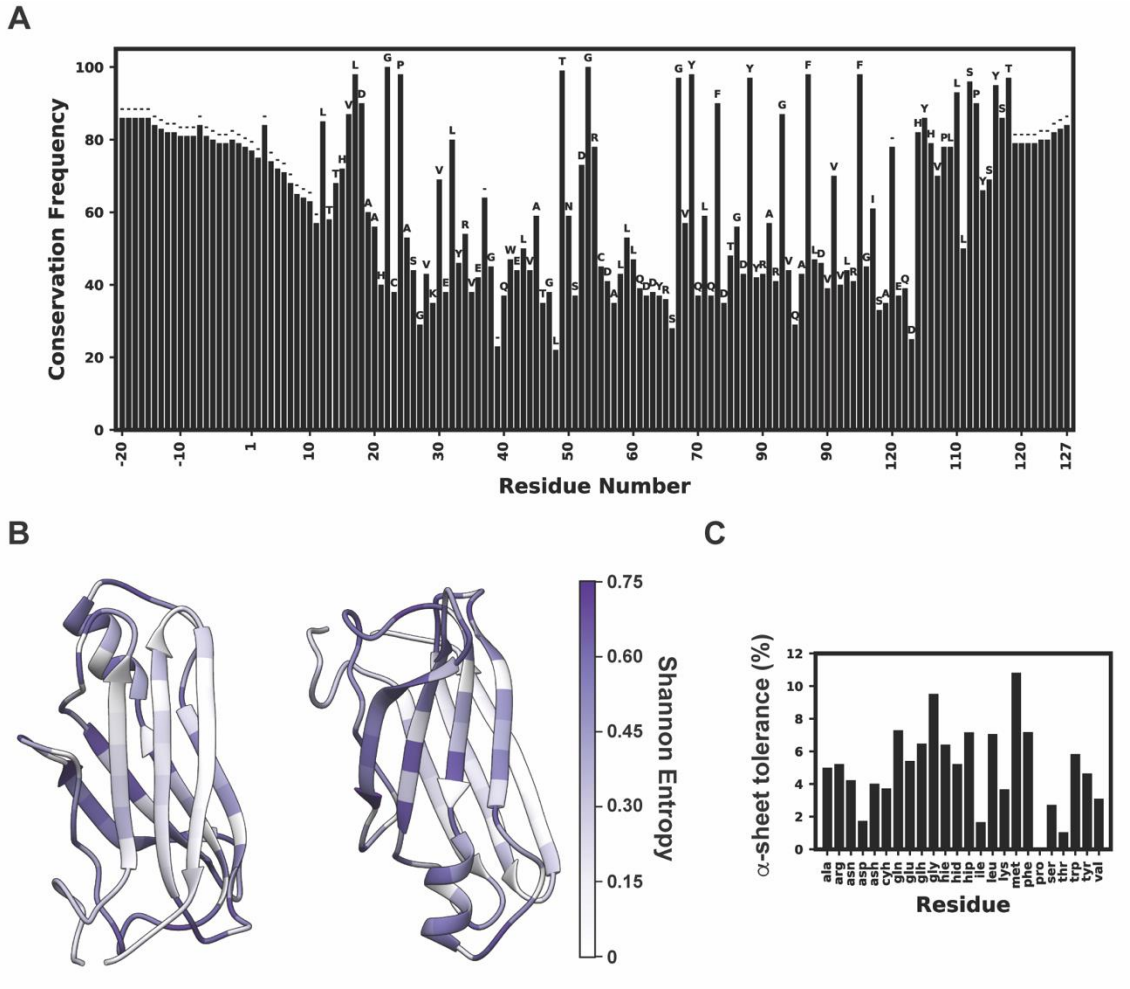


Figure 5.1. MSA Sequence Entropy and Conservation

(A) The Shannon entropy as measured in the MSA alignment of TTR sequences is mapped onto the wild type TTR structure from *H. sapiens.* Darker colors indicate higher entropy (i.e. greater variation in the MSA at a given position). (B) The frequency of the most commonly observed amino acid at each position in the wild type TTR structure from *H. sapiens*. The most commonly observed amino acid, or gap ('-'), is annotated at each position. Positions of high conservation in (B) correlate with positions of low entropy in (A). (C) The intrinsic α-sheet propensities of residues in our model pentapeptide systems is shown. The values are the sum of the left-right-left and right-left-right helical patterns associated with α-sheet.

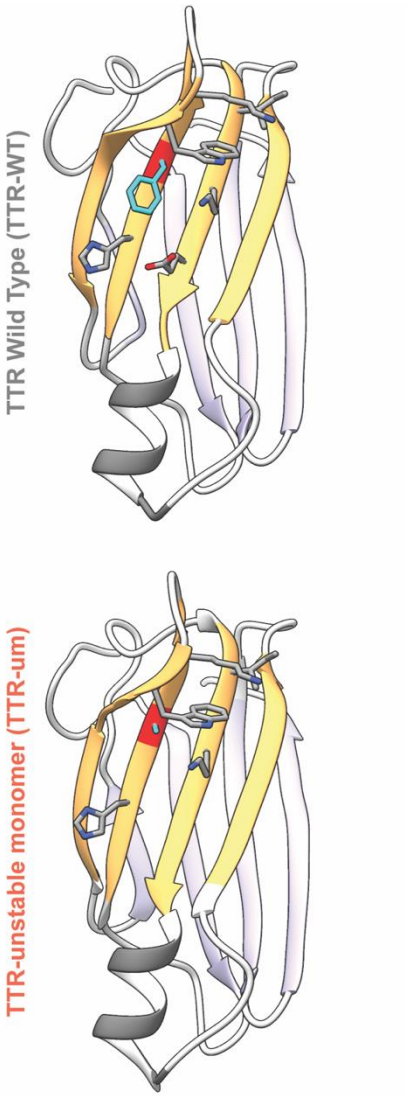


Figure 5.2. Starting structures of wild type TTR and TTR-um

The DAGH sheet is colored purple and the CBEF sheet is colored orange. The ribbons of positions of the designed mutation is colored red. Carbon atoms in the side chains of residues that were mutated in this study are colored cyan. Residues that neighbor position 33 are shown as sticks with grey carbon atoms.

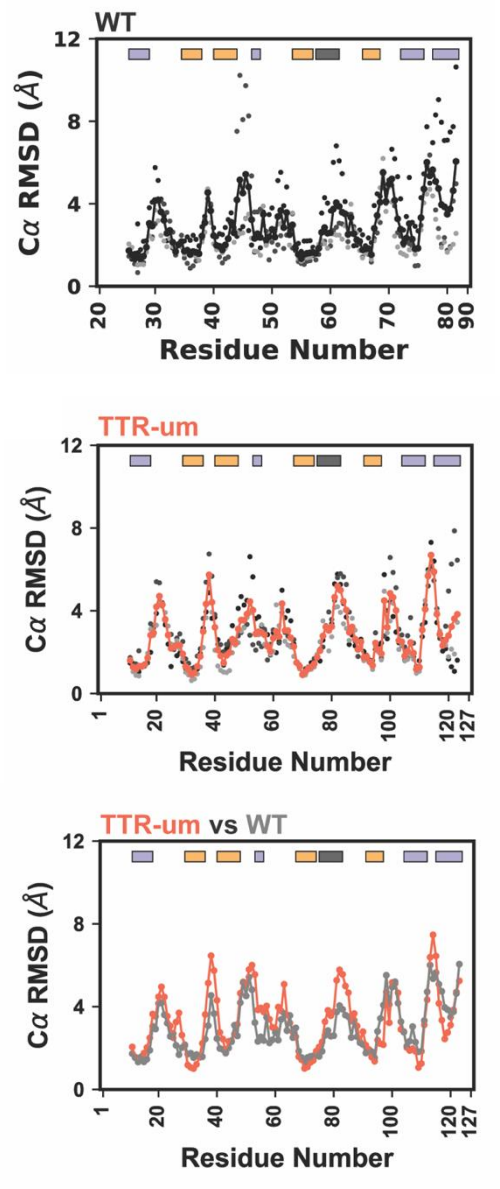


Figure 5.3. Cα RMSD as a function of time and residue number

The Cα RMSD of the TTR residues after alignment to strands A, G, B, and E is shown for runs 1 (black), 2 (grey), and 3 (light grey) is shown as a function residue number. The purple, orange, and dark grey bars denote the locations of secondary structure elements in TTR. Average results for TTR (top, black lines) and TTR-um (middle, orange lines) show trends in the RMSD data and are compared against one another in the lower plot.

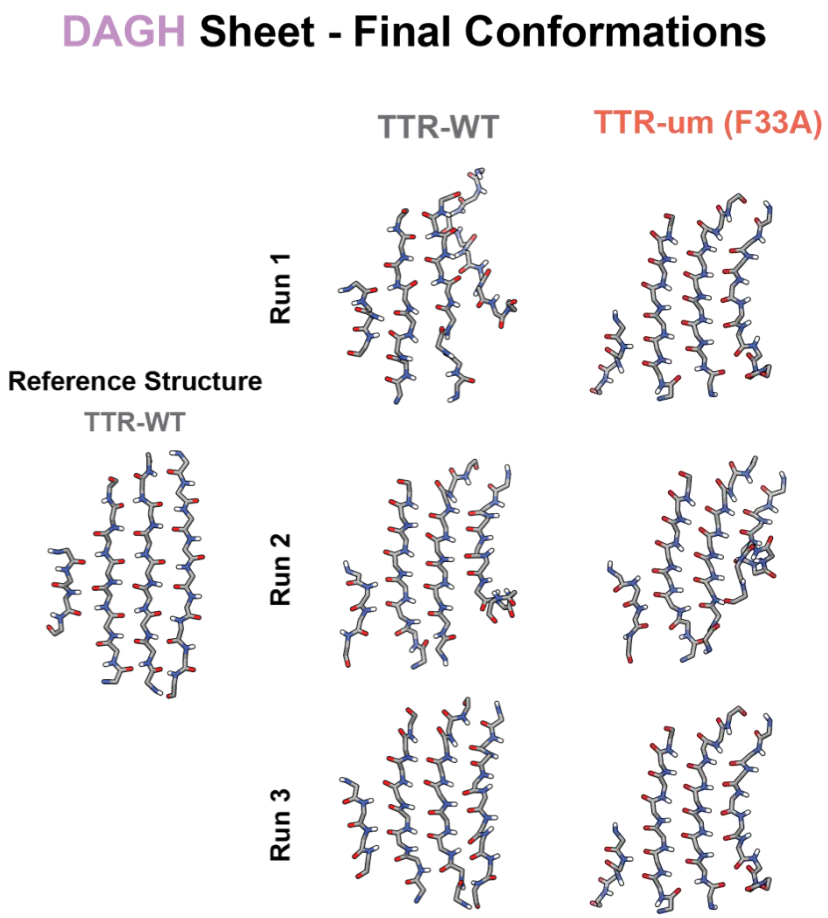


Figure 5.4. DAGH Conformations of α-sheets formed in TTR variants.

Snapshots of the main chain atoms for residues in the DAGH sheet are shown at the end of the replicate WT and TTR-um simulations. The reference starting structure corresponds to the crystallographic conformation (PDB: 1TTA) with modeled hydrogens added.

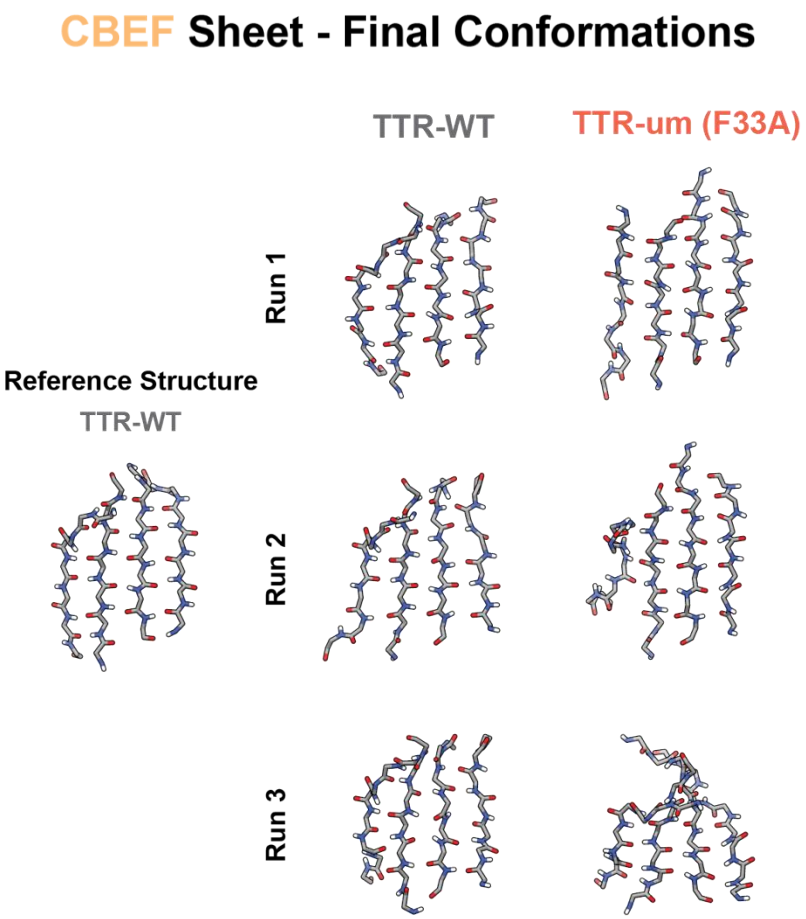


Figure 5.5. CBEF Conformations of α-sheets formed in TTR variants.

Snapshots of the main chain atoms for residues in the CBEF sheet are shown at the end of the replicate WT and TTR-um simulations. The reference starting structure corresponds to the crystallographic conformation (PDB: 1TTA) with modeled hydrogens added.

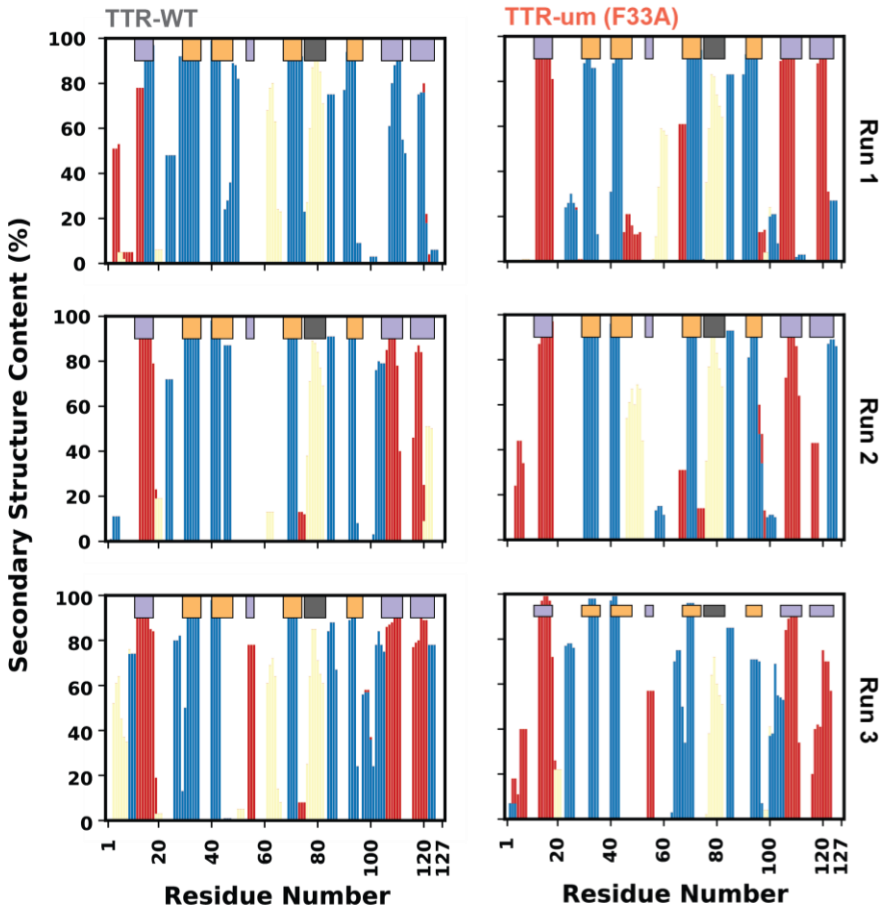


Figure 5.6. Secondary structure as a function of residue number.

The average β-sheet (blue), α-sheet (red), and α-helix (yellow) secondary structure content for individual simulations is plotted as a function of residue number over the final 25 ns of each simulation. The purple, orange, and dark grey bars denote the locations of secondary structure elements in TTR.

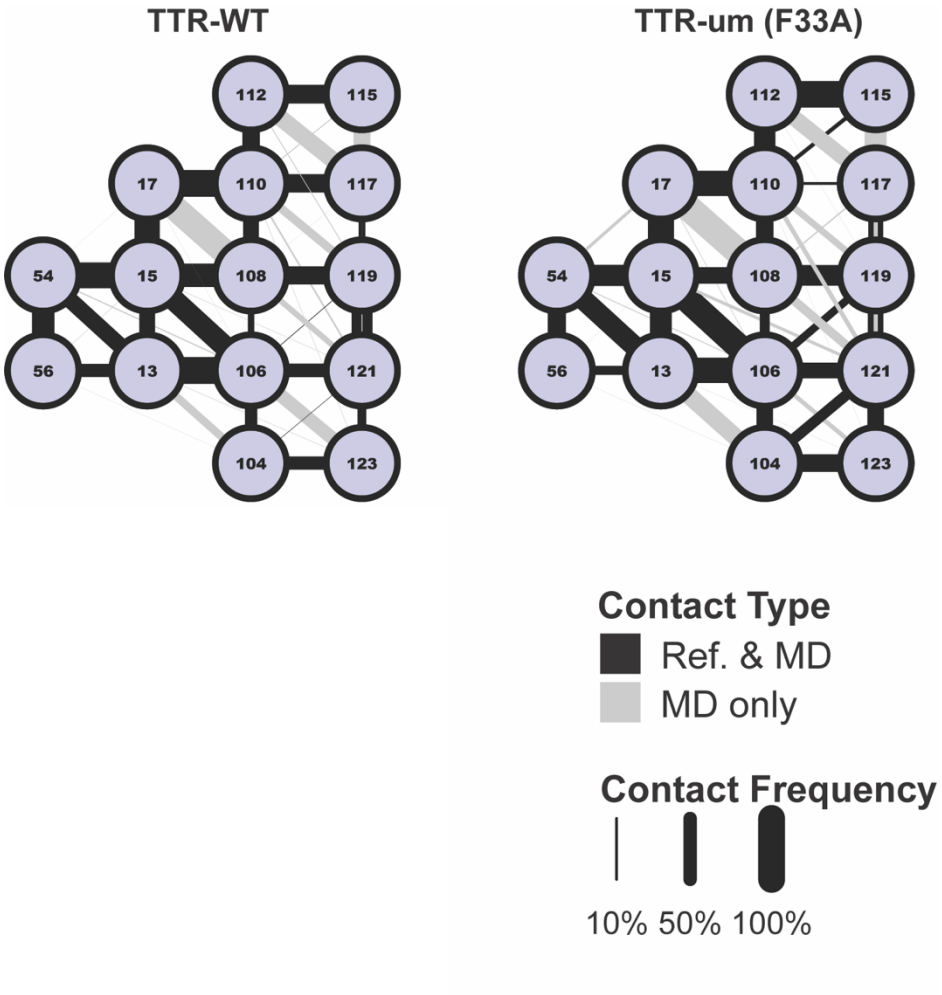


Figure 5.7. DAGH sheet solvent-exposed side chain – side chain interaction networks.
The solvent-exposed side chain interaction network composed of residues in the DAGH sheet is represented as a graph. Each residue is a node and edges indicate that the side chains of the connected residues were in contact during the simulation. The width of the edges is proportional to the percentage of simulation time that the contact was present for. Reference state contacts (i.e. those present in the WT minimized starting structure) are colored black edges and contacts only observed during MD are colored grey.

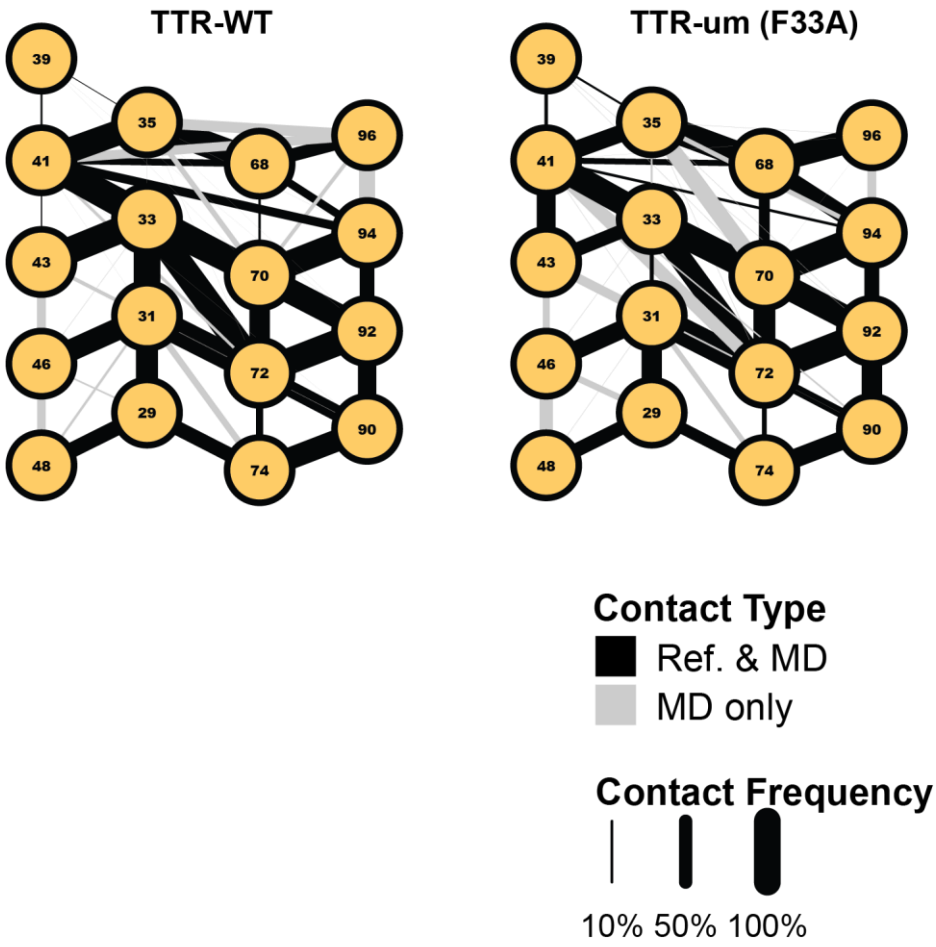


Figure 5.8. CBEF sheet solvent-exposed side chain – sidechain interaction networks.

The solvent-exposed side chain interaction network composed of residues in the CBEF sheet is represented as a graph. Each residue is a node and edges indicate that the side chains of the connected residues were in contact during the simulation. The width of the edges is proportional to the percentage of simulation time that the contact was present for. Reference state contacts (i.e. those present in the WT minimized starting structure) are colored black edges and contacts only observed during MD are colored grey.

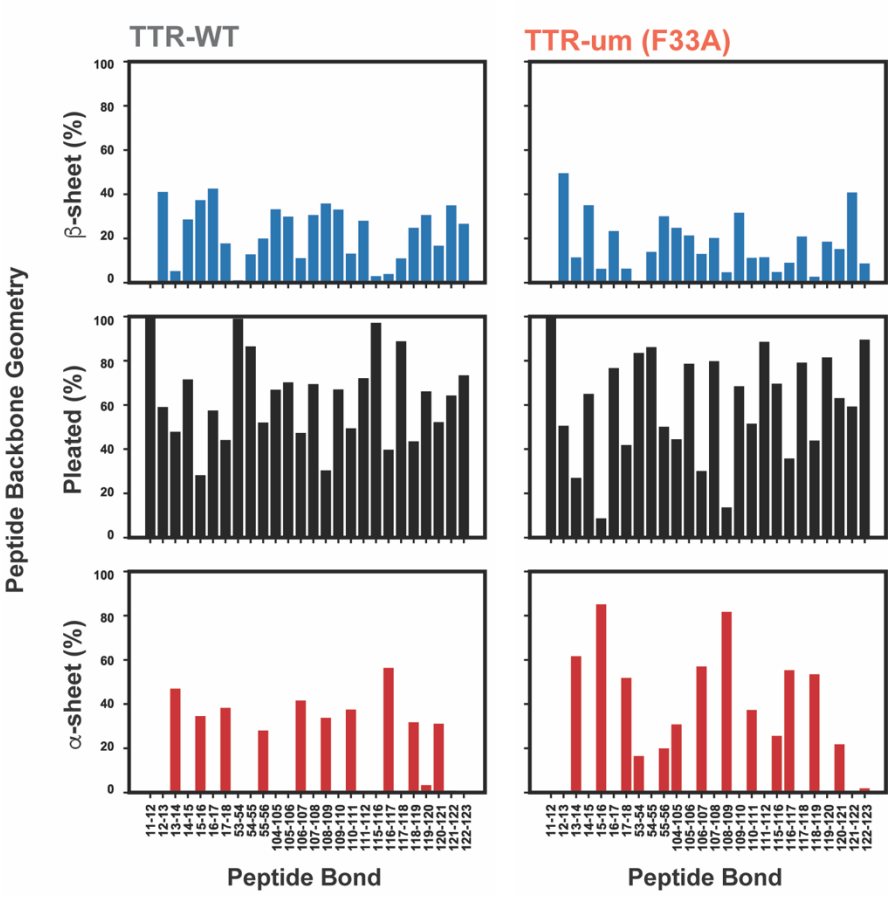


Figure 5.9. DAGH Formation of pleated main-chain conformations in TTR.

The population of β-sheet like (top, blue bars), pleated (middle, black bars), and α-sheet like (bottom, red bars) main chain geometries in the DAGH sheet of WT and TTR-um are shown as a function of peptide bond number.

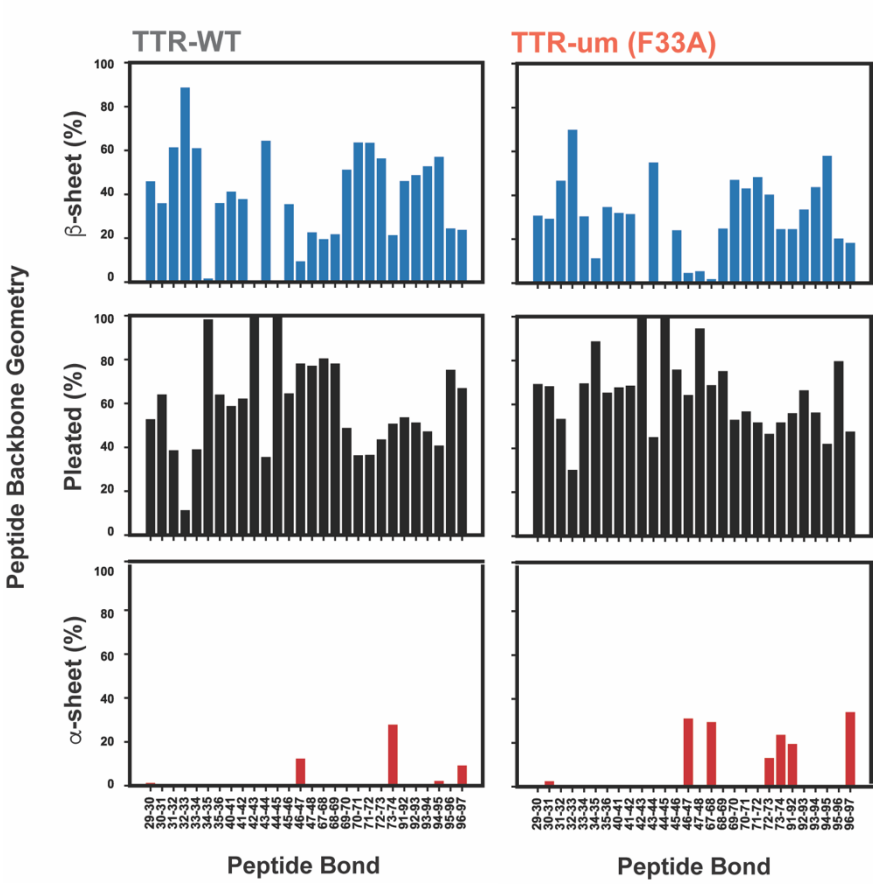


Figure 5.10. CBEF Formation of pleated main-chain conformations in TTR.

The population of β-sheet like (top, blue bars), pleated (middle, black bars), and α-sheet like (bottom, red bars) main chain geometries in the CBEF sheet of WT and TTR-um are shown as a function of peptide bond number.

Chapter 6. THE EFFECT OF CHIRALITY AND STERIC HINDRANCE ON INTRINSIC BACKBONE CONFORMATIONAL PROPENSITIES: TOOLS FOR PROTEIN DESIN

6.1 Summary

The conformational propensities of amino acids are an amalgamation of sequence effects, environmental effects and underlying intrinsic behavior. Many have attempted to investigate neighboring residue effects to aid in our understanding of protein folding and improve structure prediction efforts, especially with respect to difficult to characterize states, such as disordered or unfolded states. Host–guest peptide series are a useful tool in examining the propensities of the amino acids free from the surrounding protein structure. Here, we compare the distributions of the backbone dihedral angles (φ/ψ) of the 20 proteogenic amino acids in two different sequence contexts using the AAXAA and GGXGG host–guest pentapeptide series. We further examine their intrinsic behaviors across three environmental contexts: water at 298 K, water at 498 K, and 8 M urea at 298 K. The GGXGG systems provide the intrinsic amino acid propensities devoid of any conformational context. The alanine residues in the AAXAA series enforce backbone chirality, thereby providing a model of the intrinsic behavior of amino acids in a protein chain. Our results show modest differences in φ/ψ distributions due to the steric constraints of the Ala side chains, the magnitudes of which are dependent on the denaturing conditions. One of the strongest factors modulating φ/ψ distributions was the protonation of titratable side chains, and the largest differences observed were in the amino acid propensities for the rarely sampled α_L region.

6.2 Introduction

Amino acid conformational preferences can aid our understanding of protein conformational ensembles in the unfolded and denatured states of globular proteins (Dill and Shortle, 1991; Gillespie and Shortle, 1997; Voelz *et al.*, 2010; Meng *et al.*, 2013), the 'native' states of intrinsically disordered proteins (Fawzi *et al.*, 2005; Uversky, 2013), and in the misfolded ensembles that precede protein aggregation (Fawzi *et al.*, 2005; Bowler, 2012; Uversky, 2013). Secondary and tertiary structural interactions stabilize the ensemble of conformations present in

the native, folded state. However, the intrinsic preferences that underlie the conformational preferences of nonnative states are less well understood. In addition, our understanding of the link between an amino acid sequence and the fold topology that sequence will assume is limited (Sánchez *et al.*, 2006). This sensitivity to the amino acid sequence is illustrated by the designed heteromorphic proteins GA98 and GB98, which have all-α and α/β folds, respectively, but whose sequences differ by only a single residue (He *et al.*, 2012). Conformational propensities of amino acids depend on sequence, intrinsic behavior, and structural and chemical environment; hence, parsing the specific effects of different contexts on amino acid propensities remains a great challenge. In order to understand the dependence of amino acid conformational propensities on sequence, we must first isolate the intrinsic propensities. As many unfolding experiments are performed at high temperature or in chaotropic solvents (Gattin *et al.*, 2009; Ghosh *et al.*, 2013; Silva-Lucca *et al.*, 2013; Tischer and Auton, 2013; Roy and Bagchi, 2014; Ahmad *et al.*, 2015), we must also determine how environmental conditions modulate those propensities. A strategic approach that examines intrinsic behavior in different host systems can provide a more detailed understanding of the factors that modulate amino acid propensities, in turn providing a better understanding of protein folding, unfolding, nonfolding, and misfolding.

The sterically accessible φ/ψ regions of polypeptide chains were first estimated using simple definitions of van der Waals contact distances and bond angles and were validated by the conformations of protein structures known at that time (Ramachandran *et al.*, 1963). Since these original studies, the definitions of sterically accessible regions in φ/ψ space now include infrequently sampled conformations (Porter and Rose, 2011; Zhou *et al.*, 2011). Polymer physics models of amino acid intrinsic propensities predict that residues in the 'random coil' state sample φ/ψ space without preference for any conformational region and independently of their nearest neighbors (Flory, 1969; O'Connell *et al.*, 1999; Pappu *et al.*, 2000). In these models, the random coil state of amino acids refers to sampling where the amino acids have complete access to all sterically allowed conformations, do not participate in secondary or tertiary interactions, and have solvent exposed side chains (Toal and Schweitzer-Stenner, 2014). However, experimental and theoretical investigations demonstrate that the 20 proteogenic amino acids possess intrinsic preferences for conformational regions in φ/ψ space (Avbelj *et al.*, 2006; Beck *et al.*, 2008; Schweitzer-Stenner, 2009; Mirtič *et al.*, 2014; Schweitzer-Stenner and Toal, 2014; Toal and Schweitzer-Stenner, 2014; Towse *et al.*, 2014, 2016b).

Two major strategies are used to model the intrinsic conformational preferences of amino acids: coil libraries and small, unstructured peptides. Coil libraries excise coil regions, simply defined as sequences adopting neither α-helical nor β-sheet structure, from X-ray crystal structures to define intrinsic conformational preferences (Fitzkee *et al.*, 2005; Jha *et al.*, 2005; Tamiola *et al.*, 2010; Jiang *et al.*, 2013). However, such libraries determine φ/ψ preferences from short sequences whose conformations are biased by the structural context provided by the surrounding protein structure and, therefore, such libraries do not represent intrinsic propensities (Jiang *et al.*, 2013). Host–guest peptide series using a small number of residues circumvent these issues by preventing secondary or tertiary interactions from biasing conformational preferences. However, enough residues need to be present to adequately model the protein chain. For example, intrinsic conformational preferences obtained from dipeptide models account for only a single neighboring residue and do not reflect the intrinsic behavior of amino acids within a polypeptide chain (Oh *et al.*, 2012a-b). In addition, although the commonly used host–guest GGXGG pentapeptides are of sufficient length to provide the intrinsic propensities in the context of a chain of amino acids while allowing the greatest conformational freedom, the simple Gly-host lacks the backbone chirality present for the majority of residues in a true polypeptide chain (Avbelj *et al.*, 2006; Beck *et al.*, 2008; Vymětal *et al.*, 2013; Schweitzer-Stenner and Toal, 2014). Thus, while the GGXGG series is a good model for amino acid intrinsic propensities in the simplest polypeptide chains, the Ala-host residues in the AAXAA series provide a more realistic model of the intrinsic propensities of an amino acid within the simplest of chiral L-protein sequences. However, there has been concern that Ala induces artificially high levels of helix both in experiments and simulations. While Ala is frequently found in α-helices, its ability to induce helical content is reduced drastically in small peptides. For example, Firestine *et al*. demonstrated this in the KKA_nKKGY model system, where significant helical content was observed only when $n \geq 9$ (Firestine *et al.*, 2008). With respect to simulations, many molecular dynamics (MD) force fields induce excessive helicity. Experimental estimates are ~20% for Ala_3-, Ala_4- and Ala_5-based peptides (Firestine *et al.*, 2008; Jiang *et al.*, 2013); other common force fields produce high helical contents for such systems except for GROMOS: 13.1% (GROMOS53a6), 57.5% (CHARMM27 with CMAP), 62.3% (AMBER03), 94.2% (AMBER99), and 97.6% (AMBER94) (Best *et al.*, 2008). The average helix content for our blocked AAAAA peptide simulations is 19.4% (Towse *et al.*, 2016b), which is in very good agreement with experiment.

Challenges in modeling the intrinsic conformational propensities of amino acids are echoed in the computational design of loops and other flexible regions in proteins (Hu *et al.*, 2007). Difficulties in modeling loop regions may be attributed to several sources. First, these regions may adopt multiple conformational regions that interconvert on timescales from nanoseconds to milliseconds (Benson and Daggett, 2008; Gu *et al.*, 2015). Second, regular patterns of secondary structure do not stabilize loop structure; instead, interactions between the loop, solvent, and the surrounding protein environment all contribute to the structure and dynamics of loops (Zimmermann and Jernigan, 2012; Papaleo *et al.*, 2016). Third, many computational design strategies rely on an initial target conformation obtained from the PDB. The crystal environment necessary to solve protein structures via X-ray crystallography frequently bias loop structures and mask the true dynamic conformations(s) of the loop in solution. This results in a mismatch between the template for design and the true structure(s) (Fiser *et al.*, 2000; Sellers *et al.*, 2008; Messih *et al.*, 2015). Finally, the conformational ensemble and dynamics of loop regions are more sensitive to mutation; thus, even highly similar sequences may possess distinct conformations and dynamics, rendering the use of templates in loop design less useful (Ceruso *et al.*, 2003; Papaleo *et al.*, 2016). In such cases, the intrinsic conformational preferences of amino acids play a greater role in determining loop dynamics. As flexible loops are intimately connected to the dynamics and functions of proteins, correct modeling of flexible loops and linkers is crucial to improving the accuracy and sophistication of computationally designed proteins. Knowledge of amino acid conformational propensities and the sensitivity of those propensities to environment and backbone chirality should aid in the computational design of flexible regions in proteins.

Our two independent studies of the GGXGG and AAXAA host–guest series reveal a dependence on structural and environmental contexts on amino acid intrinsic propensities (Beck *et al.*, 2008; Towse *et al.*, 2016b). Not only are the intrinsic propensities different from the amino acid propensities within folded proteins (Beck *et al.*, 2008), but the intrinsic propensities also show sensitivity to environment (Towse *et al.*, 2016b). These initial studies were experimentally validated and demonstrate the power of MD simulations to provide atomistic detail of heterogeneous conformational ensembles (Beck *et al.*, 2008; Towse *et al.*, 2016b). Here, we performed a comparative study of the φ/ψ propensities of the guest 'X' residues in the GGXGG and AAXAA host–guest series. For this study, we extended our earlier GGXGG simulations (Beck *et al.*, 2008) by performing longer simulations and performing simulations in 8 M urea to match

conditions in our recent AAXAA study (Towse *et al.*, 2016b), allowing us to determine the environmental sensitivity of the GGXGG propensities. By comparing these new and more extensive GGXGG simulations with the AAXAA simulations, we have determined the effects of two factors on intrinsic propensities: backbone chirality and simple sterics provided by C_β methyl groups.

6.3 Methods

6.3.1 *MD simulations of host–guest pentapeptides*

GGXGG pentapeptide simulations were previously performed for 100 ns in water at 298 K. However, both the alternative protonation states of titratable amino acids and the side-chain chirality of Thr were neglected in this earlier study. The Thr residue previously studied was the rarer allo-form with chirality inverted at the C_β position (Beck *et al.*, 2008). Here, we extended the existing GGXGG simulations to be consistent in length with the AAXAA simulations, which required longer times to converge, and simulated new models for the additional protonation states and Thr. We also simulated GGXGG series peptides under thermally denaturing and chemically denaturing conditions. End-capped (N-acetylated, C-amidated) GGXGG peptides containing all 20 amino acids were built with extended conformations (φ/ψ angles set to 180° and −180°, respectively) and simulated under three conditions: water at 298 K, 8 M urea at 298 K, and water at 498 K. Where necessary, we generated additional pentapeptides for both neutral and acidic protonation states (Asp, Ash, Glu, Glh). Three individual simulations of His were performed for each possible protonation state: Hid (δH), Hie (εH), and Hip (both δH and εH protonated). Cysteine was modeled in the reduced state (-CH_2-SH, denoted Cyh). The simulation and analysis procedures used for the AAXAA series (Towse *et al.*, 2016b) were also used here to ensure a direct comparison of φ/ψ sampling between the AAXAA and GGXGG peptides.

Simulations were performed using the *in lucem* molecular mechanics (*il*mm) package (Beck *et al*. 2000–2016) with the Levitt *et al*. force field (Levitt *et al.*, 1995), the microcanonical NVE (constant number of particles, volume and energy) ensemble, and the flexible three-center (F3C) water model (Levitt *et al.*, 1997). Nonbonded interactions were treated with an 8-Å force-shifted cutoff (Beck *et al.*, 2005), and explicit solvent molecules were used for both 8 M urea (Zou *et al.*, 2002; Day and Daggett, 2005) and water (Levitt *et al.*, 1997) simulations. The F3C water

model is fully flexible and lacks a fictitious H–H bond in contrast to other commonly used models, resulting in better agreement with experiment for both the structural and dynamic properties of water (Levitt *et al.*, 1997). To obtain 8 M urea solvent boxes, water molecules in a pre-solvated peptide system were randomly substituted with urea molecules. Both water and 8 M urea solvent systems at 298 K were simulated with a box size that reproduced the experimental densities, 0.9970 g/mL and 0.7813 g/mL, respectively. For the simulations at 498 K, the density was set to the reduced density for that temperature, 0.829 g/mL (Kell, 1967).

To assess convergence, multiple simulations were performed of the GGAGG, GGGGG, and GGWGG peptides in 8 M urea and of the GGAGG peptides in water at 298 and 498 K. All simulations were performed for a minimum of time consistent with requirements to reach convergence as determined for AAXAA (Towse *et al.*, 2016b): 1 μs in 8M urea (1.5 μs for Ala, Tyr and Gly); 600 ns in water at 298 K; and 100 ns in water at 498 K (200 ns for Ala). The total simulation time for both GGXGG and AAXAA was 106 μs.

6.3.2 *Calculation of conformational propensities*

Populations were calculated for the four quadrants of the conformational φ/ψ space and for specific conformational regions, defined as α_R: $-100° \leq \varphi \leq -30°$, $-80° \leq \psi \leq -5°$; near-α_R: $-175° \leq \varphi \leq -100°$, $-55° \leq \psi \leq -5°$; α_L: $5° \leq \varphi \leq 75°$, $25° \leq \psi \leq 120°$; β: $-180° \leq \varphi \leq -50°$, $80° \leq \psi \leq -170°$; P_{IIL}: $-110° \leq \varphi \leq -50°$, $120° \leq \psi \leq 180°$; and P_{IR}: $-180° \leq \varphi \leq -115°$, $50° \leq \psi \leq 100°$. An additional β region, named non-polyproline β (nPβ), was defined as the area of the β region that does not overlap with either P_{IIL} or P_{IR}. Populations were calculated as percentages to account for the different trajectory lengths. φ/ψ frequency distributions were generated using two-dimensional histograms with 5° × 5° bins, and correlations between the distributions were taken to form similar correlation matrices.

6.3.3 *Comparison with NMR*

The various parameterizations and simulation methodologies present in modern force fields and simulation packages can affect the distributions of conformational propensities (Vymětal *et al.*, 2013). Consequently, we strive to compare our simulations with experiment whenever possible. Experimental validation of the AAXAA simulations has been reported using experimental chemical shifts acquired in 8 M urea at 298 K, pH 2.5, for the AAXAA series. Here, we assessed

the experimental agreement of the GGXGG simulations using experimental data for the GGXGG series also acquired in 8 M urea at 298 K, pH 2.5 (Schwarzinger *et al.*, 2000). Experimental GGXGG chemical shifts were obtained from the Biological Magnetic Resonance database (BMRB code: 4747) (Schwarzinger *et al.*, 2000; Ulrich *et al.*, 2008). A random 1% selection of structures (5000 structures) from the production dynamics portion of the GGXGG trajectories, previously shown to be representative of the full ensemble of conformations generated (Towse *et al.*, 2016b), were used to calculate the ^{1}H, ^{15}N and ^{13}C chemical shifts using SHIFTX2 (Han *et al.*, 2011). As no homology models exist for these pentapeptide systems, only the SHIFTX+ component of SHIFTX2 was used to calculate the chemical shifts.

6.4 Results

We previously reported the intrinsic conformational preferences of the 20 naturally occurring amino acids within the GGXGG host under native conditions (water at 298 K) for 100 ns (Beck *et al.*, 2008) and the AAXAA host under native (water at 298 K) and denaturing conditions (water at 498 K and 8 M urea at 298 K) (Towse *et al.*, 2016b). Here, we report the effect of chirality and simple steric effects of the neighboring Ala C_β methyl groups on the conformational sampling of guest residues in native (water at 298 K) and denaturing (8 M urea at 298 K and water at 498 K) conditions. To draw direct comparisons between our GGXGG and AAXAA pentapeptide systems (Beck *et al.*, 2008; Towse *et al.*, 2016b), we extended the GGXGG trajectories under native conditions from 100 ns to 600 ns and performed simulations of the GGXGG peptides under denaturing conditions. Additionally, we have performed simulations of the protonated states of glutamate (Glh), aspartate (Ash), and three protonated forms of histidine (the neutral Hid and Hie states and the diprotonated, positively charged Hip) in the Gly-host.

6.4.1 *Equilibrium sampling of conformational space*

To assess the intrinsic conformational sampling of the guest amino acids, MD simulations of the GGXGG host–guest pentapeptides, where X is any of the 20 proteogenic amino acids, were performed under native, control (water, 298 K), thermally denaturing (water, 498 K), and chemically denaturing (8 M urea, 298 K) conditions. To obtain meaningful φ/ψ statistics, we confirmed that our simulations reached an equilibrium distribution with stable population

frequencies to ensure Boltzmann sampling had been achieved. We assumed that each pair of φ/ψ dihedral angles of the three central residues can occupy four possible states that correspond to the four quadrants of Ramachandran space, which results in a total of 64 possible conformations of the three central residues. We compared the fraction of the ensemble spent in each of the four quadrants of Ramachandran space across different portions of the trajectory and between independent replicate simulations. All simulations converged with respect to the sampling of Ramachandran space (Beck *et al.*, 2008; Towse *et al.*, 2016b; Figures S6.1 and S6.2). Convergence was monitored between replicate simulations of GGAGG, GGGGG, and GGWGG as well as across different trajectory windows for all GGXGG peptide simulations (Figure 6.1, Table A.6.1). The production dynamics portion of the trajectories, after the point of convergence, was used for analysis and to compare the intrinsic sampling of the central guest residues in the Gly- and Ala-based systems across native, thermally denaturing, and chemically denaturing conditions. Comparison of the extended Gly-based peptide simulations in water at 298 K to the 100 ns trajectories previously published shows that the intrinsic conformational propensities were retained with little variation, further demonstrating Boltzmann sampling was achieved (Beck *et al.*, 2008).

To ensure that the GGXGG simulations captured experimentally valid ensembles, we calculated NMR chemical shifts for the production dynamics using SHIFTX2 and compared them against experimental data obtained of the same peptides in 8 M urea (Schwarzinger *et al.*, 2000; Han *et al.*, 2011). As the NMR experiment provides chemical shifts that are an average over all molecules root mean squared deviations (RMSD) between the simulated and experimental chemical shifts were also calculated to compare the average chemical shifts over the MD ensembles and those obtained by experiment. Previously, we showed agreement between the AAXAA peptide simulations and experimental AAXAA NMR observables, with excellent correlation coefficients between calculated and experimental chemical shifts ($R > 0.99$) (Towse *et al.*, 2016b). Here, we also obtained highly satisfactory individual correlations ($R \geq 0.94$, Table 6.1) between the GGXGG simulations and experimental data for all nuclei except for H_N, for which little dispersion in the range of values contributed to the low correlation coefficient (Table I). The correlation coefficient between the GGXGG simulations and experimental data when all nuclei were considered was excellent ($R > 0.99$, Table 6.1) with an RMSD of 0.72 ppm (Table 6.1).

6.4.2 *Neighboring residues do not alter coverage of φ/ψ space*

The coverage of φ/ψ space for the Gly guest residue (79–81%) was independent of host residue and simulation environment. Steric overlap of the backbone atoms restricts the remaining 20% of φ/ψ space as initially determined by Ramachandran (Ramachandran *et al.*, 1963; Table 6.2). The addition of heavy atoms to the side chain of guest residues reduced the coverage of φ/ψ space by 32% on average. Across all conditions, the average change in coverage of φ/ψ space between AAXAA and GGXGG was ~2%, with the largest differences observed for Hip, Tyr, Ile and Thr. This did not apply to Pro, which sampled the smallest area of φ/ψ space (<20% across all conditions). The sampling of the remaining non-Gly residues showed that they all accessed φ/ψ space to the same degree (mean 48–52%), irrespective of environment, with the β-branched residues at the lower end of the coverage range due to the increased steric clashes (Table 6.2).

Within a host series, there were no appreciable differences in the coverage of φ/ψ space except for the different protonation states of Asp and His. The negative charge on the Asp side chain that would be present at neutral pH reduced the area sampled compared with its protonated state, Ash. This was also true of the di-protonated His residue (Hip), which sampled less than the Hid and Hie forms. However, under thermally denaturing conditions, the restriction of backbone sampling by the charged Asp and Hip residues was no longer observed. These two exceptions aside, the largest difference across the environments was for residues in the AAXAA series, where a change in the coverage was 2–3% greater than that observed for residues in the Gly-based hosts.

6.4.3 *Intrinsic propensities are weakly host-dependent*

Under native-state conditions, water at 298 K, the sampling of the GGXGG and AAXAA host–guest pentapeptides was not random and the guest residues exhibited intrinsic conformational propensities in both hosts (Figure 6.2, Tables S6.2 and S6.3). To determine differences in these conformational propensities, percentage populations were calculated in the four quadrants of the φ/ψ plots and within seven defined conformational regions that correspond to elements of secondary structure in folded proteins (Figure 6.3a). We take the Gly-based peptides to reflect the 'true' intrinsic propensities of the amino acids; and, deviations from these values in the Ala-based peptides reflect the simple steric effects of the side-chain methyl group of the Ala-host residues and the backbone chirality imposed by their presence (Table A.6.4).

Correlations of the φ/ψ frequency distributions showed little change in the overall sampling trends by a given residue due to the identity of the host (Table 6.3). This was consistent with our finding that the additional steric effects of the Ala C_β methyl group did not reduce the accessibility of φ/ψ space (Table 6.2). Although similar broad biases towards particular conformational regions on the Ramachandran plots were observed (Figure 6.2), differences in sampling were immediately discernable from the φ/ψ quadrant populations (Tables S6.2-S6.4). In Ala-hosts, the greatest population resided in the $Q\alpha_R$ for all residues except Pro, the β-branched Ile and Val residues and Ser, which showed preferential sampling of Qβ. In the GGXGG series, this discrimination between these three residues and the others did not exist, all non-Pro residues predominantly sampled $Q\alpha_R$. Closer examination of the populations shows that for all residues there was a greater population in $Q\alpha_R$ when in Gly-hosts versus the Ala-hosts. The reduced steric constraints of the Gly neighbors increased the favorability of even Pro, known for its dominance of the β and P_{IIL} regions, to sample $Q\alpha_P$ regions.

Two residues that showed little difference between the hosts were Gly and Asp. In both Ala- and Gly-hosts, the quadrant populations of the Gly guest residue were almost indistinguishable. Both the broad sampling of φ/ψ space by Gly and the population of the conformational regions within the quadrants were retained; Gly showed very little response to the change in host residue (Figure 6.2, Tables S6.2-S6.4). Similarly, Asp showed a <3% difference in sampling across the quadrants as well as the individual conformational regions. In most cases, there were only marginal differences (<10%) observed in the populations of the conformational regions for the majority of guest residues in the Gly- and Ala-hosts consistent with the high correlations between the overall φ/ψ distributions (Table 6.3, Tables S6.2-S6.4).

Overall, the preferences for the α_R and near-α_R regions were robust, and the greater sampling in these regions by the Gly-based peptides, greatest in near-α_R, echoed the distinct increases in the Qα populations observed. In contrast, guest residues in Ala-based peptides had larger populations in the α_L, non-Pβ, and P_{IIL} regions. The difference in the populations of the P_{IR} and 'other' regions was negligible, suggesting that sampling in these regions was less dependent on the host residue. Bulky hydrophobic, β-branched, and aromatic residues play important roles in guiding protein folding (Frank *et al.*, 2002) and in the formation of interaction sites for intrinsically disordered proteins (Espinoza-Fonseca, 2012). The β-branched residues Ile and Val had below-average sampling of structures in the Qα quadrant in both Gly- and Ala-hosts and above-average

sampling in the β quadrant. Thr did not share this behavior with the other β-branched residues; instead, Thr had the highest sampling of the α_R conformational region in both Ala- and Gly-host peptides (Tables S6.2-S6.4). And, as the volume of the guest residue increased close to the main chain, e.g. the increased steric constraints at the C_β position for the orientation of the Ser side chain, the β-branched residues and the cyclic imino acid Pro, so did the preference for β structures (Tables S6.2-S6.4). Most residues sampled the P_{IIL} region to a greater extent in the Ala-host than in the Gly-host; however, this behavior did not apply to the aromatic residues Phe, Trp and Tyr.

To determine whether the differences in sampling between the two peptide systems was the result of altered intra-molecular contacts, we calculated ensemble averaged atomic contacts in three categories: main-chain–main-chain contacts, side-chain–side-chain contacts, and main-chain–side-chain contacts. No backbone atom pairs formed hydrogen bonds for >5% of the total simulation, and the introduction of Ala neighbors did not change the frequency of hydrogen bond formation by >2% for any individual hydrogen bond (Figure A.6.3). Similarly, there was no change in the formation or duration of contacts made between the guest residue and the peptide backbone under any condition.

The Ala-host introduces additional side-chain–side-chain contacts between the guest residues and their nearest neighbors; however, the formation of these additional contacts did not correlate with changes in φ/ψ distributions. We found that the side-chain contacts formed in the Ala peptides were equivalent under both native and thermally denaturing conditions (Figure A.6.3). Since the φ/ψ distributions between AAXAA and GGXGG were similar under thermally denaturing conditions, we concluded that intra-molecular contacts were not directly responsible for changes in φ/ψ distributions. We found no significant differences in intra-molecular hydrogen bonding, hydrophobic interactions, or nonspecific interactions.

6.4.4 *Conformational propensities are sensitive to protonation state*

Comparison of the sampling behavior of the Glu/Glh, Asp/Ash and Hie/Hid/Hip residue sets showed just how great an effect protonation could have (Figure 6.4). In the Ala-host system, protonation can appreciably change the conformational propensities of a residue. Although not so apparent with the Glu/Glh pair (R = 0.98), the effect of protonation was more substantial for Asp via a larger β population (increase ∼13%) due to the increased sampling of extended structures by Ash (Figure 6.4, Table A.6.5).

The effect of protonation was more pronounced for His ($R < 0.7$). The Hie (εH) and Hid (δH) protonation states exhibited similar φ/ψ sampling ($R = 0.96$) with only marginal changes observed in Qβ coincident with an offset between the stabilization of the near-α_R and P_{IIL} regions dependent on whether the δN or εN was protonated (Table 6.3). However, diprotonated His (Hip, δH and εH) showed a distinct mirroring effect, where Hid and Hie showed a significant preference for α_R over α_L (α_R: 27%, α_L: 8%) and Hip showed almost the exact opposite (α_R: 9%, α_L: 25%). Protonation had a lesser effect on the His and Asp sampling in the Gly-hosts, and instead showed more pronounced changes for the Glu/Glh residue pair. As observed for His in the Ala-host, a distinct change in preference for the α_L region occurred for protonated Glu (Glu: 6%, Glh: 20%) (Figure 6.3, Tables S6.2-S6.4).

6.4.5 *Steric effects of alanine neighbors under denaturing conditions*

The difference in sampling of the guest residues in GGXGG and AAXAA under chemically denaturing conditions was marginally greater than the difference in sampling under native conditions (Figure 6.3, Table A.6.9). We previously showed that in AAXAA under chemically denaturing conditions, the conformational preferences observed under native conditions are largely retained in 8 M urea at 298 K. As was observed for residues in the Ala system, residues in the Gly system under chemically denaturing conditions primarily sampled structures in $Q\alpha_P$, although more expanded helical structures were sampled in 8 M urea, reflected by the increases in sampling in the near-α_P region. As in native conditions, the guest residues in GGXGG peptides sampled $Q\alpha_R$ structure to a greater extent than in AAXAA peptides, which sampled the α_L, nPβ and P_{IIL} regions more (Tables S6.6-S6.8). Chemical denaturation destabilized the α_L region and promoted sampling of the P_{IIL} region. The nPβ and P_{IIL} regions had the greatest difference in sampling between the AAXAA and GGXGG host systems (Table A.6.8). Chemical denaturation increased the impact of neighboring Ala groups on the sampling of guest residues, leading to slightly greater differences in sampling between AAXAA and GGXGG.

The sampling of φ/ψ space increased further at high temperature, and residues had diminished preferences for one conformational region over another (Figure 6.3e). Under thermally denaturing conditions, the difference in sampling of the guest residues in GGXGG and AAXAA hosts indicated a lack of host-residue-dependent sampling propensities (Figure 6.3e, Table A.6.8). In both peptide systems, thermally denaturing conditions reduced intrinsic propensities for any

single conformational region and led to increased sampling of extended β structures. This made the φ/ψ distributions of the two peptide systems virtually indistinguishable from one another.

6.5 Discussion

The coverage of φ/ψ space for any given residue was invariant for the majority of guest residues in the Gly- and Ala-based systems, appearing insensitive to the introduction of the neighboring Ala residues, irrespective of the conditions. This led us to two conclusions. First, the additional side-chain interactions in the AAXAA system pose no great steric restriction to the regions of φ/ψ space that can be sampled. Second, there is no significant consequence from enforcing backbone chirality on the extent of φ/ψ sampling by the central guest residue. It is the relative free energy of regions of φ/ψ space that changes in response to the neighboring residues and conditions, not the area that can be sampled. Consequently, populations shift and conformational propensities can change.

Under native conditions, the populations of the seven defined conformational regions of the guest residues were weakly dependent on their host peptide. Residues with titratable side chains, however, did display some sensitivity to host sequence and environment. The protonation state of the amino acids with titratable side chains can have important consequences for electrostatic interactions in protein folding (Shen, 2010). Although the sampling of Asp was insensitive to host peptide sequence, the sampling protonated and amide counterparts of Asp (Ash and Asn, respectively) did show sensitivity to the host peptide sequence. The contrast in the sampling between Asp, Ash and Asn, which have side chains of similar geometry and size, shows how side-chain charge affected intrinsic sampling. This contrasting behavior was not as prominent for the Glu/Glh/Gln set where the functional groups are located farther away from the main chain. The distinct sampling of the Glu/Glh, Asp/Ash and Hie/Hid/Hip residue sets showed just how great an effect protonation could have, suggesting that protonation sensitive behaviors could contribute to the nature of electrostatic interactions in protein folding. A common conformational effect of protonation was to increase α_L structures and was dependent on both the identity of the titratable residue side chain and that of the neighboring residues. This shift to α_L is coincident with the increased frequency of α_L in active sites where it is believed the unusual structure at this position places residues in a functional pose (Novotny and Kleywegt, 2005). Moreover, this sensitivity to host correlates with observations that the pKa of certain residues, e.g. Glu78 and

Glu172 in *Bacillus circulans* xylanase, is dependent on the concerted action of conserved residues in active sites (Joshi *et al.*, 2001).

Access to the denatured state ensemble is commonly accomplished through the use of co-solvents such as urea (Das and Mukhopadhyay, 2009) or elevated temperatures. Under thermally denaturing conditions, guest residues were essentially insensitive to their host peptide and they sampled φ/ψ space similarly. This similar sampling of φ/ψ space suggests that the thermal energy at 498 K exerts a greater influence over the φ/ψ distributions than the other factors guiding intrinsic propensities. Under these conditions, the guest residues traverse φ/ψ space with ease as barriers separating conformational states are lowered. The guest residues also have reduced preferences for conformational regions with broad, shallow distributions. Under chemically denaturing conditions, the sampling of the guest residues was more dependent on the host peptide and varied by guest residue type. The presence of urea molecules allows for distinct solute–solvent interactions, which could lead to the stabilization of certain conformational regions via the formation of solute–solvent interaction networks. Overall, the host-dependent sampling differences, though weak, were distinct among the native, elevated temperature, and 8 M urea conditions, and led us to conclude that neighbor-dependent intrinsic propensities in natively unfolded states, under physiological conditions, differ from those under harsher denaturing conditions. The sensitivity of the sequence-dependent conformational preferences, and the conformational preferences themselves, to environmental conditions, of these simple peptide models agrees with the finding that the chemically denatured state ensembles of proteins differ from physiological unfolded states (Arcus *et al.*, 1994, 1995; Bond *et al.*, 1997; Das and Mukhopadhyay, 2009; Meng *et al.*, 2013).

6.6 Conclusions

Here we have quantified the difference in φ/ψ distributions of the naturally occurring amino acids within two sequence contexts, GGXGG and AAXAA, and across three environmental conditions. Due to the achirality of Gly, the GGXGG series is a model for the true intrinsic propensities of the amino acids and provides a baseline on which to build an understanding of how sequence contexts can modulate amino acid propensities. The AAXAA series also provides a model for the intrinsic propensities of the amino acids, but intrinsic to their tethering within a chiral

protein chain. Hence, the differences observed here reveal how the introduction of chirality and simple sterics into the polypeptide chain alters amino acid conformational propensities.

Loops in proteins frequently serve more complex functions than simply connecting elements of secondary structure. For example, loops serve as active sites in enzymes, bind small molecules, participate in protein–protein interactions, and can aid in the allosteric regulation of protein dynamics (Papaleo *et al.*, 2016). Thus, accurate design of loops may be instrumental in the success of computational and experimental protein design studies. To this end, this study provides insight into physically realistic models of intrinsic conformational preferences under different conditions. We have previously incorporated these propensities into the design of amyloid inhibiting peptides (Hopping *et al.*, 2014). We anticipate that the intrinsic propensities of neighboring group effects determined here, in combination with our fragment library (Rysavy *et al.*, 2014), may be exploited in the design of small peptides and in the design of flexible components of proteins.

6.7 Tables

Table 6.1. Correlations between experimental and calculated NMR chemical shifts for the GGXGG series in 8M urea.

Nucleus	R^1	n	RMSD (ppm)
H_N	0.77	19	0.20
$H\alpha$	0.94	20	0.07
$H\beta_1$	>0.99	4	0.06
$H\beta_2$	>0.99	16	0.10
$H\beta_3$	0.99	16	0.09
N_H	0.97	19	1.36
$C\alpha$	0.99	20	0.69
$C\beta$	>0.99	19	1.14
C’	0.95	20	0.64
Overall	>0.99	153	0.72

[1.]Experimental data for the GGXGG peptides were determined by Schwarzinger *et al.* in 8 M urea at pH 2.3 (45). Calculated NMR chemical shifts were obtained using SHIFTX2 (see Materials and Methods) for the 8M urea MD simulations.

Table 6.2. Coverage of φ/ψ space by Gly- and Ala-based pentapeptide systems in control and denaturing conditions.

	Water 298		8M Urea		Water 498 K	
Residue	GGXGG (%)	AAXAA (%)	GGXGG (%)	AAXAA (%)	GGXGG (%)	AAXAA (%)
Ala	59 ± 1[§]	61[§]	61 ± 1[§]	63 ± 2[§]	62 ± 1[§]	62[§]
Arg	48	50	52	50	50	53
Asn	46	50	46	49	52	53
Asp	38	37	37	36	51	52
Ash	43	44	45	46	48	50
Cys	52	56	54	53	54	53
Gln	48	50	49	50	53	53
Glu	48	50	50	49	51	51
Glh	49	48	50	53	53	52
Gly	79	81	80[§]	81[§]	79	79
Hid	50	50	53	51	52	55
Hie	52	51	54	52	53	54
Hip	40	52	46	53	51	50
Ile	34	39	37	39	38	39
Leu	49	50	49	51	48	52
Lys	48	50	51	48	50	52
Met	50	51	53	49	51	53
Phe	49	51	48	49	51	50
Pro	16	13	16	15	19	18
Ser	53	51	53	55	55	54
Thr	38	39	35	36	41	42
Trp	50	50	50 ± 2[§]	53 ± 1[§]	52	53
Tyr	47	53	50	44	51	53
Val	37	38	34	40	41	40
Mean*	48 ± 7	49 ± 11	49 ± 7	52 ± 13	51 ± 6	52 ± 10

*Mean coverage is calculated excluding glycine and proline values. [§]Averages across triplicate simulations; standard deviations omitted where they were <0.5%

Table 6.3. Correlation coefficients* between AAXAA and GGXGG ϕ/ψ frequency distributions.

Guest Residue	Water 298K	8M Urea 298 K	Water 498 K
Ala	0.94	0.90	0.93
Arg	0.96	0.97	0.96
Asn	0.97	0.91	0.96
Asp	0.99	0.99	0.97
Ash	0.98	0.96	0.97
Cys	0.96	0.88	0.96
Gln	0.88	0.96	0.96
Glu	0.97	0.96	0.96
Glh	0.97	0.89	0.96
Gly	0.97	0.90	0.88
Hid	0.99	0.99	0.96
Hie	0.96	0.95	0.96
Hip	0.74	0.94	0.96
Ile	0.94	0.68	0.97
Leu	0.96	0.95	0.98
Lys	0.95	0.94	0.97
Met	0.92	0.97	0.97
Phe	0.96	0.97	0.97
Pro	0.94	0.98	0.99
Ser	0.91	0.96	0.96
Thr	0.94	0.96	0.97
Trp	0.98	0.95	0.96
Tyr	0.92	0.99	0.97
Val	0.89	0.94	0.94

*Pearson's correlation coefficient was calculated as the correlation of population frequencies in each of the 5148 (72 x 72) bins in ϕ/ψ space.

6.8 Figures

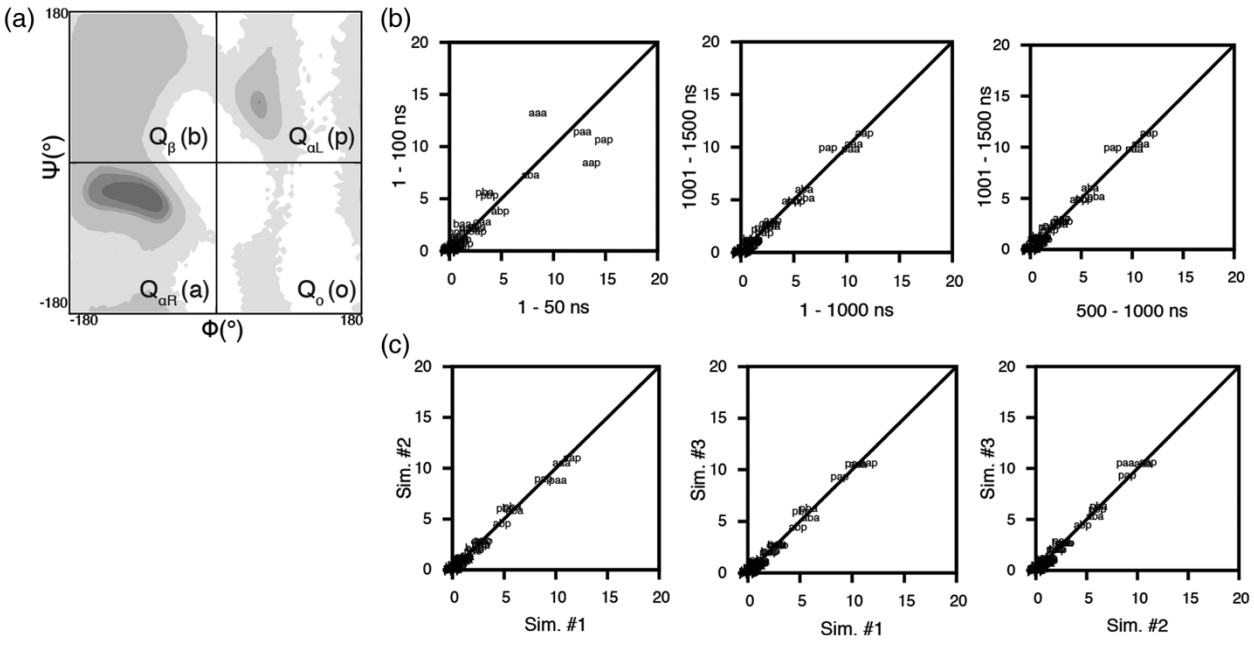


Figure 6.1. Convergence of the population of conformational states sampled by the three central residues shown for GGAGG in 8M urea at 298 K.

(a) Quadrants of Ramachandran space used to define the 64 conformational states: **a** ($Q\alpha_R$, right-handed α-helical) –φ, -ψ; **b** ($Q\beta$, β) –φ, +ψ; **p** ($Q\alpha_L$, left-handed α-helical) +φ, +ψ; **o** (Q_o, other) +φ, -ψ. (b) Comparison of the sampling of the 64 conformational states of run #1 of GGAGG across different portions of the trajectory. (c) Comparison of the sampling of the 64 conformational states of all three simulations of GGAGG.

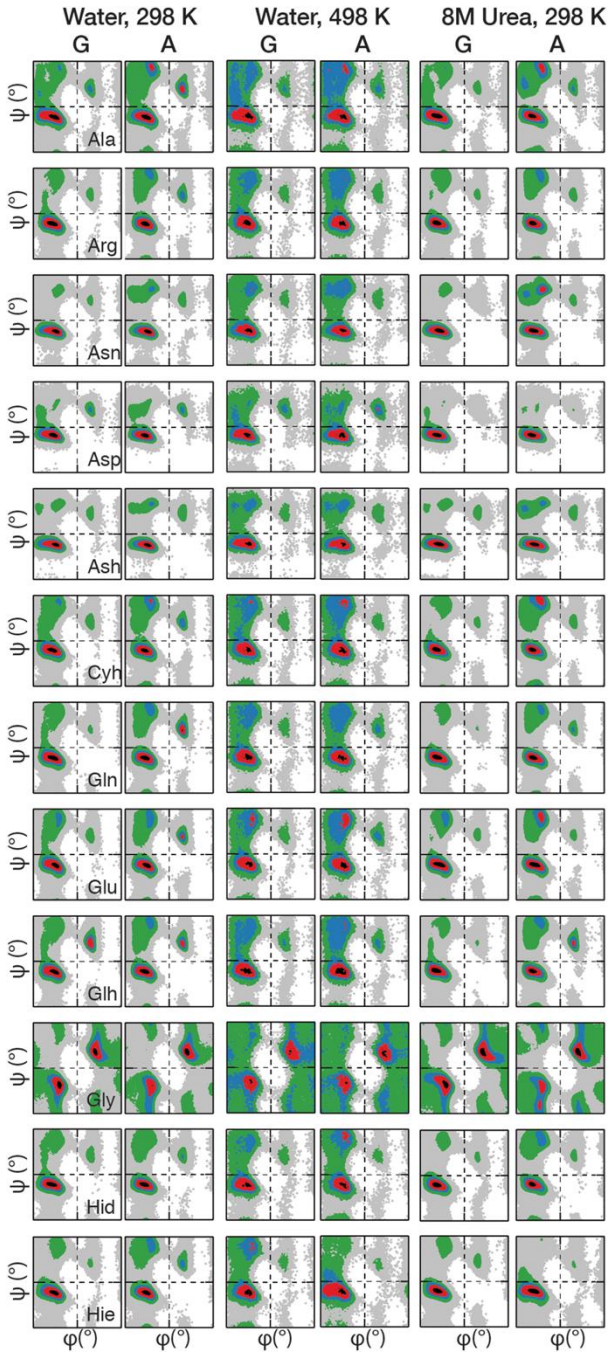


Figure 6.2.Ramachandran plots of the guest residues ('X') in the GGXGG (G) and AAXAA (A) hosts. Plots have been prepared for all peptides under three environmental conditions: water at 298 K, water at 498 K, and 8 M urea at 298 K. Plots have been normalized to the maximally populated bin of each plot and colored by increasing percentage population from gray to black: 0 = white; 0 < grey < 0.05; 0.05 ≤ green < 0.2; 0.2 ≤ blue < 0.4; 0.4 ≤ red < 0.8; 0.8 ≤ black. (Continued on next page)

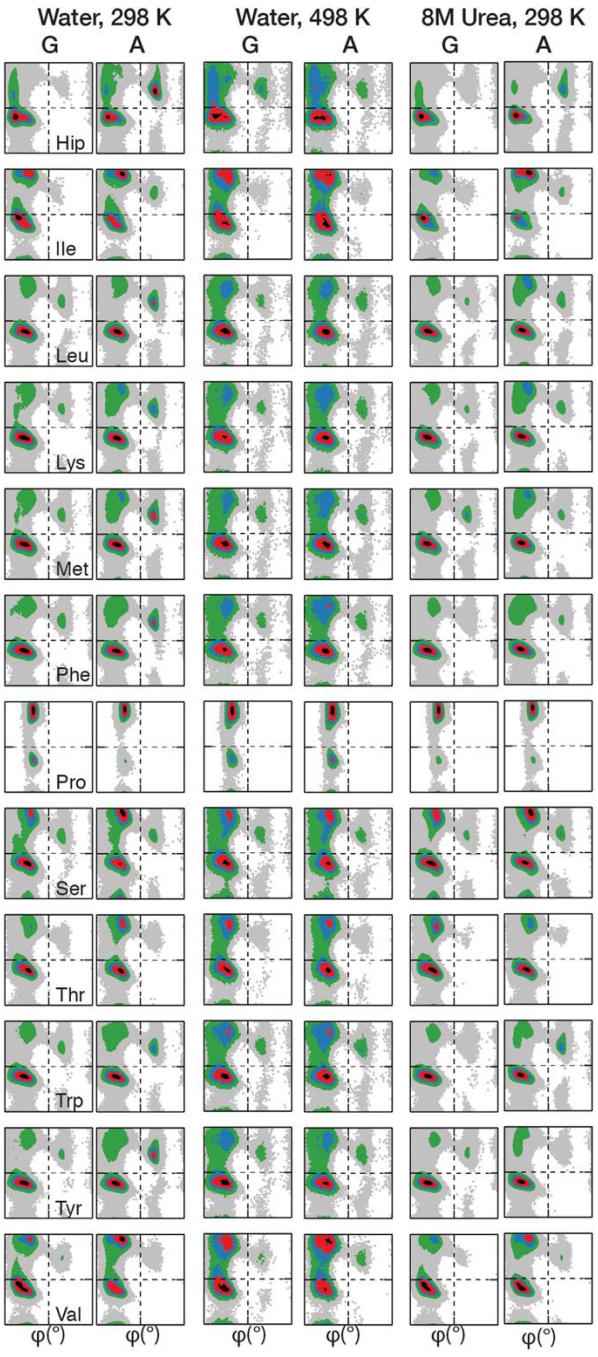


Figure 6.2 (Continued)

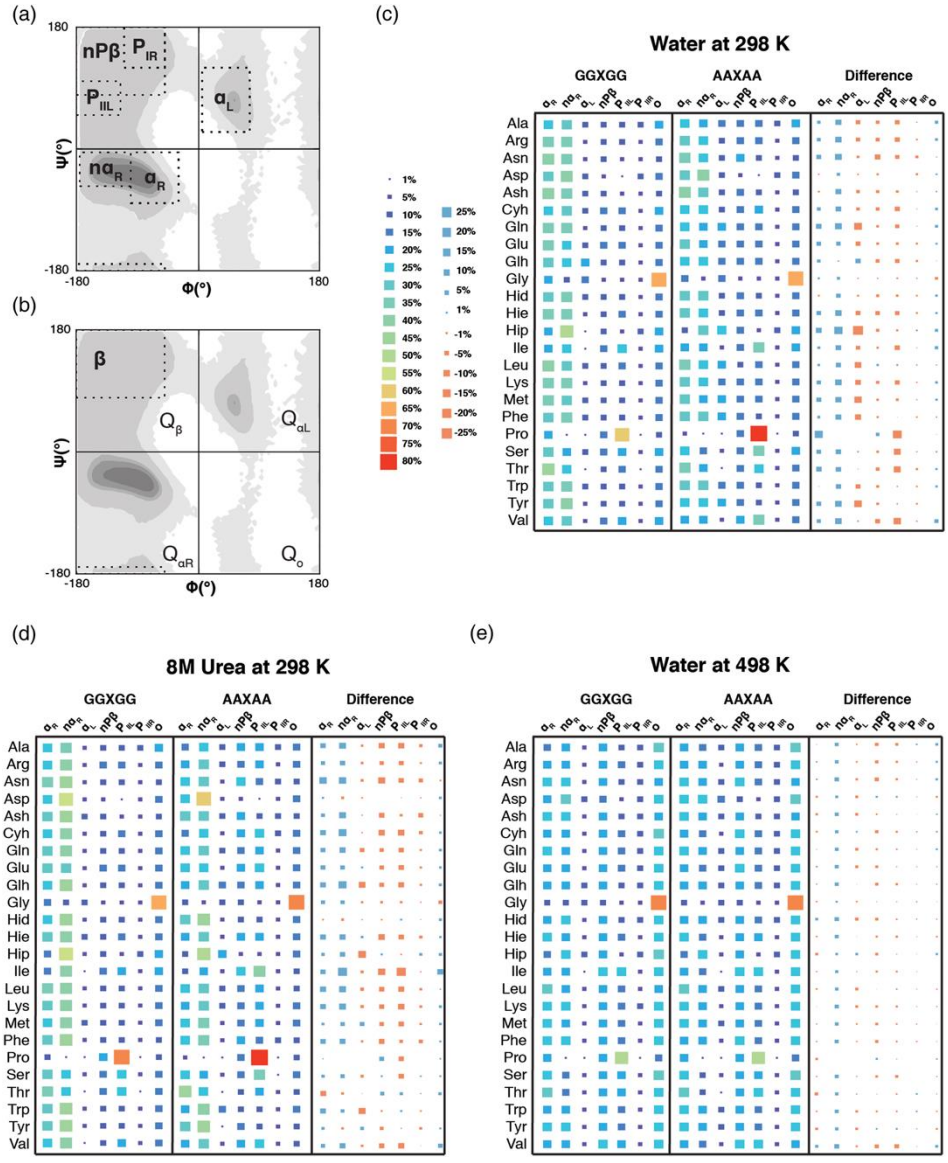


Figure 6.3. Differences in the fractional populations of conformational regions between GGXGG and AAXAA peptides in three environmental contexts.

(**a**) Ramachandran plots are provided that outline the seven conformational regions as defined in Materials and Methods: α_R, near-α_R, α_L, non-Pβ, P_{IIL}, P_{IR}, and o, and (**b**) the four quadrants and broadly defined β region. (**b–d**) Hinton diagram of the population of conformational states of the guest residues in the GGXGG and AAXAA peptides as well as the differences in sampling between the two peptide series under native (c), chemically denaturing (d), and thermally denaturing (**e**) conditions.

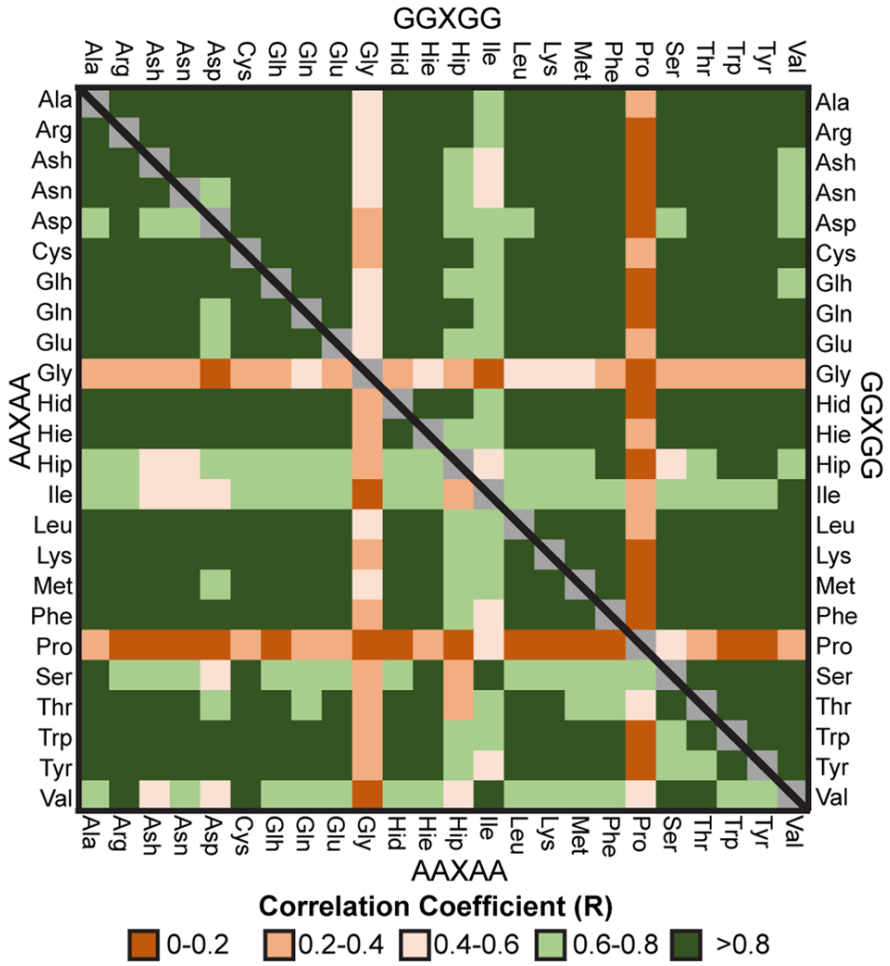


Figure 6.4. **Correlation matrix for GGXGG and AAXAA simulations in native conditions (water, 298 K).**

The Pearson correlation coefficients was calculated between φ/ψ frequency distributions for the central guest residue. Legend for the matrix element colors is inset.

Chapter 7. MOLECULAR DYNAMICS-DERIVED ROTAMER LIBRARIES FOR D-AMINO ACIDS WITHIN HOMOCHIRAL AND HETEROCHIRAL POLYPEPTIDES

7.1 Summary

Computational resources have contributed to the design and engineering of novel proteins by integrating genomic, structural and dynamic aspects of proteins. Non-canonical amino acids, such as D-amino acids, expand the available sequence space for designing and engineering proteins; however, the rotamer libraries for D-amino acids are usually constructed as the mirror images of L-amino acid rotamer libraries, an assumption that has not been tested. To this end, we have performed molecular dynamics (MD) simulations of model host–guest peptide systems containing D-amino acids. Our simulations systematically address the applicability of the mirror image convention as well as the effects of neighboring residue chirality. Rotamer libraries derived from these systems provide realistic rotamer distributions suitable for use in both rational and computational design workflows. Our simulations also address the impact of chirality on the intrinsic conformational preferences of amino acids, providing fundamental insights into the relationship between chirality and biomolecular dynamics. While D-amino acids are rare in naturally occurring proteins, they are used in designed proteins to stabilize a desired conformation, increase bioavailability or confer favorable biochemical and physical attributes. Here, we present D-amino acid rotamer libraries derived from MD simulations of alanine-based host–guest pentapeptides and show how certain residues can deviate from mirror image symmetry. Our simulations directly model D-amino acids as guest residues within the chiral L-Ala and D-Ala pentapeptide series to explicitly incorporate any contributions resulting from the chiralities of neighboring residues.

7.2 Introduction

Of the 20 standard proteinogenic amino acids, all except Gly have a chiral center at the Cα backbone atom and thus are present as pairs of stereoisomers. Although their occurrence in natural proteins is rare, D-amino acids can be incorporated into natural proteins via non-ribosomal peptide

synthetases (Caboche *et al.*, 2008), post-translational conversion of L-amino acids by isomerases (Bai *et al.*, 2009; Ollivaux *et al.*, 2014) or racemization (Fujii *et al.*, 2011). D-amino acids have been observed in antibacterial and antifungal peptides produced by bacteria (Radkov and Moe, 2014); peptide toxins in the venoms of spiders (*Agelenopsis aperta*) (Shikata *et al.*, 1995), platypuses (*Ornithorynchus anatinus*) (Torres *et al.*, 2005) and snails (*Achatina fulica*) (Kamatani *et al.*, 1989) and opiate-like peptides in frogs (*Phyllomedusa sauvagei*) (Montecucchi *et al.*, 1981). The potent biological functions demonstrated by these peptides has inspired the incorporation of D-amino acids into designed proteins.

There are two primary objectives for the incorporation of D-amino acids into designed and engineered proteins. First, D-amino acids can stabilize specific conformational states via their ability to adopt backbone φ and ψ angles that are unfavorable for L-amino acids. D-amino acids have been incorporated into designed proteins to stabilize α-helices (Rodriguez-Granillo *et al.*, 2011), turns (Imperiali *et al.*, 1992), anti-amyloid peptides (Hopping *et al.*, 2014; Kellock *et al.*, 2016), β-hairpins (Makwana and Mahalakshmi, 2016), novel peptide topologies (Rana *et al.*, 2004, 2005), metal-binding sites (Peacock *et al.*, 2009) and large topologies (Valiyaveetil *et al.*, 2004). Second, incorporation of D-amino acids can confer favorable biological and chemical properties to designed proteins. For example, the incorporation of D-amino acids into therapeutic peptides has been shown to increase their resistance to proteolysis, prolonging their biological half-lives (Zhou *et al.*, 2002). Although several successful designs have been reported, random incorporation of a D-amino acid into a protein does not guarantee that design goals will be met (Rodriguez-Granillo *et al.*, 2011). For example, in a systematic study of helix-stabilizing mutations to the Trp-Cage mini protein, Rodriguez-Granillo *et al.* found that increases in stability were modulated by the identity of the D-amino acid. Also key is the interplay between the backbone and side chain dihedral angles, which are coupled for some residues. Thus, consideration of the accessible conformations and intrinsic flexibilities of both the backbones and side chains of D-amino acids is necessary to make wise design choices.

Sterically allowed combinations of the backbone (φ and ψ) and side chain ($\chi_1 \ldots \chi_n$) dihedral angles define the accessible conformational space for a given amino acid. However, this conformational landscape is not evenly populated; instead, amino acid side chains preferentially adopt well-defined conformational states, known as rotamers (Chandrasekaran and Ramachandran, 1970; Bahar and Jernigan, 1996). These intrinsic conformational preferences can

be summarized as probability distributions in rotamer libraries. These rotamer libraries provide quick access to the most probable rotamer conformations and have widespread applications in molecular modeling (Krivov *et al.*, 2009), protein design (Ota *et al.*, 2001; Renfrew *et al.*, 2014) and protein structure prediction (Gordon *et al.*, 2003; Ryu *et al.*, 2016). In rational protein design applications, rotamer libraries list favorable conformations (ranked by probability) and allow users to identify potential mutations. In automated design applications, algorithms sort through the favorable conformations in rotamer libraries to search for minimum energy sequences that may adopt a desired structure. To this end, many L-amino acid rotamer libraries have been derived from high-resolution static structures deposited in the Protein Data Bank (PDB, www.rcsb.org) (Berman *et al.*, 2000) as well as from solvated atomistic molecular dynamics (MD) simulations of representatives of essentially all protein folds in the Dynameomics database (Dunbrack and Karplus, 1993, 1994; Dunbrack, 2002; Scouras and Daggett, 2011; Hintze *et al.*, 2016; Towse et al., 2016a).

In contrast to the abundance of side chain rotamer information available for L-amino acids, there are limited data available for D-amino acids due to their much rarer occurrence in natural proteins. In a survey of the PDB, Mitchell and Smith found 492 instances of D-amino acid residues distributed among 148 unique PDB entries, which corresponds to 0.6–0.7% (http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total) of the structures available in the PDB at that time (Mitchell and Smith, 2003). Due to the lack of high-quality experimental data for D-amino acids, their conformations and dynamics have been assumed to adhere to the mirror image convention, as evidenced by many organic molecules (Ota *et al.*, 2001; Makwana and Mahalakshmi, 2016). The mirror image convention assumes that mirrored configurational symmetry extends to mirrored dynamic symmetry. That is, the backbone-dependent probability distribution of the side chain dihedral angles for a D-amino acid (P_D) is the mirror image of the same distribution for its enantiomer (P_L) (Equation 7.1).

$$P_D(\chi_1 \dots \chi_N|\varphi,\psi) = P_L(-\chi_1 \dots -\chi_N|-\varphi,-\psi)$$

EQUATION 7.1

For a residue with one side chain dihedral angle, this corresponds to a reflection across $\chi 1 = 0°$. For a residue with two dihedral angles, this corresponds to a 180° rotation of the $\chi 1$ vs $\chi 2$ landscape. Previously, it has been stated that the mirror image convention does not apply to Ile and Thr (Gfeller *et al.*, 2013). We wish to clarify that for residues with multiple chiral centers, the mirror image convention can only be applied when the chirality is inverted at all centers. However, if the chirality is inverted at only one site, the mirror image convention cannot be applied. For example, L-Thr (absolute stereochemistry: 2S, 3R) and D-Thr (2R, 3S) have mirrored configurations but L-Thr and D-allo-Thr (2R, 3R) do not. This mirrored behavior has served as the theoretical foundation for the structural refinement and engineering of D-amino acid containing proteins. However, the assumption that amino acid enantiomers within peptide side chains mirror one another ignores the possibility that deviations from symmetry could result from the backbone dipole or the chirality of neighboring residues.

To our knowledge, there are two rotamer libraries for the D-enantiomer counterparts of the standard amino acids, one available through the SwissSidechain database (Gfeller *et al.*, 2013) and another via the Rosetta molecular Modeling Suite (Renfrew *et al*., 2012). Gfeller *et al.* constructed the Swiss side chain rotamer library using complementary physics- and knowledge-based approaches (Gfeller *et al.*, 2012, 2013). In this library, they performed four 50 ns MD simulations of L-Ala host–guest tripeptides (Ac-AXA-Nme, where X = the 20 amino acids) with implicit solvation, mapped propensities between structurally homologous residues and normalized the resulting probability distributions. For residues with achiral side chains, L-amino acid probability distributions were used to derive D-amino acid distributions via the mirror image convention. The second rotamer library for D-amino acids was constructed by (Renfrew *et al.* 2012, 2014) and is available in the Rosetta Molecular Modeling Suite (Das and Baker 2008). For this library, Renfrew *et al.* designed an algorithm that generates rotamer probability distributions for amino acids by iteratively seeding potential conformations followed by energy minimization. This approach used multiple structures of a dipeptide model of the amino acid that were seeded by making periodic rotations about the backbone and side chain dihedral angles. The methods used to generate these rotamer libraries exclude any effects from the chirality of neighboring residues. This is a key point that needs to be addressed given that designed proteins incorporating D-amino acids are almost always heterochiral polypeptide chains.

Although the mirror image convention is expected to apply for molecules that are exact enantiomers, it is not necessarily expected to apply in other cases, notably heterochiral molecules. In these instances, context-specific data are required. Here, we expand on prior work in computationally deriving rotamer libraries for D-amino acids by performing MD simulations of D-amino acid guest residues within two host pentapeptides series: AAxAA and aaxaa, where upper case characters refer to L-amino acids and lower case to D-amino acids. In addition, we perform a comparative analysis with rotamer libraries obtained for L-amino acid guest residues (AAXAA (Childers *et al.*, 2016; Towse *et al.*, 2016b) and aaXaa and the backbone-independent (BBIND) rotamer library derived from protein simulations in the Dynameomics database. These new MD-derived rotamer libraries complement existing ones by directly simulating the dynamics of D-amino acids and by considering the effects that the chirality of neighboring residues has on guest residue rotamer distributions. Our results provide empirical predictions of the impact of neighboring residue chirality on intrinsic rotamer distributions. In addition, our analysis confirmed that the mirror image convention describes the dynamics of most residues, but several residues demonstrated deviation from mirror image symmetry.

7.3 Methods

7.3.1 *MD simulations*

To comprehensively examine the effects of chirality on side chain rotamers, four host–guest peptide systems (AAXAA, AAxAA, aaXaa and aaxaa) were simulated as end-capped (N-acetylated, C-amidated) pentapeptides with extended starting conformations (φ and ψ angles set to 180° and −180°, respectively). The initial backbone conformations (φ and ψ) for Pro residues were −80° and −180°, respectively. Of the 20 naturally occurring amino acids, two—Ala and Gly—do not have rotameric side chain dihedral angles and these systems are not discussed here. In addition, we have simulated peptides corresponding to the neutral and acidic pH protonation states of Asp (Ash) and Glu (Glh) as well as the three protonation states of His: Hid (δH), Hie (εH) and Hip (both δH and εH). Cysteine was modeled in the reduced state (–CH_2–SH, denoted Cyh). This yields 22 guest residues (Table A.7.2) placed into four host peptides for a net total of 88 simulations (Table A.7.1). Simulations were performed using the *in lucem* molecular mechanics (*il*mm) package (Beck *et al.*, 2000–2018) with Levitt *et al.*'s force field (1995), the microcanonical NVE

(constant **n**umber of particles, **v**olume and **e**nergy) ensemble, and the flexible three-center (F3C) water model (Levitt *et al.*, 1997). Non-bonded interactions were treated with an 8 Å force-shifted cutoff (Beck *et al.*, 2005). Simulations were performed at 298 K with explicit water molecules using a box size that reproduced the experimental density at that temperature (0.9970 g/ml). The initial starting conformations were prepared for simulation by performing 1000 rounds of steepest descent energy minimization, followed by solvation in a periodic box extending 10 Å beyond any protein atom. The solvated systems were then subjected to iterative rounds of minimization. Next, production runs were performed for 600 ns each, corresponding to the lengths of our other prior Ala host simulations (Childers *et al.*, 2016; Towse *et al.*, 2016b), and the final 580 ns of each simulation was used to calculate rotamer distributions. The initial 20 ns of each simulation was excluded from rotamer library calculations to allow for movement away from the initial fully extended starting structure, which was used to avoid biasing the conformational sampling. In total, the results from 88 simulations are presented here with an aggregate simulation time >51 μs (Tables S7.1-S7.2).

7.3.2 *Mining the Dynameomics database*

The Dynameomics database (Beck *et al.*, 2008a; van der Kamp *et al.*, 2010) contains systematic MD simulations of representatives of essentially all known protein folds (97%) as determined by our Consensus Domain Dictionary (Schaeffer *et al.*, 2011). By analyzing all entries in this database, it is possible to construct distributions of structural and dynamical properties that serve as benchmark distributions (Jonsson *et al.*, 2009). One such reference distribution is the BBIND rotamer library constructed using all Dynameomics entries, which we compare to the rotamer distributions observed in our AAXAA simulations to determine whether rotamer distributions are similar in folded proteins and unstructured pentapeptides. In addition, it is possible to extract conformations for individual proteins in the native and temperature-unfolded states. For example, ubiquitin (PDB ID: 1UBQ) was chosen as the representative for the β-grasp, ubiquitin-like fold. Conformations of thermally unfolded ubiquitin (PDB ID: 1UBQ, Vijay-Kumar *et al.*, 1987) were obtained from the Dynameomics database to compare the side chain χ_1 dihedral angles observed in simulation and experiment. Per the Dynameomics protocol (van der Kamp *et al.*, 2010), two 50-ns high-temperature NVE simulations were performed at 498 K using the *il*mm (Beck *et al.*, 2000–2018) package and Levitt *et al.* force field (Levitt *et al.*, 1995).

7.3.3 *Definitions of specific conformational regions*

The intrinsic backbone sampling preferences were calculated by binning Ramachandran space into six specific conformational regions. For L-amino acids, these regions were defined as follows: α_R: $-100° \leq \phi \leq -30°$, $-80° \leq \psi \leq -5°$; near-α_R: $-175° \leq \phi \leq -100°$, $-55° \leq \psi \leq -5°$; α_L: $5° \leq \phi \leq 75°$, $25° \leq \psi \leq 120°$; β: $-180° \leq \phi \leq -50°$, $80° \leq \psi \leq -170°$; P_{IIL}: $-110° \leq \phi \leq -50°$, $120° \leq \psi \leq 180°$; P_{IR}: $-180° \leq \phi \leq -115°$, $50° \leq \psi \leq 100°$. An additional region, termed npβ, describes a subset of the β region that does not overlap with either of the polyproline regions (i.e. the **n**on-**p**olyproline **β** region). For D-amino acids, the corresponding regions were defined according to mirror image symmetry: $D\alpha_R$: $30° \leq \phi \leq 100°$, $5° \leq \psi \leq 80°$; near-$D\alpha_R$: $100° \leq \phi \leq 175°$, $5° \leq \psi \leq 55°$; $D\alpha_L$: $-75° \leq \phi \leq -5°$, $-120° \leq \psi \leq -25°$; Dβ: $50° \leq \phi \leq 180°$, $-80° \leq \psi \leq 170°$; DP_{IIL}: $50° \leq \phi \leq 110°$, $-180° \leq \psi \leq -120°$; DP_{IR}: $115° \leq \phi \leq 180°$, $-100° \leq \psi \leq -50°$ (Figure 7.1). Note that the naming convention conveys the mirror image symmetry of the enantiomers, i.e. the $D\alpha_R$ region corresponds geometrically to a left-handed α-helix. Side chain dihedral angles were classified into rotamers using the canonical bin definitions with the *gauche*+ (*g*+), *trans* (*t*) and *gauche*− (*g*−) nomenclature (Figure 7.1, Table A.7.3).

7.3.4 *Assessment of convergence*

To assess convergence, the rotamer populations were compared over different portions of the trajectories (0–6, 0–60 and 0–600 ns). In the case of short production runs, e.g. 6-ns, convergence as not reached in that the rotamer populations for the two halves of the trajectory were distinct (0–3 ns vs 3–6 ns). The simulations are deemed to be converged when the difference between the two halves of the trajectory is minimal.

7.3.5 *Calculation of correlation coefficients*

The Pearson correlation coefficient was calculated for each guest residue's rotamer distribution as follows. First, the production portion of a trajectory was randomly divided into two halves. An *n*-dimensional histogram (*n* = number of rotameric side chain dihedral angles) was constructed for each half of the trajectory by splitting each dihedral angle into 36 10° bins. The correlation coefficient was calculated between the bin counts of the resulting histograms. The reported correlation coefficients are the average of 50 iterations of this procedure.

7.3.6 *Calculation of NMR scalar coupling constants*

We checked our simulations against experiment by comparing calculated nuclear magnetic resonance (NMR) scalar couplings for the pentapeptides and thermally unfolded ubiquitin to experimental values obtained for chemically unfolded lysozyme, ubiquitin and protein G. We calculated $^3J_{H\alpha,H\beta}$, $^3J_{N,H\beta}$, $^3J_{C,H\beta}$, $^3J_{C,C\gamma}$ and $^3J_{N,C\gamma}$ scalar coupling constants from MD simulations using the Karplus Relation (Equation 7.2):

$$^3J(\theta) = C_0 + C_1 \cos\theta + C_2 \cos^2\theta$$

Equation 7.2

where $\theta = \chi_1$ by taking the average of the coupling constants calculated for each frame in the simulation. Residue-specific Karplus coefficients (C_0, C_1, and C_2) were obtained from Pérez *et al.* (2001) and are reproduced in Table A.7.4. Experimental data for urea-unfolded ubiquitin and protein G were obtained from the Biological Magnetic Resonance Bank (BMRB, Ulrich *et al.*, 2008) entries 16626 and 16627, respectively (Vajpai *et al.*, 2010). Experimental data for urea-unfolded hen egg white lysozyme (HEWL) were obtained from the supplementary materials accompanying Hennig *et al.* (Hennig *et al.*, 1999).

7.3.7 *Comparison of rotamer distributions*

Rotamer populations were calculated as the fraction of the MD ensemble that lies within the defined rotamer bins. The rotamer distribution for each residue corresponds to an *n*-dimensional (*n* = the number of rotameric states) vector and each element in the vector is the fraction of the ensemble that populates that bin. We employed a population displacement metric (Scouras and Daggett, 2011) to quantify the similarity between matched rotamer distribution vectors $\hat{X}$ and $\hat{Y}$. To address the extent to which the mirror image convention applies, comparisons between L- and D- enantiomers were done after matching the mirror image conformational states as in Table B.7.1. The similarity was calculated as the sum of the fraction of the population that does not change between matched rotamer bins x_i and y_i (Equation 7.3).

$$Similarity = \sum_i^n \min(x_i, y_i)$$

Equation 7.3

7.3.8 *Rotamer library construction and availability*

After establishing convergence of the simulations and their agreement with experiment, the simulations were analyzed and rotamer distributions were systematically cataloged. Four separate rotamer libraries were constructed, each corresponding to the guest residues in the four systems: AAXAA, aaxaa, AAxAA and aaXaa. Rotamer populations were calculated as the fraction of the MD ensemble that lies within the defined rotamer bins (Figure 7.1B, Table A.7.3). For each rotameric state, the modal angles for each side chain dihedral were also recorded. Comparisons between enantiomers were done after the mirror image rotamer conformations were matched (as in (Table B.7.1). Full rotamer libraries are provided both in tabular form in the Table B.7.1.

7.4 Results & Discussion

7.4.1 *Conformational sampling and convergence*

Prior to rotamer library construction, we confirmed that all simulations had achieved an equilibrium distribution with stable population frequencies. Convergence of these distributions was assessed by dividing the production trajectories into different ranges of time, e.g. 0–300 and 300–600 ns and comparing the populations of rotameric states between the different portions (Figure 7.2). After the point of convergence, the two halves of the production run should have nearly identical populations of rotameric states. Comparison of the rotamer populations over the initial stages of the simulation showed that the two rotamer distributions were not converged, demonstrated by the 0–3 ns vs 3–6 ns data in Figure 7.2. However, with longer trajectories, the population in the latter halves of the production runs became more similar, demonstrated by the 0–300 ns vs 300–600 ns data in Figure 7.2. Residues with longer side chains took longer to converge due to the increased degrees of freedom; however, these results show that 300 ns of production time is sufficient to achieve convergence even for the most complex residues, Arg and Lys. Our rotamer libraries have been constructed with nearly twice the requisite time to achieve convergence, yielding Boltzmann sampling for better estimates of conformational probability distributions. Extension of the trajectories an additional order of magnitude (i.e. 6 μs) is expected to yield only minimal changes to the population frequencies (Figure 7.2).

7.4.2 *Comparison with NMR coupling constants*

Quantitative spectroscopic measurements of intrinsic conformational distributions are challenging to obtain. Furthermore, the values can vary widely when different techniques are employed. For example, experimental studies have placed the polyproline (P_{II}) content of tri-alanine at various populations ranging between 50% and 92% (Woutersen *et al.*, 2002; Eker *et al.*, 2003; Graf *et al.*, 2007; Schweitzer-Stenner, 2009; Oh *et al.*, 2010; Sharma and Asher, 2010). Despite this, we strive to make quantitative comparisons with experiment whenever possible and have historically obtained good agreement with experiment. In our investigation of the GGXGG system, we showed that MD simulations of the GGAGG peptide reproduced the vicinal spin–spin coupling constant ($^3J_{NHC\alpha}$) within experimental error (Beck *et al.*, 2008b). In our investigation of the AAXAA system (Towse *et al.*, 2016b), we showed that the average helix content for the AAAAA system (19.4%) was in closer agreement with experimental estimates (10–20%, Firestine *et al.*, 2008; Jiang *et al.*, 2013) than the same system simulated with other force fields (Best *et al.*, 2008): 57.5% (CHARMM27), 62.3% (AMBER03), 94.2% (AMBER99) and 97.6% (AMBER94). In our comparative analysis of the GGXGG and AAXAA systems, we showed that our force field reproduces experimentally derived proton and heavy atom chemical shifts for the GGXGG system obtained in 8 M urea, 298 K, pH 2.5 (Childers *et al.*, 2016). Finally, we showed that our force field also reproduces experimental S^2 side chain order parameters for side chains with a methyl group (Ala, Ile, Leu, Met, Thr and Val) from a variety of globular proteins (Towse *et al.*, 2016a; Scouras and Daggett, 2011).

While no systematic data exist for the measurement of intrinsic side chain populations, NMR scalar coupling data has been used to gain insight into the χ_1 distributions in the urea-unfolded states of HEWL, ubiquitin and protein G (Hennig *et al.*, 1999; Vajpai *et al.*, 2010). To assess the correspondence between the intrinsic scalar couplings observed in pentapeptide systems to those in unfolded proteins, we calculated $^3J_{H\alpha,H\beta}$, $^3J_{N,H\beta}$, $^3J_{C,H\beta}$, $^3J_{C',C\gamma}$ and $^3J_{N,C\gamma}$ scalar coupling constants obtained from our MD simulations using the Karplus relation and the residue-specific Karplus coefficients as determined by Pérez *et al.* (Pérez *et al.*, 2001). The resulting comparison shows good agreement between simulation and experiment for most residues (Figure 7.3A). We also compared the $^3J_{H\alpha,H\beta}$, $^3J_{N,H\beta}$ and $^3J_{C,H\beta}$ scalar coupling constants obtained in our MD simulations of unfolded ubiquitin to its corresponding experimental values (Figure 7.3B). More than one point is shown for residues for which stereospecific assignments were made (note that

the different rotamers partition above and below 6 Hz in the top panel of Figure 7.3B). The resulting comparison shows an improved correspondence relative to the pentapeptide–protein comparison in Figure 7.3A. Overall, the best agreement was found for Phe, Pro, Thr, Trp and Tyr, while the worst agreement was found for Asn, Ash and Ile. The disagreement for Asn and Ash may result from the overpopulation of *trans* conformations in our simulations or the stabilization of *gauche* conformations within the experimental data relative to the intrinsically dominant conformations. No comparison could be made for Cys or the unprotonated forms of Asp, Glu and His, as these residues were not in the experimental data. The experimental results were obtained for globular proteins in 8 M urea at pH 2.5, hence we can only present a qualitative analysis as solvents (Bennion and Daggett, 2003; Li *et al.*, 2011; Childers *et al.*, 2016; Towse *et al.*, 2016b), potential residual structure (Aznauryan *et al.*, 2016) in the unfolded state and neighboring residue effects (Jung *et al.*, 2014) are known to modulate intrinsic sampling preferences. Finally, experimental data are available for only three proteins and are restricted to the populations of χ_1 dihedrals and for two proteins (ubiquitin and protein G). Nevertheless, although qualitative, the correspondence between the experimental and computational results is encouraging.

7.4.3 *Rotamer distributions in folded and unfolded states*

After establishing that our simulations qualitatively reproduced side chain distributions observed in unfolded states, we next compared the distributions observed in our peptides to the distributions observed in MD simulations of globular proteins in the folded state. We compared the intrinsic rotameric preferences sampled in pentapeptides with rotameric preferences within globular proteins by comparing rotamer distributions for L-amino acids derived from AAXAA simulations to the Dynameomics BBIND rotamer library obtained from the Dynameomics data set as described in the Methods section (Scouras and Daggett, 2011; Towse *et al.*, 2016a). This comparison tests whether rotamer distributions observed in folded, globular proteins are recapitulated in the unfolded state as modeled by simple pentapeptides. We found that the two distributions were qualitatively similar but distinct (Figure 7.4, Table B.7.1). Residues with short side chains, such as Ser and Pro, had the most similar distributions between the two states. For example, the distribution of Ser rotamers was 23% (*gauche*+, *g*+)/ 2% (*trans*, *t*)/ 75% (*gauche*−, *g*−) in the Dynameomics BBIND Library and 18% (*g*+)/ 1% (*t*)/ 81% (*g*−) in the AASAA peptide. Residues with bulkier side chains and charged residues showed greater deviations between the

AAXAA and BBIND libraries. These are the residues likely to form more complex interactions in specific structural environments. The residues with the greatest difference for a rotameric state between the folded and unfolded models were the *g+*, *t* rotamers of Ile (BBIND: 43%, AAIAA: 75%) and the *t*, *g+* rotamers of Asp (BBIND: 66%, AADAA: 86%) (Figure 7.4, Table B.7.1). The largest differences in sampling were restricted to the dominantly populated rotamers in the BBIND library. In other words, the differences between the AAXAA series and BBIND library were not the result of a rotamer switching between a low population in BBIND and high population in the pentapeptides (or vice versa). Instead, the population of dominant rotamers in the BBIND library increased in the AAXAA systems, implying that the dominant rotamers in peptides and proteins are similar and that intrinsically dominant rotamers tend to be preserved in the native state. Moreover, the presence of secondary and tertiary interactions in folded protein structures can increase the population of rotamers that are not intrinsically preferred. No comparison could be made for the protonated states of Asp (Ash), Glu (Glh) or diprotonated His (Hip) as the Dynameomics data set contains simulations at neutral pH.

7.4.4 *Symmetry in rotamer dynamics*

We previously confirmed through MD simulations of GGXGG and GGxGG that the mirror image convention applies for the backbone dihedral angles ϕ and ψ in an achiral host (Towse *et al.*, 2016b). To investigate to what extent the mirror image convention applies to the side chain dihedral angles of L- and D- guest residue pairs, we calculated the similarity between matched rotamer distributions for enantiomeric pairs (i.e. AAXAA and aaxaa; AAxAA and aaXaa) (Table 7.1). To investigate the impact that the chirality of neighboring residues had on guest residue rotamer distributions, we calculated the similarity between the rotamer distributions for the diastereoisomeric pairs for a given guest residue chirality (i.e. AAXAA and aaXaa; AAxAA and aaxaa) (Table 7.2). A similarity score of 100% indicates that the mirror image convention applies while lower scores can be attributed to deviation from the mirror image convention. The average similarity value calculated between three replicate simulations of the GGWGG peptide was 96.9 ± 1.4%. This calculation estimates that percent similarity differences of up to 4.5% can be attributed to dynamic variability among the simulations. Deviation from mirror image symmetry was predicted for residue pairs with less than 91% similarity. This stringent threshold (twice that predicted by replicate simulations) was chosen to minimize false positives.

Adherence to the mirror image convention was anticipated for enantiomeric pairs (AAXAA and aaxaa; AAxAA and aaXaa). However, in the homochiral enantiomers (AAXAA and aaxaa) the β-branched residues (Ile, Thr and Val) all deviated from mirror image symmetry (Figure 7.4, Table 7.1, Table B.7.1). The magnitude of the deviation from mirror image symmetry was dependent on the side chain: Ile (44% deviation) > Thr (33%) > Val (26%). For the heterochiral enantiomers (AAxAA and aaXaa) (Figure 7.4, Table 7.2, Table B.7.1) the results also showed a symmetry deviation for Ile and Thr (Figure 7.4). Again, the magnitude of this deviation was dependent on the side chain, with Ile displaying a greater deviation (46%) than Thr (21%). Heterochiral configurations reduced the magnitude of the deviation for Thr and resulted in mirror image symmetry for Val (Figure 7.4, Table II). Our results suggest that β-branched configurations can result in deviations from mirror image symmetry.

7.4.5 *Intrinsic sampling in heterochiral diasteroisomers*

After confirming mirrored behavior in enantiomeric pairs, we examined the extent to which host residue chirality affected the rotamer distributions of the guest residues by calculating the similarity values for heterochiral diastereoisomer pairs (AAXAA vs aaXaa and aaxaa vs AAxAA). Deviations from 100% similarity indicate that guest residue rotamer distributions are affected by the chirality of neighboring residues. This sensitivity to neighboring residue chirality was predicted for pairs with less than 91% similarity. For most of the guest residues, the rotamer distributions were insensitive to the chirality of the neighboring host residues (Table 7.2). However, for L-amino acid guest residues, Arg, Hie, Hip, Ile, Lys and Val exhibited sensitivity to the chirality of neighboring residues (Figure 7.4, Table 7.2, Table B.7.1). For D-amino acid guest residues, Arg, Hie, Hip, Ile, Lys, Thr, Trp and Val exhibited sensitivity to the chirality of neighboring residues (Figure 7.4, Table 7.2, Table B.7.1). Modulations in the rotamer populations due to host chirality are exemplified by the *t, Ng+* (7% populated in AAHAA, 14% populated in aaHaa) and *g−, Cg−* (22% in AAHAA, 13% in aaHaa) rotamers of L-Hip. Our rotamer libraries have been developed using the simplest chiral side chain – Ala. Greater deviations in rotamer similarity are anticipated for host residues with more complex side chains.

7.4.6 *pH-dependent rotameric preferences*

To our knowledge, there are no libraries that compare the rotamer distributions for alternate protonation states of Asp (Ash), Glu (Glh) and His (Hie, Hid and Hip). A common protein design goal is to develop a system with structural and/or functional properties that are linked with changes in pH. In these circumstances, the introduction or removal of ionizable residues is key (Guranda *et al.*, 2004; Leone and Picone, 2016). Therefore, we examined the effect that pronation could have on rotamer populations for the titratable residues and witnessed a focusing of the rotamer populations (Figure A.7.1). The populations in the dominant rotameric states of the neutral forms of Asp, Glu and His increased for the charged forms of these residues; the other minor rotamer populations correspondingly decreased. This finding indicates that the charge on an amino acid affects its intrinsic side chain conformational propensities as well as backbone conformational propensities (Childers *et al.*, 2016). The charged residues also had the greatest difference in sampling between the AAXAA systems and the Dynameomics BBIND rotamer library, suggesting that tertiary interactions within folded proteins affect the side chain sampling of titratable residues. The tertiary structure influence is expected given the specific nature of charged interactions in salt bridges, hydrogen bond networks and functional sites.

7.4.7 *Breakdown in valine's mirror image symmetry*

As described above, the β-branched residues had the greatest deviation from mirror image symmetry. This behavior was unanticipated for Ile, Thr and Val. In the SwissSidechain library, deviations from mirrored behavior were anticipated for Ile and Thr as D-*allo*-Ile and D-*allo*-Thr were simulated in that library (Gfeller *et al.*, 2013). To investigate the molecular origins of this deviation, we performed an analysis of the dynamics of Val guest residues. We first determined whether deviation from mirror image symmetry was observed in achiral host systems, i.e. the GGVGG and GGvGG peptides. Previously, we showed that in achiral hosts, L- and D-enantiomers have mirrored backbone propensities (Towse *et al.*, 2014). As a control, we extended the GGVGG and GGvGG simulations to 900 ns and our prior results were maintained in longer simulations (Figure 7.5). Furthermore, in the achiral host peptide simulations, L-Val and D-Val rotamer distributions were mirrored (Table B.7.1). Thus, deviations in mirror image symmetry for L-Val and D-Val occur only when the host residues enforce backbone chirality and impose steric

constraints on the guest residue. When deviations in mirror image symmetry occur, they must be accompanied by a change in the probability of transition between rotameric states and/or changes in interactions that contribute to the stabilization of different rotameric states. Consequently, we calculated the populations of specific conformational regions for the backbone dihedral angles for the Val peptides. In an achiral host, L-Val and D-Val have mirror image backbone propensities. In our chiral hosts, however, Val demonstrated context-dependent sampling of specific conformational regions in the backbone (Figure 7.5). In the all-D system, the nPβ and P_{IIL} regions were sampled (~20 and ~35%, respectively) to a larger extent than in the all-L system (~15 and ~30%, respectively) (Figure 7.5). The increased sampling of conformations in the β quadrant was met with decreased sampling of α_R conformations in the all-D system (~10%) relative to the all-L system (~20%). Although both residues have low sampling of the α_L conformation, D-Val had an order of magnitude lower sampling of this region (0.2%) than L-Val (2.7%) (Figure 7.5). Closer inspection of the dominant helical basins for L-Val and D-Val showed that the energy landscapes spanning the α_R and Dα_R basins were asymmetric (Figure 7.5). As a control, we extended the AAVAA and aavaa simulations to 900 ns (Figure 7.5). The asymmetric sampling of backbone conformations was maintained in the longer simulations, demonstrating that these deviations cannot be attributed to errors in conformational sampling. Similar results were obtained for the other β-branched residues, Ile and Thr (Figure 7.5).

Next, we examined the lifetimes and transition probabilities for Val. If mirror image convention extends to dynamics, then we should observe a mirroring of dynamic properties such as the probability of transition between rotamers and the lifetimes within rotamers. This was not the case, instead, the four peptides displayed system-dependent dynamics. For example, while the lifetimes for the *g*+ and *g*− rotamers were mirrored, the average lifetime of the *t* rotamer was significantly lower in aavaa than in AAVAA (Table 7.3). This suggests a destabilization of the *trans* conformation in aavaa as the source of the breakdown in mirror image symmetry.

Prior rotamer libraries have shown that the rotamers of the β-branched residues' rotamers have a strong dependence on the backbone conformation (Dunbrack and Karplus, 1993; Towse *et al.*, 2016a). We analyzed the relationship between the backbone and side chain dihedrals for Val by constructing backbone Ramachandran plots when the side chain was in each of the three rotameric states (Figure 7.6). We found that the *trans* rotamer conformation was the dominant rotamer in the α_L basin as well as in the α_R and Dα_R basins below $\psi \approx -20°$ and above $\psi \approx 20°$,

respectively (Figure 7.6). Thus, asymmetry in the population of α_R and $D\alpha_R$ helical structures, as well as a substantial decrease in the sampling of $D\alpha_L$ conformations for D-Val, contributed to the breakdown in the mirror image symmetry for Val. Similar results were obtained for the other β-branched residues, Ile and Thr (Figure 7.6).

The modal angles supplied with rotamer libraries define the default placement of a specific rotamer in modeling applications. In the AAVAA and aavaa peptides, the modal angles were not mirrored, demonstrating further deviation from mirror image symmetry. For example, the most populated angle in the *g*− rotamer for AAVAA was −71°, while the most populated angle in the *g*+ rotamer for aavaa was 81°. Collectively, these results predict that while the configurations of the AAVAA and aavaa peptides are mirrored, their dynamics are imperfectly mirrored. Because the side chain conformation for Val is coupled with the backbone conformation, the deviations in the α_R and $D\alpha_R$ basins translate to deviations in the rotamer populations. The molecular origins of this deviation remain elusive; though two possibilities present themselves. First, deviations in the solvation of the AAVAA and aavaa peptides may modulate the intrinsic conformational propensities. Second, minute steric interactions may be imperfectly mirrored in the two systems. Moreover, this effect may become more pronounced in more complex sequences than have been probed in these simple Ala-based peptides (Table 7.4).

Our simulations predict that at convergence (Table 7.4) the dynamics of several residues are incompletely described by the mirror image convention. Interestingly, the set of residues identified here include those that are known to induce the neighboring residue effect, by which a residue modulates the φ/ψ distribution of its nearest neighbors (Avbelj and Baldwin, 2004). This set also includes residues that have a strong coupling between the side chain and backbone dihedral angles (Dunbrack and Karplus; 1994; Towse *et al.*, 2016a). Our results contribute to mounting evidence that the unique rotamer dynamics for these residues may play important roles in determining protein folding pathways, the population of heterogeneous structures by IDPs and in protein–protein interaction sites.

While many assume that mirror image symmetry must be obeyed, deviation from mirror image symmetry for the amino acids is not unheard of; indeed, several studies have reported deviations from mirror image symmetry at various levels of structural organization. The Shinitzky group showed that the L- and D-enantiomers of 24-residue polyglutamate and polylysine peptides have energetic differences in the helix–coil transition, which they attributed to solvation

differences between D- and L-amino acids (Scolnik *et al.*, 2005). X-ray crystal structures of L- and D-monellin (PDB ID: 1KRL) show several significant differences at the dimer interface, which results in a 0.91–1.02 Å main chain root-mean-square deviation (RMSD) (Hung *et al.*, 1999). In that study the authors also found that asymmetry in the crystal structures was accompanied by asymmetric primary solvation layers around the L- and D-monellin monomers. Thus, minute deviations from mirror image symmetry present at the level of a single residue can be propagated to quaternary structure.

7.4.8 *Expanding protein design space*

Natural proteins are not restricted to the 20 proteinogenic amino acids coded for by DNA. Both prokaryotes and eukaryotes can incorporate non-canonical amino acids, including selenocysteine via a SECIS (selenocysteine insertion sequence) element (Su *et al.*, 2005) and some methanogenic prokaryotes can incorporate pyrrolysine (Borrel *et al.*, 2014). In total, over 140 amino acids have been observed in natural proteins (Ambrogelly *et al.*, 2007) and over 50 non-natural amino acids have been engineered into proteins (Young and Schultz, 2010). Thus, D-amino acids represent only one class of non-standard amino acids that can be used in protein design. A thorough description of the structural and dynamical features of these amino acids is necessary before their routine incorporation into designs. Our approach shows that MD simulations of simple systems can yield predictions for conformational propensities as well as insights into the dynamics that contribute to the stabilization of specific conformations. We have shown that intrinsic conformational propensities are sensitive to structural context and that it is not always possible to simply map conformational propensities from one amino acid to a homologous residue. Our group is committed to an exploration of intrinsic structural propensities and we regularly update our Structural Library of Intrinsic Residue Preferences (SLIRP) (publicly available at www.dynameomics.org), which is comprised of intrinsic backbone and side chain propensities as well as conformational behaviors of short fragments of protein structure (Beck *et al.*, 2008b; Scouras and Daggett, 2011; Rysavy *et al.*, 2014; Towse *et al.*, 2014, 2016a,b; Childers *et al.*, 2016).

7.5 Conclusions

The relatively few instances of non-standard amino acids in high-resolution protein structures precludes the generation of empirical data sets to define heuristic rules for protein design with these residues. However, computational methods can be employed to obtain conformational propensities for non-standard amino acids. Here, we have systematically studied the impact of backbone chirality on intrinsic rotamer distributions for L- and D-amino acids. In the studies presented here, the rotamer propensities of D-amino acids did not always mirror those of L-amino acids, although the mirror image convention was upheld for most residues. Deviation from symmetry in the propensities was most prominent for the β-branched residues. In some cases, the preferred D-amino acid side chain conformations exhibited sensitivity to the chirality of neighboring residues, highlighting that chain configuration is one of the many cumulative factors contributing to the observed conformational preferences. We anticipate that future protein design efforts in which both D- and L-amino acids are used will benefit from our libraries, which consider the impact of chain and residue chirality on rotamer distributions.

7.6 Tables

Table 7.1.The similarity percentage between enantiomeric peptide pairs reflects the extent to which the mirror image convention describes the rotamer populations for L- and D- amino acid pairs

Residue	Homochiral Similarity (AAXAA vs aaxaa)	Heterochiral Similarity (AAxAA vs aaXaa)
Arg	92.9	94.2
Asn	96.7	97.8
Asp	99.6	99.5
Ash	99.7	99.5
Cyh	98.8	99.2
Gln	98.7	98.2
Glu	98.6	99.2
Glh	99.4	99.3
Hid	96.5	97.7
Hie	98.1	96.4
Hip	94.0	95.1
Ile	56.0	54.2
Leu	97.7	97.3
Lys	95.6	95.8
Met	98.2	98.7
Phe	96.4	99.4
Pro	>99.9	99.3
Ser	99.6	99.6
Thr	67.0	78.8
Trp	94.4	96.8
Tyr	98.6	98.6
Val	74.2	96.9

Table 7.2. The similarity percentage between heterochiral diastereoisomeric pairs measure the impact of host residue chirality on the rotamer populations on L- and D- guest residues.

Residue	L-AA Guest (AAXAA vs aaXaa)	D-AA Guest (AAxAA vs aaxaa)
Arg	86.4	88.2
Asn	95.4	93.6
Asp	98.1	98.4
Ash	98.9	99.3
Cyh	96.9	96.0
Gln	94.3	93.5
Glu	97.4	98.1
Glh	96.1	96.8
Hid	95.6	93.8
Hie	88.1	88.1
Hip	84.2	86.5
Ile	89.2	86.3
Leu	95.4	98.1
Lys	88.5	89.9
Met	94.3	94.4
Phe	95.0	91.8
Pro	97.3	96.6
Ser	99.9	99.3
Thr	97.2	88.0
Trp	91.5	89.9
Tyr	94.6	96.2
Val	90.4	61.6

Table 7.3.The average rotameric state lifetimes for rotameric states of Valine in homochiral valine enantiomeric peptides

	AAVAA	**aavaa**	**aaVaa**	**AAvAA**
Rotamer	**Lifetime (ps)**	**Lifetime (ps)**	**Lifetime (ps)**	**Lifetime (ps)**
g+	118 ± 161	202 ± 283	111 ± 130	181 ± 304
t	145 ± 238	23 ± 56	180 ± 241	187 ± 260
g-	259 ± 413	114 ± 144	214 ± 268	101 ± 136

Table 7.4. The correlation coefficients and root mean squared difference calculated over randomly sampled trajectories indicates converged rotamer distributions.

	AAXAA		aaxaa		aaXaa		AAxAA	
Res	R[a]	RMSD[b]	R[a]	RMSD[b]	R[a]	RMSD[b]	R[a]	RMSD[b]
Arg	0.986	$1.25 \cdot 10^{-4}$	0.970	$1.54 \cdot 10^{-4}$	0.952	$1.82 \cdot 10^{-4}$	0.972	$1.42 \cdot 10^{-4}$
Asn	> 0.999	$3.49 \cdot 10^{-3}$	> 0.999	$3.48 \cdot 10^{-3}$	> 0.999	$4.63 \cdot 10^{-3}$	> 0.999	$3.73 \cdot 10^{-3}$
Asp	> 0.999	$3.68 \cdot 10^{-3}$	> 0.999	$3.87 \cdot 10^{-3}$	> 0.999	$3.85 \cdot 10^{-3}$	> 0.999	$3.34 \cdot 10^{-3}$
Ash	> 0.999	$3.38 \cdot 10^{-3}$	> 0.999	$3.78 \cdot 10^{-3}$	> 0.999	$3.55 \cdot 10^{-3}$	> 0.999	$3.01 \cdot 10^{-3}$
Cyh	> 0.999	$2.52 \cdot 10^{-2}$	> 0.999	$2.20 \cdot 10^{-2}$	> 0.999	$1.42 \cdot 10^{-2}$	> 0.999	$1.84 \cdot 10^{-2}$
Gln	0.997	$6.71 \cdot 10^{-4}$	0.996	$7.41 \cdot 10^{-4}$	0.996	$7.4 \cdot 10^{-4}$	0.996	$6.83 \cdot 10^{-4}$
Glu	0.997	$7.4 \cdot 10^{-4}$	0.996	$9.81 \cdot 10^{-4}$	0.996	$8.46 \cdot 10^{-4}$	0.997	$8.09 \cdot 10^{-4}$
Glh	0.997	$6.56 \cdot 10^{-4}$	0.996	$7.07 \cdot 10^{-4}$	0.997	$6.27 \cdot 10^{-4}$	0.997	$6.26 \cdot 10^{-4}$
Hid	0.999	$3.69 \cdot 10^{-3}$	> 0.999	$3.34 \cdot 10^{-3}$	0.999	$3.93 \cdot 10^{-3}$	> 0.999	$3.18 \cdot 10^{-3}$
Hie	0.999	$3.57 \cdot 10^{-3}$	0.999	$3.56 \cdot 10^{-3}$	0.999	$3.67 \cdot 10^{-3}$	> 0.999	$3.37 \cdot 10^{-3}$
Hip	> 0.999	$4.35 \cdot 10^{-3}$	> 0.999	$3.14 \cdot 10^{-3}$	> 0.999	$3.54 \cdot 10^{-3}$	> 0.999	$3.68 \cdot 10^{-3}$
Ile	> 0.999	$3.74 \cdot 10^{-3}$	> 0.999	$3.41 \cdot 10^{-3}$	> 0.999	$3.55 \cdot 10^{-3}$	> 0.999	$3.74 \cdot 10^{-3}$
Leu	> 0.999	$3.79 \cdot 10^{-3}$	> 0.999	$3.46 \cdot 10^{-3}$	> 0.999	$3.03 \cdot 10^{-3}$	> 0.999	$3.85 \cdot 10^{-3}$
Lys	0.987	$1.26 \cdot 10^{-4}$	0.988	$1.23 \cdot 10^{-4}$	0.985	$1.57 \cdot 10^{-4}$	0.979	$2.33 \cdot 10^{-4}$
Met	0.996	$7.95 \cdot 10^{-4}$	0.997	$6.51 \cdot 10^{-4}$	0.997	$7.34 \cdot 10^{-4}$	0.997	$7.12 \cdot 10^{-4}$
Phe	> 0.999	$3.85 \cdot 10^{-3}$	> 0.999	$3.72 \cdot 10^{-3}$	0.999	$3.92 \cdot 10^{-3}$	> 0.999	$3.41 \cdot 10^{-3}$
Pro	0.995	$1.88 \cdot 10^{-1}$	> 0.999	$6.31 \cdot 10^{-2}$	0.945	$5.5 \cdot 10^{-1}$	0.987	$2.74 \cdot 10^{-1}$
Ser	> 0.999	$2.72 \cdot 10^{-2}$	> 0.999	$1.43 \cdot 10^{-2}$	> 0.999	$1.89 \cdot 10^{-2}$	> 0.999	$2.20 \cdot 10^{-2}$
Thr	> 0.999	$1.56 \cdot 10^{-2}$	> 0.999	$2.34 \cdot 10^{-2}$	> 0.999	$2.07 \cdot 10^{-2}$	> 0.999	$1.14 \cdot 10^{-2}$
Trp	0.979	$2.27 \cdot 10^{-2}$	0.988	$1.79 \cdot 10^{-2}$	0.979	$2.27 \cdot 10^{-2}$	0.996	$1.02 \cdot 10^{-2}$
Tyr	> 0.999	$3.79 \cdot 10^{-3}$	> 0.999	$3.59 \cdot 10^{-3}$	> 0.999	$3.51 \cdot 10^{-3}$	> 0.999	$3.75 \cdot 10^{-3}$
Val	> 0.999	$1.66 \cdot 10^{-2}$	> 0.999	$2.33 \cdot 10^{-2}$	> 0.999	$1.81 \cdot 10^{-2}$	> 0.999	$1.20 \cdot 10^{-2}$

a. Pearson's correlation coefficient, averaged over 50 iterations, comparing each 10° x 10° bins
b. RMS Difference between two sets, averaged over 50 iterations, this is average RMSD per bin (%)

7.7 Figures

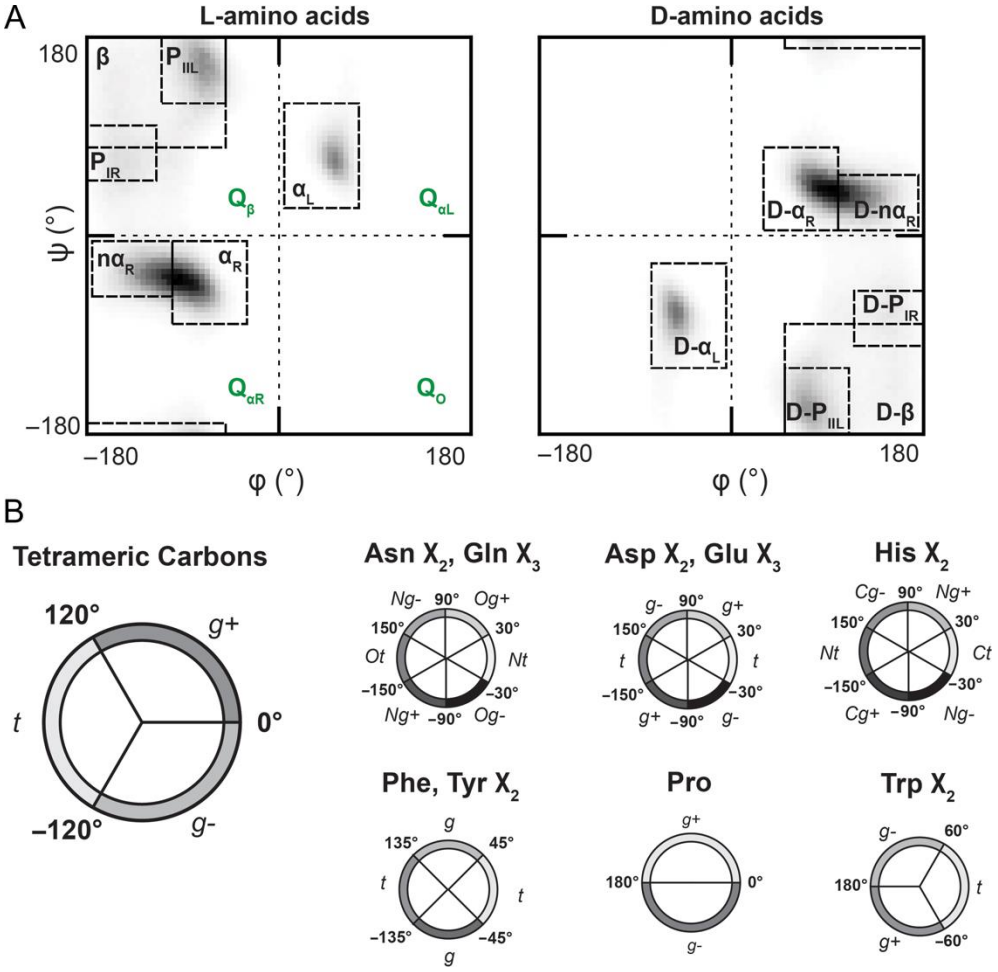


Figure 7.1. Definitions of specific conformational regions (backbone dihedrals) and rotameric states (side chain dihedrals).

Intrinsic conformational propensities are determined by measuring the population of specific conformational regions that are defined as basins in dihedral angle space. (**A**) For intrinsic backbone propensities, the regions corresponding to the six specific conformational regions are outlined on Ramachandran plots for L-amino acids (left) and D-amino acids (right). Adapted from Towse *et al*. (2014). (**B**) For intrinsic side chain propensities, the bins corresponding to the *gauche*+ (*g*+), *trans* (*t*) and *gauche*− (*g*−) rotameric states are defined for the tetrameric carbons (left), which applies to most side chain dihedral angles. For the remaining dihedral angles, residue-specific bin definitions are supplied (right).

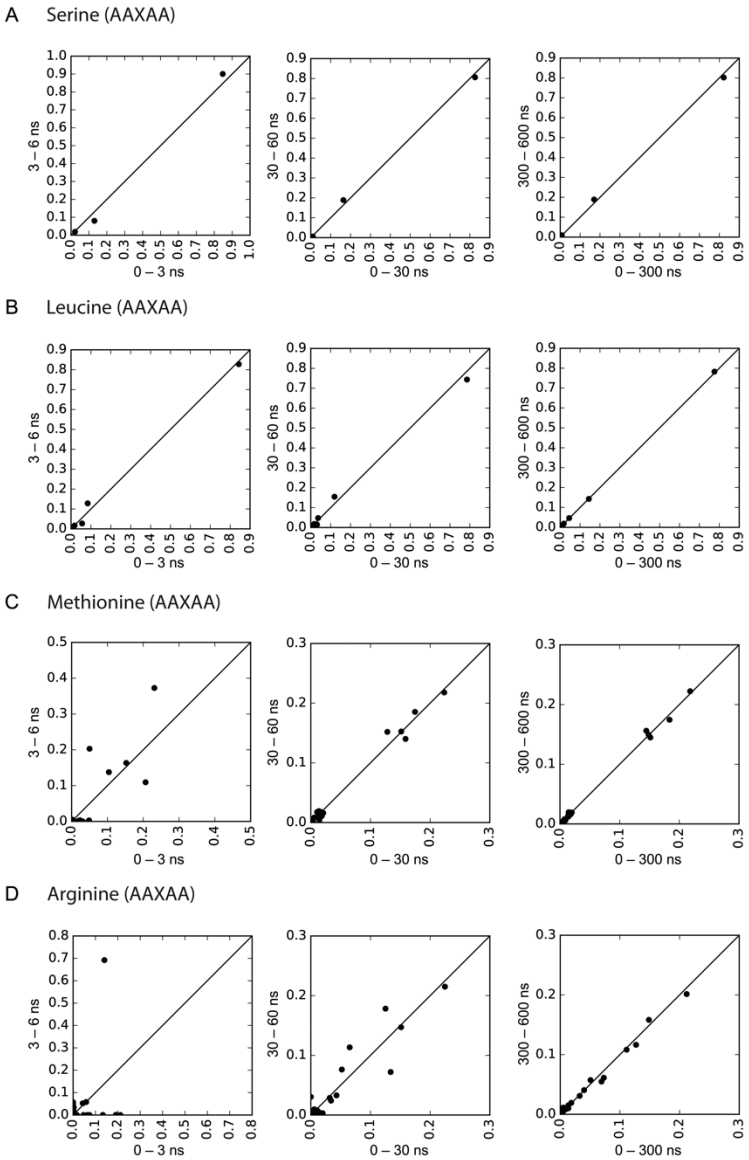


Figure 7.2. Convergence of the population of rotameric states sampled by Ser, Leu, Met and Arg in the AAXAA system.

In each plot, a single point corresponds to a rotameric state. The population of that state is compared between the first and second halves of the trajectories for the designated times in the simulations. Plots are shown for three different portions of the trajectories: 0–6 ns (left), 0–60 ns (center) and 0–600 ns (right) for four different residues: serine (**A**, one side chain dihedral), leucine (**B**, two side chain dihedrals), methionine (**C**, three side chain dihedrals) and arginine (**D**, four side chain dihedrals). Residues with longer side chains converge on slower timescales due to the increased accessible conformational space introduced by additional dihedral angles.

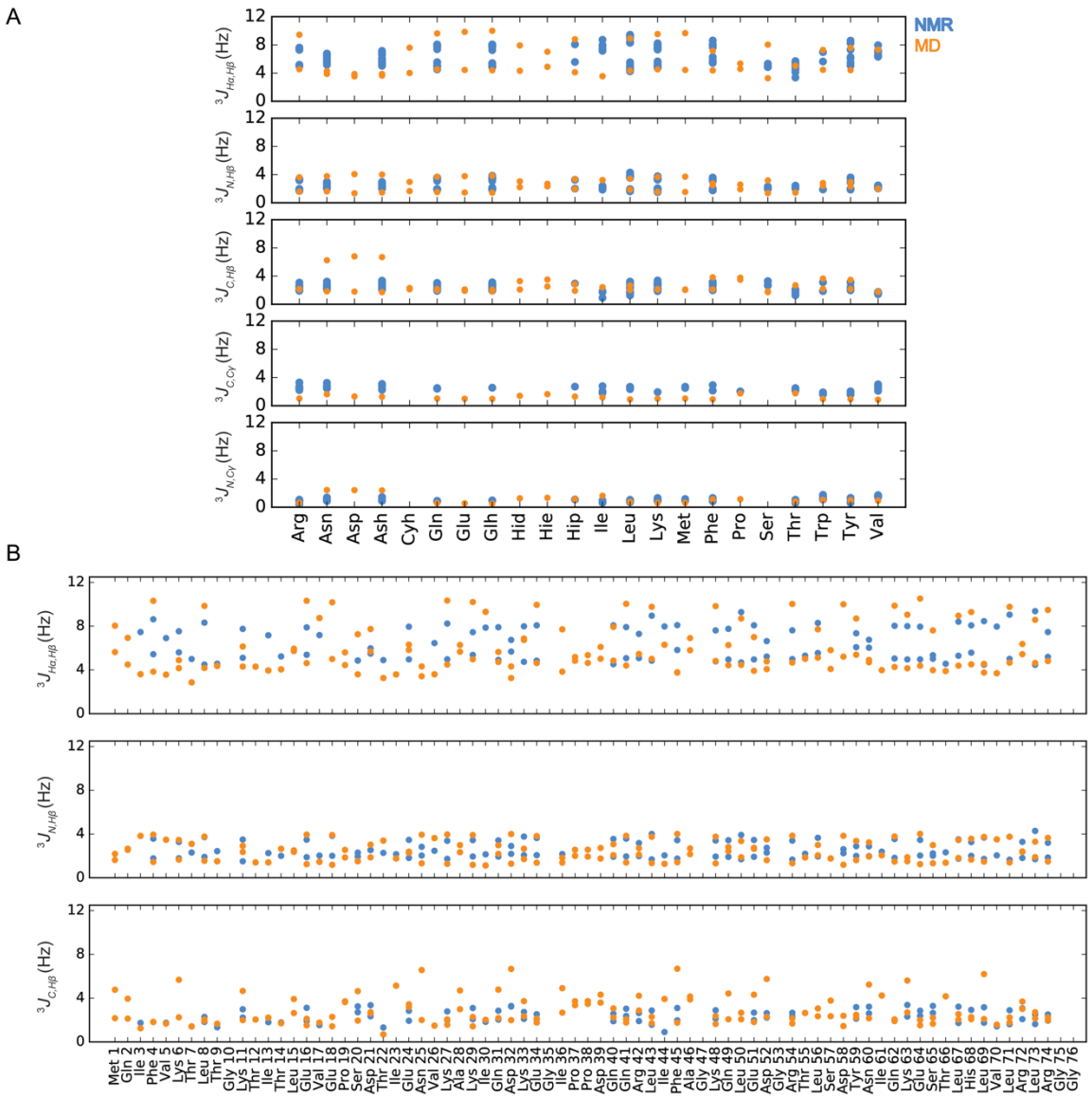


Figure 7.3. Validation of peptide dynamics via a comparison with experimental ^{3}J coupling constants.

(**A**) Experimental $^3J_{Hα,Hβ}$, $^3J_{N',Hβ}$, $^3J_{C',Hβ}$, $^3J_{C',Cγ}$ and $^3J_{N',Cγ}$ scalar coupling constants for urea-unfolded ubiquitin, protein G and HEWL (blue points) are presented. The experimental data for ubiquitin and protein G contain stereospecific assignments for the scalar coupling constants of prochiral $H_{β1}$ and $H_{β2}$ atoms. The corresponding values over the MD ensembles for the AAXAA series (orange points) were obtained using the Karplus relation (Equation 2) and coefficients as determined by Pérez *et al.* (2001). Error bars denote standard deviations. (**B**) Comparison of experimental (blue) and MD-derived (orange) NMR coupling constants for urea-unfolded (blue) and temperature-unfolded (orange) ubiquitin.

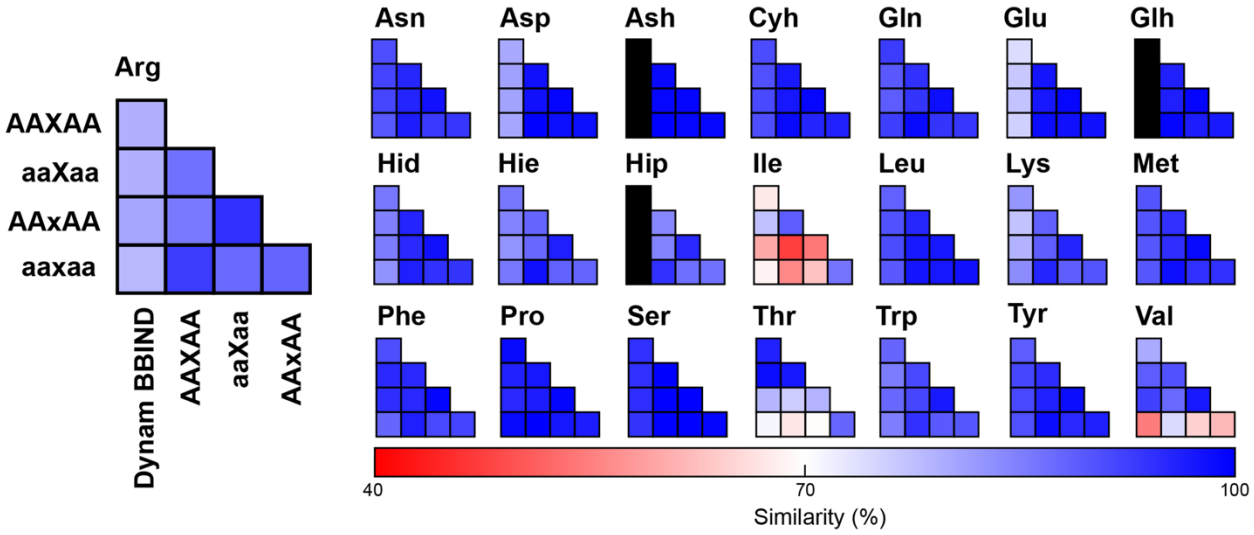


Figure 7.4. Similarity plots of rotamer distributions derived from the Dynameomics BBIND library and pentapeptide simulations.

Each matrix reports the similarity coefficients for the rotamer distributions of a given guest residue across each of the four pentapeptide systems plus the Dynameomics BBIND library. For comparisons between L- and D-amino acids, comparisons were made between the matched rotamer distributions. High-similarity coefficients (blue) indicate that two rotamer distributions are similar, whereas low-similarity coefficients (red) indicate that the two rotamer distributions are distinct. Black squares indicate that no valid comparison is available. Adapted from Scouras and Daggett (2011), Fig. 3.

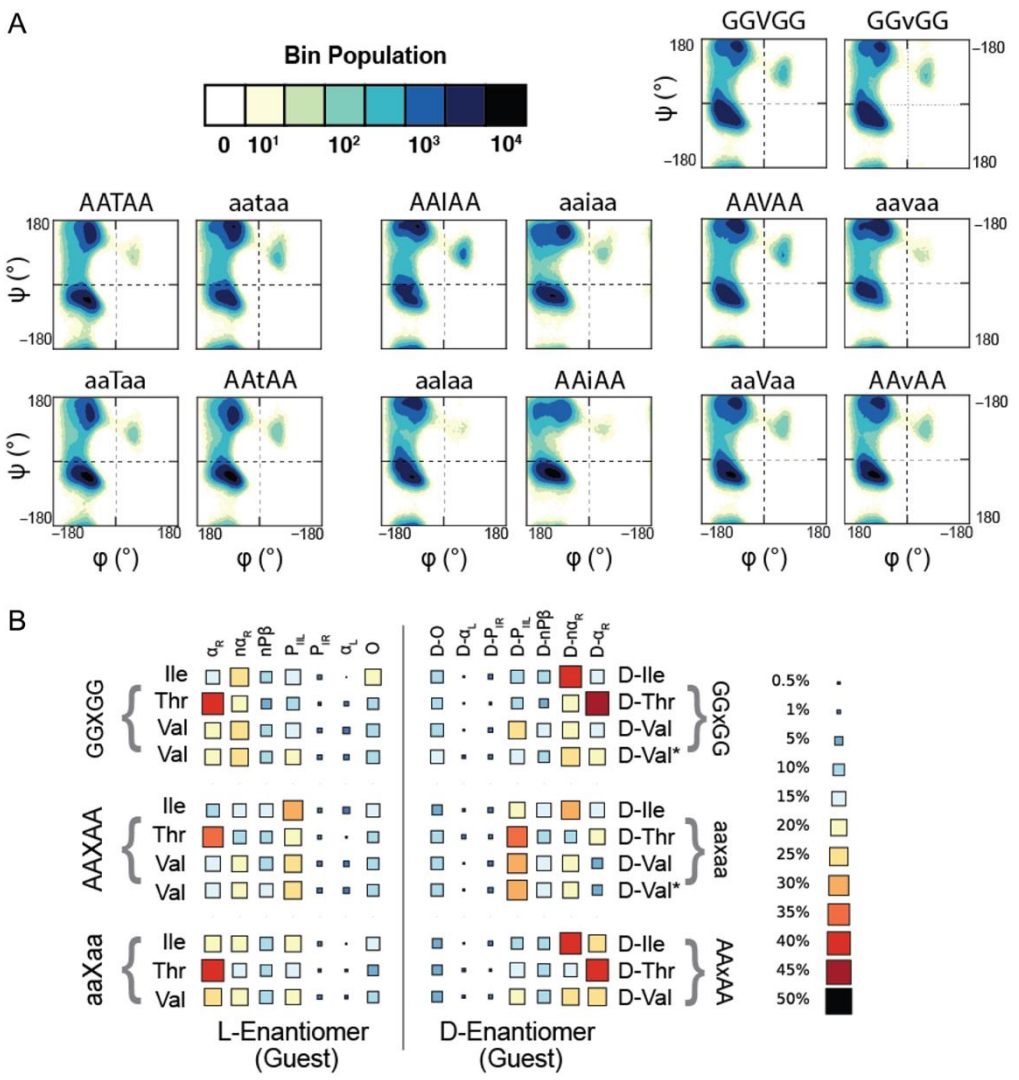


Figure 7.5. Enantiomeric β-branched amino acids sample backbone conformations asymmetrically within chiral host peptides.

(**A**) Ramachandran plots of the three β-branched residues within each of the four host peptides. Plots for D-amino acids have been rotated by 180° to facilitate comparisons with L-amino acids. (**B**) In these Hinton plots, each row corresponds to a single guest residue and reflects the distribution of the trajectory among each of the seven specific conformational regions (columns). For the Val guest residues, the upper boxes correspond to 600 ns of production time and the lower boxes correspond to 900 ns of production time, marked with an asterisk.

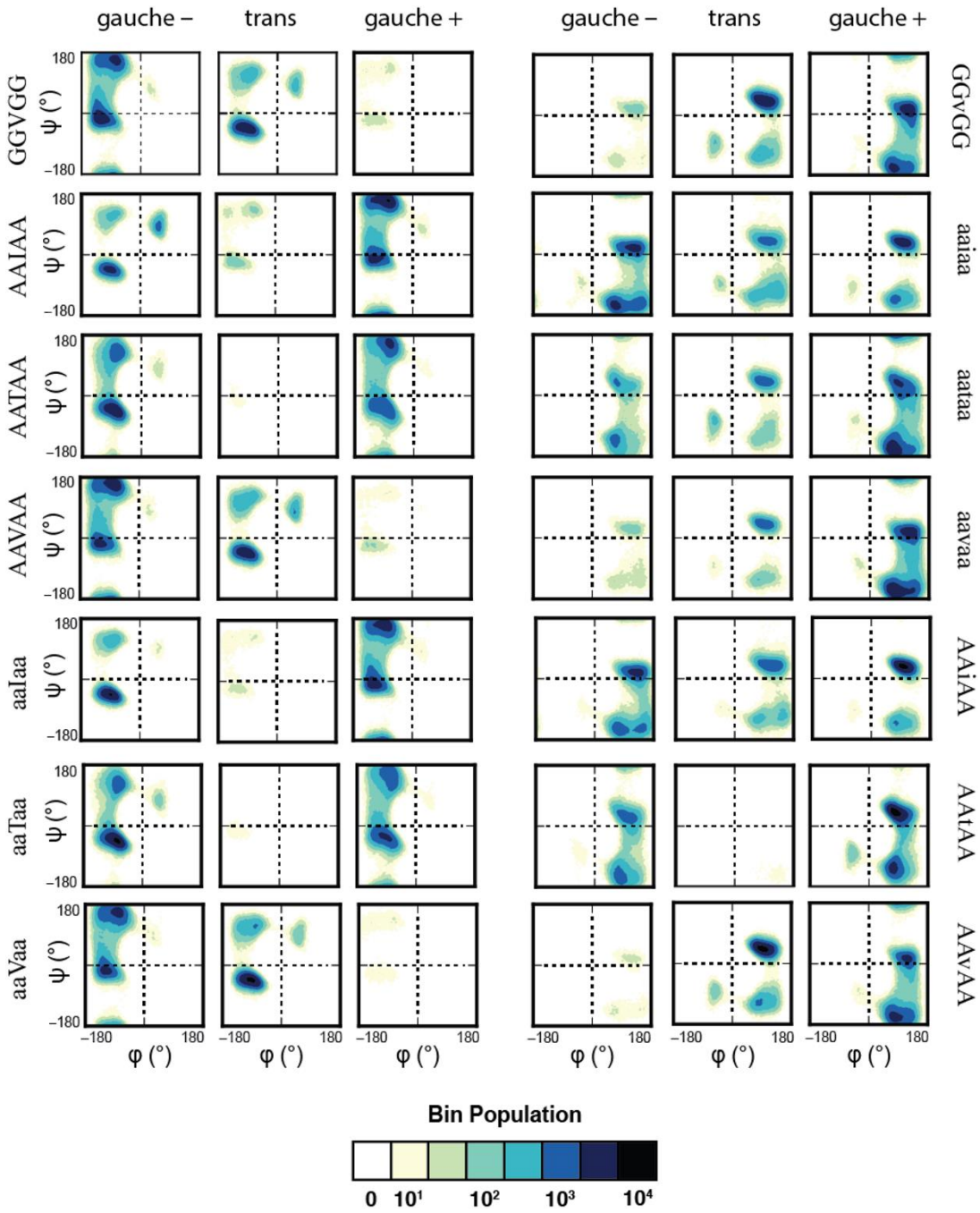


Figure 7.6. Side chain dependent Ramachandran plots reflect backbone-dependent sampling of rotameric states for β-branched residues.

Sets of three Ramachandran plots are shown for each β-branched guest residue in each of the four host peptides. Separate Ramachandran plots were constructed for each rotameric state of the side chain to illustrate the backbone-dependent sampling of rotameric states. Furthermore, when the β-

branched guest residues are placed within a chiral host system, asymmetry in the sampling of backbone conformations (Figure 7.5) results in asymmetric sampling of side chain conformations.

BIBLIOGRAPHY

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., and Lindahl, E. (2015) *SoftwareX*, 19–25.

Adams, D., Cauquil, C. and Labeyrie, C. (2017) *Curr. Opin. Neurol.*, 30, 481–489.

Ahmad, B., Muteeb, G., Alam, P., Varshney, A., Zaidi, N., Ishtikhar, M., Badr, G., Mahmoud, M. H., and Khan, R. H. (2015) *Int. J. Biol. Macromol.*, 75, 447–452.

Almeida, M.R., Alves, I. L., Terazaki, H., Ando,Y., and Saraiva, M. J. (2000) *Biochem. Biophys. Res. Commun.*, 270, 1024–1028.

Alonso, D. O. V. and Daggett, V. (2000) *Proc. Natl. Acad. Sci. 97*, 133-138.

Alshehri, B., D'Souza, D. G., Lee, J. Y., Petratos, S., and Richardson, S. J. (2015) *J. Neuroendocrinol.*, 27, 303–323.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) *J. Mol. Biol.*, 215, 403–410.

Ambrogelly, A., Palioura, S., and Söll, D. (2007) *Nat. Chem. Biol.*, 3, 29–35.

Arcus, V. L., Vuilleumier, S., Freund, S. M., Bycroft, M., and Fersht, A. R. (1994) *Proc. Natl. Acad. Sci. U. S. A.*, 91, 9412–9416.

Arcus, V. L., Vuilleumier, S., Freund, S. M., Bycroft, M., and Fersht, A. R. (1995) *J. Mol. Biol.*, 254, 305–321.

Armen, R. S., Alonso, D. O. V., and Daggett, V. (2004a) *Structure*, 12, 1847–1863.

Armen, R. S., DeMarco, M. L., Alonso, D. O. V., and Daggett,V. (2004b) *Proc. Natl. Acad. Sci. U. S. A.*, 101, 11622–11627.

Avbelj, F. and Baldwin, R. L. (2004) *Proc. Natl. Acad. Sci. U. S. A.*, 101, 10967–10972.

Avbelj, F., Grdadolnik, S. G., Grdadolnik, J., and Baldwin, R. L. (2006) *Proc. Natl. Acad. Sci. U. S. A.*, 103, 1272–1277.

Azinas, S., Colombo, M., Barbiroli, A., Santambrogio, C., Giorgetti, S., Raimondi, S., Bonomi, F., Grandori, R., Bellotti, V., Ricagno, S., and Bolognesi, M. (2011) *FEBS J.*, 278, 2349–2358.

Aznauryan, M., Delgado, L., Soranno, A., Nettels, D., Huang, J. R., Labhardt, A. M., Grzesiek, S., and Schuler, B. (2016) *Proc. Natl. Acad. Sci. U. S. A.*, 113, E5389-98.

Bahar, I. and Jernigan, R. L. (1996) *Fold Des*, 1, 357–370.

Bai, L., Sheeley, S., and Sweedler, J. V (2009) *Bioanal. Rev.*, 1, 7–24.

Banerjee, A., Bairagya, H. R., Mukhopadhyay, B. P., Nandi, T. K., and Bera, A. K. (2010) *Indian J. Biochem. Biophys.*, 47, 197–202.

Beauchamp, K. A., Lin, Y-S., Das, R., and Pande, V. S. (2012) *J. Chem. Theory Comput.*, 8, 1409-1414.

Beberg, A. L., Ensign, D. L., Jayachandran, G., Khaliq, S., and Pande, V. S (2009) *2009 IEEE International Symposium on Parallel & Distributed Processing*, IEEE, 1–8.

Beck, D. A. C., McCully, M. E., Alonso, D. O. V., and Daggett, V. (2000-2019) *il*mm -- *in lucem* molecular mechanics, Computer program, University of Washington, Seattle, WA.

Beck, D. A. C., Armen, R. S., and Daggett, V. (2005) *Biochemistry*, 44, 609–616.

Beck, D. A. C., Alonso, D. O. V, Inoyama, D., and Daggett, V. (2008) *Proc. Natl. Acad. Sci. U. S. A.*, 105 12259–12264.

Beck, D. A. C. and Daggett, V. (2004) *Methods*, *34*, 112–120.

Beck, D. A. C., Jonsson, A. L., Schaeffer, R. D., Scott, K. A., Day, R., Toofanny, R. D., Alonso, D. O. V., and Daggett, V. (2008) *Protein Eng. Des. Sel.*, 21, 353–368.

Bennion, B. J. and Daggett, V. (2003) *Proc. Natl. Acad. Sci. U. S. A.*, 100, 5142–5147.

Benson, N. C. and Daggett, V. (2008) *Protein Sci.*,17 2038–2050.

Benson, N. C. and Daggett, V. (2012) *J. Phys. Chem. B*, *116* (29), 8722–8731.

Benson, M. D., Buxbaum, J. N., Eisenberg, D. S., Merlini, G., Saraiva, M. J., Sekijima, Y., Sipe, J. D., and Westermark, P. (2018) *Amyloid*, 25, 215–219.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) *Nucleic Acids Res.*, 28, 235–242.

Bernardi, R. C., Cann, I., and Schulten, K. (2014) *Biotechnol. Biofuels*, *7*, 83.

Best, R. B., Buchete, N. V., and Hummer, G. (2008) *Biophys J*, 95, L07-9.

Bi, T. M. and Daggett, V. (2018) *Yale J. Biol. Med.*, 91, 247–255.

Blake, C. C. F., Geisow, M. J., Oatley ,S .J., Rérat, B. and Rérat, C. (1978) *J. Mol. Biol.*, 121, 339–356.

Blake, C. C. F., Swan, I. D. A., Rerat, C., Berthou, J., Laurent, A., and Rerat, B. (1971) *J. Mol. Biol.*, 61, 217–224.

Blasio, B. Di, Saviano, M., Fattorusso, R., Lombardi, A., Pedone, C., Valle, V., and Lorenzi, G.P. (1994) *Biopolymers*, 34, 1463–1468.

Bleem, A. and Daggett, V. (2017) *Biotechnol. Bioeng.*, 114, 7–20.

Bleem, A ., Francisco, R., Bryers, J. D., and Daggett, V. (2017) *npj Biofilms Microbiomes*. 3, 16.

Bond, C. J., Wong, K. B. B., Clarke, J., Fersht, A. R., and Daggett, V. (1997) *Proc. Natl. Acad. Sci. U. S. A.*, 94, 13409–13413.

Booth, D. R., Booth, S. E., Persey, M. R., Tan, S. Y., Madhoo, S., Pepys, M. B., and Hawkins, P. N. (1996) *Neuromuscul. Disord.*, 6, S20.

Borrel, G., Gaci, N., Peyret, P., O'Toole, P. W., Gribaldo, S., and Brugère, J. F. (2014) *Archaea*, 2014, 1–11.

Bowers, K. J., Dror, R. O., and Shaw, D. E (2007) *J. Comput. Phys.*, *221*, 303–329.

Bowler, B. E. (2012) *Curr. Opin. Struct. Biol.*, 22, 4–13.

Bowman, G.R. (2016) *Journal of Computational Chemistry*, 37, 558-566.

Bulawa, C. E., Connelly, S., Devit, M., Wang, L., Weigel, C., Fleming, J. A., Packman, J., Powers, E. T., Wiseman, R. L. Foss, T. R., Wilson, I. A., Kelly, J. W., and Labaudiniere, R. (2012) *Proc. Natl. Acad. Sci. U. S. A.*, 109, 9629–9634.

Caboche, S., Pupin, M., Leclère, V., Fontaine, A., Jacques, P., and Kucherov, G. (2008) *Nucleic Acids Res.*, 36, D326-31.

Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. J. (2005) *J. Comput. Chem.*, 26, 1668–1688.

Caves, L. S., Evanseck, J. D., and Karplus, M. (1998) *Protein Sci.*, 7, 649–666.

Cao, H., Sun, Y., Wang, L., Zhao, C., Fu, J. and Zhang, A. (2017) *Mol. Biosyst.*, 13, 736–749.

Ceruso, M. A., Grottesi, A., and Nola, A. Di (2003) *Proteins*, 50, 222–229.

Chandrasekaran, R. and Ramachandran, G. N. (1970) *Int. J. Protein Res.*, 2, 223-233.

Chaudhuri, R., Tang, S., Zhao, G., Lu, H., Case, D. A., and Johnson, M. E. (2011) *J. Mol. Biol.*, 414, 272–288.

Chen, I. J., Yin, D., and MacKerell, A. D. (2002) *J. Comput. Chem.*, 23, 199–213.

Childers, M. C. and Daggett, V. (2017) *Mol. Sys. Des. & Engr*, 2, 9–33.

Childers, M. C., Towse, C-L., and Daggett, V. (2016) *Protein Eng Des Sel*, 29, 271–280.

Childers, M. C. and Daggett, V. (2018) *J. Phys. Chem. B*, 122, 6673-6689.

Childers, M. C., Towse, C-L., and Daggett, V. (2018) *Protein Eng. Des. Sel.*, 31, 191–204.

Cianci, M., Folli, C., Zonta, F., Florio, P., Berni, R., and Zanotti, G. (2015) *Acta Crystallogr. Sect. D Biol. Crystallogr.*, 71, 1582–1592.

Clarke, N. D., Kissinger, C. R., Desjarlais, J., Gilliland, G. L., and Pabo, C. O. (1994) *Protein Sci.*,

3, 1779–1787.

Cohen, E. (2012) *Rambam Maimonides Med. J.*, 3, e0021.

Collet, O. (2011) *J. Chem. Phys.* 34, 085107.

Colon, W. and Kelly, J. W. (1992) *Biochemistry*, 31, 8654–8660.

Daggett, V., Li, A., Itzhaki, L. S., Otzen, D. E., and Fersht, A. R. (1996) *J. Mol. Biol.*, 257, 430–440.

Daggett, V. (2006) *Acc. Chem. Res.* 39, 594–602.

Daggett, V., Li, A., Itzhaki, L. S., Otzen, D. E., and Fersht, A. R. (1996) *J. Mol. Biol.* 257, 430–440.

Das, A. and Mukhopadhyay, C. (2009) *J. Phys. Chem. B*, 113, 12816–12824.

Das, R. and Baker, D. (2008) *Annu. Rev. Biochem.*, 77, 363–82.

Das, J. K., Mall, S. S., Bej, A., and Mukherjee, S. (2014) *Angew. Chem. Int. Ed. Engl.*, 53, 12781–12784.

Dasari, A. K. R., Hughes, R. M., Wi, S., Hung, I., Gan, Z., Kelly, J. W., and Lim, K. H. (2019) *Sci. Rep.* 9, 33.

Davis, P. J., Handwerger, B. S., and Gregerman, R. I. (1972) *J. Clin. Invest.*, 51, 515–521.

Day, R. and Daggett, V. (2005) *Protein Sci.*,14, 1242–1252.

Debiec, K. T., Gronenborn, A. M., and Chong, L. T (2014) *J. Phys. Chem. B*, 118, 6561-6569.

Dill, K. A. and Shortle, D. (1991) *Annu. Rev. Biochem.*, 60, 795–825.

Dodson, G. G., Lane, D. P., and Verma, C. S. (2008) *EMBO Rep.*, 9, 144–150.

Doncheva, N.T., Assenov, Y., Domingues, F.S., and Albrecht, M. (2012) *Nature Protocols*, 7, 670-685.

Duan, Y. and Kollman, P. A. (1998) *Science*, 282, 740–744.

Dunbrack, R. L. (2002) *Curr Opin Struct Biol*, 12, 431–440.

Dunbrack, R. L. and Karplus, M. (1993) *J. Mol. Biol.*, 230, 543–574.

Dunbrack, R. L. and Karplus, M. (1994) *Nat Struct Biol*, 1, 334–340.

Eker, F., Griebenow, K., and Schweitzer-Stenner, R. (2003) *J. Am Chem. Soc.*, 125, 8178-8185.

Eneqvist, T., Andersson, K., Olofsson, A., Lundgren, E., and Sauer-Eriksson, A. E. (2000) *Mol. Cell*, 6, 1207–1218.

Ercolessi, F. and Adams, J. B. (1994) *Europhys. Lett.*, 26, 583–588.

Espinoza-Fonseca, L. M. (2012) *Mol. Biosyst.*, 8, 237–46.

Esposito, G., Ricagno, S., Corazza, A., Rennella, E., Gumral, D., Mimmi, M. C., Betto, E., Pucillo, C. E., Fogolari, F., Viglino, P., Raimondi, S., Giorgetti, S., Bolognesi, B., Merlini, G., Stoppini, M., Bolognesi, M., and Bellotti, V. (2008) *J. Mol. Biol.*, 378, 885–895.

Fawzi, N. L., Chubukov, V., Clark, L. A., Brown, S., and Head-Gordon, T. (2005) *Protein Sci.*, 14, 993–1003.

Fersht, A. R. (1993) *FEBS Lett.*, 325, 5–16.

Fersht, A. R., Bycroft, M., Horovitz, A., Kellis, J. T., Matouschek, A., and Serrano, L. (1991) *Philos. Trans. R. Soc. B Biol. Sci.*, 332, 171–176.

Firestine, A. M., Chellgren, V. M., Rucker, S. J., Lester, T. E., and Creamer, T. P. (2008) *Biochemistry*, 47, 3216–3224.

Fiser, A., Do, R. K., and Sali, A. (2000) *Protein Sci.*, 9, 1753–1773.

Fitzkee, N. C., Fleming, P. J., and Rose, G. D. (2005) *Proteins*,58 852–4.

Flory J.P. (1969) *Statistical Mechanics of Chain Molecules*. Wiley, New York.

Foss, T.R., Wiseman, R.L., and Kelly, J.W. (2005) *Biochemistry*, 44, 15525–15533.

Frank, B. S., Vardar, D., Buckley, D. A., and McKnight, C. J. (2002) *Protein Sci.*, 11, 680–7.

Freddolino, P. L., Arkhipov, A. S., Larson, S. B., McPherson, A., and Schulten, K. (2006) *Structure*, 14, 437–449.

Frigerio, R., Fabrizi, G. M., Ferrarini, M., Cavallaro, T., Brighina, L., Santoro, P., Agostoni, E., Cavaletti, G., Rizzuto, N., and Ferrarese, C. (2004) *Amyloid*, 11, 121–124.

Fujii, N., Kaji, Y., and Fujii, N. (2011) *J. Chromatogr. B*, 879, 3141–3147.

Gattin, Z., Riniker, S., Hore, P. J., Mok, K. H., and Gunsteren, W. F. van (2009) *Protein Sci.*,18 2090–2099.

Gfeller, D., Michielin, O., and Zoete, V. (2012) *J Comput Chem*, 33, 1525–1535.

Gfeller, D., Michielin, O., and Zoete, V. (2013) *Nucleic Acids Res*, 41, D327-332.

Ghosh, R., Roy, S., and Bagchi, B. (2013) *J. Phys. Chem. B*, 117, 15625–15638.

Gianni, S., Guydosh, N. R., Khan, F., Caldas, T. D., Mayor, U., White, G. W. N., DeMarco, M. L., Daggett, V., and Fersht, A. R. (2003) *Proc. Natl. Acad. Sci. U. S. A.*, *100* (23), 13286–13291.

Gillespie, J. R. and Shortle, D. (1997) *J. Mol. Biol.*, 268, 158–169.

Glabe, C. G. and Kayed, R. (2006) Neurology, 66, S74–78.

Gordon, D. B., Hom, G. K., Mayo, S. L., and Pierce, N. A. (2003) *J. Comput. Chem.*, 24, 232–

243.

Graf, J., Nguyen, P. H., Stock, G., and Schwalbe, H. (2007) *J. Am. Chem. Soc.*, 129, 1179–1189.

Gregory, M. E., Carey, M., Hawkins, P. N., Banerjee, S., and Gillmore, J. D. (2008) *Br. J. Ophthalmol.*, 92, 34–35.

Grossfield, A. and Zuckerman, D. M. (2009) *Annu. Rep. Comput. Chem.*, 5, 23–48.

Gu, Y., Li, D. W., and Brüschweiler, R. (2015) *J. Chem. Theory Comput.*, 11, 1308–1314.

Gunsteren, W. F. van, Daura, X., Hansen, N., Mark, A., Oostenbrink, C., Riniker, S., and Smith, L. (2018) *Angew. Chemie Int. Ed.*, 57, 884–902.

Guranda, D. T., Volovik, T. S., and Svedas, V. K. (2004) *Biochem. (Moscow)*, 69, 1386–1390.

Haar, L., Gallagher, J. S., and Kell, G. S. (1984) National Standard Reference Data System (U.S.). *NBS/NRC Steam Tables : Thermodynamic and Transport Properties and Computer Programs for Vapor and Liquid States of Water in SI Units*, Washington, D.C.

Habeck, M., Rieping, W., and Nilges, M. (2005) *J. Magn. Reson.*, 177, 160–165.

Halgren, T. A. (1995) *Curr. Opin. Struct. Biol.*, 5, 205–210.

Halgren, T. A. (1996a) *J. Comput. Chem.*, 17, 490–519.

Halgren, T. A. (1996b) *J. Comput. Chem.*, 17, 520–552.

Halgren, T. A. (1996c) *J. Comput. Chem.*, 17, 553–586.

Halgren, T. A. (1996d) *J. Comput. Chem.*, 17, 616–641.

Halgren, T. A. and Nachbar, R. B. (1996) *J. Comput. Chem.*, 17, 587–615.

Hammarström, P., Jiang, X., Hurshman, A. R., Powers, E. T., and Kelly, J. W. (2002) *Proc. Natl. Acad. Sci. U. S. A.*, 99, 16427–16432.

Hammarström, P., Schneider, F., and Kelly, J. W. (2001) *Science*, 293, 2459–2462.

Hammarström, P., Sekijima, Y., White, J. T., Wiseman, R. L., Lim, A., Costello, C. E., Altland, K., Garzuly, F., Budka, H., and Kelly, J. W. (2003) *Biochemistry*, 42, 6656-6663.

Han, B., Liu, Y., Ginzinger, S. W., and Wishart, D. S. (2011) *J. Biomol. NMR*, 50, 43–57.

Harding, J., Skare, J., and Skinner, M. (1991) *BBA - Mol. Basis Dis.*, 1097, 183–186.

Harrison, H. H., Gordon, E. D., Nichols, W. C., and Benson, M. D. (1991) *Am. J. Med. Genet.* 39, 442–452.

Hayward,S. (2008) *Protein Sci.*, 10, 2219–2227.

Hayward, S. and Milner-White, J. E. (2011) *Proteins Struct. Funct. Bioinforma.*, 79, 3193–3207.

He, Y., Chen, Y., Alexander, P. A., Bryan, P. N., and Orban, J. (2012) *Structure*, 20, 283–291.

Hennebry, S. C. (2009) *FEBS J.*, 276, 5367–5379.

Hennig, M., Bermel, W., Spencer, A., Dobson, C. M., Smith, L. J., and Schwalbe, H. (1999) *J. Mol. Biol.*, 288, 705–723.

Henriques, J., Cragnell, C., and Skepö, M. (2015) *J Chem Theory Comput*., 11, 3420–3431.

Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. G. E. M. (1997) *J. Comput. Chem.*, 18, 1463–1472.

Hilaire, M. R., Ding, B., Mukherjee, D., Chen, J., and Gai, F. (2018) *J. Am. Chem. Soc.* 140, 629–635.

Hintze, B. J., Lewis, S. M., Richardson, J. S., and Richardson, D. C. (2016) *Proteins Struct. Funct. Bioinforma.*, 84, 1177–1189.

Holmgren, G., Hellman, U., Jonasson, J., Lundgren, H. E., Westermark, P., and Suhr, O. B. (2005) *Amyloid*, 12, 189–192.

Hopping, G., Kellock, J., Barnwal, R. P., Law, P., Bryers, J., Varani, G., Caughey, B., and Daggett, V. (2014) *Elife*, 3, e01681.

Horn, H. W., Swope, W. C., Pitera, J. W., Madura, J. D., Dick, T. J., Hura, G. L., and Head-Gordon, T. (2004) *J. Chem. Phys.*, 120, 9665–9678.

Hörnberg, A., Eneqvist, T., Olofsson, A., Lundgren, E., and Sauer-Eriksson, A. E. (2000) *J. Mol. Biol.*, 302, 649–669.

Hornstrup, L. S., Frikke-Schmidt, R., Nordestgaard, B. G., and Tybjcrg-Hansen, A. (2013) *Arterioscler. Thromb. Vasc. Biol.*, 33, 1441-1447.

Hu, J.-S. H. and Bax, A. (1997) *J. Am. Chem. Soc.*, 119, 6360-6368.

Hu, X., Wang, H., Ke, H., and Kuhlman, B. (2007) *Proc. Natl. Acad. Sci U. S. A.*, 104, 17668–17673.

Huang, J. and MacKerell, A. D. (2013) *J Comput Chem*, 34, 2135–2145.

Hung, L. W., Kohmura, M., Ariyoshi, Y., and Kim, S. H. (1999) *J. Mol. Biol.*, 285, 311–321.

Hurshman, A. R., White, J. T., Powers, E. T., and Kelly, J. W. (2004) *Biochemistry*, 43, 7365–7381.

Imperiali, B., Fisher, S. L., Moats, R. A., and Prins,T .J. (1992*) J. Am. Chem. Soc.*, 114, 3182–3188.

Imperiali, B., Moats, R. A., Fisher, S. L., and Prins, T. J. (2002). *J. Am. Chem. Soc.*, 114, 3182-88.

Jaarsveld, P. P. van, Edelhoch, H., Goodman, D. S., and Robbins, J. (1973) *J. Biol. Chem.*, 248, 4698–4705.

Jacobson, D. R. and Buxbaum, J. N. (1994) *Hum. Mutat.*, 3, 254–260.

Jacobson, D. R., Gorevic, P. D., and Buxbaum, J. N. (1990) *Am. J. Hum. Genet.*, 47, 127–136.

Janowski, P. A., Liu, C., Deckman, J., and Case, D. A. (2016) *Protein Sci.*, 25, 87–102.

Jha, A. K., Colubri, A., Zaman, M. H., Koide, S., Sosnick, T. R., and Freed, K. F. (2005) *Biochemistry*, 44, 9691–9702.

Jiang, F., Han, W., and Wu, Y. D. (2013) *Phys. Chem. Chem. Phys.*, 15, 3413–28.

Jiang, X., Smith, C. S., Petrassi, H. M., Hammarström, P., White, J. T., Sacchettini, J. C., and Kelly, J.W. (2001) *Biochemistry*, 40, 11442–11452.

João, M. and Saraiva, M. (1995) *Hum. Mutat.*, 5, 191–196.

Johnson, S. M., Connelly, S., Fearns, C., Powers, E. T., and Kelly, J. W. (2012) *J. Mol. Biol.*, 421, 185–203.

Jonsson, A. L., Scott, K. A., and Daggett, V. (2009) *Biophys. J.*, 97, 2958–2966.

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983) *J. Chem. Phys.*, 79, 926–935.

Joshi, M. D., Sidhu, G., Nielsen, J. E., Brayer, G. D., Withers, S. G., and McIntosh, L. P. (2001) *Biochemistry*, 40, 10115–10139.

Jung, Y. S., Oh, K. I., Hwang, G. S., and Cho, M. (2014) *Chirality*, 26, 83–92.

Kamatani, Y., Minakata, H., Kenny, P. T. M., Iwashita, T., Watanabe, K., Funase, K., Sun, X. P., Yongsiri, A., Kim, K. H., Novales-Li, P., Novales, E. T., Kanapi, C. G., Takeuchi, H., and Nomoto, K. (1989) *Biochem. Biophys. Res. Commun.*, 160, 1015–1020.

Kamp, M. W. V. D., Schaeffer, R. D., Jonsson, A. L., Scouras, A. D., Simms, A. M., Toofanny, R. D., Benson, N. C., Anderson, P. C., Merkley, E. D., Rysavy, S., Bromley, D., Beck, D. A. C., and Daggett, V. (2010) *Structure*, 18, 423–435.

Kanaya, S., Katsuda-Nakai, C., and Ikehara, M. (1991) *J. Biol. Chem.*, 266, 11621–11627.

Karplus, M. (1963) *J. Am. Chem. Soc.*, 85, 2870–2871.

Katayanagi, K., Miyagawa, M., Matsushima, M., Ishikawa, M., Kanaya, S., Nakamura, H., Ikehara, M., Matsuzaki, T., and Morikawa, K. (1992) *J. Mol. Biol.*, *223* (4), 1029–1052.

Kayed, R. and Glabe, C. G. (2006) *Methods Enzymol.*, 413, 326–344.

Kayed, R., Head, E., Thompson, J. L., McIntire, T. M., Milton, S. C., Cotman, C. W., and Glabe, C. G. (2003) *Science*, 300, 486–489.

Kazmirski S., Li A., and Daggett, V. (1999) *J. Mol. Biol.*, 290, 283–304.

Kazmirski, S. L. and Daggett, V. (1998) *J. Mol. Biol.*, 284, 793–806.

Kell, G. S. (1967) *J. Chem. Eng. Data*, 12, 66–69.

Kellock, J., Hopping, G., Caughey, B., and Daggett, V. (2016) *J. Mol. Biol.*, 428, 2317–2328.

Kelly, J. W. (1998) *Curr. Opin. Struct. Biol.*, 8, 101–106.

Kihara, M., Chatani, E., Iwata, K., Yamamoto, K., Matsuura, T., Nakagawa, A., Naiki, H., and Goto, Y. (2006) *J. Biol. Chem.*, 281, 31061–31069.

Kim, J. H., Oroz, J., and Zweckstetter, M. (2016) *Angew. Chemie - Int. Ed.*, 55, 16168–16171.

Kimanius, D., Pettersson, I., Schluckebier, G., Lindahl, E., and Andersson, M. (2015) *J. Chem. Theory Comput.*, 11, 3491–3498.

Kiss, G., Çelebi-Ölçüm, N., Moretti, R., Baker, D., and Houk, K. N. (2013) *Angew. Chem. Int. Ed. Engl.* 52 , 5700–5725.

Kitani, T., Yoda, K., Ogawa, T., and Okazaki, T. (1985) *J. Mol. Biol.*, 184, 45–52.

Knowles, T. P. J., Vendruscolo, M., and Dobson, C. M. (2014) *Nat. Rev. Mol. Cell Biol.*, 15, 384–396.

Korber, B. T., Farber, R. M., Wolpert, D. H., and Lapedes, A. S. (1993) *Proc. Natl. Acad. Sci.*, 90, 7176–7180.

Krivov, G. G., Shapovalov, M. V., and Dunbrack, R. L. (2009) *Proteins*, 77, 778–795.

Kruskal, J. B. (1964) *Psychometrika*, 29, 1–27.

Lai, Z., Colón, W. and Kelly, J. W. (1996) *Biochemistry*, 35, 6470–6482.

Lai, Z., McCulloch, J., Lashuel, H.A., and Kelly, J.W. (1997). *Biochemistry*, 36, 10230–10239.

Larson, S. M., Snow, C. D., Shirts, M., and Pande, V. S. (2009), *arXiv:*0901.866.

Lashuel, H. A., Lai, Z., and Kelly, J. W. (1998) *Biochemistry*, 37, 17851–17864.

Leach, B. I., Zhang, X., Kelly, J. W., Dyson, H. J., and Wright, P. E. (2018) *Biochemistry*. 57, 4421–4430.

Lee, E. H., Hsin, J., Sotomayor, M., Comellas, G., and Schulten, K. (2009) *Structure*, 17, 1295–1306.

Lee, M. and Na, S. (2016) *Chem. Phys. Chem.*, 17, 425–432.

Leone, S. and Picone, D. (2016) *PLoS One*, 11, e0158372.

Levitt, M., Hirshberg, M., Sharon, R., and Daggett, V. (1995) *Comput. Phys. Commun.*,91 215–231.

Levitt, M., Sharon, R., Laidig, K. E., and Daggett, V. (1997) *J. Phys. Chem. B*, 5647, 5051–5061.

Li, A. and Daggett, V. (1994) *Proc. Natl. Acad. Sci.*, 91, 10430–10434.

Li, A. and Daggett, V. (1996) *J. Mol. Biol.*, 257, 412–429.

Li, A. and Daggett, V. (1998) *J. Mol. Biol.*, 275, 677–694.

Li, W., Qin, M., Tie, Z., and Wang, W. (2011) *Phys. Rev. E*, 84, 041933.

Li, Y., Hou, N., Iok, U., and Leong, W. (2008) *Chinese Med. Sci. J.*, 23, 230–233.

Lim, K. H., Dasari, A. K. R., Hung, I., Gan, Z., Kelly, J. W., Wright, P. E., and Wemmer, D. E. (2016a) *Biochemistry*, 55, 5272–5278.

Lim, K. H., Dasari, A. K. R. R., Hung, I., Gan, Z., Kelly, J. W., and Wemmer, D. E. (2016b) *Biochemistry*, 55, 1941–1944.

Lim, K. H., Dyson, H. J., Kelly, J. W., and Wright, P. E. (2013) *J. Mol. Biol.*, 425, 977–988.

Lindorff-Larsen, K., Maragakis, P., Piana, S., Eastwood, M. P., Dror, R. O., and Shaw, D. E. (2012) *PLoS One*, 7, e32131.

Lindorff-Larsen, K., Piana, S., Dror, R. O., and Shaw, D. E. (2011) *Science*, 334, 517–520.

Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., and Shaw, D. E. (2010) *Proteins Struct. Funct. Bioinforma.*, 78, 1950–1958.

Liu, K., Cho, H. S., Hoyt, D. W., Nguyen, T. N., Olds, P., Kelly, J. W., and Wemmer, D. E. (2000a) *J. Mol. Biol.*, 303, 555–565.

Liu, K., Cho, H. S., Lashuel, H. A., Kelly, J. W., and Wemmer, D. E. (2000b) *Nat. Struct. Biol.*, 7, 754–757.

Lopes, P. E. M., Guvench, O., and MacKerell, A. D. (2015) *Methods Mol. Biol. Biology*, 1215, 47–71.

Ludvigsen, S., Andersen, K. V., and Poulsen, F. M. (1991) *J. Mol. Biol.*, 217, 731–736.

MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F.T.K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D.T., Prodhom, B., Reiher, W.E., Roux, B., Schlenkrich, M., Smith, J.C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D., and Karplus, M. (1998) *J. Phys. Chem. B*, 102, 3586–3616.

Mackerell, A. D., Feig, M., and Brooks, C. L. (2004) *J. Comput. Chem.*, 25, 1400–1415.

Makwana, K. M. and Mahalakshmi, R. (2016) *Biopolymers*, 106, 260–266.

Marcos-Alcalde, I., Setoain, J., Mendieta-Moreno, J. I., Mendieta, J., and Gómez-Puertas, P. (2015) *Bioinformatics*, btv453.

Maris, N. L., Shea, D., Bleem, A., Bryers, J. D., and Daggett,V. (2018) *Biochemistry*, 57, 507–510.

Massy, B. de, Fayet, O., Kogoma, T. (1984) *J. Mol. Biol.*, 178, 227–236.

Matsuzaki, T., Akasaki, Y., Olmer, M., Alvarez-Garcia, O., Reixach, N., Buxbaum, J. N., and Lotz, M. K. (2017) *Aging Cell*, 16, 1313–1322.

Mayor, U., Guydosh, N. R., Johnson, C. M., Grossmann, J.G., Sato S., Jas, G. S., Freund, S. M. V., Alonso, D. O. V., Daggett V., and Fersht A. R. (*2*003) *Nature,* 421, 863–867.

Mayor, U., Johnson, C. M., Daggett, V., and Fersht, A. R. (2000) *Proc. Natl. Acad. Sci. U. S. A.*, 97, 13518–13522.

McCammon, J. A., Gelin, B. R., and Karplus, M. (1977) *Nature*, 267, 585–590.

McCutchen, S. L. and Kelly, J. W. (1993) *Biochem. Biophys. Res. Commun.*, 197, 415–421.

McCutchen, S. L., Colon, W., and Kelly, J. W. (1993) *Biochemistry*, 32, 12119–12127.

McCutchen, S. L., Lai, Z., Miroy, G. J., Kelly, J. W., and Colon, W. (1995) *Biochemistry*, 34, 13527–13536.

Meng, W., Luan, B., Lyle, N., Pappu, R. V., and Raleigh, D. P. (2013) *Biochemistry*, 52 2662–2671.

Messih, M. A., Lepore, R., and Tramontano, A. (2015) *Bioinformatics*, 31, 3767–3772.

Mirtič, A., Merzel, F., and Grdadolnik, J. (2014) *Biopolymers*, 101, 814–818.

Mitchell, J. B. O. and Smith, J. (2003) *Proteins Struct. Funct. Bioinforma.*, 50, 563–571.

Miyamoto, S. and Kollman, P. A. (1992) *J. Comput. Chem.*, 13, 952–962.

Montecucchi, P. C., Castiglione, R. de, Piani, S., Gozzini, L., and Erspamer, V. (1981) *Int. J. Pept. Protein Res.*, 17, 275–83.

Myers, T. J., Kyle, R. A., and Jacobson, D. R. (1998) *Am. J. Hematol.*, 59, 249–251.

Nakamura, H., Oda, Y., Iwai, S., Inoue, H., Ohtsuka, E., Kanaya, S., Kimura, S., Katsuda, C., Katayanagi, K., and Morikawa, K. (1991) *Proc. Natl. Acad. Sci. U. S. A.*, 88, 11535–11539.

Nakazato, M., Kangawa, K., Minamino, N., Tawara, S., Matsuo, H., and Araki, S. (1984) *Biochem. Biophys. Res. Commun.*, 123, 921–928.

Nettleton, E. J., Sunde, M., Lai, Z., Kelly, J. W., Dobson, C. M., and Robinson, C. V. (1998) *J.*

Mol. Biol., 281, 553–564.

Nichols, W. C., Liepnieks, J. J., McKusick, V. A., and Benson, M. D. (1989) *Genomics*, 5, 535–540.

Novotny, M. and Kleywegt, G. J. (2005) *J. Mol. Biol.*, 347, 231–241.

O'Connell, T. M., Wang, L., Tropsha, A., and Hermans, J. (1999) *Proteins*, 36, 407–18.

Oda, Y., Iwai, S., Ohtsuka, E., Ishikawa, M., Ikehara, M., and Nakamura, H. (1993) *Nucleic Acids Res.*, 21, 4690–4695.

Oh, K. I., Jung, Y. S., Hwang, G. S., and Cho, M. (2012a) *J. Biomol. NMR*, 53, 25–41.

Oh, K. I., Lee, K. K., Park, E. K., Jung, Y., Hwang, G. S., and Cho, M. (2012b) *Proteins*, 80, 977–990.

Oh, K. I., Lee, K. K., Park, E. K., Yoo, D. G., Hwang, G. S., and Cho, M. (2010) *Chirality*, 22, E186–E201.

Okal, A., Cornillie, S., Matissek, S. J., Matissek, K. J., Cheatham, T. E., and Lim, C. S. (2014) *Mol. Pharm.*, 11, 2442–2452.

Ollivaux, C., Soyez, D., and Toullec, J. Y. (2014) *J. Pept. Sci.*, 20, 595–612.

Olofsson, A., Ippel, J. H., Wijmenga, S. S., Lundgren, E., and Öhman, A. (2004) *J. Biol. Chem.* 279, 5699–5707.

Oroz, J., Kim, J. H., Chang, B. J., and Zweckstetter, M. (2017) *Nat. Struct. Mol. Biol.*, 24, 407–413.

Ota, M., Isogai, Y., and Nishikawa, K. (2001) *Protein Eng.*, 14, 557–564.

Palaninathan, S.K. (2012) *Curr. Med. Chem.*, 19, 2324–2342.

Páll, S. and Hess, B.(2013) *Comput. Phys. Commun.*, 184, 2641–2650.

Pantelopulos, G. A., Mukherjee, S., and Voelz, V. A. (2015) *Proteins*, 83, 1665–1676.

Papaleo, E., Saladino, G., Lambrughi, M., Lindorff-Larsen, K., Gervasio, F. L., and Nussinov, R. (2016) *Chem. Rev.* 116, 6391–6423.

Pappu, R. V, Srinivasan, R., and Rose, G. D. (2000) *Proc. Natl. Acad. Sci. U. S. A.*, 97, 12565–12570.

Paranjapye, N. and Daggett, V. (2018) *J. Mol. Biol.* 430, 3764-3773.

Pardi, A., Billeter, M., and Wüthrich, K. (1984) *J. Mol. Biol.*, 180, 741–751.

Pauling, L. and Corey, R. B. (1951) *Proc. Natl. Acad. Sci. U. S. A.*, 37, 251–256.

Pavone, V., Lombardi, A., Saviano, M., NAstri, F., Zaccaro, L., Maglio, O., Pedone, C., Omote,

Y., Yamanaka, Y., and Yamada, T. (1998*) J. Pept. Sci.*, 4, 21-32.

Peacock, A. F. A., Stuckey, J. A., and Pecoraro, V. L. (2009) *Angew. Chemie Int. Ed.*, 48, 7371–7374.

Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham, T. E., DeBolt, S., Ferguson, D., Seibel, G., and Kollman, P. (1995) *Comput. Phys. Commun.*, 91, 1–41.

Pérez, C., Löhr, F., Rüterjans, H., and Schmidt, J. M. (2001) *J. Am. Chem. Soc.*, 123, 7081–7093.

Petsko, G. A. (1996) *Nat. Struct. Mol. Biol.*, 3, 565–566.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) *J. Comput. Chem.*, 25, 1605–1612.

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., and Schulten, K. (2005) *J. Comput. Chem.*, 26, 1781–1802.

Piana, S., Lindorff-Larsen, K., and Shaw, D. E. (2011) *Biophys. J.*, 100, L47–L49.

Planté-Bordeneuve, V., Kerschen, P., Bordeneuve, V., and Kerschen, P. (2013) *Handb. Clin. Neurol.*, 115, 643–658.

Porter, L. L. and Rose, G. D. (2011) *Proc. Natl. Acad. Sci. U. S. A.*, 108, 109–113.

Prinsen, B. H. C. M. and de Sain-Van Der Velden, M. G. (2004) *Clin. Chim. Acta*, 347, 1–14.

Prinz, J.-H., Wu, H., Sarich, M., Keller, B., Senne, M., Held, M., Chodera, J. D., Schütte, C., and Noé, F. (2011) *J. Chem. Phys.*, 134, 174105.

Radkov, A. D. and Moe, L. A. (2014) *Appl. Microbiol. Biotechnol.*, 98, 5363–5374.

Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. (1963) *J. Mol. Biol.*, 7, 95–99.

Rana, S., Kundu, B., and Durani, S. (2004) *Chem. Commun. (Camb).* 2462–2463.

Rana, S., Kundu, B., and Durani, S. (2005) *Chem. Commun. (Camb).* 207–209.

Ranløv, I., Alves, I. L., Ranløv, P. J., Husby, G., Costa, P. P., and Saraiva, M. J. (1992) *Am. J. Med.*, 93, 3–8.

Reixach, N., Deechongkit, S., Jiang, X., Kelly, J. W., and Buxbaum, J. N. (2004) *Proc. Natl. Acad. Sci. U. S. A.*, 101, 2817–2822.

Religa, T. L. (2008) *J. Biomol. NMR*, 40, 189–202.

Religa, T. L., Markson, J. S., Mayor, U., Freund, S. M. V., and Fersht, A. R. (2005) *Nature*, 437, 1053–1056.

Renfrew, P. D., Choi, E. J., Bonneau, R., and Kuhlman, B. (2012) *PLoS One*, 7, e32637.

Renfrew, P. D., Craven, T. W., Butterfoss, G. L., Kirshenbaum, K., and Bonneau, R. (2014) *J. Am.*

Chem. Soc., 136, 8772–8782.

Richardson, J. S. (1977) *Nature*, 268, 495–500.

Richardson, J. S. (1981) *Adv. Protein Chem.*, 34, 167-339.

Richardson, J. S. and Richardson, D. C. (2002) *Proc. Natl. Acad. Sci. U. S. A.*, 99, 2754–2759.

Rodrigues, J. R., Simões, C. J. V., Silva, C. G., and Brito, R. M. M. (2010) *Protein Sci.*, 19, 202–219.

Rodriguez-Granillo, A., Annavarapu, S., Zhang, L., Koder, R. L., and Nanda, V. (2011) *J. Am. Chem. Soc.*, 133, 18750–18759.

Roy, S. and Bagchi, B. (2014) *J. Phys. Chem. B*, 118, 5691–5697.

Ruberg, F. L., Judge, D. P., and Maurer, M. S. (2009) *J. Card. Fail.*, 15, 464.

Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. (1977) *J. Comput. Phys.*, 23, 327–341.

Rysavy, S. J., Beck, D. A. C., and Daggett, V. (2014) *Protein Sci.*, 23, 1584–1595.

Ryu, J., Lee, M., Cha, J., Laskowski, R. A., Ryu, S. E., and Kim, D. S. (2016) *Nucleic Acids Res.*, 44, W416-W423.

Saldaño, T. E., Zanotti, G., Parisi, G., and Fernandez-Alberti, S. (2017) *PLoS One*, 12, e0181019.

Salomon-Ferrer, R., Case, D. A., and Walker, R. C. (2013) *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 3, 198–210.

Salvi, F., Pastorelli, F., Plasmati, R., Bartolomei, I., Dall'Osso, D., and Rapezzi, C. (2012) *Amyloid*, 19, 58–60.

Sánchez, I. E., Tejero, J., Gómez-Moreno, C., Medina, M., and Serrano, L. (2006) *J. Mol. Biol.*, 363, 422–432.

Sant'Anna, R., Braga, C., Varejão, N., Pimenta, K. M., Grana-Montes, R., Alves, A., Cortines, J., Cordeiro, Y., Ventura, S., and Foguel, D. (2014) *J. Biol. Chem.*, 289, 28324–28337.

Saraiva, M. J. (2001) *Hum. Mutat.*, 17, 493–503.

Saraiva, M. J., Birken, S., Costa, P. P., and Goodman, D. S. (1984) *J. Clin. Invest.*, 74, 104–119.

Saraiva, M.J., Costa, P. P., and Goodman, D. S. (1983) *J. Lab. Clin. Med.*, 102, 590–603.

Sawle, L. and Ghosh, K. (2016) *J. Chem. Theory Comput.*, 12, 861–869.

Schaeffer, R. D., Jonsson, A. L., Simms, A. M., and Daggett, V. (2011) *Bioinformatics*, 27, 46–54.

Schmidt, J. M., Blümel, M., Löhr, F., and Rüterjans, H. (1999) *J. Biomol. NMR*, 14 , 1–12.

Schneider, F., Hammarström, P., and Kelly, J. W. (2001) *Protein Sci.*, 10, 1606–1613.

Schwarzinger, S., Kroon, G. J. A. A., Foss, T. R., Wright, P. E., and Dyson, H. J. (2000) *J. Biomol. NMR*, 18, 43–48.

Schweitzer-Stenner, R. (2009) *J. Phys. Chem. B*, 113, 2922–2932.

Schweitzer-Stenner, R., and Toal, S. E. (2014) *Phys. Chem. Chem. Phys.*, 16, 22527–22536.

Scolnik, Y., Portnaya, I., Cogan, U., Tal, S., Haimovitz, R., Fridkin, M., Elitzur, A. C., Deamer, D. W. and Shinitzky, M. (2005). *Phys. Chem. Chem. Phys.*, 8, 333-339.

Scott, K. A., Alonso, D. O. V, Pan, Y., and Daggett, V. (2006) *Biochemistry*, 45, 4153–4163.

Scouras, A. D. and Daggett, V. (2011) *Protein Sci.*, 20, 341–352.

Sekijima, Y. (2015) *J. Neurol. Neurosurg. Psychiatry*, 86, 1036–1043.

Sekijima, Y., Dendle, M. T., Wiseman, R .L., White, J. T., D'Haeze, W., and Kelly, J. W. (2006) *Amyloid*, 13, 57–66.

Sekijima, Y., Wiseman, R. L., Matteson, J., Hammarström, P., Miller, S. R., Sawkar, A. R., Balch, W. E,. and Kelly, J. W. (2005) *Cell*, 121, 73–85.

Sellers, B. D., Zhu, K., Zhao, S., Friesner, R. A., and Jacobson, M. P. (2008) *Proteins Struct. Funct. Genet.*, 72, 959–971.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) *Genome Res.*, 13, 2498–2504.

Sharma, B. and Asher, S. A. (2010) *J. Phys. Chem. B*, 114, 6661–6668.

Shaw, D. E., Grossman, J. P., Bank, J. A., Batson, B., Butts, J. A., Chao, J. C., Deneroff, M. M., Dror, R. O., Even, A., Fenton, C. H., Forte, A., Gagliardo, J., Gill, G., Greskamp, B., Ho, C. R., Ierardi, D. J., Iserovich, L., Kuskin, J. S., Larson, R. H., Layman, T., Lee, L.-S., Lerer, A. K., Li, C., Killebrew, D., Mackenzie, K. M., Mok, S. Y.-H., Moraes, M. A., Mueller, R., Nociolo, L. J., and Peticolas, J. L. (2014) *SC14: International Conference for High Performance Computing, Networking, Storage and Analysis*, 41–53.

Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., Bank, J. A., Jumper, J. M., Salmon, J. K., Shan, Y., and Wriggers, W. (2010) *Science*, 330, 341–346.

Shea, D., Hsu, C. C., Bi, T. M., Paranjapye, N., Childers, M., Cochran, J., Tomberlin, C. P., Wang, L., Paris, D., Zonderman, J., Varani, G., Link, C., Mullan, M., and Daggett, V. (2019) *Proc. Natl. Acad. Sci. U. S. A.*, 116, 8895–8900.

Shen, J. K. (2010) *Biophys. J.*, 99, 924–32.

Shen, Y., Koh, K. D., Weiss, B., and Storici, F. (2012) *Nat. Struct. Mol. Biol.*, 19, 98–104.

Sheridan, R. P., Lee, R. H., Peters, N., and Allen, L. C. (1979) *Biopolymers*, 18, 2451–2458.

Shikata, Y., Watanabe, T., Teramoto, T., Inoue, A., Kawakami, Y., Nishizawa, Y., Katayama, K., and Kuwada, M. (1995) *J. Biol. Chem.*, 270, 16719–16723.

Shimamoto, T., Shimada, M., Inouye, M., and Inouye, S. (1995) *J. Bacteriol.*, 177, 264–267.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J. Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011) *Mol. Syst. Biol.*, 7, 539.

Silva-Lucca, R. A., Andrade, S. S., Ferreira, R. S., Sampaio, M. U., and Oliva, M. L. V (2013) *Molecules*, 19, 233–246.

Simone, G. de, Lombardi, A., Galdiero, S., Nastri, F., Costanzo, L. Di, Gohda, S., Sano, A., Yamada, T., and Pavone, V. (2000) *Biopolymers*, 53, 182–188.

Sipe, J. D., Benson, M. D., Buxbaum, J. N., Ikeda, S., Merlini, G., Saraiva, M.J., and Westermark, P. (2016) *Amyloid*, 23, 209–213.

Sipe, J. D., Benson, M. D., Buxbaum, J. N., Ikeda, S., Merlini, G., Saraiva, M.J., and Westermark, P. (2014) *Amyloid*, 21, 221–224.

Smith, L. J., Sutcliffe, M. J., Redfield, C., and Dobson, C. M. (1991) *Biochemistry*, 30, 986–996.

Southwell, B. R., Tu, G. F., Duan,W., Achen, M., Harms, P. J., Aldred, A. R., Richardson, S. J., Thomas, T., Pettersson, T. M., and Schreiber, G. (1992) *Acta Med. Austriaca*, 19, 28–31.

Stafford, K. A., Trbovic, N., Butterwick, J. A., Abel, R., Friesner, R. A., and Palmer, A. G. (2015) *J. Mol. Biol.*, 427, 853–866.

Stanley, N., Esteban-Martín, S., and De Fabritiis, G. (2015) *Prog. Biophys. Mol. Biol.*, 119, 47–52.

Steward, R. E., Armen, R. S., and Daggett, V. (2008) *Protein Eng. Des. Sel.*, 21, 187–95.

Stollar, E. J., Mayor, U., Lovell, S. C., Federici, L., Freund, S. M. V., Fersht, A. R., and Luisi, B. F. (2003) *J. Biol. Chem.*, 278, 43699–43708.

Su, D., Li, Y. and Gladyshev, V. N. (2005) *Nucleic Acids Res.*, 33, 2486–2492.

Susuki, S., Sato, T., Miyata, M., Momohara, M., Suico, M. A., Shuto, T., Ando, Y., and Kai, H. (2009) *J. Biol. Chem.* 284, 8312–8321.

Tachibana, N., Tokuda, T., Yoshida, K., Taketomi, T., Nakazato, M., Li, F., Masuda, Y., and Ikeda, S. I. (1999) *Amyloid*, 6, 282–288.

Tamiola, K., Acar, B., and Mulder, F. A. A. (2010) *J. Am. Chem. Soc.*, 132, 18000–18003.

Terazaki, H., Ando, Y., Misumi, S., Nakamura, M., Ando, E., Matsunaga, N., Shoji, S., Okuyama, M., Ideta, H., Nakagawa, K., Ishizaki, T., Ando, M., and Saraiva, M. J. M. (1999) *Biochem. Biophys. Res. Commun.*, 264, 365–370.

Tischer, A. and Auton, M. (2013) *Protein Sci.*, 22, 1147–1160.

Toal, S. and Schweitzer-Stenner, R. (2014) *Biomolecules*, 4, 725–73.

Torii, H. (2008) *J. Phys. Chem. B*, 112, 8737–8743.

Torres, A. M., Tsampazi, C., Geraghty, D. P., Bansal, P. S., Alewood, P. F., and Kuchel, P. W. (2005) *Biochem. J.*, 391, 215–220.

Towse, C.-L., Hopping, G., Vulovic, I., and Daggett, V. (2014) *Protein Eng. Des. Sel.*, 27, 447–455.

Towse, C.-L., Rysavy, S. J., Vulovic, I. M., and Daggett, V. (2016a) *Structure*, 24, 187–199.

Towse, C. L., Vymetal, J., Vondrasek, J., and Daggett, V. (2016b) *Biophys. J.*, 110, 348–361.

Tsegaye, B., Mekasha, A., and Genet, S. (2017) *Biochem. Res. Int.*, 2017, 9196538.

Tung-Chen,Y. and Arnau, M.Á. (2018) *Acta Clin. Belg.*, 73, 460–461.

Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ionnidis, Y. E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., Nakatani, E., Schulte, C. F., Tolmie, D. E., Wenger, R. K., Yao, H., and Markley, J. L. (2008) *Nucleic Acids Res.*, 36, D402–D408.

Unan, H., Yildirim, A., and Tekpinar, M. (2015) *J Comput Aided Mol Des*, 29, 655–665.

Uversky, V. N. (2013) *Biochim. Biophys. Acta - Proteins Proteomics*, 1834, 932–951.

Vajpai, N., Gentner, M., Huang, J. rong, Blackledge, M., and Grzesiek, S. (2010) *J. Am. Chem. Soc.*, 132, 3196–3203.

Valiyaveetil, F. I., Sekedat, M., Mackinnon, R., and Muir, T. W. (2004) *Proc. Natl. Acad. Sci. U. S. A.*, 101, 17045–17049.

Vieira, M. and Saraiva, M. J. (2014) *Biomol. Concepts.* 5, 45–54.

Vijay-Kumar, S., Bugg, C. E., and Cook, W. J. (1987) *J. Mol. Biol.*, 194, 531–44.

Vlachakis, D., Bencurova, E., Papangelopoulos, N., and Kossida, S. (2014) *Adv. Protein Chem. Struct. Biol.*, 94, 269–313.

Voelz, V. A., Singh, V. R., Wedemeyer, W. J., Lapidus, L. J., and Pande, V. S. (2010) *J. Am. Chem. Soc.*, 132, 4702–4709.

Voter, A. F. (1998) *Phys. Rev. B*, 57, R13985–R13988.

Vuister, G. W. and Bax, A. (1993) *J. Am. Chem. Soc.*, 115, 7772–7777.

Vymetal, J. and Vondrasek, J. (2013) *J. Chem. Theor. Comp.* 9, 441–451.

Wagner, G., Pardi, A., and Wuethrich, K. (1983) *J. Am. Chem. Soc.*, 105, 5948–5949.

Waldher, B., Kuta, J., Chen, S., Henson, N., and Clark, A. E. (2010) *J. Comput. Chem.*, 31, 2307–2316.

Wei, C., Swan, A. J., Makover, H. B., and Kendall, P. C. (2017) *Child Psychiatry Hum. Dev.*, 48, 1001–1009.

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S., and Weiner, P. (1984) *J. Am. Chem. Soc.*, 106, 765–784.

Willis, B. T. M., Bertram T. M., and Pryor, A. W. (1975) *Thermal Vibrations in Crystallography*, Cambridge University Press.

Wong, K.-B. and Daggett, V. (1998) *Biochemistry.*, 37, 11182–11192.

Wong, K.-B., Clarke, J., Bond, C. J., Neira, J. L., Freund, S .M. V., Fersht, A. R., and Daggett, V. (2000) *J. Mol. Biol.* 296, 1257–1285.

Woutersen, S., Pfister, R., Hamm, P., Mu, Y., Kosov, D. S., and Stock, G. (2002) *J. Chem. Phys.*, 117, 6833–6840.

Xu, S. (2007) *Amyloid*. 14, 119–131.

Xue, Q., Zheng, Q. C., Zhang, J. L., Cui, Y. L., Chu, W. T., and Zhang, H. X. (2014) *Biophys. Chem*., 189, 8–15.

Yamazaki, T., Yoshida, M., Kanaya, S., Nakamura, H., and Nagayama, K. (1991) *Biochemistry*, 30, 6036–6047.

Yang, M., Lei, M., Bruschweiler, R., and Huo, S. (2005) *Biophys. J.*, 89, 433–443.

Yang, M., Lei, M., and Huo, S. (2003) *Protein Sci.*, 12, 1222–1231.

Yang, M., Lei, M., Yordanov, B., and Huo, S. (2006a) *J. Phys. Chem. B*, 110, 5829–5833.

Yang, M., Yordanov, B., Levy, Y., Brüschweiler, R., and Huo, S. (2006b) *Biochemistry*, 45, 11992–12002.

Yazaki, M. and Higuchi, K. (2014) *Brain Nerve*, 66, 817–826.

Yokoyama, T., Hanawa, Y., Obita, T., and Mizuguchi, M. (2017) *FEBS Lett.*, 591, 1862–1871.

Yokoyama, T., Mizuguchi, M., Nabeshima, Y., Kusaka, K., Yamada, T., Hosoya, T., Ohhara, T., Kurihara, K., Tomoyori, K., Tanaka, I., and Niimura, N. (2012) *J. Struct. Biol.*, 177, 283–290.

Young, T. S. and Schultz, P. G. (2010) *J. Biol. Chem.*, 285, 11039–11044.

Zanotti, G., Vallese, F., Ferrari, A., Menozzi, I., Saldaño, T. E., Berto, P., Fernandez-Alberti, S., and Berni, R. (2017) *PLoS One*, 12, e0187716.

Zawadzke, L. E. and Berg, J. M. (1993) *Proteins Struct. Funct. Genet.*, 16, 301–305.

Zhang, H., Neal, S., and Wishart, D. (2003) *J. Biomol. NMR*, 25, 173–195.

Zhou, A. Q., O'Hern, C. S., and Regan, L. (2011) *Protein Sci.*, 20, 1166–1171.

Zhou, N., Gupta, K., Yao, J., Ye, K., Panda, D., Giannakakou, P., and Joshi, H. C. (2002) *J. Biol. Chem.*, 277, 17476–17485.

Zimmermann, M. T. and Jernigan, R. L. (2012) *Entropy*, 14, 687–700.

Ziskin, J. L., Greicius, M. D., Zhu, W., Okumu, A. N., Adams, C. M., and Plowey, E. D. (2015) *Acta Neuropathol. Commun.*, 3, 43.

Zou, Q., Bennion, B. J., Daggett, V., and Murphy, K. P. (2002) *J. Am. Chem. Soc.*, 124, 1192–1202.

APPENDIX A: SUPPLEMENTAL FIGURES AND TABLES

Chapter 2 Supplemental Tables

Table A.2.1. Hydrogen Bond Occupancy for residue K17 of EnHD

Donor	Acceptor	AMBER	GROMACS	*il*mm	NAMD
17@N	13@O	96.9 %	97.7 %	97.5 %	98.6 %
17@N	14@O	13.5 %	14.2 %	24.4 %	0.0 %

Table A.2.2. Satisfaction of NOE restraints for EnHD

		Satisfaction (%)				Violations	
MD Package	**Run #**	**Short**	**Medium**	**Long**	**All**	**Total[a]**	**Severe[b]**
X-ray	**-**	**97.9**	**94.0**	**91.6**	**95.1**	**32 (0.9 Å)**	**4 (2.7 Å)**
AMBER	**1**	96.6	91.3	92.2	93.9	40 (0.7 Å)	3 (2.8 Å)
	2	95.5	92.3	91.6	93.6	42 (0.7 Å)	3 (2.5 Å)
	3	96.2	90.7	92.7	93.7	41 (0.7 Å)	2 (2.6 Å)
	Ensemble	**95.5**	**93.4**	**93.3**	**94.3**	**37 (0.5 Å)**	**2 (2.6 Å)**
GROMACS	**1**	97.9	95.1	93.3	95.9	27 (0.8 Å)	2 (2.7 Å)
	2	97.3	95.1	93.3	95.6	29 (1.1 Å)	6 (2.9 Å)
	3	97.3	93.9	92.7	95.1	32 (1.0 Å)	4 (3.3 Å)
	Ensemble	**97.9**	**95.6**	**94.9**	**96.5**	**23 (0.7 Å)**	**2 (2.6 Å)**
***il*mm**	**1**	98.3	92.9	86.5	93.6	42 (0.5 Å)	1 (2.7 Å)
	2	97.6	94.5	92.2	95.3	31 (0.4 Å)	2 (2.4 Å)
	3	97.3	95.1	91.1	95.0	33 (0.5 Å)	2 (2.4 Å)
	Ensemble	**98.9**	**96.2**	**93.3**	**96.7**	**22 (0.5 Å)**	**1 (2.7 Å)**
NAMD	**1**	98.3	89.1	91.6	93.9	40 (1.5 Å)	12 (3.7 Å)
	2	97.6	93.4	93.8	95.4	30 (0.9 Å)	5 (3.0 Å)
	3	98.3	94.0	95.0	96.2	25 (0.8 Å)	3 (3.0 Å)
	Ensemble	**98.3**	**92.3**	**95.0**	**95.7**	**28 (0.8 Å)**	**4 (3.0 Å)**

a. **The total number of violations (Mean violation distance)**
b. **The total number of violations > 2 Å (Mean violation distance)**

Table A.2.3. Satisfaction of NOE restraints for RNase H

		Satisfaction (%)				Violations	
Force Field	**Run #**	**Short**	**Medium**	**Long**	**All**	**Total[a]**	**Severe[b]**
X-ray	**-**	99.3	98.1	96.9	98.3	24 (0.7 Å)	2 (3.1 Å)
AMBER	**1**	97.1	94.9	96.7	97.1	41 (0.7 Å)	3 (3.1 Å)
	2	98.1	96.3	95.4	96.9	44 (0.8 Å)	4 (3.5 Å)
	3	98.2	96.3	95.2	96.9	44 (0.7 Å)	2 (4.0 Å)
	Ensemble	**98.2**	**95.8**	**96.2**	**97.2**	**40 (0.7 Å)**	**3 (3.5 Å)**
GROMACS	**1**	98.8	95.4	95.2	97.1	42 (0.6 Å)	2 (3.1 Å)
	2	98.8	96.3	96.1	97.5	36 (0.8 Å)	2 (3.7 Å)
	3	98.5	94.9	95.9	97.1	42 (0.9 Å)	4 (3.1 Å)
	Ensemble	**98.6**	**96.3**	**96.3**	**97.5**	**36 (0.6 Å)**	**2 (3.5 Å)**
***Il*mm**	**1**	97.9	92.6	85.3	92.9	102 (1.2 Å)	18 (3.1 Å)
	2	97.3	94.0	84.1	92.4	109 (1.1 Å)	17 (2.8 Å)
	3	97.3	91.7	88.2	93.3	95 (0.8 Å)	6 (3.6 Å)
	Ensemble	**98.2**	**94.4**	**89.9**	**94.8**	**74 (0.9 Å)**	**5 (2.9 Å)**
NAMD	**1**	98.6	95.4	95.4	97.1	40 (1.2 Å)	4 (6.6 Å)
	2	98.2	96.8	96.3	97.3	38 (0.9 Å)	4 (3.9 Å)
	3	98.5	95.4	95.7	97.1	42 (0.9 Å)	4 (4.7 Å)
	Ensemble	**98.9**	**96.3**	**95.9**	**97.5**	**36 (0.9 Å)**	**4 (4.3 Å)**

a. **The total number of violations (Mean violation distance)**
b. **The total number of violations > 2 Å (Mean violation distance)**

Table A.2.4. Satisfaction of NOE restraints associated with L26 of EnHD

NOE	Atom 1	Atom 2	X-Ray[a]	NMR[a]	AMBER[a]	GROMACS[a]	*il*mm[a]	NAMD[a]
324	26@HD11/2/3	32@H	0	0	0	0	0	0
207	19@HB2/3	26@HD21/2/3	0	1	1	1	1	0
288	23@HB2/3	26@HD21/2/3	0	1	1	0	1	0
289	23@HD21/2	26@HD21/2/3	0	1	1	0	0	0
309	25@H	26@HG	0	1	0	0	1	0
329	26@HD11/2/3	46@HA	0	1	1	1	1	1
330	26@HD11/2/3	46@H	0	1	0	1	0	0
334	26@HD11/2/3	49@H	0	1	0	1	1	0
335	26@HD21/2/3	27@H	0	1	1	1	1	0
336	26@HD21/2/3	30@HD2/3	0	1	1	1	1	0
338	26@HD21/2/3	30@H	0	1	0	0	0	0
318	26@HD11/2/3	30@H	1	0	1	1	0	1
327	26@HD11/2/3	34@H	1	0	1	1	0	1
347	26@HD21/2/3	46@HA	1	0	1	1	1	1
348	26@HD21/2/3	48@HE3	1	0	0	0	0	0
315	26@HB2/3	31@H	1	1	1	1	1	0
# of L26 NOEs not satisfied			11	5	6	6	7	12

a. A value of 0 indicates that the specified NOE restraint is not satisfied. A value of 1 indicates that the specified NOE restraint is satisfied.

Table A.2.5. Populations of rotameric states sampled by residue L26 of EnHD

Rotamer	1ENH	2JWT	AMBER	GROMACS	*il*mm	NAMD
t, *g*+	100 %	0 %	71 %	67 %	44 %	97 %
g-, *t*	0 %	100 %	20 %	33 %	54 %	3 %
Other	0 %	0 %	9 %	< 0.1 %	2 %	< 0.1 %

Table A.2.6. Lifetimes associated with the primary rotameric states of residue L26 of EnHD

Rotamer	AMBER	GROMACS	*il*mm	NAMD
t, *g*+	208 ps	16,152 ps	319 ps	22,787 ps
g-, *t*	1,386 ps	8,570 ps	158 ps	615 ps

Table A.2.7. Effects of empirically derived parameter sets on the level of agreement between MD-derived and experimental coupling constants

		Parameter Set						
		Habeck	**Hu**	**Ludvig**	**Pardi**	**Schmidt**	**Smith**	**Vuister**
AMBER	**RMSD**	0.90	0.90	0.97	0.93	1.21	1.01	0.97
	R	0.84	0.84	0.84	0.83	0.84	0.83	0.83
GROMACS	**RMSD**	0.80	0.82	0.75	0.85	0.84	1.04	1.01
	R	0.89	0.89	0.89	0.89	0.89	0.88	0.88
ilmm	**RMSD**	0.98	0.99	1.02	1.03	1.16	1.13	1.11
	R	0.801	0.80	0.80	0.79	0.82	0.78	0.77
NAMD	**RMSD**	0.85	0.85	0.94	0.86	1.24	0.91	0.88
	R	0.86	0.86	0.86	0.85	0.86	0.85	0.85
X-Ray	**RMSD**	0.91	0.92	0.93	0.90	1.19	1.00	1.00
	R	0.84	0.84	0.84	0.83	0.84	0.83	0.83

Table A.2.8. Hydrogen bond occupancy of residue N41 of EnHD

Donor	Acceptor	AMBER	GROMACS	*il*mm	NAMD
45@N	41@O	80.9 %	56.9 %	69.3 %	76.7 %
44@N	41@O	26.4 %	34.5 %	34.6 %	13.7 %
44@N	41@OD1	44.2 %	28.2 %	45.7 %	22.1 %
43@N	41@OD1	26.4 %	36.5 %	37.8 %	10.2 %

Table A.2.9. Comparison of ϕ– and S- values for EnHD transition state

a. R = Pearson's correlation coefficient

			S-Value											
			AMBER			GROMACS			*il*mm			NAMD		
Mutant	**Location**	ϕ_F	**Run 1**	**Run 2**	**Run 3**	**Run 1**	**Run 2**	**Run 3**	**Run 1**	**Run 2**	**Run 3**	**Run 1**	**Run 2**	**Run 3**
F8A	N-term.	0.42	0.45	0.24	0.20	0.53	0.83	0.64	0.37	0.52	0.45	0.22	0.38	0.62
L13A	HI	0.51	0.84	0.45	0.87	0.55	0.56	0.57	0.64	0.63	0.61	0.73	0.66	0.41
A14G	HI	0.79	0.99	0.55	0.91	0.69	0.77	0.72	0.97	0.84	0.88	0.94	0.91	0.73
L16V	HI	0.39	0.94	0.70	0.82	0.55	0.70	0.86	0.62	0.38	0.55	0.53	0.83	0.74
F20A	HI	0.36	0.70	0.29	0.35	0.56	0.58	0.44	0.43	0.28	0.30	0.40	0.40	0.72
Y25G	Loop	0.28	0.62	0.23	0.18	0.58	0.65	0.39	0.10	0.12	0.13	0.50	0.13	0.49
A25G	Loop	0.17	0.62	0.23	0.18	0.58	0.65	0.39	0.10	0.12	0.13	0.50	0.13	0.49
L26A	Loop	0.46	0.56	0.29	0.02	0.54	0.62	0.54	0.53	0.14	0.50	0.56	0.88	1.00
L38A	HII	0.48	0.82	0.73	0.38	0.41	0.36	0.38	0.68	0.46	0.43	0.34	0.65	0.47
L38V	HII	0.83	0.82	0.73	0.38	0.41	0.36	0.38	0.68	0.46	0.43	0.34	0.65	0.47
G39A	Turn	0.92	0.66	0.46	0.48	0.67	0.87	0.94	0.73	0.16	0.75	0.42	0.74	0.35
L40A	Turn	0.95	0.83	0.63	0.75	0.84	1.02	0.81	0.57	0.56	0.59	0.86	0.77	0.66
A43G	HIII	1.05	0.87	0.83	0.91	0.87	0.76	0.77	0.82	0.84	0.81	0.63	0.71	0.82
I45V	HIII	0.69	0.91	0.48	0.71	0.72	0.48	0.73	0.56	0.54	0.58	0.80	0.71	0.66
A54G	HIII	0.62	0.02	0.18	0.49	0.52	0.88	0.64	0.37	0.15	0.52	0.23	0.42	0.78
R[a]	-	-	0.26	0.61	0.58	0.60	0.31	0.59	0.73	0.56	0.79	0.35	0.62	0.07

Table A.2.10. χ^2 values calculated for EnHD and RNase H

Protein	Data Type	AMBER	GROMACS	*il*mm	NAMD
RNase H	**CS (N)**	1.51	1.25	2.47	1.35
	CS (Cα)	0.75	0.66	1.68	0.61
	CS (Hα)	1.58	1.05	2.03	1.16
	CS (H)	1.40	1.23	1.94	1.15
	S^2	0.99	0.80	1.20	0.81
	Total	6.23	5.00	9.32	5.07
EnHD	**CS (N)**	1.27	1.18	1.50	1.09
	CS (C)	0.85	1.04	0.59	0.63
	CS (Cα)	0.41	0.42	0.52	0.28
	CS (Cβ)	0.53	0.71	0.99	0.50
	CS (Hα)	1.31	1.38	2.37	1.59
	CS (H)	1.72	1.55	1.38	1.35
	$^3J_{HN,H\alpha}$	0.89	1.00	1.12	0.78
	S^2	0.45	0.38	0.41	0.60
	Total	7.43	7.67	8.88	6.82

Chapter 2 Supplemental Figure Captions

Figure A.2.1 Core C_α RMSD. C_α RMSD versus time for MD simulations of the engrailed homeodomain (A) and ribonuclease H (B).

Figure A.2.2 Minor excursions from the crystal structure conformation of EnHD. Simulations of EnHD in all force fields experienced some level of departure from the native state. For *il*mm, this corresponded to a rotation of HIII. For AMBER, GROMACS, and NAMD, this corresponded to partial unfolding of the C-terminus of HIII.

Figure A.2.3 HI-HII loop dynamics in EnHD. Stollar *et al.* have demonstrated that the HI-HII undergoes conformational exchange on timescales greater than probed by these simulations. Both *il*mm and AMBER demonstrated small perturbations to the loop structure that may precede a larger conformational change.

Figure A.2.4 Core C_α RMSF of EnHD. Larger fluctuations in the HII-HIII loop (see arrow) contributed to the lower correspondence between RMSF and B-Factors for GROMACS and NAMD.

Figure A.2.5 MD simulations yield structures with chemical shifts consistent with experiment. A comparison of MD-derived (x-axis) and experimental (y-axis) chemical shifts shows that all force fields can reproduce experimentally determined chemical shifts for EnHD.

Figure A.2.6 MD simulations yield structures with chemical shifts consistent with experiment. A comparison of MD-derived (x-axis) and experimental (y-axis) chemical shifts shows that all force fields can reproduce experimentally determined chemical shifts for RNase H.

Figure A.2.7 Force field dependent agreement of K17 chemical shifts. Each plot shows the experimental and MD-derived chemical shifts of K17 for a given nucleus type. In all plots, the red line denotes the value obtained in solution and the bars correspond to the values obtained from SHIFTX2 for the X-ray crystal structure (black) and MD simulations. For many nuclei (N, H, H_α, and C_β), MD produced chemical shifts that were more consistent with the experimental data than the crystal structure alone.

Figure A.2.8 Force field dependent agreement of the $J_{HN,H\alpha}$ coupling constant of N41. The MD-derived distributions for the N41 $J_{HN,H\alpha}$ coupling constant were force field dependent. (A) For each distribution in panel A, the vertical black line corresponds to the experimentally determined value, the vertical red line corresponds to the value inferred from the X-ray structure, and the colored vertical line corresponds to the average MD-derived value. (B) Ramachandran plots for residue N41 in each of the force fields are shown, with the value in the X-ray structure denoted by a red point. Better agreement with the experimentally determined coupling constant was associated with the population of more positive values of ψ (see *il*mm and GROMACS).

Figure A.2.9 EnHD Order Parameter Absolute Errors The absolute value of the error between simulation and experiment is shown as a function of residue number for the four different MD/force field combinations.

Figure A.2.10 Selection of putative transition state ensembles via conformational clustering. (A) First, an *n* x *n* C_α RMSD matrix is constructed (where n = the number of time points in the simulation). In total, this matrix describes the conformational similarity among all conformations in the simulation. (B) This matrix is then reduced into three-dimensional space and the transition state ensemble is chosen to be 5 structures surrounding the first major conformational change that occurs during the simulation.

Figure A.2.11 The effect of ensemble sub-sampling on chemical shift agreement. We calculated the chemical shifts for the N, Ca, Ha, and H nuclei of the first 100 nanoseconds of run 1 of RNase H using the AMBER software package at three different granularities: 1 ps (every frame was used in the chemical shift calculations), 10 ps (every tenth frame), and 100 ps (every 100^{th} frame). The different granularities produced little variation in the final predicted chemical shifts and show that 100 ps granularity is sufficient for the calculation of chemical shifts.

Figure A.2.12 Distribution of the absolute errors between the experimental and MD-derived values for 8 observables. In each plot, the absolute values of the difference between the MD and experimental values for a specific data type have been allocated into 25 bins. The threshold for excluding data for calculation of the χ^2 value was set at twice the mean value.

Chapter 2 Supplemental Figures

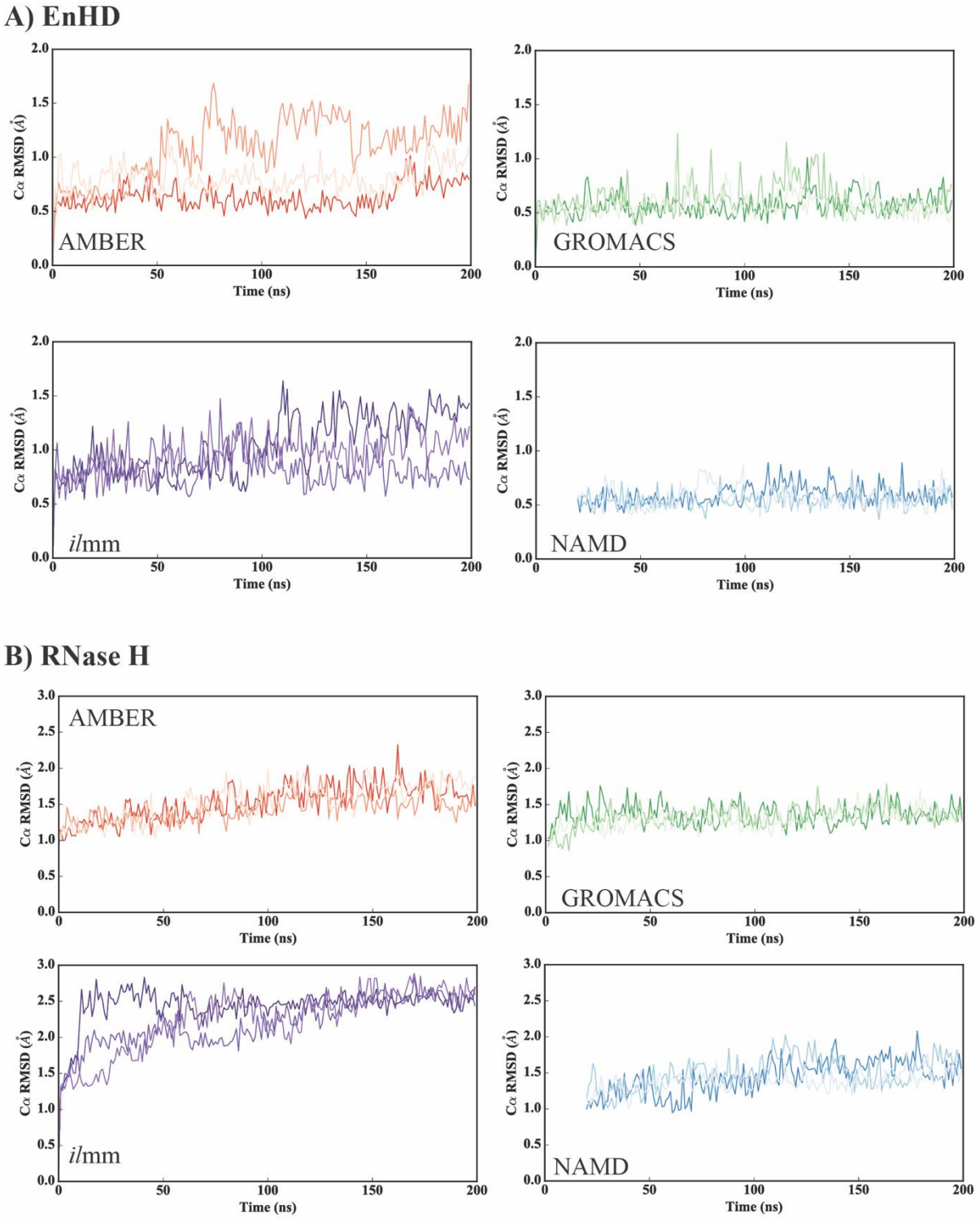


Figure A.2.1

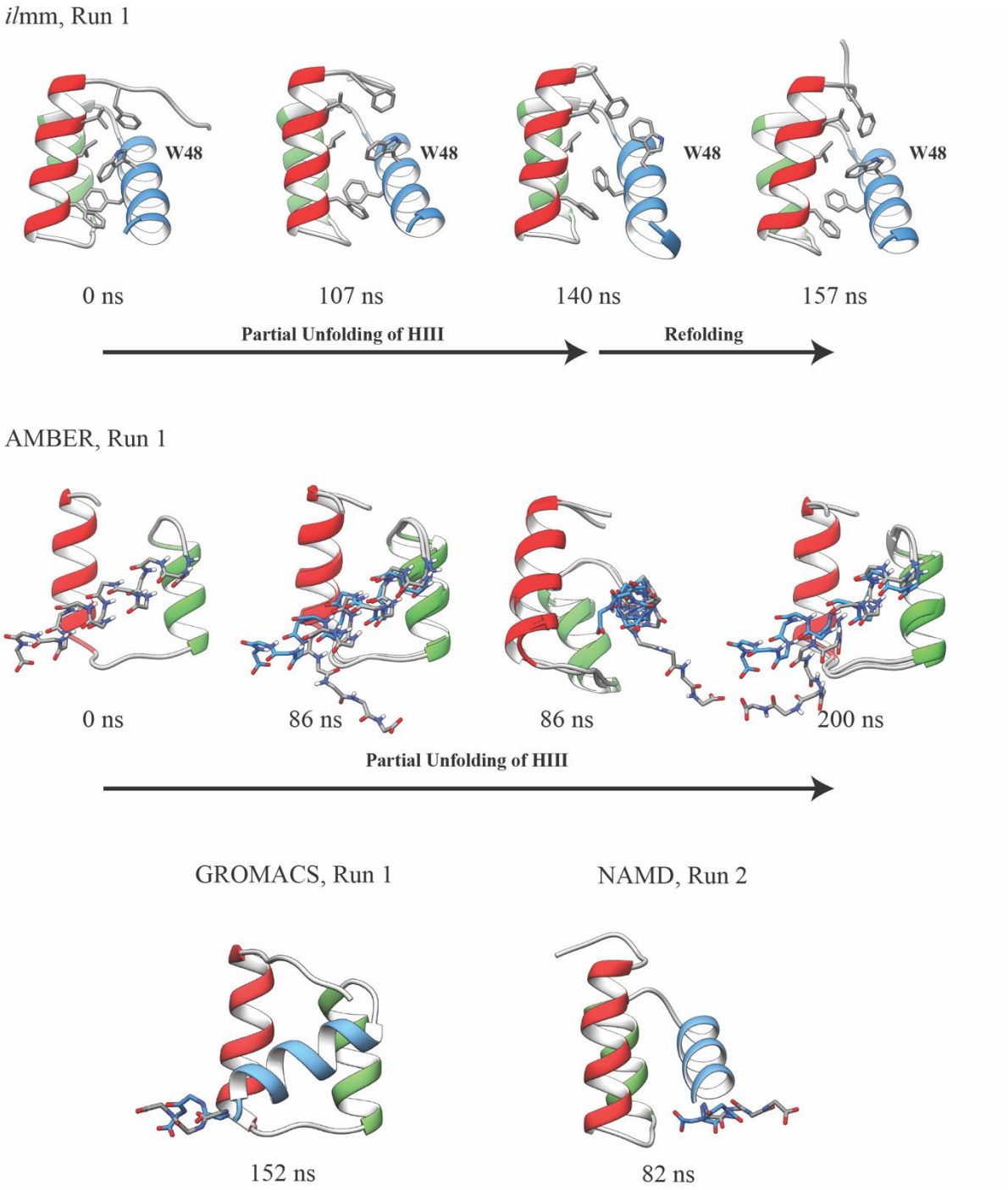
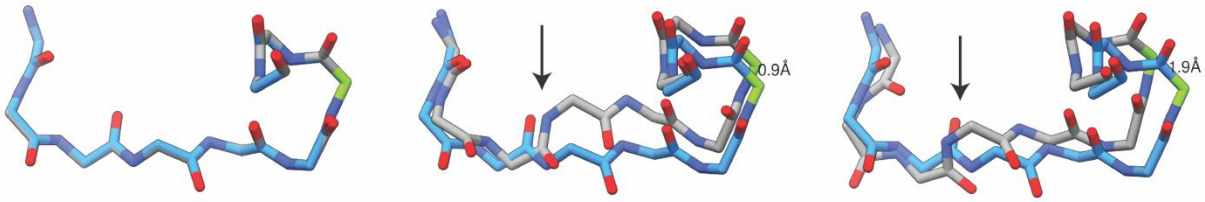

*il*mm, Run 1
W48
W48
W48
W48
0 ns
107 ns
140 ns
157 ns
Partial Unfolding of HIII
Refolding
AMBER, Run 1
0 ns
86 ns
86 ns
200 ns
Partial Unfolding of HIII
GROMACS, Run 1
NAMD, Run 2
152 ns
82 ns

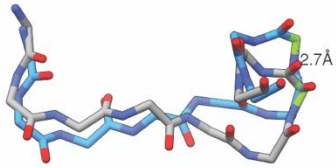
Figure A.2.2

*il*mm, Run 3



0 ns 61 ns 199 ns

AMBER, Run 2



199 ns

Figure A.2.3

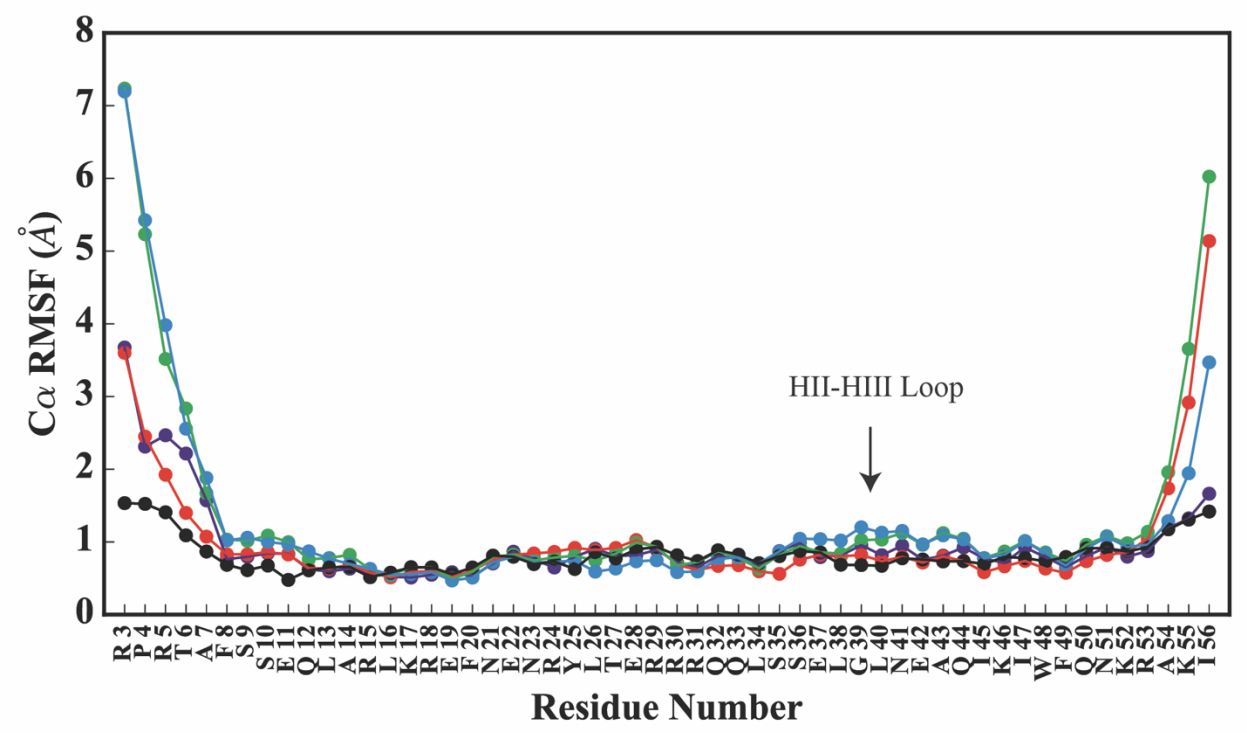

Cα RMSF (Å)
0
1
2
3
4
5
6
7
8
HII-HIII Loop
R 3
P 4
R 5
T 6
A 7
F 8
S 9
S 10
E 11
Q 12
L 13
A 14
R 15
L 16
K 17
R 18
E 19
F 20
N 21
E 22
N 23
R 24
Y 25
L 26
T 27
E 28
R 29
R 30
R 31
Q 32
Q 33
L 34
S 35
S 36
E 37
L 38
G 39
L 40
N 41
E 42
A 43
Q 44
I 45
K 46
I 47
W 48
F 49
Q 50
N 51
K 52
R 53
A 54
K 55
I 56
Residue Number

Figure A.2.4

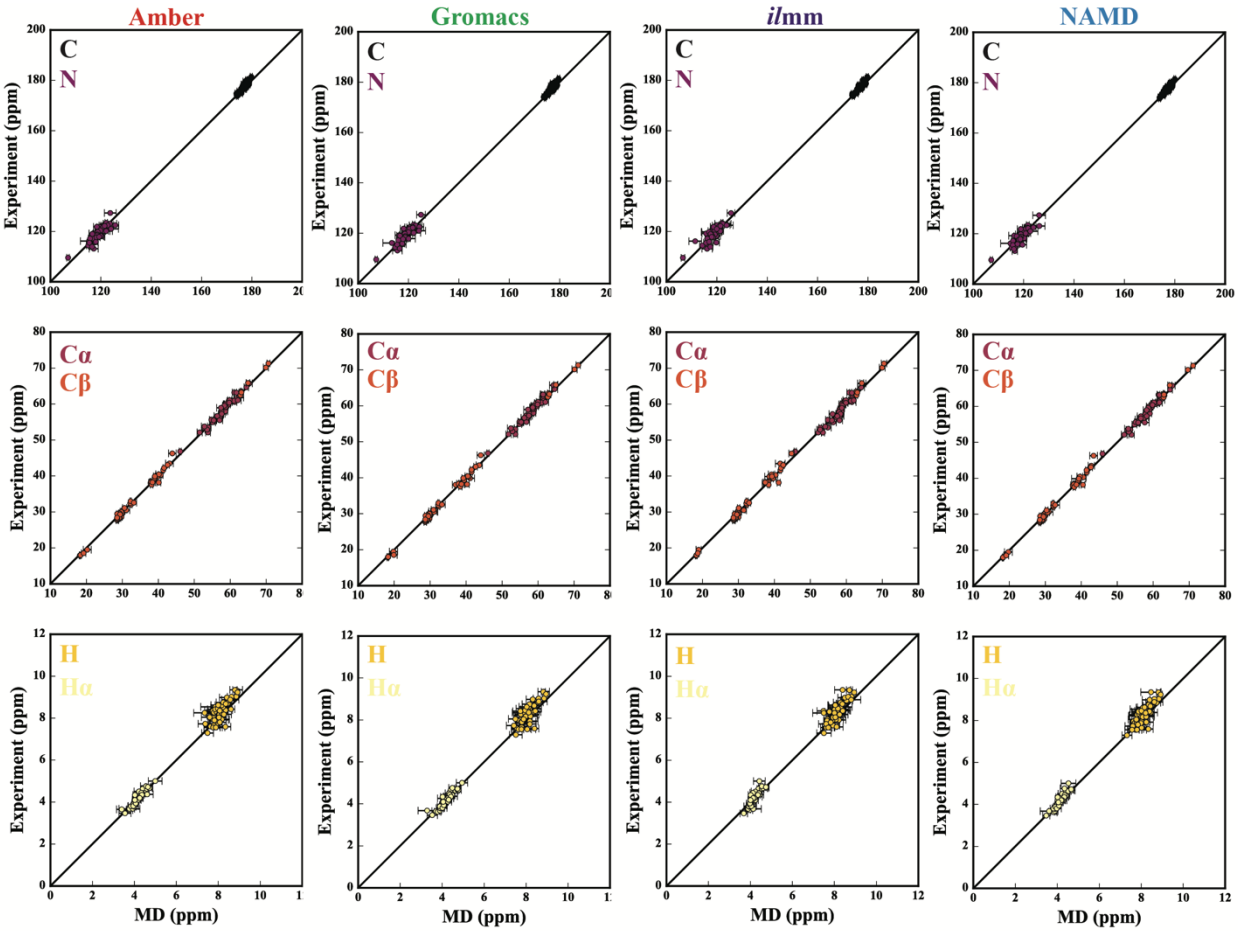


Figure A.2.5

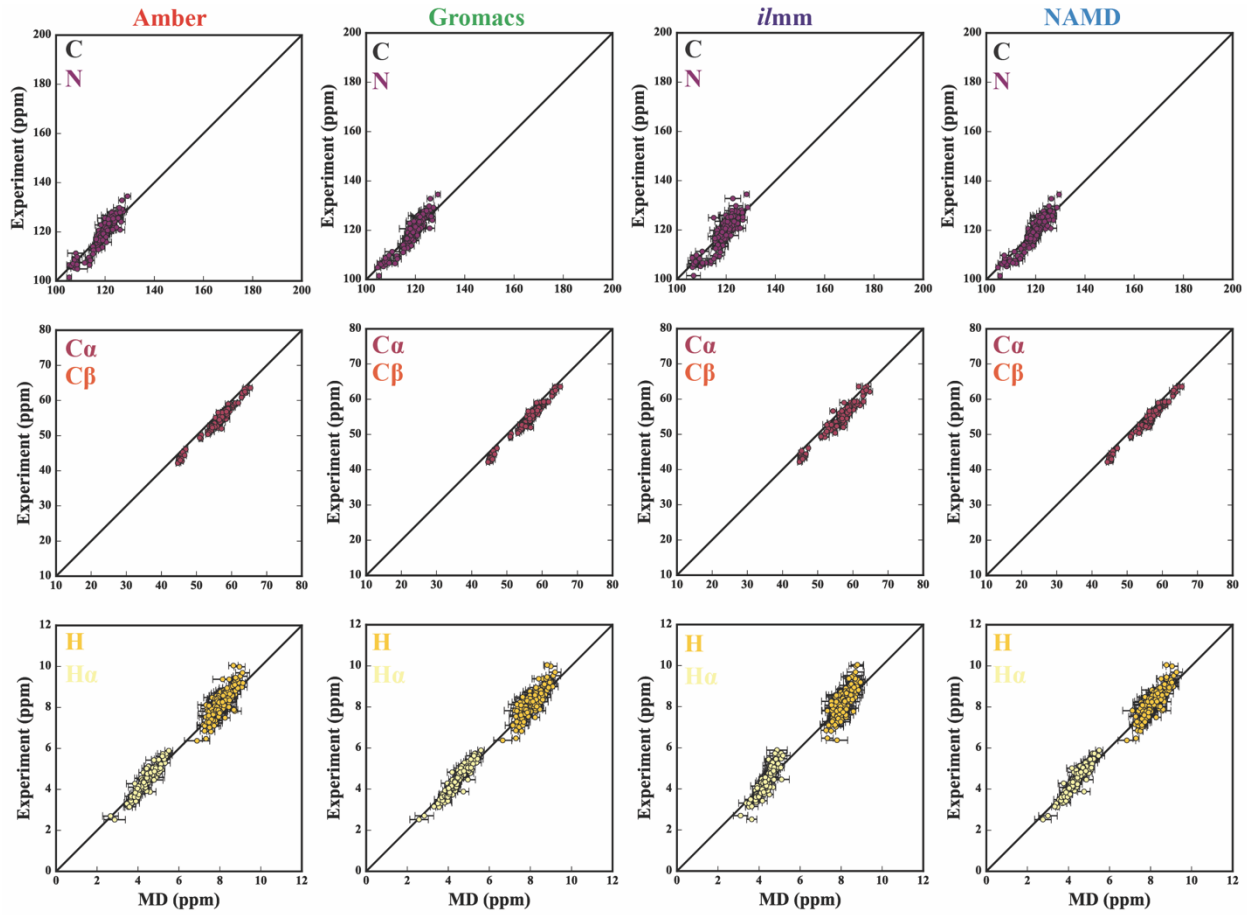


Figure A.2.6

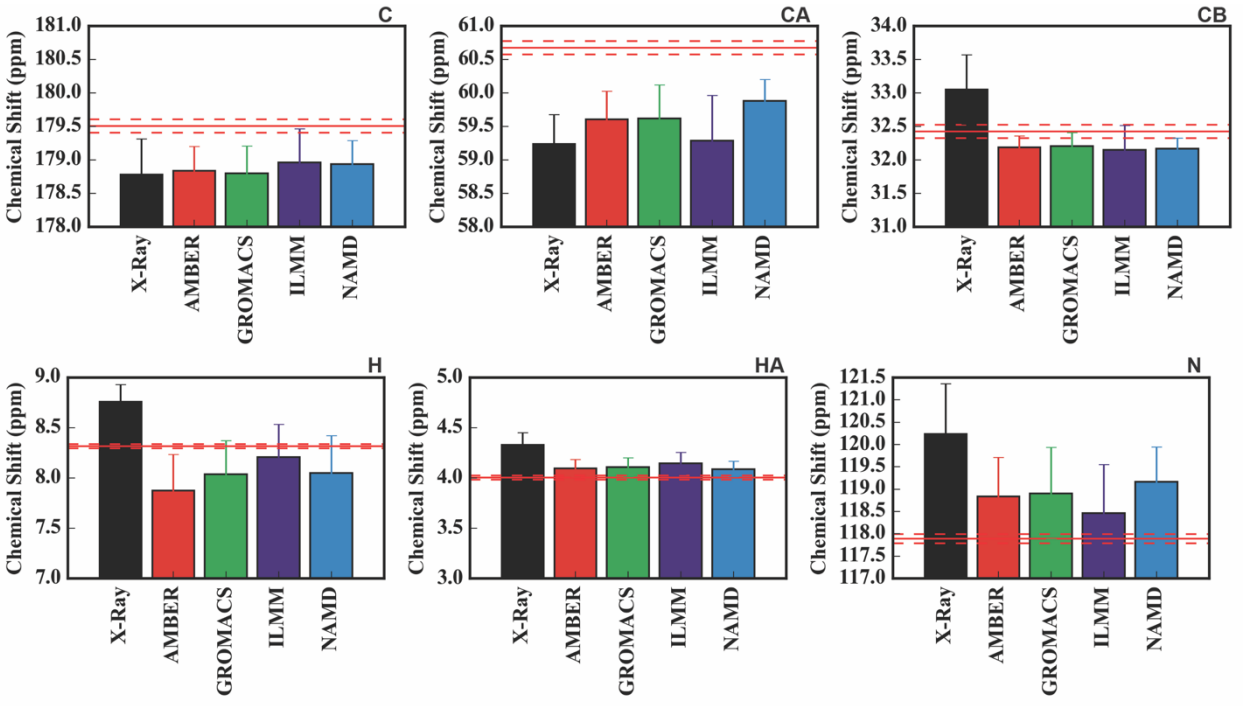

C
Chemical Shift (ppm)
181.0
180.5
180.0
179.5
179.0
178.5
178.0
X-Ray
AMBER
GROMACS
ILMM
NAMD
CA
Chemical Shift (ppm)
61.0
60.5
60.0
59.5
59.0
58.5
58.0
X-Ray
AMBER
GROMACS
ILMM
NAMD
CB
Chemical Shift (ppm)
34.0
33.5
33.0
32.5
32.0
31.5
31.0
X-Ray
AMBER
GROMACS
ILMM
NAMD
H
Chemical Shift (ppm)
9.0
8.5
8.0
7.5
7.0
X-Ray
AMBER
GROMACS
ILMM
NAMD
HA
Chemical Shift (ppm)
5.0
4.5
4.0
3.5
3.0
X-Ray
AMBER
GROMACS
ILMM
NAMD
N
Chemical Shift (ppm)
121.5
121.0
120.5
120.0
119.5
119.0
118.5
118.0
117.5
117.0
X-Ray
AMBER
GROMACS
ILMM
NAMD

Figure A.2.7

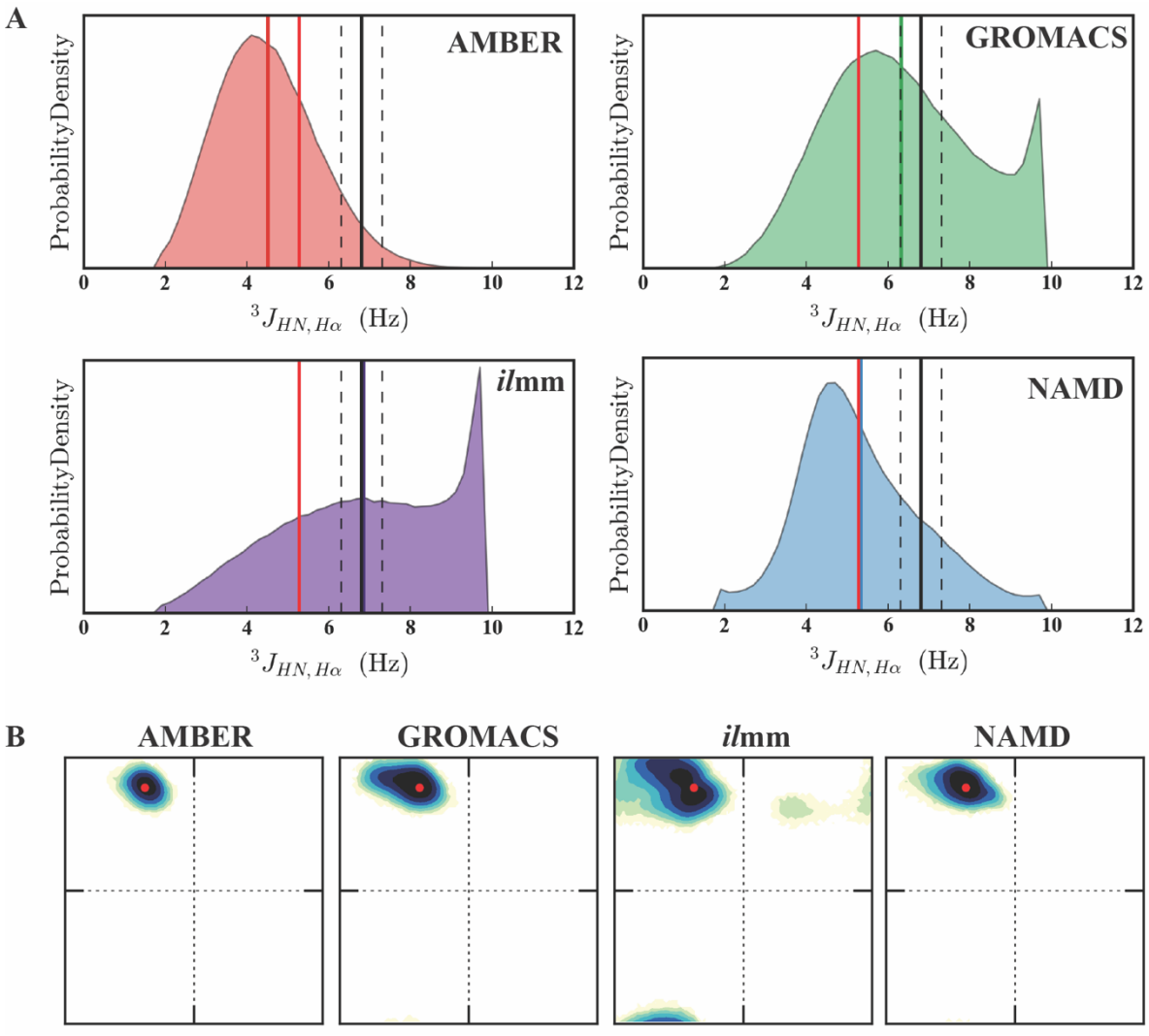

A
AMBER
GROMACS
ilmm
NAMD
ProbabilityDensity
$^3J_{HN,H\alpha}$ (Hz)
0
2
4
6
8
10
12
B
AMBER
GROMACS
ilmm
NAMD

Figure A.2.8

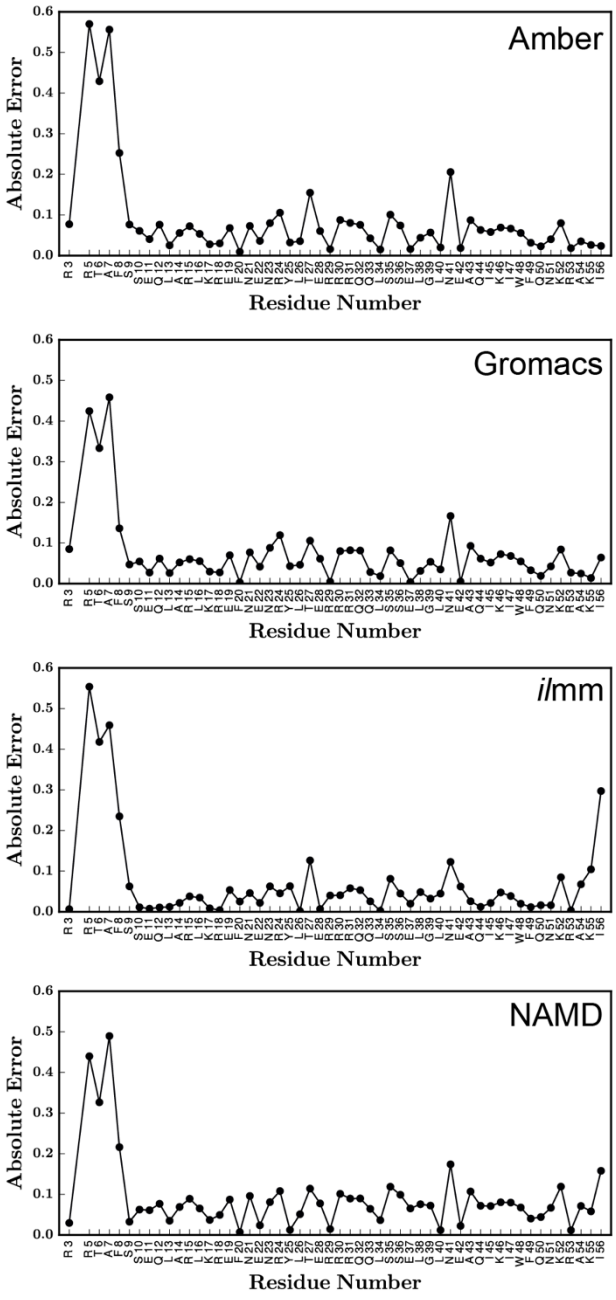


Figure A.2.9

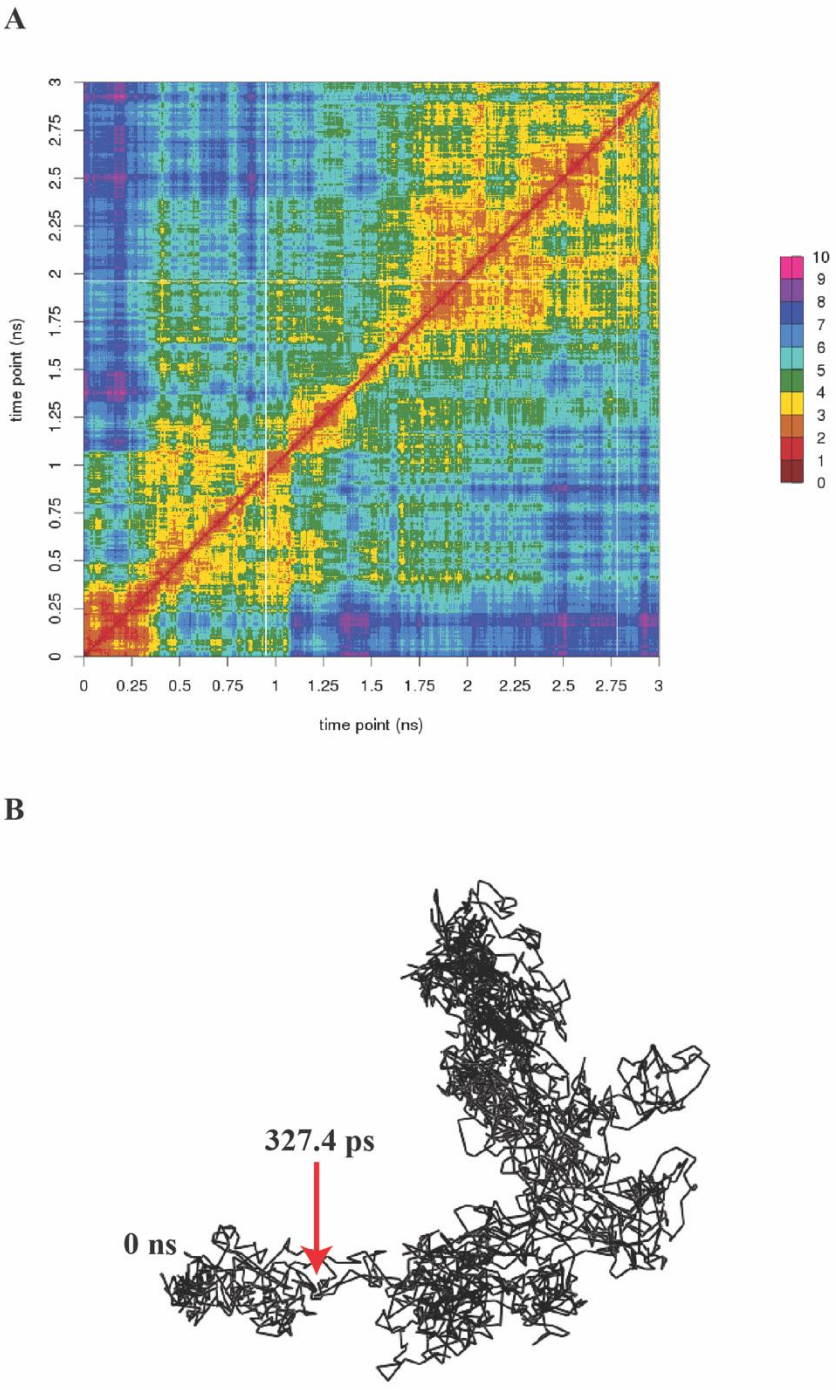

A
time point (ns)
0
0.25
0.5
0.75
1
1.25
1.5
1.75
2
2.25
2.5
2.75
3
time point (ns)
10
9
8
7
6
5
4
3
2
1
0
B
327.4 ps
0 ns

Figure A.2.10

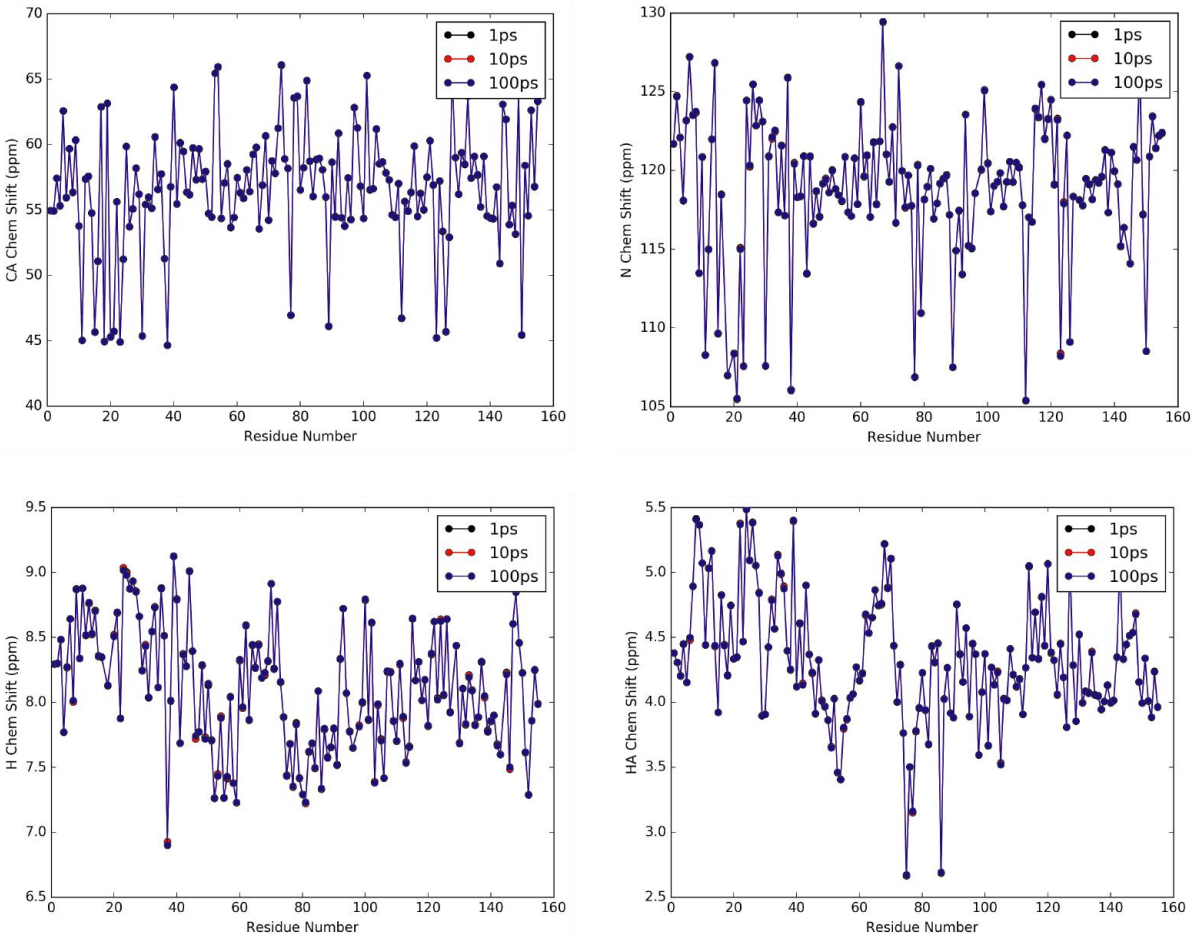

1ps
10ps
100ps
CA Chem Shift (ppm)
N Chem Shift (ppm)
H Chem Shift (ppm)
HA Chem Shift (ppm)
Residue Number

Figure A.2.11

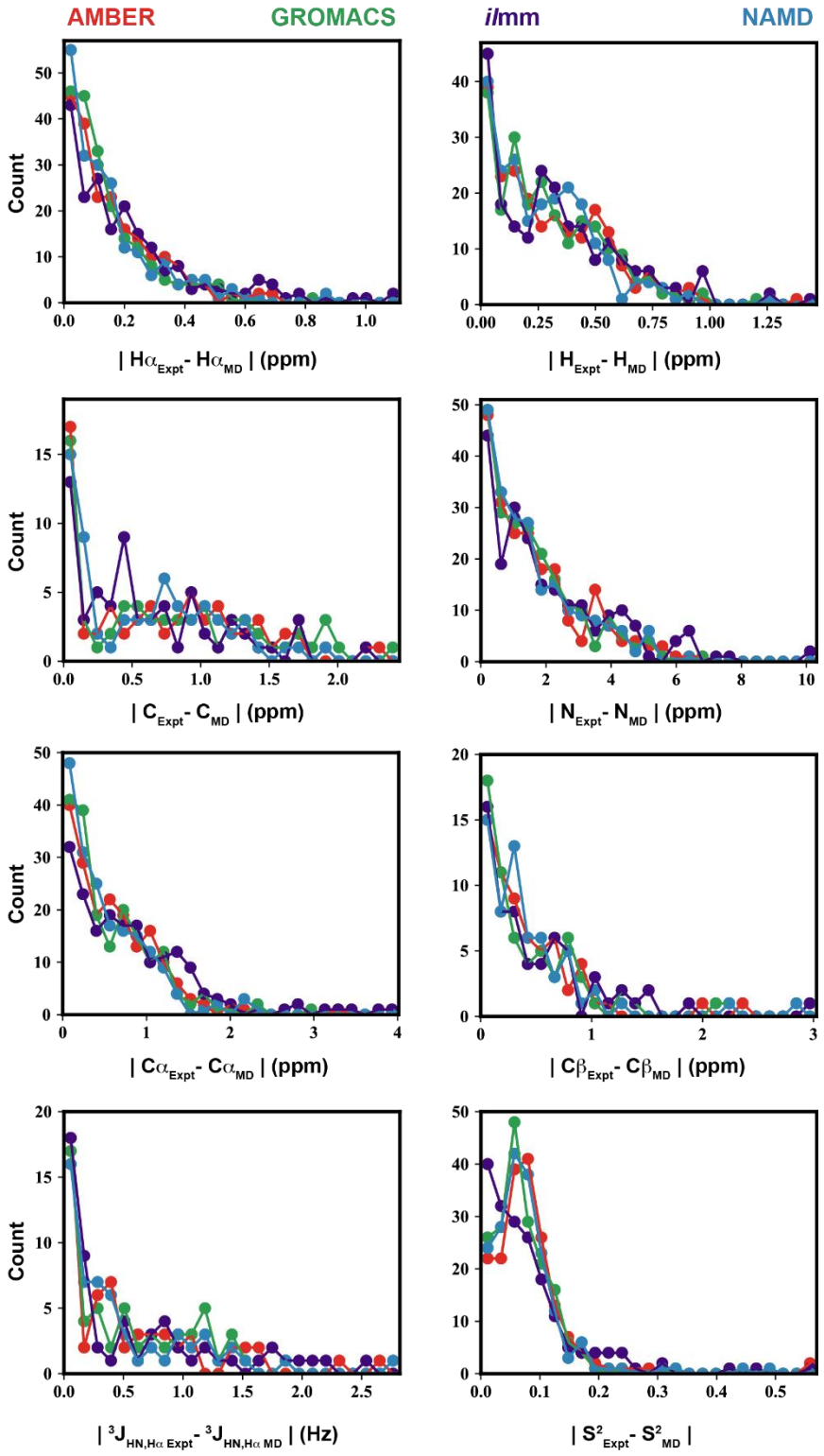

AMBER
GROMACS
ilmm
NAMD
Count
| Hα Expt - Hα MD | (ppm)
| H Expt - H MD | (ppm)
| C Expt - C MD | (ppm)
| N Expt - N MD | (ppm)
| Cα Expt - Cα MD | (ppm)
| Cβ Expt - Cβ MD | (ppm)
| 3J HN,Hα Expt - 3J HN,Hα MD | (Hz)
| S2 Expt - S2 MD |

Figure A.2.12

Chapter 3 Supplemental Figure Captions

Figure A.3.1. Cα RMSD vs. time for individual TTR simulations. The average Cα RMSD of all low pH TTR simulations (21 total simulations) as a function of time. The alignment was performed on strands A,G,B, and E, and the Cα RMSD was reported for the 'core' residues (11-123). The three replicate runs for each mutant are plotted together (run 1 – dark grey, run 2 – medium grey, run 3 – light grey).

Figure A.3.2. Cα RMSD vs. residue number for individual TTR simulations. The average Cα RMSD of all low pH TTR simulations (21 total simulations) as a function of residue number. The alignment was performed on strands A,G,B, and E, and the Cα RMSD was reported for residues 11-123. The three replicate runs for each mutant are plotted together (run 1 – dark grey, run 2 – medium grey, run 3 – light grey). The large black circles and thick black line indicate the average values over three replicate simulations. The colored bars denote the positions of the strands that compose the DAGH (purple) and CBEF (orange) sheets.

Figure A.3.3. Secondary structure content as a function of residue number for individual TTR simulations. The average β-sheet (blue), α-sheet (red), and α-helix (yellow) secondary structure content averaged between 475 – 500 ns for individual trajectories plotted as a function of residue number.

Figure A.3.4. Secondary structure content as a function of time for individual TTR simulations. The average α-sheet secondary structure in the DAGH (purple) and CBEF (orange) sheets as a function of time for individual trajectories.

Figure A.3.5. The DAGH sheet formed α-sheet secondary structure to a greater extent than the CBEF sheet. Sampling of α-sheet secondary structure was averaged over the final 25ns of each (n = 21) simulation for the DAGH (purple) and CBEF (orange) sheets. The sampling of α sheet was expressed as the fraction of time during this portion of the simulations where α−sheet secondary structure was observed. Error bars denote the standard deviation.

Figure A.3.6. The DAGH sheet sampled non-native main chain geometries to a greater extent than the CBEF sheet. The extent of sampling of β-sheet like (left), pleated (center), and α−sheet like (right) main chain geometries was averaged over the final 475ns of each (n = 21) simulation for the DAGH (purple) and CBEF (orange) sheets. The sampling of these main chain geometries was expressed as the fraction of time during this portion of the simulations where the specific main chain geometries were observed. Error bars denote the standard deviation.

Figure A.3.7. Native hydrogen bonding patterns are stable in the CBEF sheet. The main-chain atoms of the peptide main chain for residues in the CBEF sheet are represented as circles in the conformation present in the X-ray structure. The disruption in the main chain geometry of strand C due to residues 43 and 44 distorts the representation of the sheet in 2D. The hydrogen bonds are represented as dashed blue lines and are only shown for bonds present after hydrogen atoms were modeled onto the X-ray structure (i.e. reference state hydrogen bonds). The color and width of the hydrogen bonds are proportional to the average percentage of time during which these hydrogen bonds were present.

Figure A.3.8 The DAGH sheet had less stable native main chain hydrogen bonds than the CBEF sheet. The population of native main chain hydrogen was averaged over the final 475ns of each (n = 21) simulation for the DAGH (purple) and CBEF (orange) sheets. The sampling of these hydrogen bonds was expressed as the fraction of time during this portion of the simulations where the hydrogen bonds were observed. Three separate sets of hydrogen bond sampling were calculated: sampling for all main chain to main chain hydrogen bonds in the DAGH and CBEF sheets (left), sampling of main chain to main chain hydrogen bonds between the interior strands of the sheets (strands AG and BE, center), and sampling of main chain to main chain hydrogen bonds involving edge strands (DA and GH, CB and EF, right). Error bars denote the standard deviation.

Figure A.3.9. Representative example of the reorganization of the side-chain – side-chain interaction in the DAGH sheet. The side chains atoms for solvent-exposed residues in the DAGH sheet are represented as balls and sticks. Inter-residue atomic contacts are represented as green bars. Over the course of a single trajectory (WT TTR, run 3) show that the side-chains of these residues continually sample alternate conformations.

Figure A.3.10. Dynamic reorganization of the solvent-exposed side-chain interaction network in the CBEF sheet. The solvent-exposed side-chain interaction network composed of residues in the CBEF sheet is represented as a graph. Each residue is a node and edges indicate that the side-chains of the connected residues were in contact during the simulation. The width of the edges is proportional to the percentage of simulation time that the contact was present for. Reference state contacts (i.e. those present in the minimized starting structure) are colored black

edges and contacts only observed during MD are colored grey. The network on the left-hand side shows the interaction network present in the reference structure (since there is only one conformation for this model, all interactions have been set to their max value – 100%). The network on the right-hand side shows the same interaction network using data averaged from all MD simulations.

Figure A.3.11 The solvent exposed side chain interaction networks of DAGH and CBEF sheets were dynamically distinct. Four network-level properties of the DAGH (purple) and CBEF (orange) sheets were calculated using data averaged over the final 475ns of each (n = 21) simulation. Error bars denote the standard deviation. The normalized network centralization reports on the degree of connectivity within the residue interaction network (top left). The network clustering coefficient is the average of all clustering coefficients for all nodes, which measures the extent to which residues in the network form interactions with their neighbors (top right). The network heterogeneity reflects the presence of hub nodes in a network (bottom left). The edges per node reports the average number of contacts formed by each residue within the network.

Figure A.3.12. The solvent exposed side chain interaction network of the CBEF sheet was stronger than that of the DAGH sheet. There were differences in average 'strengths' of the solvent exposed interaction networks in the DAGH (purple) and CBEF (orange) sheets using data from the final 475ns of each (n = 21) simulation. Error bars denote the standard deviation. Comparisons of the interaction network strength were made for all contacts (left), native contacts (i.e. those contacts present in the crystallographic structure, center), and non-native contacts (i.e. those contacts only observed in MD simulations, right).

Chapter 3 Supplemental Figures

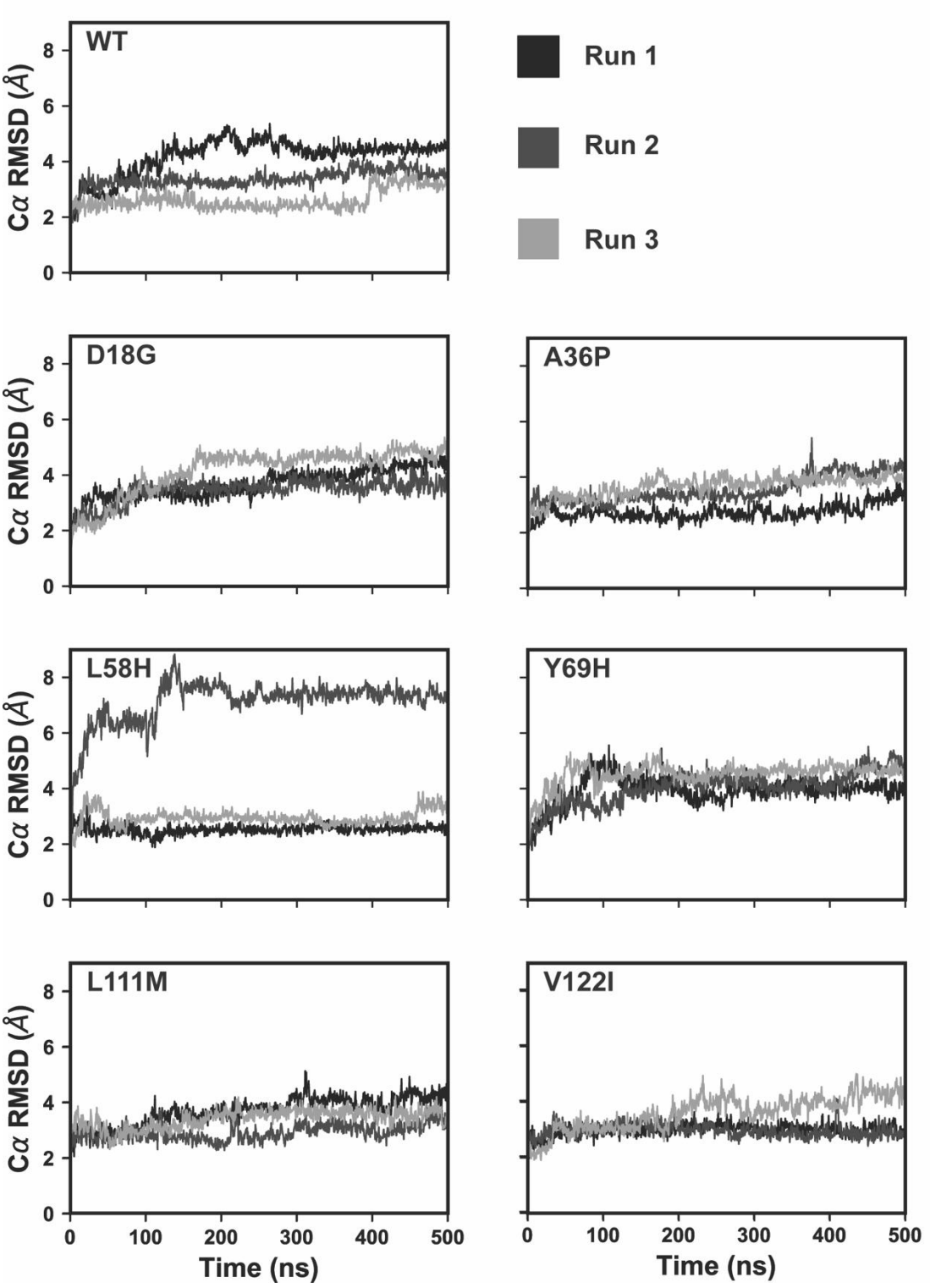


Figure A.3.1

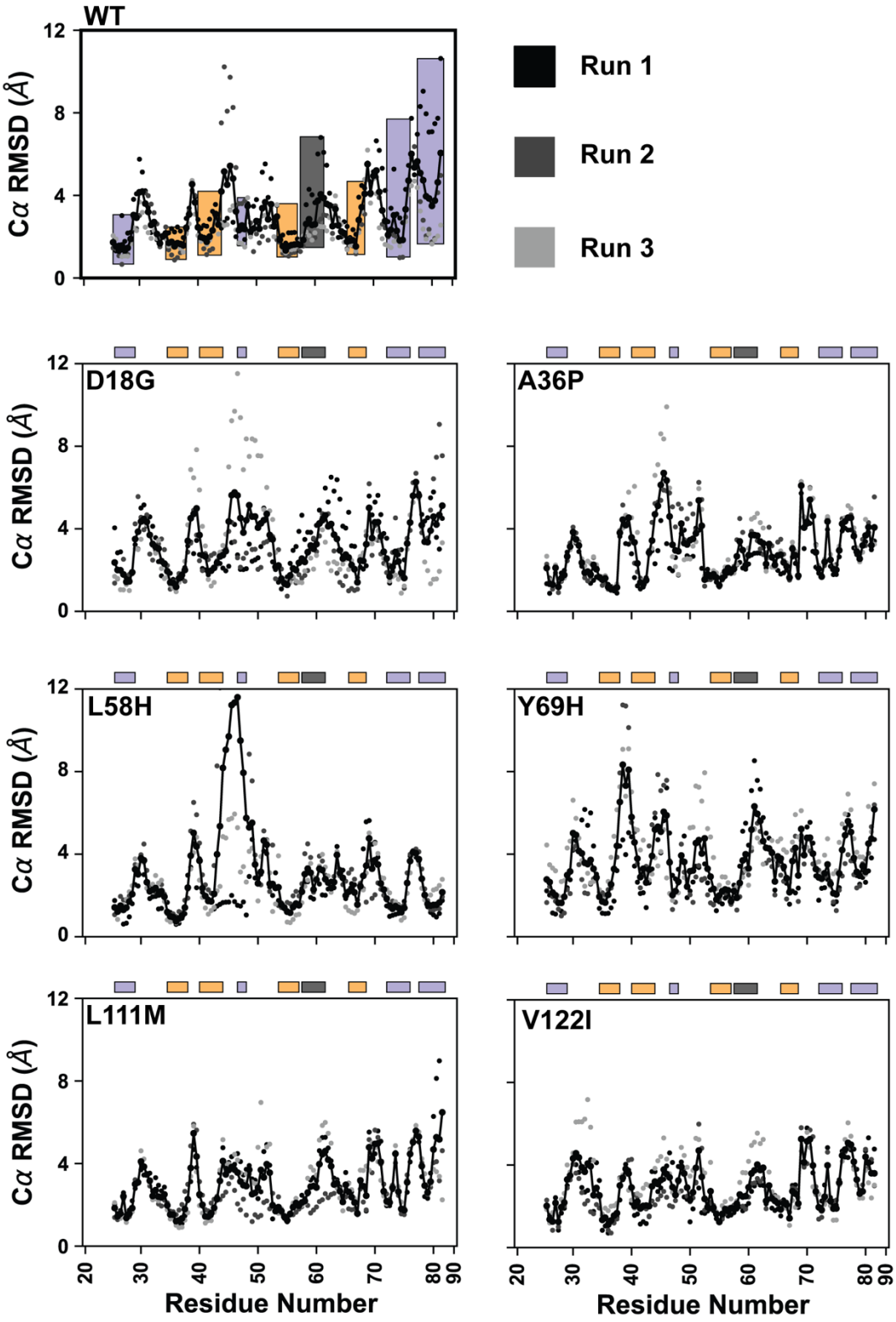

WT
Run 1
Run 2
Run 3
D18G
A36P
L58H
Y69H
L111M
V122I
Cα RMSD (Å)
Residue Number
0
4
8
12
20
30
40
50
60
70
80
90

Figure A.3.2

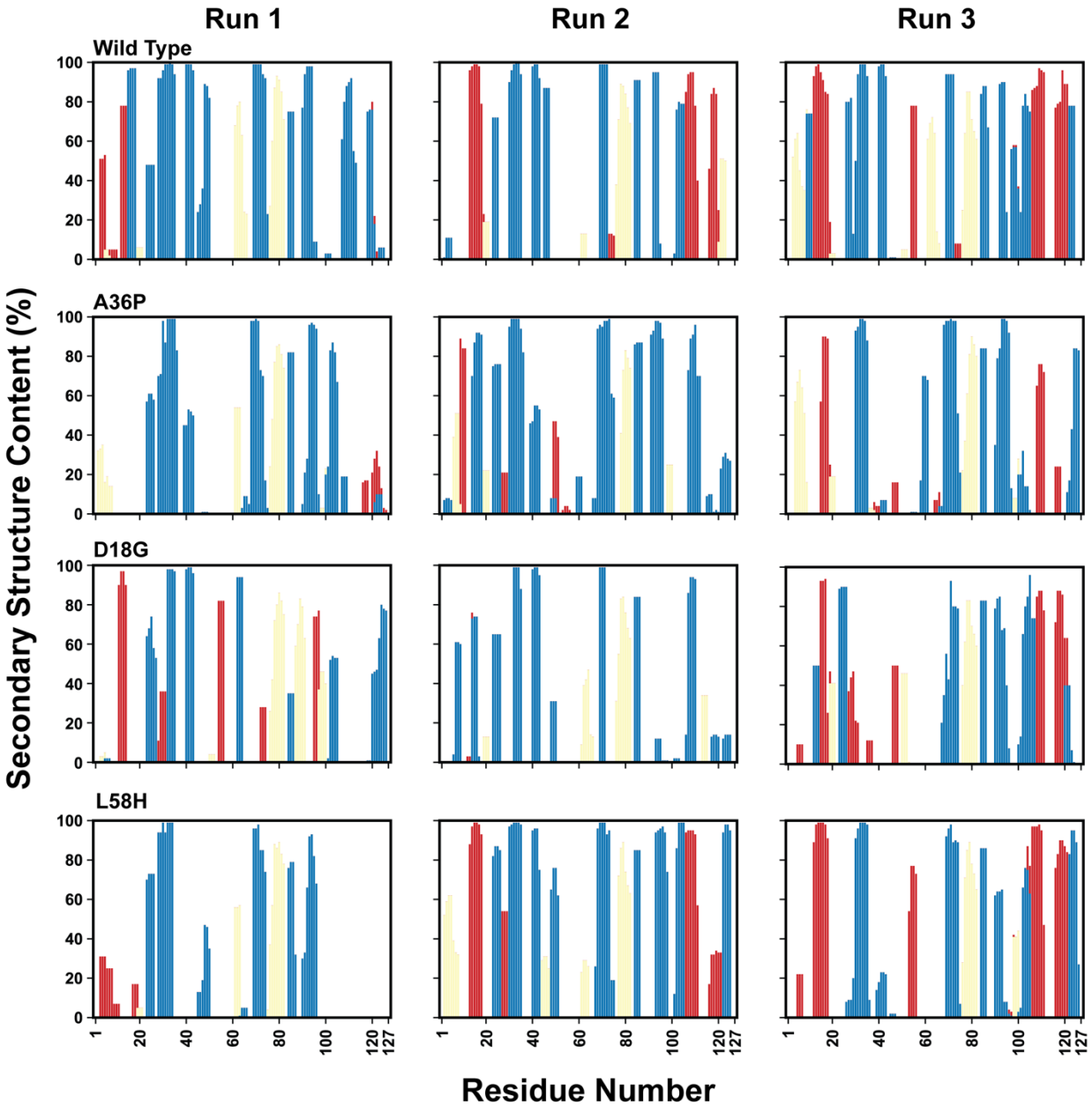

Run 1
Run 2
Run 3
Wild Type
A36P
D18G
L58H
Secondary Structure Content (%)
Residue Number
0
20
40
60
80
100
1
20
40
60
80
100
120
127

Figure A.3.3

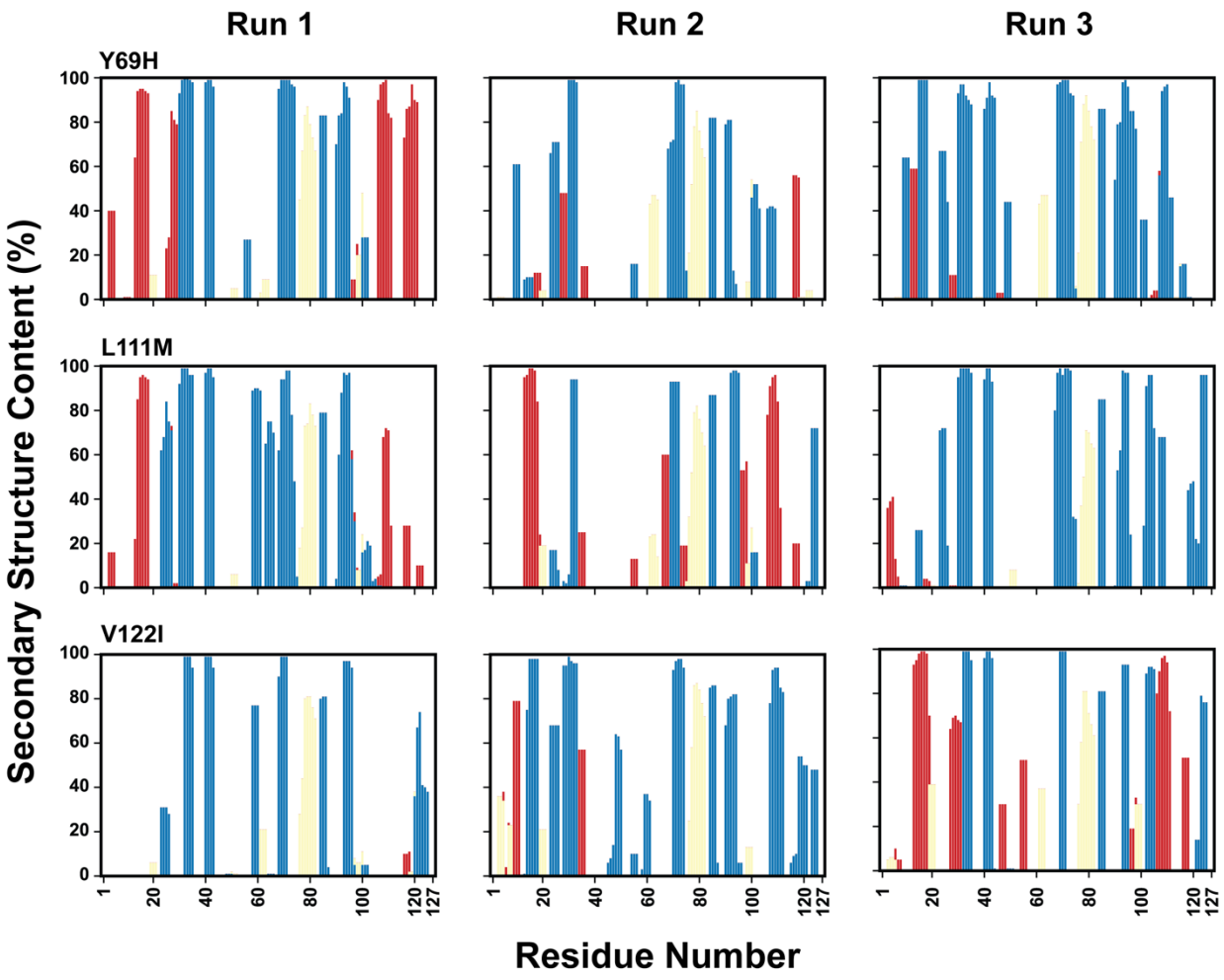


Figure A.3.3, continued

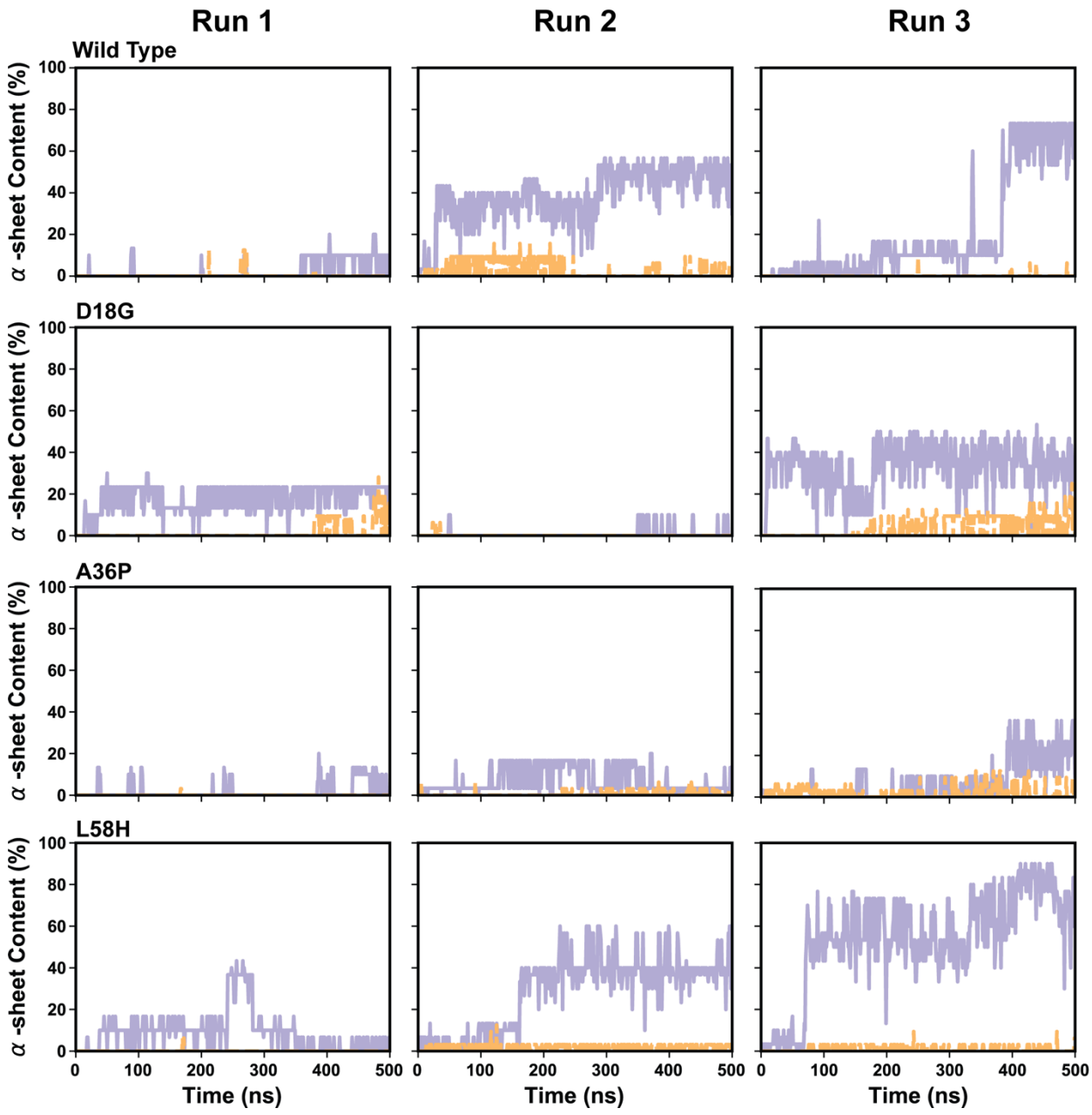

Run 1
Run 2
Run 3
Wild Type
D18G
A36P
L58H
α -sheet Content (%)
0
20
40
60
80
100
0
100
200
300
400
500
Time (ns)

Figure A.3.4

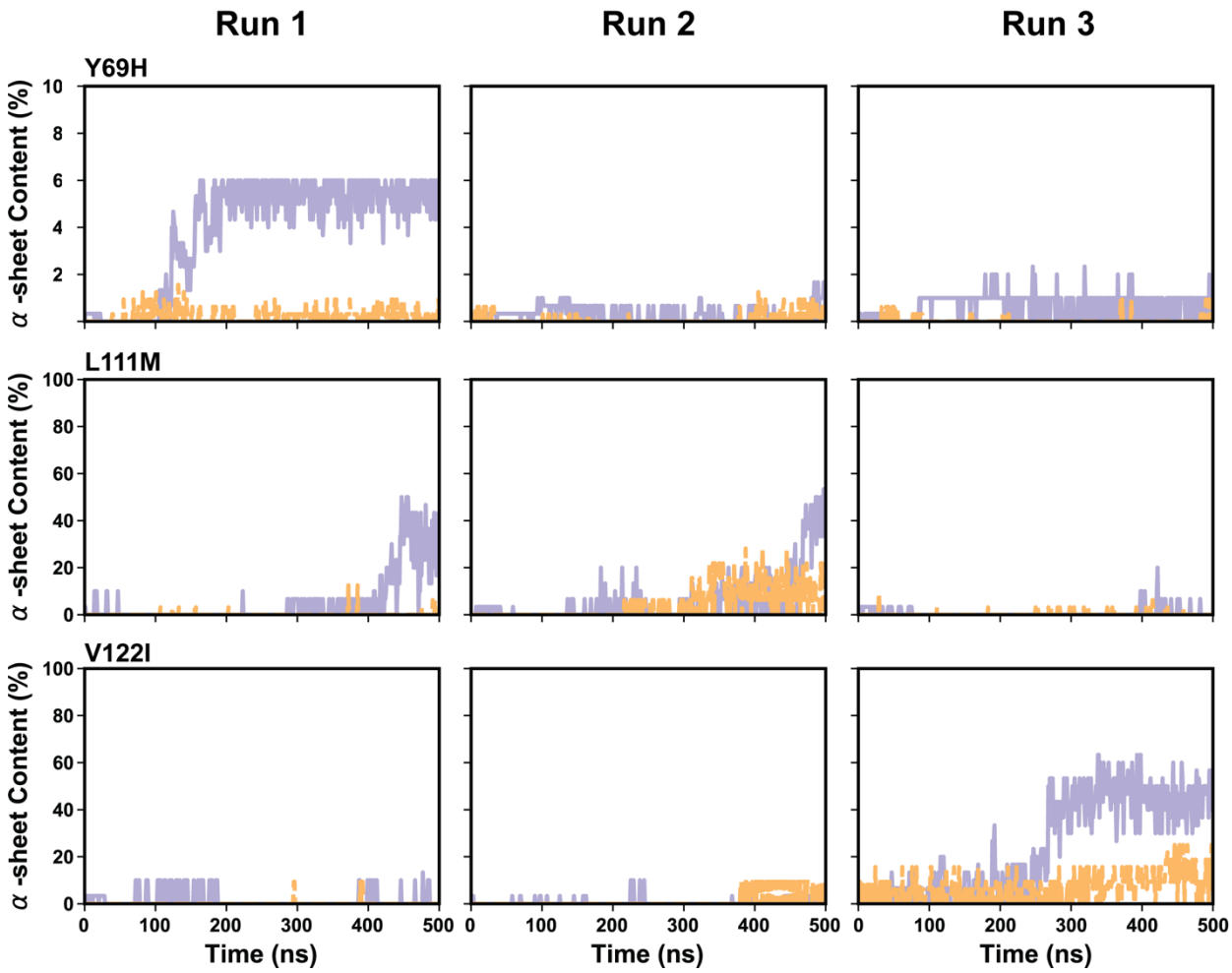


Figure A.3.4, continued

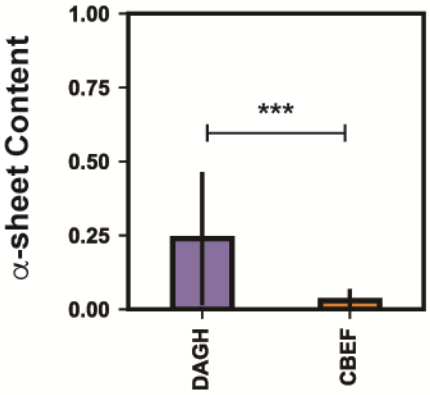


Figure A.3.5

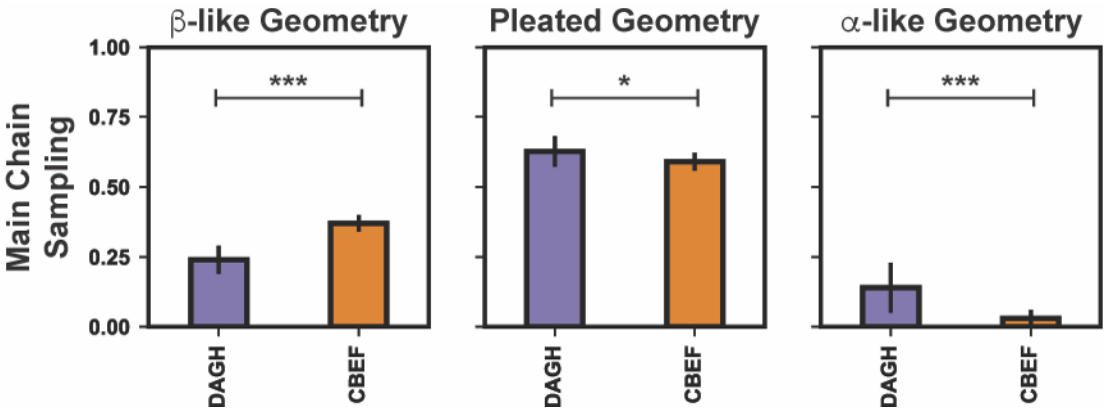

β-like Geometry
Pleated Geometry
α-like Geometry
Main Chain Sampling
1.00
0.75
0.50
0.25
0.00
***
*
***
DAGH
CBEF
DAGH
CBEF
DAGH
CBEF

Figure A.3.6

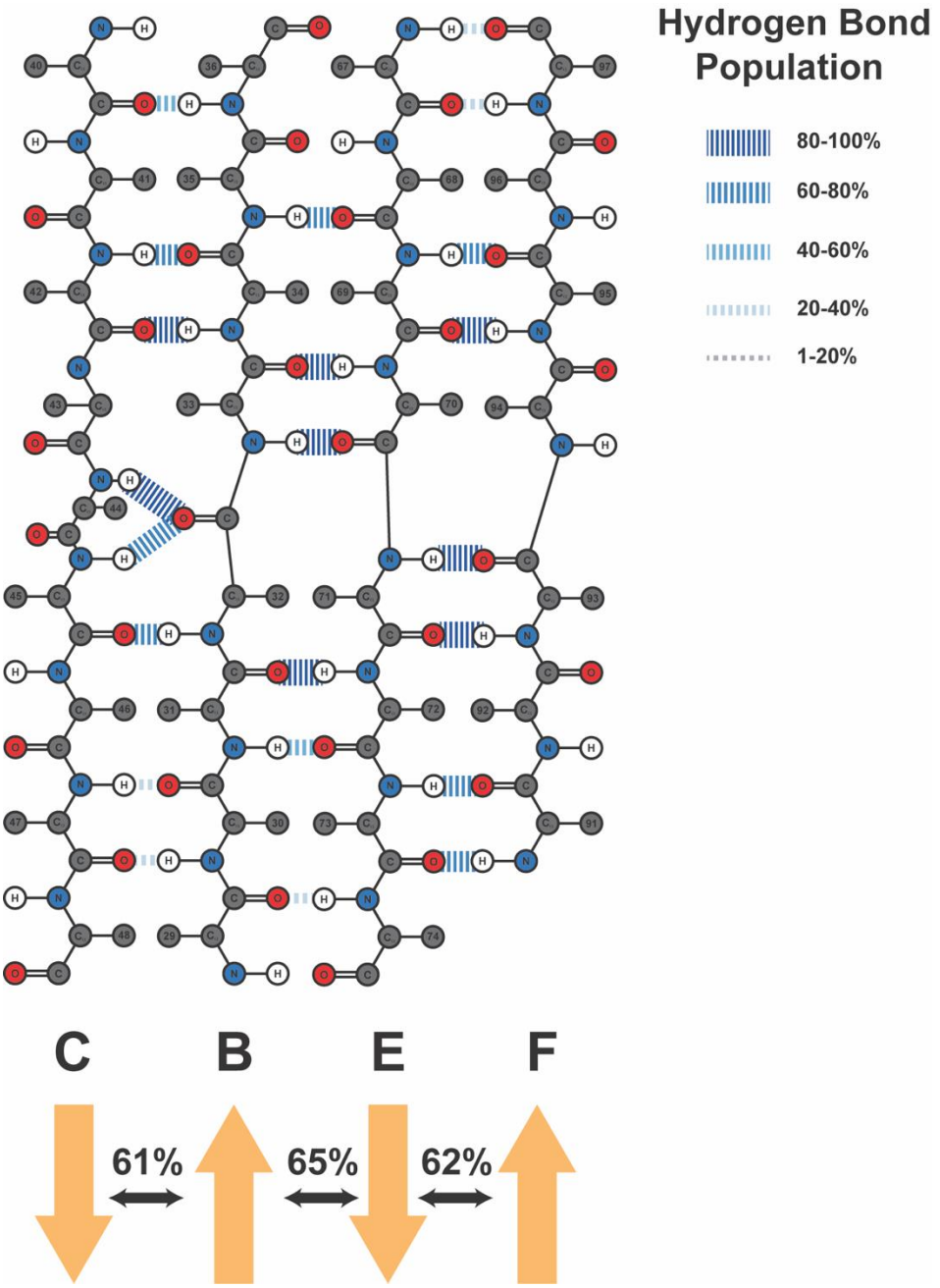

Hydrogen Bond Population
80-100%
60-80%
40-60%
20-40%
1-20%
C
B
E
F
61%
65%
62%

Figure A.3.7

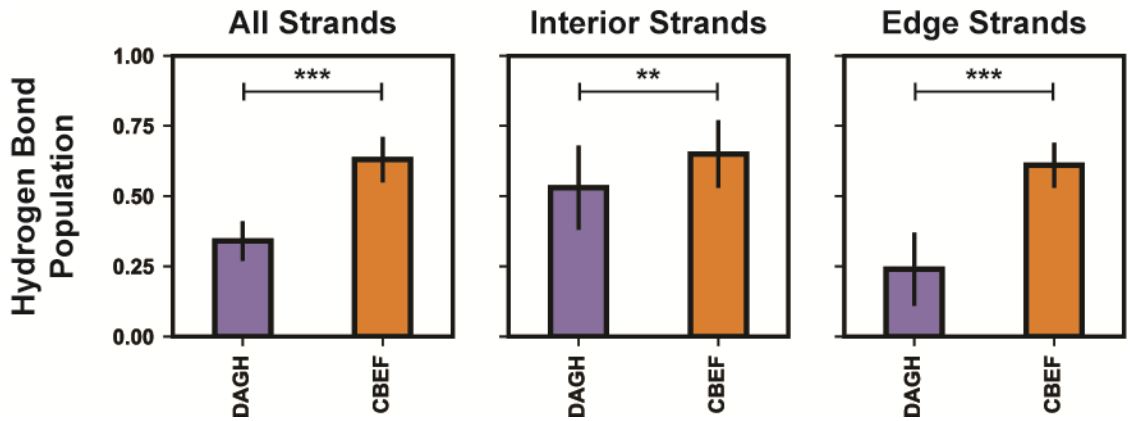

All Strands
Interior Strands
Edge Strands
Hydrogen Bond Population
1.00
0.75
0.50
0.25
0.00
***
**
***
DAGH
CBEF
DAGH
CBEF
DAGH
CBEF

Figure A.3.8

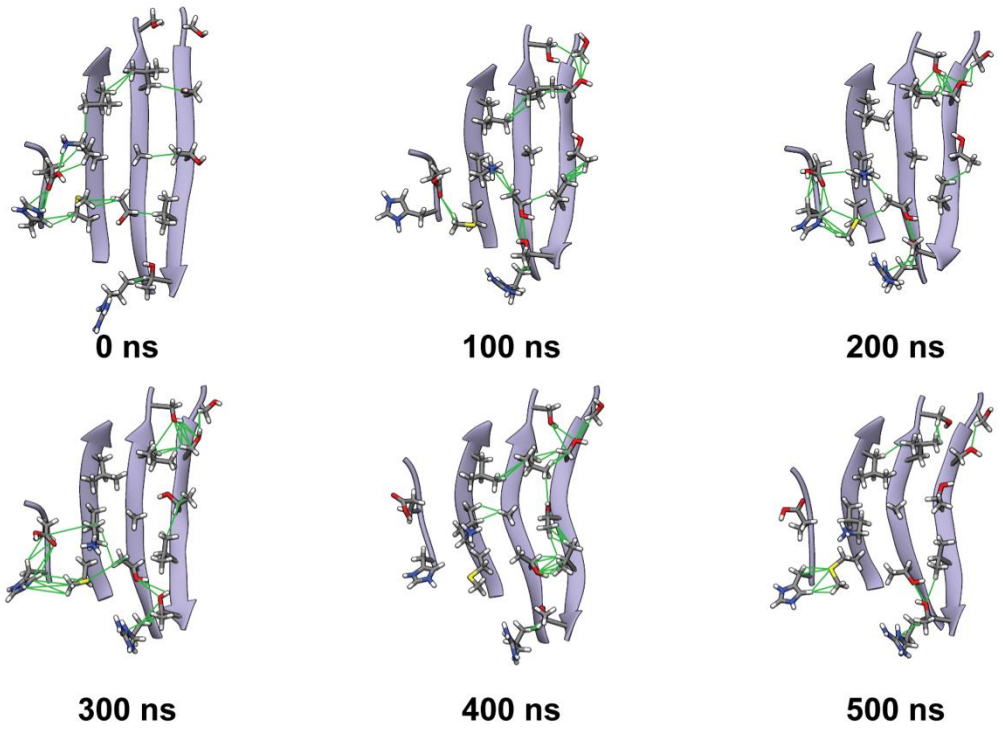


Figure A.3.9

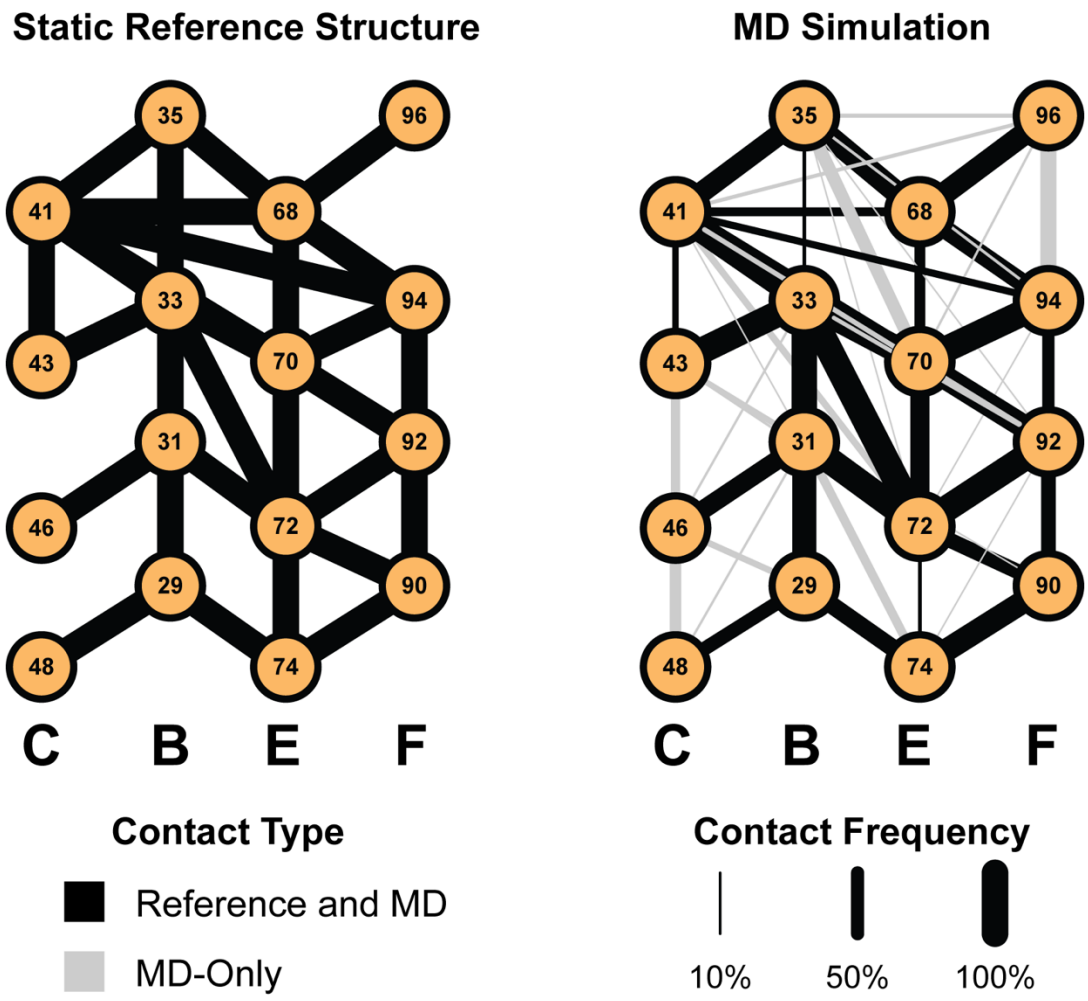

Static Reference Structure
MD Simulation
35
96
41
68
33
94
43
70
31
92
46
72
29
90
48
74
C
B
E
F
Contact Type
Reference and MD
MD-Only
Contact Frequency
10%
50%
100%

Figure A.3.10

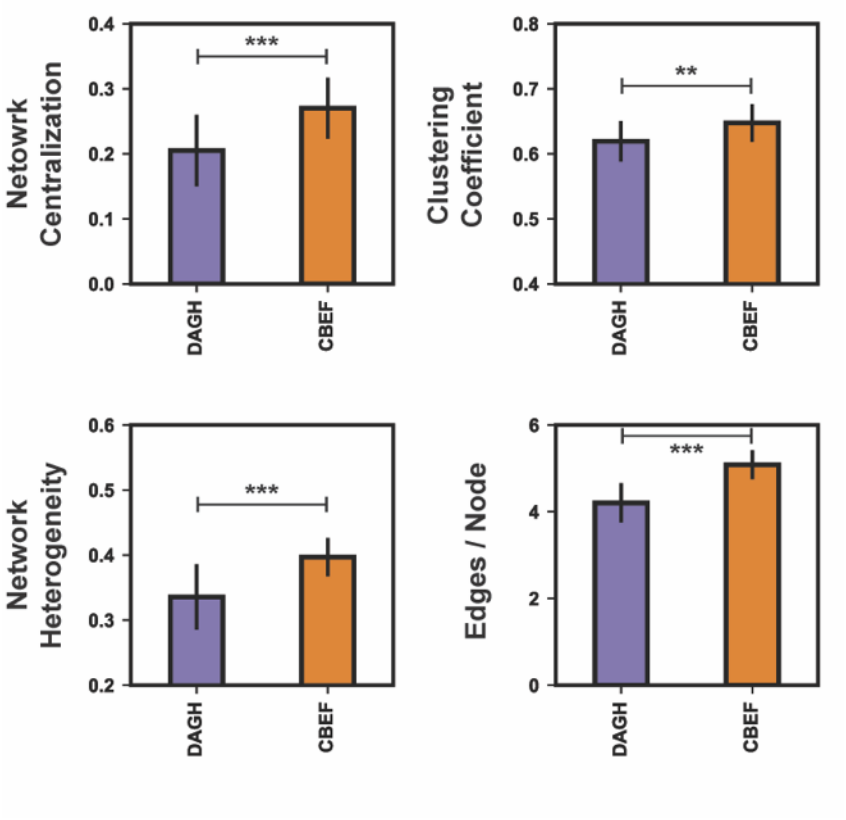

Netowrk Centralization
0.4
0.3
0.2
0.1
0.0
***
DAGH
CBEF
Clustering Coefficient
0.8
0.7
0.6
0.5
0.4
**
DAGH
CBEF
Network Heterogeneity
0.6
0.5
0.4
0.3
0.2
***
DAGH
CBEF
Edges / Node
6
4
2
0
***
DAGH
CBEF

Figure A.3.11

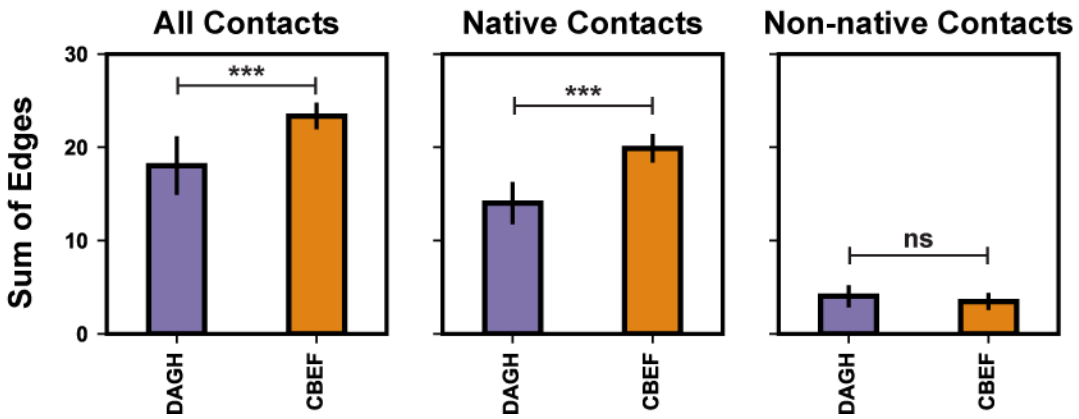

All Contacts
Native Contacts
Non-native Contacts
Sum of Edges
30
20
10
0
***
***
ns
DAGH
CBEF
DAGH
CBEF
DAGH
CBEF

Figure A.3.12

Chapter 4 Supplemental Figure Captions

Figure A.4.1. C_α RMSD vs. time for individual TTR simulations. The average C_α RMSD of all low pH TTR simulations (21 total simulations) as a function of time. The alignment was performed on strands A,G,B, and E, and the C_α RMSD was reported for the 'core' residues (11-123). The three replicate runs for each mutant are plotted together (run 1 – dark grey, run 2 – medium grey, run 3 – light grey).

Figure A.4.2. C_α RMSD vs. time for individual TTR simulations. The average C_α RMSD of all low pH TTR simulations (21 total simulations) as a function of time. The alignment was performed on strands A,G,B, and E, and the C_α RMSD was reported for the 'core' residues (11-123). The three replicate runs for each mutant are plotted together (run 1 – dark grey, run 2 – medium grey, run 3 – light grey).

Chapter 4 Supplemental Figures

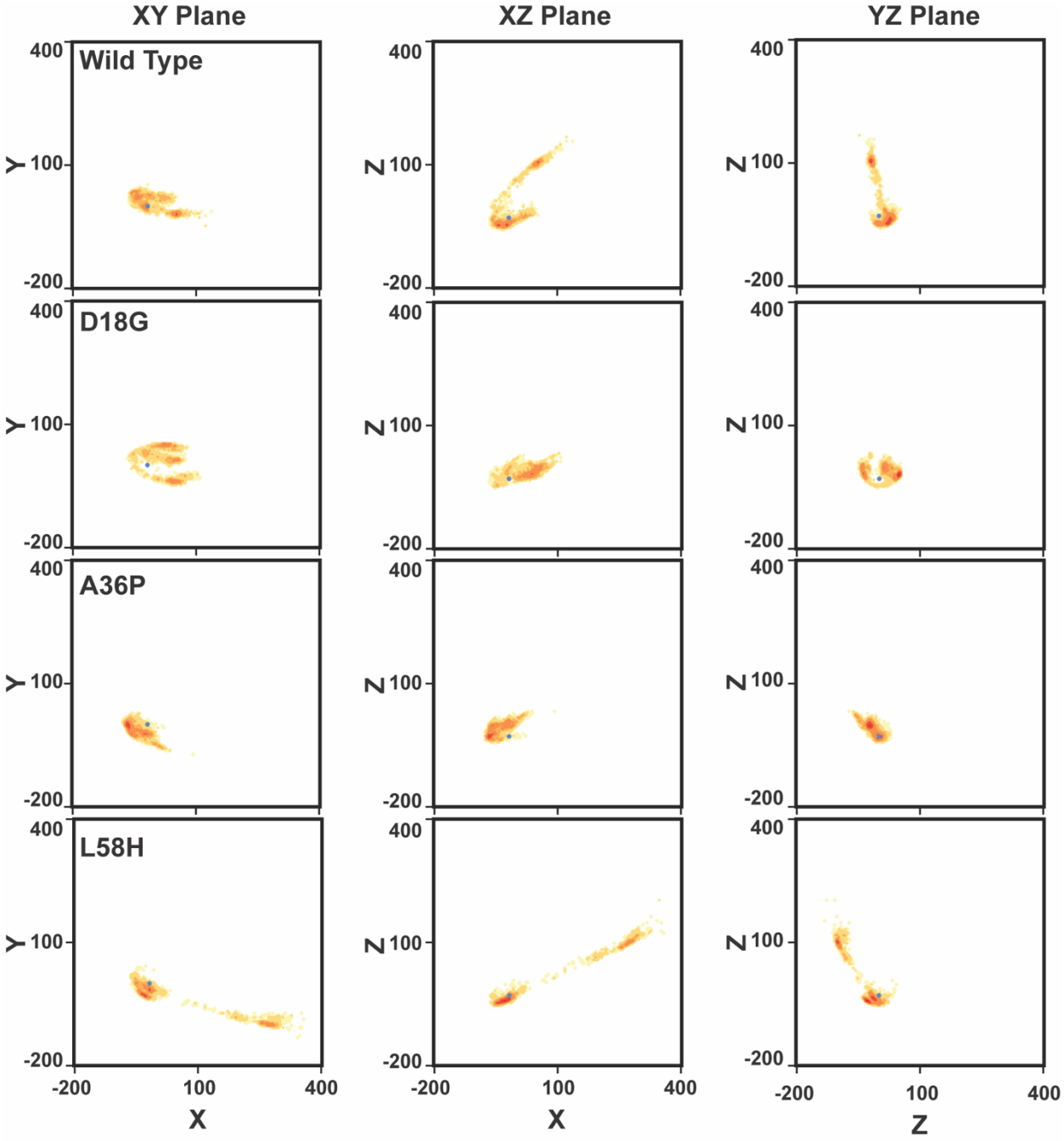


Figure A.4.1

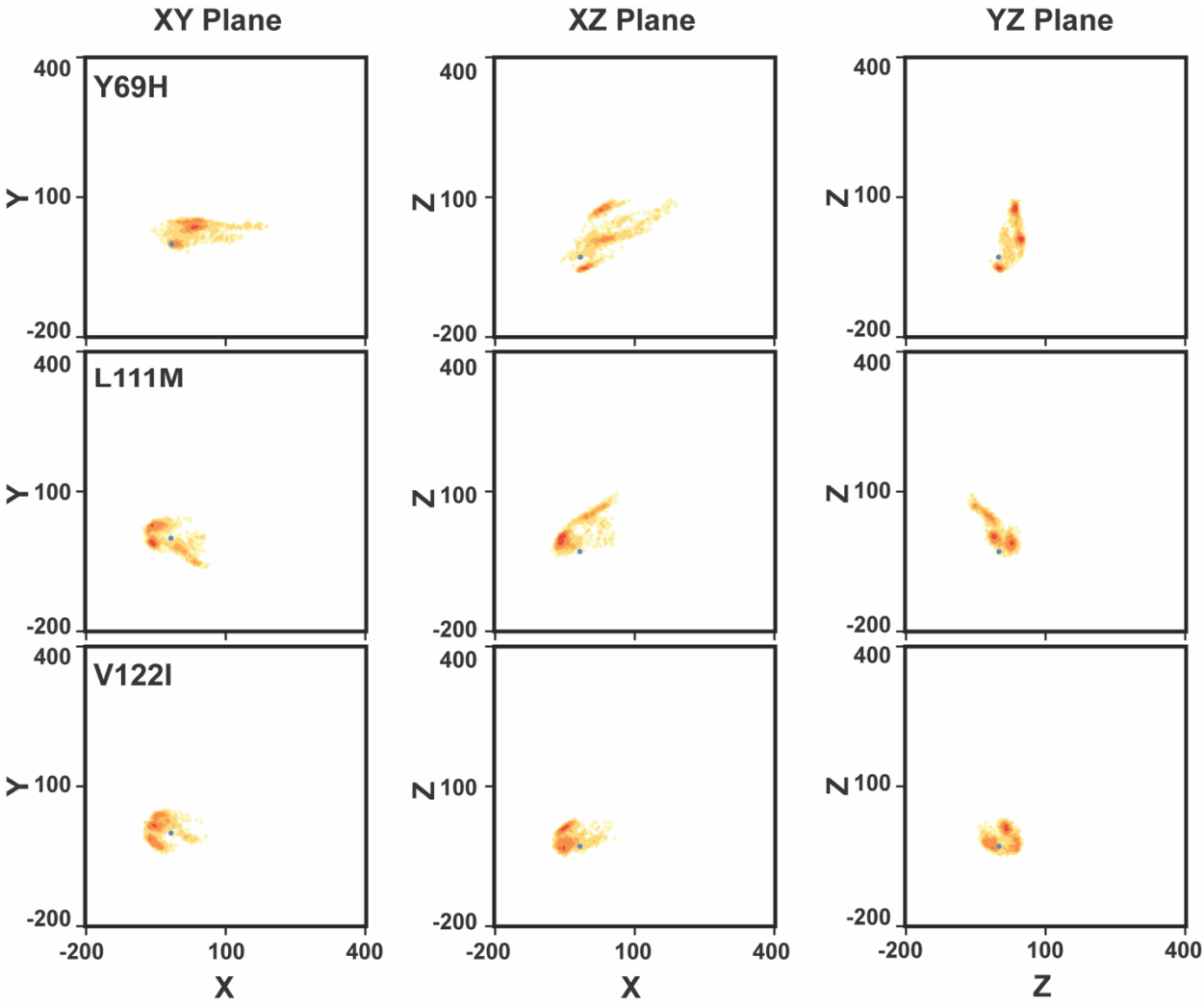


Figure A.4.1 continued

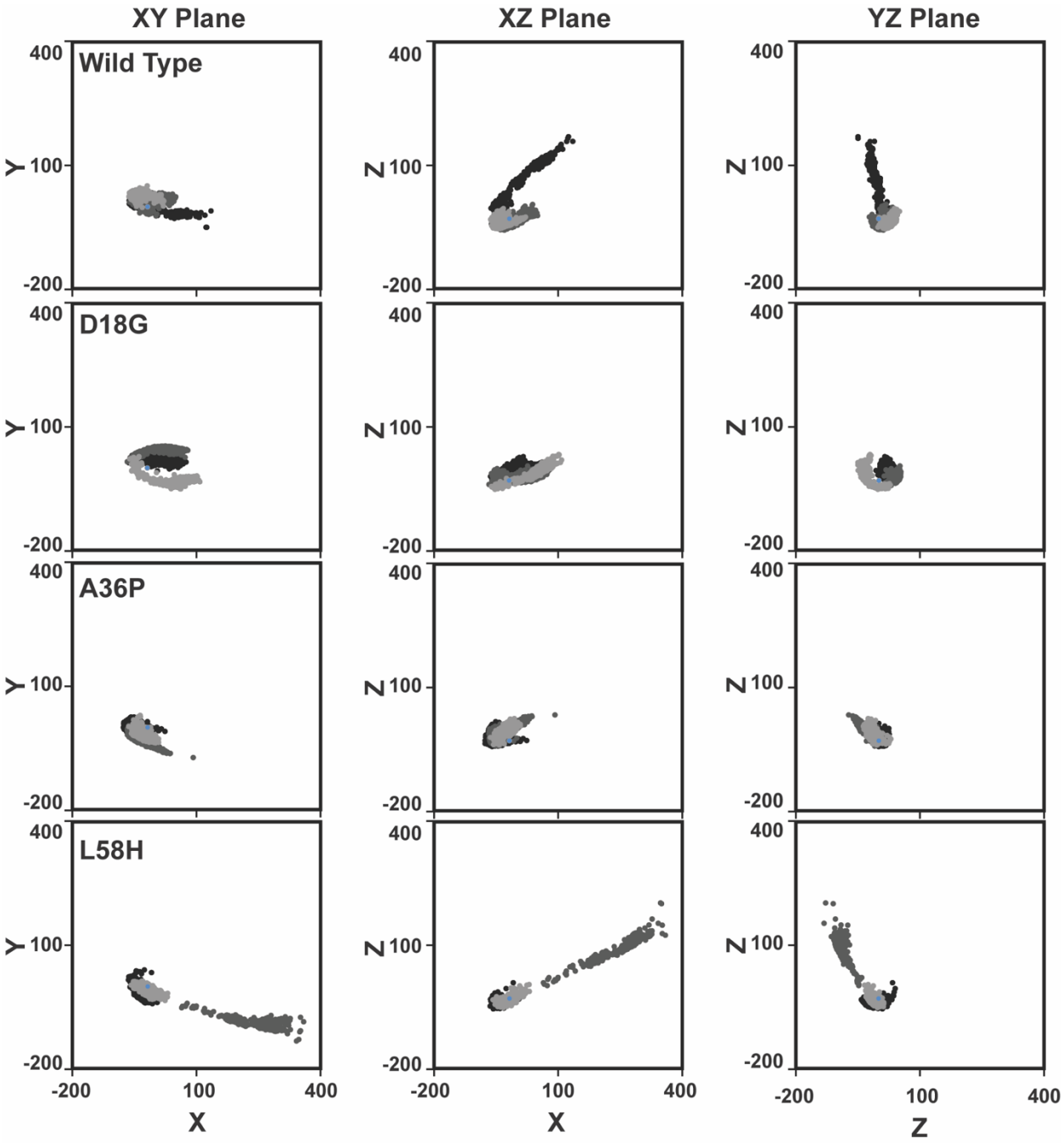

XY Plane
XZ Plane
YZ Plane
Wild Type
D18G
A36P
L58H
400
100
-200
X
Y
Z

Figure A.4.2

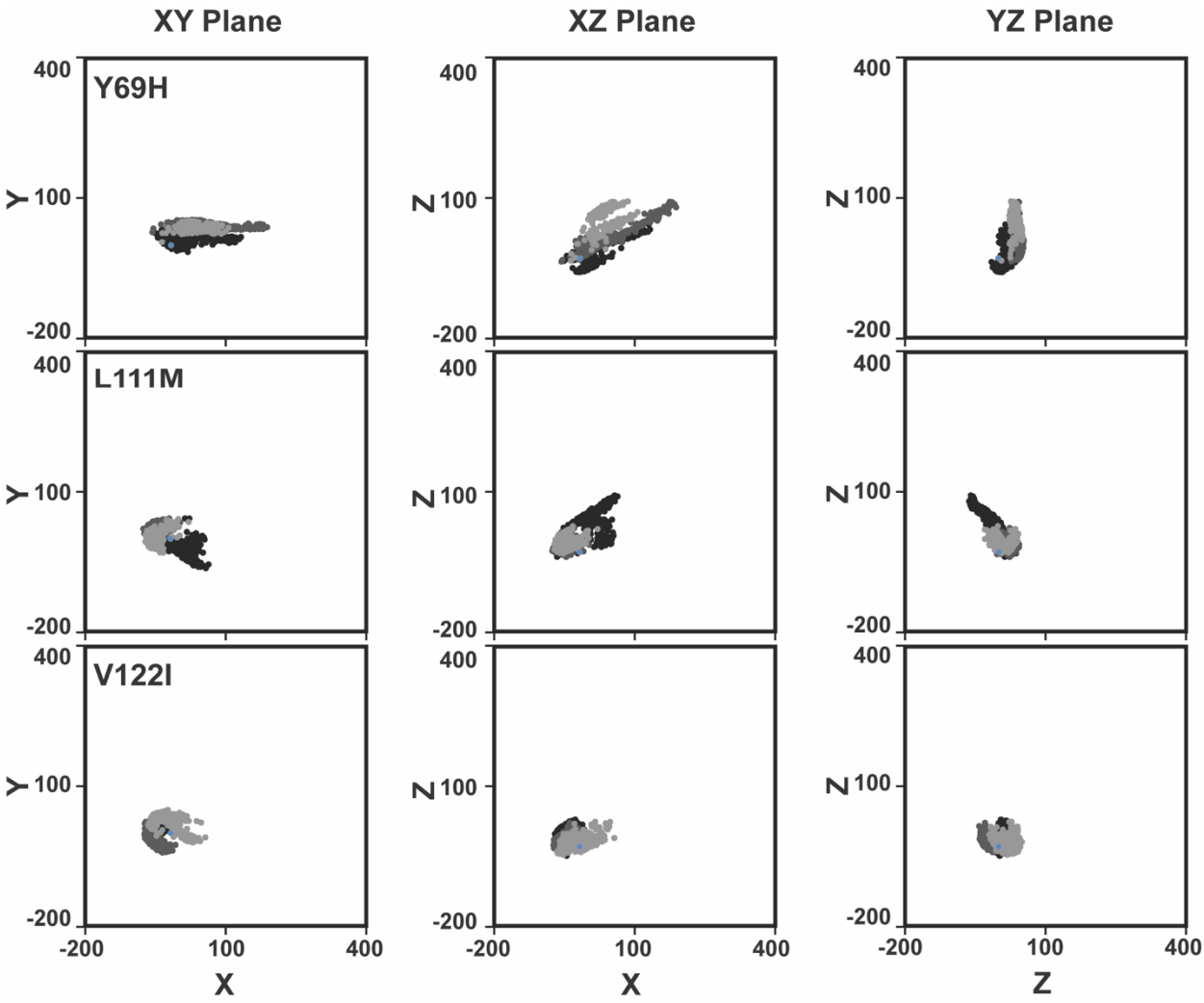


Figure A.4.2, continued

Chapter 4 Supplemental Tables

Table A.4.1 Edge weights in the side chain interaction network

Residue I	Residue J	Reference Structure Weight	MD Ensemble Weight
V32	V71	100	100
V71	V93	100	100
V14	I107	100	99
V30	I73	100	99
T75	W79	100	99
R34	F44	100	98
Y105	I107	100	98
I73	A91	100	97
V32	F44	100	97
V71	I73	100	97
I73	T75	100	97
L12	V14	100	97
V32	Y69	100	96
V14	V16	100	95
L12	Y105	100	95
V30	V71	100	94
V32	A45	100	93
V16	A109	100	92
V16	L111	100	89
V71	F95	100	88
Y69	F95	100	88
R34	Y69	100	88
I73	L111	100	87
V14	L55	100	87
Y116	T118	100	87
V93	F95	100	86
F95	Y105	100	86
H88	Y116	100	86
F95	I107	100	84
A109	L111	100	84
V30	L55	100	83
V28	V30	100	83
I73	V93	100	83
L111	Y116	100	83
V30	V32	100	81
A91	Y116	100	80
A109	T118	100	78
V28	I73	100	77
F44	Y69	100	77
T75	L111	100	73

T49	L55	100	72
F44	T59	100	71
Y105	V122	100	71
L12	F95	100	70
V30	T49	100	70
V16	V28	100	66
A97	Y105	100	66
D18	Y78	100	66
V28	Y78	100	65
F64	Y69	100	64
T75	Y78	100	64
I107	T118	100	62
V93	T118	100	62
L55	L58	100	61
Y69	A97	100	60
L12	F64	100	60
T75	A91	100	59
V16	I73	100	58
I107	L120	100	57
V16	T49	100	56
T75	H88	100	55
V28	T49	100	55
L12	V71	100	55
D18	W79	100	55
F64	Y105	100	54
L12	V32	100	53
F64	A97	100	51
A45	L58	100	50
L12	L58	100	50
A91	L111	100	46
V16	V30	100	45
V32	L58	100	44
L120	V122	100	42
T59	F64	100	41
F95	A97	100	40
W79	L111	100	40
L12	I107	100	39
V14	V30	100	33
F44	F64	100	27
L12	F44	100	26
F64	F95	100	26
V14	V71	100	26
L12	T59	100	24
V28	L111	100	21
I73	A109	100	19
V30	L58	100	19

N98	Y105	100	19
Y78	L111	100	18
V71	I107	100	15
V14	I73	100	15
V93	I107	100	14
V14	L58	100	11
F64	N98	100	10
V14	T49	100	7
I73	I107	100	6
R34	E42	0	89
H88	A91	0	81
Y69	V71	0	75
A91	V93	0	75
L12	L55	0	71
V93	Y116	0	70
I107	A109	0	68
D18	L111	0	68
V14	Y105	0	58
L12	Y69	0	57
V14	A109	0	55
V28	T75	0	54
H88	L111	0	53
W79	H88	0	52
V32	L55	0	51
V14	F95	0	50
F95	T118	0	46
Y105	L120	0	43
F44	L58	0	40
V16	L55	0	37
V93	L111	0	35
F95	L120	0	34
V16	D18	0	31
L111	T118	0	30
L55	V71	0	29
I107	V122	0	28
A45	L55	0	26
F95	A109	0	22
A45	T59	0	21
T59	Y69	0	21
E42	F44	0	20
Y69	Y105	0	20
I73	H88	0	18
T118	L120	0	18
V32	R34	0	18
V16	I107	0	16
D18	T75	0	15

I73	Y116	0	15
V32	F64	0	14
L58	F64	0	13
V32	T59	0	13
L58	Y69	0	13
V93	A109	0	11
Y69	N98	0	10
D18	V28	0	10

Chapter 6 Supplemental Tables

Table A.6.1. Correlation coefficients (R) between initial and latter portions of the GGXGG trajectories for the 64 conformational states populated by the three central residues.

Guest X Residue	Water 298 K*	8M Urea 298 K†	Water 498 K§
Ala-1	0.99	1.00	0.99
Ala-2	0.99	0.98	1.00
Ala-3	0.97	0.99	0.99
Arg	0.99	0.99	0.99
Asn	0.99	0.99	1.00
Asp	0.99	0.99	0.99
Ash	0.98	0.99	0.99
Cyh	0.98	0.98	0.99
Gln	0.97	0.97	0.99
Glu	0.99	0.98	1.00
Glh	0.95	0.99	1.00
Gly-1	0.99	0.99	0.99
Gly-2	0.99	0.98	0.95
Gly-3	0.99	0.99	0.98
Hid	0.97	0.99	0.98
Hie	0.99	0.98	0.99
Hip	0.97	0.98	0.99
Ile	0.99	0.98	0.99
Leu	0.99	0.96	0.99
Lys	0.99	0.99	0.99
Met	0.98	0.98	0.99
Phe	1.00	0.99	0.99
Pro	0.98	0.99	0.99
Ser	1.00	0.99	0.99
Thr	0.99	0.97	1.00
Trp-1	0.98	0.98	0.99
Trp-2	0.99	0.96	-
Trp-3	0.99	0.97	-
Tyr	0.97	0.99	0.99
Val	0.99	0.98	0.99
Average	**0.99**	**0.98**	**1.00**

The latter portions were that part of the trajectory (production dynamics) used for analysis. The start position and length of the production dynamics for each set of simulations was defined by the convergence behavior as exhibited in Figure X. *last 300 ns, †last 500 ns, §last 50 ns.

Table A.6.2. Conformational propensities of the guest residues (X) in GGXGG in water at 298 K

	Population Frequency (%)											
	By quadrant				By specific conformational region							
X	**Q_α**	**Q_β**	**$Q_{\alpha L}$**	**Q_o**	**α_R**	**Near α_R**	**α_L**	**β**	**Non-P β**	**P_{IIL}**	**P_{IR}**	**Other**
Ala	58.5	32.8	8.2	0.4	23.2	26.8	6.7	19.3	9.3	8.0	6.2	19.6
Ala	56.3	34.6	8.7	0.4	23.3	25.2	7.2	20.9	10.0	8.9	6.6	18.9
Ala	54.9	33.0	11.7	0.4	21.8	25.5	10.0	19.9	9.6	8.4	6.3	18.4
Arg	70.0	24.5	5.4	0.1	30.2	31.7	4.7	17.1	7.9	8.3	2.7	14.6
Asn	77.0	17.9	5.0	0.1	39.6	32.8	4.4	13.6	8.1	4.2	2.7	8.3
Asp	70.7	19.5	9.6	0.1	28.8	38.9	9.1	5.3	3.4	0.4	5.2	14.3
Ash	73.1	19.8	6.9	0.2	35.6	33.6	6.1	15.1	9.6	3.6	3.6	7.9
Cyh	59.2	34.0	6.7	0.1	24.3	27.8	5.8	23.3	10.7	11.1	4.1	16.2
Gln	68.8	28.2	2.9	0.1	30.6	30.3	2.5	20.2	9.2	9.8	3.1	14.5
Glu	60.9	32.8	6.2	0.1	31.0	22.9	5.7	24.5	10.8	12.5	3.0	14.2
Glh	54.4	23.5	21.9	0.2	25.8	22.6	20.0	16.9	7.9	8.0	2.9	12.8
Gly	40.3	9.6	40.4	9.7	15.2	5.6	11.1	6.9	3.3	3.1	1.6	60.0
Gly	41.1	9.1	41.1	8.7	16.1	5.4	11.6	6.6	3.1	3.1	1.4	59.3
Gly	39.2	9.4	42.3	9.1	15.5	5.3	11.8	6.9	3.2	3.2	1.5	59.6
Hid	67.0	25.9	7.0	0.1	26.5	33.4	6.2	19.3	9.4	7.8	3.3	13.4
Hie	70.3	25.9	3.7	0.1	33.2	30.0	3.0	18.6	8.3	10.1	2.1	13.3
Hip	70.1	29.5	0.4	< 0.1	16.9	45.0	0.2	12.8	7.8	2.0	8.9	19.1
Ile	50.8	49.1	0.1	<< 0.1	17.5	24.3	0.1	38.2	13.0	24.8	1.6	18.7
Leu	70.3	23.7	5.9	0.1	35.2	27.7	5.2	20.1	8.2	11.2	1.6	11.0
Lys	71.6	24.3	4.1	< 0.1	31.7	31.3	3.7	17.7	8.3	8.3	2.8	13.8
Met	68.0	26.0	5.9	0.2	31.3	28.7	5.3	19.5	8.6	9.9	2.5	13.7
Phe	68.2	24.5	7.2	0.1	28.4	33.5	6.4	19	9.5	8.3	2.6	11.3
Pro	21.0	79.0	-	-	19.2	<< 0.1	-	73.3	14.4	58.9	-	7.5
Ser	54.9	39.3	5.6	0.2	29.0	17.0	5.0	31.5	11.0	19.7	2.1	16.1

Thr	69.4	29.9	0.7	0.0	41.8	22.7	0.6	22.8	8.7	13.7	0.9	11.5
Trp	70.1	24.1	5.6	0.2	28.7	34.6	5.0	18.2	8.7	8.5	2.5	12.0
Trp	67.2	21.8	10.9	0.1	27.1	34.2	9.8	16.3	7.9	7.4	2.2	11.4
Trp	72.4	19.9	7.4	0.2	30.4	36.1	6.7	12.5	6.9	6.4	2.3	11.2
Tyr	71.4	24.4	4.1	0.1	29.8	35.2	3.6	18.2	9.0	7.9	2.9	11.6
Val	57.5	40.7	1.7	-	21.0	29.9	1.6	29.8	11.0	18.1	1.8	16.5

Table A.6.3. Conformational propensities of the guest residues (X) in AAXAA in water at 298 K

	Population Frequency (%)											
	By quadrant				By specific conformational region							
X	**Q_α**	**Q_β**	**$Q_{\alpha L}$**	**Q_o**	**α_R**	**Near α_R**	**α_L**	**β**	**Non-P β**	**P_{IIL}**	**P_{IIR}**	**Other**
Ala	46.2	39.8	13.6	0.4	20.3	18.1	11.6	28.9	12.8	13.7	6.6	16.9
Ala	41.7	41.8	16.2	0.4	17.7	16.1	14.0	32.2	14.4	15.4	6.1	16.2
Ala	45.6	41.7	12.4	0.3	20.3	17.3	10.4	31.3	13.7	15.2	6.5	16.6
Arg	55.6	33.8	10.3	0.2	26.6	21.7	9.2	28	12.0	14.5	3.4	12.6
Asn	59.2	33.4	7.3	0.1	34.0	21.6	6.3	27.7	16.0	8.7	4.9	8.5
Asp	67.8	21.8	10.3	0.1	29.2	36.2	9.7	6.9	4.6	0.4	6.5	13.5
Ash	65.4	24.5	10.0	0.1	36.5	25.4	9.1	20	12.3	5.4	3.9	7.5
Cyh	48.2	41.4	10.0	0.4	20.8	20.3	8.8	33.2	14.3	17.3	4.3	14.3
Gln	48.2	32.2	19.4	0.2	22.9	18.8	17.7	26.4	11.0	14.1	3.2	12.2
Glu	50.6	36.8	12.4	0.2	25.9	18.1	11.4	28.9	12.7	14.5	4.2	13.2
Glh	51.4	33.9	14.6	0.1	22.7	22.2	13.3	26.4	11.7	13.0	4.2	12.9
Gly	41.2	9.2	39.4	10.2	14.2	3.2	11.5	8	3.7	3.9	1.2	62.4
Hid	62.3	29.6	8.0	0.1	27.3	28.3	7.2	23.1	11.8	9.5	4.0	11.9
Hie	54.6	35.1	10.2	0.1	27.9	20.0	8.9	29.5	13.1	14.9	3.0	12.2
Hip	42.3	27.1	30.1	0.5	8.8	28.2	24.7	16.6	10.3	3.2	8.2	16.6
Ile	39.6	55.7	4.7	<< 0.1	15.5	17.0	4.4	46.1	14.2	31.4	1.8	15.7
Leu	61.1	21.0	17.6	0.3	33.7	21.2	15.8	18	6.9	10.4	1.3	10.7
Lys	55.2	32.1	12.5	0.2	27.2	21.1	11.3	26.3	10.9	14.0	3.3	12.2
Met	51.0	28.7	20.0	0.3	25.3	19.7	18.0	23	10.0	11.7	2.8	12.4
Phe	56.8	25.8	17.2	0.2	26.4	25.1	15.5	20.4	10.5	8.3	3.2	10.9
Pro	5.1	94.9	-	-	3.1	-	-	89.4	11.2	78.2	-	7.6
Ser	45.0	49.9	5.0	0.1	24.0	10.2	4.5	44.6	12.7	31.3	1.7	15.5

Thr	55.0	44.3	0.7	< 0.1	35.77	14.2	0.7	38.1	12.3	25.5	1.2	10.6
Trp	62.3	26.0	11.6	0.2	28.6	28.0	10.5	20.5	10.8	8.1	3.2	10.8
Tyr	54.0	25.7	20.0	0.3	24.0	24.4	18.0	20.1	10.9	7.5	3.7	11.6
Val	42.3	56.5	1.2	<< 0.1	16.8	19.0	1.1	49.4	17.0	31.4	2.3	12.3

Table A.6.4. Difference in the sampling of conformational regions (%) of GGXGG relative to AAXAA in water at 298 K

	Population Frequency (%)											
	By quadrant				By specific conformational region							
X	**Q_α**	**Q_β**	**$Q_{\alpha L}$**	**Q_o**	**α_R**	**Near α_R**	**α_L**	**β**	**Non-P β**	**P_{IIL}**	**P_{IR}**	**Other**
Ala	12.3	-6.9	-5.5	0.1	2.9	8.7	-4.9	-9.6	-3.5	-5.6	-0.3	2.7
Arg	14.3	-9.3	-4.9	-0.1	3.6	10.0	-4.5	-10.9	-4.1	-6.2	-0.7	2.0
Asn	17.8	-15.5	-2.3	< 0.1	5.6	11.2	-1.9	-14.1	-8.0	-4.5	-2.2	-0.2
Asp	3.0	-2.3	-0.7	< 0.1	-0.5	2.7	-0.6	-1.6	-1.1	<< 0.1	-1.3	0.8
Ash	7.7	-4.7	-3.1	0.1	-0.9	8.1	-2.9	-4.8	-2.7	-1.8	-0.3	0.4
Cyh	11.0	-7.4	-3.4	-0.2	3.5	7.5	-3.0	-10.0	-3.5	-6.2	-0.2	1.9
Gln	20.7	-4.0	-16.5	-0.1	7.6	11.4	-15.2	-6.2	-1.7	-4.3	-0.2	2.3
Glu	10.3	-4.0	-6.2	-0.1	5.1	4.7	-5.7	-4.4	-1.9	-1.9	-1.2	1.0
Glh	3.0	-10.4	7.3	0.1	3.1	0.4	6.7	-9.5	-3.8	-5.0	-1.3	-0.1
Gly	-0.8	0.4	1.0	-0.5	1.1	2.4	-0.3	-1.1	-0.4	-0.7	0.3	-2.4
Hid	4.7	-3.8	-0.9	< 0.1	-0.8	5.1	-1.0	-4.5	-2.4	-1.7	-0.7	1.5
Hie	15.8	-9.3	-6.5	< 0.1	5.3	10.0	-5.8	-10.2	-4.8	-4.8	-0.9	1.1
Hip	27.7	2.4	-29.7	-0.4	8.1	16.9	-24.5	-3.8	-2.5	-1.2	0.7	2.5
Ile	11.1	-6.6	-4.6	< -0.1	2.0	7.3	-4.3	-7.9	-1.2	-6.6	-0.2	3.0
Leu	9.2	2.8	-11.8	-0.1	1.6	6.5	-10.6	2.1	1.3	0.7	0.3	0.3
Lys	16.4	-7.8	-8.4	-0.1	4.6	10.3	-7.6	-8.7	-2.6	-5.8	-0.5	1.6
Met	17.0	-2.7	-14.2	-0.1	6.0	9.0	-12.8	-3.6	-1.4	-1.9	-0.3	1.3
Phe	11.4	-1.3	-10.0	< -0.1	2.0	8.5	-9.2	-1.4	-1.0	-0.1	-0.6	0.4
Pro	16.0	-16.0	-	-	16.1	<< 0.1	-	-16.1	3.2	-19.3	-	-0.1
Ser	9.9	-10.6	0.6	0.1	5.0	6.8	0.5	-13.2	-1.6	-11.6	0.4	0.6
Thr	14.4	-14.4	0.0	0.0	6.0	8.5	-0.1	-15.3	-3.4	-11.8	-0.3	1.0
Trp	7.8	-1.9	-5.9	<< 0.1	0.1	6.6	-5.5	-2.2	-2.1	0.4	-0.7	1.2
Tyr	17.3	-1.3	-15.9	-0.2	5.8	10.8	-14.4	-2.0	-1.8	0.4	-0.8	< 0.1
Val	15.3	-15.8	0.5	<< -0.1	4.1	10.8	0.5	-19.7	-6.0	-13.3	-0.5	4.3

Table A.6.5. Similarity correlation matrices for the ϕ/ψ frequency distributions of the guest residues in the GGXGG and AAXAA peptides in water at 298 K.

AAXAA water 298 K	ALA	ARG	ASH	ASN	ASP	CYS	GLH	GLN	GLU	GLY	HID	HIE	HIP	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL	GGXGG water 298 K
ALA		0.94	0.93	0.92	0.87	0.96	0.90	0.95	0.94	***0.43***	0.95	0.95	0.85	0.69	0.92	0.94	0.94	0.95	***0.21***	0.87	0.87	0.95	0.95	0.83	ALA
ARG	0.95		0.94	0.95	0.84	0.98	0.91	1.00	0.97	***0.41***	0.99	0.99	0.80	0.70	0.98	1.00	1.00	0.99	***0.15***	0.87	0.90	0.99	0.99	0.87	ARG
ASH	0.87	0.91		0.99	0.82	0.91	0.87	0.94	0.93	***0.40***	0.95	0.96	0.78	0.54	0.94	0.94	0.94	0.96	***0.15***	0.84	0.90	0.95	0.96	0.73	ASH
ASN	0.88	0.92	0.99		0.79	0.91	0.87	0.95	0.94	***0.42***	0.95	0.97	0.75	0.56	0.96	0.95	0.95	0.96	***0.16***	0.84	0.91	0.95	0.96	0.74	ASN
ASP	0.80	0.83	0.78	0.76		0.89	0.80	0.83	0.81	***0.27***	0.86	0.85	0.78	0.65	0.78	0.82	0.82	0.85	***0.03***	0.77	0.82	0.87	0.86	0.79	ASP
CYS	0.96	0.98	0.85	0.87	0.83		0.91	0.98	0.97	***0.37***	0.98	0.97	0.82	0.78	0.95	0.97	0.97	0.97	***0.21***	0.91	0.91	0.98	0.97	0.91	CYS
GLH	0.96	0.98	0.89	0.89	0.83	0.97		0.89	0.92	***0.49***	0.91	0.89	0.68	0.61	0.91	0.90	0.92	0.91	***0.16***	0.84	0.83	0.90	0.89	0.77	GLH
GLN	0.96	0.96	0.86	0.87	0.79	0.95	0.99		0.98	***0.41***	0.99	0.99	0.80	0.72	0.98	1.00	0.99	0.99	***0.18***	0.89	0.92	0.99	0.99	0.87	GLN
GLU	0.96	0.98	0.91	0.93	0.78	0.96	0.98	0.97		***0.46***	0.95	0.98	0.72	0.72	0.98	0.97	0.98	0.96	***0.27***	0.94	0.95	0.96	0.95	0.86	GLU
GLY	***0.38***	***0.35***	***0.33***	***0.33***	***0.20***	***0.31***	***0.37***	***0.40***	***0.40***		***0.39***	***0.42***	***0.22***	***0.19***	***0.44***	***0.42***	***0.43***	***0.39***	0.11	***0.40***	***0.39***	***0.38***	***0.37***	***0.26***	GLY
HID	0.92	0.98	0.93	0.94	0.86	0.94	0.96	0.92	0.95	***0.31***		0.98	0.84	0.68	0.97	0.99	0.98	1.00	***0.14***	0.85	0.88	1.00	0.99	0.86	HID
HIE	0.96	0.99	0.93	0.95	0.81	0.97	0.97	0.95	0.98	***0.37***	0.96		0.78	0.71	0.98	0.98	0.99	0.98	***0.21***	0.91	0.95	0.98	0.98	0.87	HIE
HIP	0.73	0.67	0.59	0.57	0.68	0.68	0.77	0.75	0.67	***0.38***	0.69	0.64		0.56	0.70	0.79	0.76	0.82	***0.02***	0.60	0.63	0.82	0.82	0.72	HIP
ILE	0.73	0.73	***0.48***	0.52	0.52	0.80	0.70	0.69	0.69	***0.18***	0.64	0.72	***0.40***		0.67	0.69	0.70	0.67	***0.30***	0.78	0.72	0.69	0.68	0.95	ILE
LEU	0.92	0.97	0.94	0.93	0.80	0.92	0.96	0.96	0.96	***0.41***	0.95	0.97	0.65	0.61		0.99	0.99	0.98	***0.21***	0.90	0.93	0.97	0.97	0.83	LEU
LYS	0.95	0.99	0.92	0.93	0.81	0.97	0.99	0.97	0.99	***0.37***	0.97	0.99	0.67	0.70	0.98		1.00	0.99	***0.15***	0.87	0.90	0.99	0.99	0.85	LYS
MET	0.95	0.97	0.89	0.89	0.80	0.94	0.98	0.99	0.97	***0.40***	0.94	0.96	0.74	0.64	0.98	0.98		0.99	***0.18***	0.89	0.92	0.99	0.99	0.86	MET
PHE	0.93	0.97	0.94	0.93	0.84	0.93	0.98	0.97	0.96	***0.36***	0.97	0.96	0.75	0.60	0.98	0.98	0.98		***0.15***	0.86	0.89	1.00	1.00	0.84	PHE
PRO	***0.30***	***0.19***	***0.08***	***0.13***	***-0.02***	***0.29***	***0.18***	***0.20***	***0.22***	***0.06***	***0.10***	***0.25***	***-0.01***	***0.49***	***0.12***	***0.19***	***0.15***	***0.09***		***0.45***	***0.30***	***0.14***	***0.14***	***0.25***	PRO
SER	0.81	0.78	0.63	0.67	0.54	0.84	0.74	0.75	0.77	***0.30***	0.68	0.81	***0.34***	0.86	0.71	0.77	0.71	0.66	0.61		0.95	0.86	0.86	0.85	SER
THR	0.84	0.87	0.82	0.85	0.69	0.88	0.81	0.80	0.87	***0.30***	0.81	0.91	***0.36***	0.75	0.83	0.86	0.79	0.78	***0.42***	0.90		0.89	0.90	0.84	THR
TRP	0.93	0.98	0.94	0.94	0.88	0.94	0.97	0.94	0.95	***0.33***	0.99	0.97	0.71	0.62	0.97	0.98	0.96	0.99	***0.09***	0.67	0.80		1.00	0.86	TRP
TYR	0.93	0.96	0.91	0.90	0.84	0.93	0.98	0.97	0.95	***0.37***	0.96	0.94	0.79	0.59	0.97	0.96	0.98	0.99	***0.07***	0.63	0.74	0.98		0.85	TYR
VAL	0.79	0.81	0.58	0.63	0.59	0.86	0.77	0.75	0.76	***0.18***	0.73	0.80	***0.43***	0.98	0.68	0.78	0.70	0.67	***0.48***	0.88	0.82	0.71	0.66		VAL
	ALA	ARG	ASH	ASN	ASP	CYS	GLH	GLN	GLU	GLY	HID	HIE	HIP	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL	

AAXAA water 298 K

Table A.6.6. Conformational propensities of the guest residues (X) in GGXGG in urea at 298 K.

	Population Frequency (%)											
	By quadrant				By specific conformational region							
X	**Q_α**	**Q_β**	**$Q_{\alpha L}$**	**Q_o**	**α_R**	**Near α_R**	**α_L**	**β**	**Non-P β**	**P_{IIL}**	**P_{IR}**	**Other**
Ala	63.8	30.7	5.1	0.4	20.3	34.7	3.9	19.4	9.1	8.5	5.5	18.2
Ala	64.8	31.5	3.4	0.3	20.3	35.9	2.5	19.6	9.1	8.7	5.5	18.1
Ala	61.7	31.4	6.6	0.3	19.4	33.9	5.2	20.2	9.3	9.0	5.6	17.6
Arg	67.6	27.9	4.4	0.1	23.6	36.3	3.8	22.0	10.0	11.0	2.6	12.6
Asn	78.9	19.2	1.8	0.1	33.4	40.7	1.5	15.5	8.7	5.5	2.5	7.6
Asp	81.5	16.5	1.9	0.1	24.8	53.4	1.6	5.2	3.4	0.4	4.3	12.1
Ash	73.6	18.3	8.0	0.1	28.8	41.0	7.1	14.3	8.8	3.9	3.0	7.5
Cyh	66.1	29.4	4.3	0.2	20.9	38.1	3.7	21.7	9.6	11.2	2.6	14.0
Gln	74.6	22.8	2.6	< 0.1	27.8	38.9	2.2	17.1	7.9	8.3	2.2	12.8
Glu	64.3	31.0	4.6	0.1	27.9	28.9	4.1	25.3	10.4	14.1	2.1	12.7
Glh	73.0	24.3	2.7	< 0.1	23.8	41.4	2.3	16.5	8.0	7.3	3.2	14.0
Gly	39.9	10.4	39.9	9.8	12.6	8.0	8.4	7.6	3.7	3.4	1.6	62.2
Gly	41.3	9.3	39.7	9.7	13.0	8.4	7.8	6.7	3.2	3.0	1.6	63.1
Gly	39.9	10.0	40.8	9.4	12.0	8.2	8.3	7.3	3.5	3.3	1.6	63.1
Hid	73.8	21.7	4.3	0.1	24.4	42.3	3.8	15.6	7.6	7.0	2.4	12.5
Hie	70.1	23.5	6.1	0.2	25.1	38.0	5.2	18.2	7.2	10.3	1.7	12.6
Hip	75.1	23.1	1.8	0.1	13.0	54.7	1.5	9.4	5.8	1.5	6.6	16.9
Ile	58.3	41.6	0.2	<< 0.1	13.3	36.3	0.2	29.9	10.6	18.8	1.3	19.5
Leu	77.2	19.0	3.7	0.1	32.0	37.4	3.3	16.4	6.5	9.4	1.1	10.2
Lys	73.3	23.2	3.5	0.1	27.1	38.5	3.1	17.7	8.2	8.5	2.4	12.2
Met	67.4	22.3	10.1	0.2	24.1	36.3	9.2	17.1	7.3	8.9	2.1	12.2
Phe	76.3	21.4	2.2	0.1	27.2	42.9	2.0	16.8	8.5	7.3	2.2	10.0
Pro	7.2	92.8	-	-	5.2	<< 0.1	-	86.1	16.1	70.0	-	8.7
Ser	55.5	41.7	2.7	0.1	26.2	20.2	2.4	35.7	11.1	24.1	1.5	14.5
Thr	62.0	38.0	< 0.1	<< 0.1	33.8	23.4	< 0.1	31.3	10.8	20.3	0.6	11.1
Trp	76.6	21.4	1.9	0.1	27.6	42.4	1.6	16.4	8.4	7.0	2.1	10.8

Trp	75.5	19.1	5.4	0.1	26.7	42.5	4.7	14.4	6.9	6.6	1.9	10.7
Trp	73.7	20.3	5.8	0.2	25.5	41.9	5.2	15.7	7.6	7.1	2.2	10.6
Tyr	77.0	18.2	4.7	0.1	28.4	42.7	3.9	13.8	6.9	6.0	1.9	10.2
Val	65.2	34.7	< 0.1	<< 0.1	20.5	38.0	< 0.1	25.4	9.6	15.3	1.7	14.9

Table A.6.7. Conformational propensities of the guest residues (X) in AAXAA in urea at 298 K

	Population Frequency (%)											
	By quadrant				By specific conformational region							
X	**Q_α**	**Q_β**	**$Q_{\alpha L}$**	**Q_o**	**α_R**	**Near α_R**	**α_L**	**β**	**Non-P β**	**PIIL**	**PIR**	**Other**
Ala	45.1	48.1	6.7	0.2	13.1	23.6	5.3	39.7	18.2	17.6	8.1	14.0
Ala	50.4	40.4	8.8	0.3	16.2	26.3	7.0	32.7	13.8	16.3	5.9	14.6
Ala	40.6	48.5	10.6	0.3	13.6	19.3	8.9	39	16.5	19.0	8.3	14.4
Arg	55.8	37.9	6.2	0.1	19.2	29.7	5.7	33.2	14.6	16.8	3.6	10.4
Asn	50.9	44.7	4.3	< 0.1	20.1	27.2	3.7	37.8	20.9	12.5	7.2	8.3
Asp	81.7	15.8	2.4	< 0.1	23.7	55.5	2.2	5.4	3.4	0.3	4.7	10.3
Ash	57.9	34.5	7.6	0.1	25.0	29.7	6.8	28	16.5	5.8	8.9	7.2
Cyh	44.0	50.9	5.0	0.1	13.0	22.7	4.2	44.4	19.7	21.9	6.1	12.4
Gln	58.8	32.8	8.3	0.1	21.9	29.5	7.5	28.8	12.2	15.0	3.3	10.7
Glu	53.6	42.9	3.4	0.1	22.0	23.7	3.0	38.6	15.6	21.3	3.4	11.1
Glh	48.3	36.2	15.4	0.2	15.9	25.2	13.5	31.2	13.5	15.6	4.5	11.9
Gly	35.2	12.1	39.1	13.5	8.5	3.7	9.8	11.3	5.4	5.0	1.9	65.7
Gly	37.8	12.3	38.4	11.5	9.5	4.7	8.9	11.5	5.3	5.5	1.6	64.6
Gly	39.8	11.1	36.8	12.4	9.3	4.6	8.7	10.4	5.1	4.5	1.9	66.0
Hid	75.9	19.8	4.1	0.2	25.1	44.2	3.5	15.4	8.1	6.1	2.6	10.4
Hie	53.7	42.1	4.2	< 0.1	16.9	29.1	3.5	37.4	17.4	17.6	4.3	11.1
Hip	63.3	14.7	21.5	0.5	10.0	47.7	16.6	8.6	4.7	2.1	4.4	14.5
Ile	29.6	67.6	2.9	< 0.1	6.8	15.6	2.6	64.2	23.5	39.8	1.7	10.0
Leu	55.1	36.8	7.8	0.3	22.9	25.7	6.9	33.3	13.5	18.2	3.1	9.8
Lys	53.5	39.7	6.7	< 0.1	19.2	27.2	6.0	35.3	14.5	18.7	4.3	10.1
Met	60.6	34.8	4.5	0.1	21.9	31.5	4.0	30.4	12.6	16.1	3.5	10.4
Phe	60.1	35.9	3.9	< 0.1	20.2	34.5	3.4	30.1	16.3	10.4	6.2	9.0
Pro	7.1	92.9	-	-	5.0	-	-	86.9	10.3	76.6	-	8.1
Ser	49.0	45.0	5.7	0.3	23.5	15.4	4.9	41.8	10.3	31.2	0.9	13.7
Thr	71.3	28.6	0.1	< 0.1	42.1	24.2	0.1	24.9	7.1	17.6	0.6	8.4

Trp	64.2	20.5	15.0	0.3	19.4	38.9	13.1	16.7	9.3	5.9	2.9	10.5
Trp	63.2	30.0	6.6	0.2	22.0	35.0	5.6	25.3	14.3	8.6	4.5	10.0
Trp	60.9	29.8	9.1	0.2	22.0	33.3	7.9	25.3	13.3	9.4	4.6	9.4
Tyr	76.3	22.8	0.9	< 0.1	25.8	44.2	0.8	18	10.4	5.7	3.9	9.2
Val	51.5	46.6	1.8	< 0.1	16.1	29.3	1.7	42.4	16.8	24.7	1.8	9.6

Table A.6.8. Difference in the sampling of conformational regions of GGXGG relative to AAXAA in urea at 298 K

	Population Frequency (%)											
	By quadrant				By specific conformational region							
X	**Q_α**	**Q_β**	**$Q_{\alpha L}$**	**Q_o**	**α_R**	**Near α_R**	**α_L**	**β**	**Non-P β**	**P_{IIL}**	**P_{IR}**	**Other**
Ala	18.7	-17.4	-1.5	0.2	7.2	11.2	-1.5	-20.3	-9.2	-9.2	-2.7	4.2
Arg	11.8	-10.0	-1.8	0.1	4.4	6.7	-1.8	-11.1	-4.6	-5.9	-1.0	2.2
Asn	28.0	-25.5	-2.6	< 0.1	13.3	13.4	-2.2	-22.3	-12.2	-7.0	-4.6	-0.7
Asp	-0.2	0.7	-0.5	0.1	1.1	-2.1	-0.6	-0.2	<< 0.1	0.1	-0.4	1.8
Ash	15.9	-16.2	0.3	0.1	3.8	11.3	0.3	-13.7	-7.7	-1.9	-5.9	0.3
Cyh	21.9	-21.2	-0.7	0.1	7.8	15.4	-0.5	-22.7	-10.1	-10.7	-3.5	1.6
Gln	16.0	-10.0	-5.9	-0.1	5.9	9.4	-5.3	-11.7	-4.3	-6.7	-1.1	2.1
Glu	10.7	-12.0	1.3	<< -0.1	5.9	5.2	1.1	-13.4	-5.2	-7.3	-1.3	1.6
Glh	24.9	-12.0	-12.7	-0.1	7.9	16.2	-11.2	-14.7	-5.5	-8.2	-1.3	2.2
Gly	4.6	-1.7	0.7	-3.7	4.2	4.3	-1.4	-3.7	-1.7	-1.6	-0.4	-3.5
Hid	-2.1	1.9	0.2	< -0.1	-0.7	-1.9	0.3	0.2	-0.6	1.0	-0.2	2.2
Hie	16.3	-18.4	1.9	0.2	8.1	8.8	1.7	-19.2	-10.3	-7.3	-2.6	1.4
Hip	11.8	8.4	-19.7	-0.4	3.0	7.0	-15.1	0.9	1.1	-0.6	2.2	2.4
Ile	28.8	-26.1	-2.7	<< -0.1	6.5	20.7	-2.4	-34.3	-12.9	-21.0	-0.4	9.5
Leu	22.1	-17.9	-4.0	-0.2	9.1	11.8	-3.5	-16.9	-7.0	-8.8	-1.9	0.4
Lys	19.9	-16.8	-3.2	0.1	7.9	11.3	-2.9	-17.6	-6.2	-10.2	-1.9	2.2
Met	6.6	-12.4	5.6	0.2	2.1	4.8	5.2	-13.3	-5.3	-7.2	-1.4	1.7
Phe	16.2	-14.5	-1.7	< 0.1	7.0	8.3	-1.4	-13.4	-7.8	-3.1	-4.0	1.0
Pro	0.2	-0.2	-	-	0.3	<< 0.1	-	-0.9	5.8	-6.7	-	0.6
Ser	6.7	-3.5	-3.0	-0.2	2.6	4.7	-2.5	-6.1	0.8	-7.0	0.6	0.8
Thr	-9.3	9.4	< 0.1	<<0.1	-8.3	-0.8	< -0.1	6.4	3.7	2.7	< 0.1	2.7
Trp	12.5	0.9	-13.1	-0.2	8.2	3.6	-11.5	-0.3	-0.9	1.1	-0.8	0.3
Tyr	0.5	-4.4	3.8	0.1	2.6	-1.6	3.2	-4.2	-3.5	0.3	-2.0	1.0
Val	13.7	-11.9	-1.8	<< -0.1	4.4	8.6	-1.7	-17.0	-7.2	-9.4	-0.1	5.3

Table A.6.9. Conformational propensities of the guest residues (X) in GGXGG in water at 498 K.

	Population Frequency (%)											
	By quadrant				By specific conformational region							
X	**Q_α**	**Q_β**	**$Q_{\alpha L}$**	**Q_o**	**α_R**	**Near α_R**	**α_L**	**β**	**Non-P β**	**PIIL**	**PIR**	**Other**
Ala	44.0	46.5	8.0	1.5	14.7	17.2	4.5	29.5	17.7	9.0	7.5	29.4
Ala	42.1	47.6	8.6	1.7	14.2	16.2	5.0	30.4	18.2	9.4	7.6	29.4
Ala	44.1	45.2	9.1	1.6	15.2	17.1	5.4	28.3	16.5	9.1	7.2	29.4
Arg	49.3	44.7	5.5	0.5	20.0	19.0	4.0	30.2	17.2	10.9	5.2	23.8
Asn	51.4	42.3	5.5	0.8	22.2	20.6	3.7	31.1	18.6	9.9	5.3	19.8
Asp	49.7	38.2	11.2	0.9	19.2	25.0	8.4	17.1	10.8	3.2	8.3	25.0
Ash	51.4	41.0	6.8	0.8	21.6	23.2	4.3	28.2	18.5	6.3	7.1	19.0
Cyh	45.1	48.2	5.9	0.8	17.6	17.2	4.2	32.4	18.1	12.3	5.4	25.2
Gln	48.8	44.7	5.7	0.8	20.0	18.6	4.1	30.5	17.5	10.8	5.3	23.7
Glu	45.7	48.3	5.6	0.4	19.9	15.9	4.1	33.8	18.2	13.6	5.1	23.3
Glh	47.4	44.4	7.2	0.9	18.0	19.8	5.4	29.3	17.5	9.3	6.2	23.9
Gly	31.5	17.9	33.0	17.6	8.6	5.2	5.4	11.6	7.1	3.5	3.0	67.3
Hid	50.1	43.2	6.0	0.7	19.4	20.8	4.2	29.5	17.5	9.6	5.7	22.9
Hie	49.9	44.8	4.5	0.7	21.8	17.1	3.0	33.7	18.1	14.1	3.4	22.6
Hip	46.4	45.2	7.8	0.7	12.5	24.6	5.4	25.6	17.3	4.7	9.5	26.0
Ile	41.3	57.9	0.8	< 0.1	17.0	15.5	0.6	43.0	20.8	21.0	3.0	22.0
Leu	54.9	40.7	4.0	0.4	25.3	19.5	2.9	30.3	16.7	11.7	3.9	19.9
Lys	52.3	43.0	4.2	0.5	21.4	20.1	3.0	29.5	17.1	10.3	5.1	23.0
Met	52.0	41.3	6.0	0.7	22.3	19.2	4.5	28.6	16.0	10.7	4.4	23.0
Phe	49.4	44.8	5.2	0.6	20.3	20.3	3.7	32.9	18.7	12.1	4.6	20.4
Pro	21.9	78.1	-	-	17.7	< 0.1	-	68.0	19.8	48.3	<< 0.1	14.3
Ser	47.1	47.2	4.7	0.9	21.3	13.7	3.5	33.2	16.4	15.3	3.8	26.0
Thr	48.8	49.9	1.2	0.1	25.4	14.1	1.0	36.6	17.4	18.1	2.7	21.3
Trp	47.6	46.5	5.2	0.6	19.6	18.7	3.7	34.0	19.1	12.8	4.7	21.4
Tyr	49.6	44.7	5.2	0.6	20.5	20.5	3.6	32.5	18.7	11.6	4.8	20.3
Val	42.7	55.3	1.8	0.2	17.7	17.1	1.5	40.5	20.6	18.4	3.6	21.1

Table A.6.10. Conformational propensities of the guest residues (X) in AAXAA in water at 498 K

	Population Frequency (%)											
	By quadrant				By specific conformational region							
X	**Q_α**	**Q_β**	**$Q_{\alpha L}$**	**Q_o**	**α_R**	**Near α_R**	**α_L**	**β**	**Non-P β**	**P_{IIL}**	**P_{IIR}**	**Other**
Ala	41.0	48.7	8.5	1.9	15.7	14.0	5.1	33.6	19.8	10.6	6.9	27.9
Ala	41.9	48.4	8.2	1.4	15.8	14.1	5.1	33.5	20.3	10.1	7.0	27.5
Ala	40.1	49.3	8.9	1.7	15.0	13.6	5.4	34.8	20.6	10.3	7.1	28.0
Arg	44.2	48.9	6.0	0.9	19.4	15.7	4.6	34.4	19.9	12.1	5.5	22.7
Asn	45.0	47.0	7.3	0.7	21.5	15.9	5.0	36.6	22.5	11.1	5.5	18.4
Asp	48.5	37.1	13.3	1.1	20.4	23.0	10.3	15.1	9.8	2.4	8.2	25.9
Ash	49.0	40.9	9.0	1.1	23.0	19.5	6.5	29.2	19.3	6.1	7.2	18.5
Cyh	43.2	51.2	5.1	0.6	18.4	14.4	3.5	36.1	20.6	13.3	5.1	24.6
Gln	44.6	46.1	8.5	0.8	20.1	15.4	6.4	32.3	18.7	11.3	5.1	22.9
Glu	40.8	50.7	8.0	0.5	19.3	12.4	6.4	35.8	20.1	13.4	5.4	23.1
Glh	43.2	48.9	7.2	0.7	17.7	16.2	5.4	33.4	20.0	10.7	6.4	23.6
Gly	31.5	17.1	32.0	19.4	9.2	4.2	6.0	11.5	7.2	3.3	2.7	67.4
Hid	42.4	47.2	9.4	1.0	17.4	16.4	6.9	33	20.0	10.2	6.0	23.2
Hie	44.3	48.9	6.1	0.7	22.4	13.5	3.0	36.1	19.2	15.4	3.3	23.3
Hip	47.2	46.6	5.6	0.5	14.5	23.0	3.7	28.4	19.4	4.9	9.6	24.8
Ile	39.2	59.7	1.0	0.1	17.3	13.2	0.8	45	21.9	21.9	2.9	22.0
Leu	52.2	39.9	6.8	1.1	26.0	16.7	5.1	29.9	16.7	11.3	3.8	20.3
Lys	47.1	45.5	6.6	0.9	21.4	16.0	5.0	32.2	18.6	11.3	4.9	22.7
Met	47.6	44.4	6.7	1.3	22.0	16.0	5.1	31.3	17.9	11.4	4.7	22.9
Phe	45.0	49.0	5.5	0.5	19.9	16.5	4.1	36.5	21.3	12.8	5.0	20.5
Pro	23.5	76.5	< 0.1	-	19.1	<< 0.1	-	67.1	18.7	48.5	<< 0.1	13.7
Ser	43.9	51.2	4.2	0.8	21.8	10.0	3.2	37	17.9	17.5	3.8	25.7
Thr	48.6	50.2	1.1	0.1	28.1	11.1	0.9	37.3	17.6	18.8	2.4	21.2
Trp	44.4	47.9	6.9	0.8	20.2	15.2	5.3	35.3	20.0	13.4	4.1	21.9
Tyr	48.7	43.2	7.1	1.0	22.3	17.8	5.4	30.9	18.0	10.7	4.5	21.2
Val	33.9	63.1	3.0	< 0.1	15.4	11.2	2.5	49.9	24.8	23.5	3.5	19.1

Table A.6.11. Difference in the sampling of conformational regions of GGXGG relative to AAXAA in water at 498 K

	Population Frequency (%)											
	By quadrant				By specific conformational region							
X	**Q_α**	**Q_β**	**$Q_{\alpha L}$**	**Q_o**	**α_R**	**Near α_R**	**α_L**	**β**	**Non-P β**	**P_{IIL}**	**P_{IR}**	**Other**
Ala	3.0	-2.2	-0.5	-0.4	-1	3.2	-0.6	-4.1	-2.1	-1.6	0.6	1.5
Arg	5.1	-4.1	-0.5	-0.5	0.6	3.3	-0.6	-4.2	-2.8	-1.2	-0.3	1.1
Asn	6.5	-4.7	-1.7	< 0.1	0.6	4.7	-1.3	-5.5	-4.0	-1.3	-0.2	1.4
Asp	1.2	1.1	-2.1	-0.2	-1.1	2.0	-1.9	2.0	1.1	0.8	0.1	-0.9
Ash	2.4	0.1	-2.2	-0.2	-1.4	3.7	-2.1	-1.0	-0.7	0.1	< -0.1	0.5
Cyh	1.9	-3.0	0.8	0.2	-0.8	2.8	0.6	-3.7	-2.6	-1.0	0.4	0.6
Gln	4.2	-1.5	-2.8	< 0.1	-0.1	3.1	-2.3	-1.8	-1.2	-0.5	0.1	0.8
Glu	5.0	-2.4	-2.4	-0.1	0.6	3.5	-2.3	-2.0	-2.0	0.2	-0.2	0.2
Glh	4.2	-4.4	0.1	0.2	0.3	3.5	-0.1	-4.1	-2.5	-1.4	-0.2	0.3
Gly	< -0.1	0.8	1.0	-1.7	-0.6	0.9	-0.6	0.1	-0.1	0.2	0.3	-0.1
Hid	7.7	-4.0	-3.4	-0.3	2.0	4.5	-2.7	-3.4	-2.5	-0.6	-0.3	-0.3
Hie	2.6	-2.9	0.4	-0.1	-0.6	3.6	< 0.1	-2.4	-1.1	-1.3	0.1	-0.7
Hip	-0.9	-1.5	2.2	0.1	-2.0	1.6	1.7	-2.8	-2.2	-0.2	-0.1	1.2
Ile	2.1	-1.8	-0.2	-0.1	-0.4	2.4	-0.2	-2.0	-1.1	-0.8	0.1	<< -0.1
Leu	2.6	0.8	-2.8	-0.7	-0.7	2.7	-2.2	0.4	< 0.1	0.4	0.1	-0.4
Lys	5.2	-2.5	-2.4	-0.4	< 0.1	4.1	-1.9	-2.6	-1.5	-1.1	0.2	0.3
Met	4.4	-3.1	-0.7	-0.6	0.3	3.2	-0.7	-2.7	-1.9	-0.7	-0.3	< 0.1
Phe	4.5	-4.2	-0.3	0.1	0.4	3.8	-0.4	-3.6	-2.5	-0.7	-0.4	-0.1
Pro	-1.6	1.6	<< -0.1	-	-1.4	<< 0.1	-	0.9	1.1	-0.2	< -0.1	0.6
Ser	3.3	-3.9	0.5	0.1	-0.5	3.7	0.3	-3.9	-1.6	-2.2	< -0.1	0.3
Thr	0.2	-0.3	0.1	0.0	-2.7	3.0	0.1	-0.7	-0.2	-0.6	0.3	0.1
Trp	3.3	-1.4	-1.7	-0.2	-0.6	3.5	-1.5	-1.3	-0.9	-0.6	0.6	-0.5
Tyr	0.9	1.5	-1.9	-0.5	-1.8	2.7	-1.8	1.7	0.7	0.9	0.2	-0.9
Val	8.8	-7.8	-1.2	0.2	2.3	5.8	-1.0	-9.5	-4.2	-5.1	0.1	2.0

Chapter 6 Supplemental Figure Captions

Figure A.6.1. Convergence of guest residue sampling across different portions of the trajectories and in replicate simulations under native conditions. Comparison of fraction of time spent in each of the 64 possible conformational states over different portions of the trajectory (above) and across replicate simulations (below) for GGAGG (A), GGGGG (B), and GGWGG (C) in pure water at 298 K.

Figure A.6.2. Convergence of guest residue sampling across different portions of the trajectories and in replicate simulations under chemically denaturing conditions. Comparison of fraction of time spent in each of the 64 possible conformational states over different portions of the trajectory (above) and across replicate simulations (below) for GGAGG (A), GGGGG (B), and GGWGG (C) in 8M urea at 298 K.

Figure A.6.3. Intramolecular interactions do not change in response to host peptide sequence or environment. (A) Hydrogen bond formation in GGAGG, water 298 (left) and in AAAAA, water 298 (right). (B) Hydrophobic interactions in AAAAA, water 298 (left) and in water 498 (right).

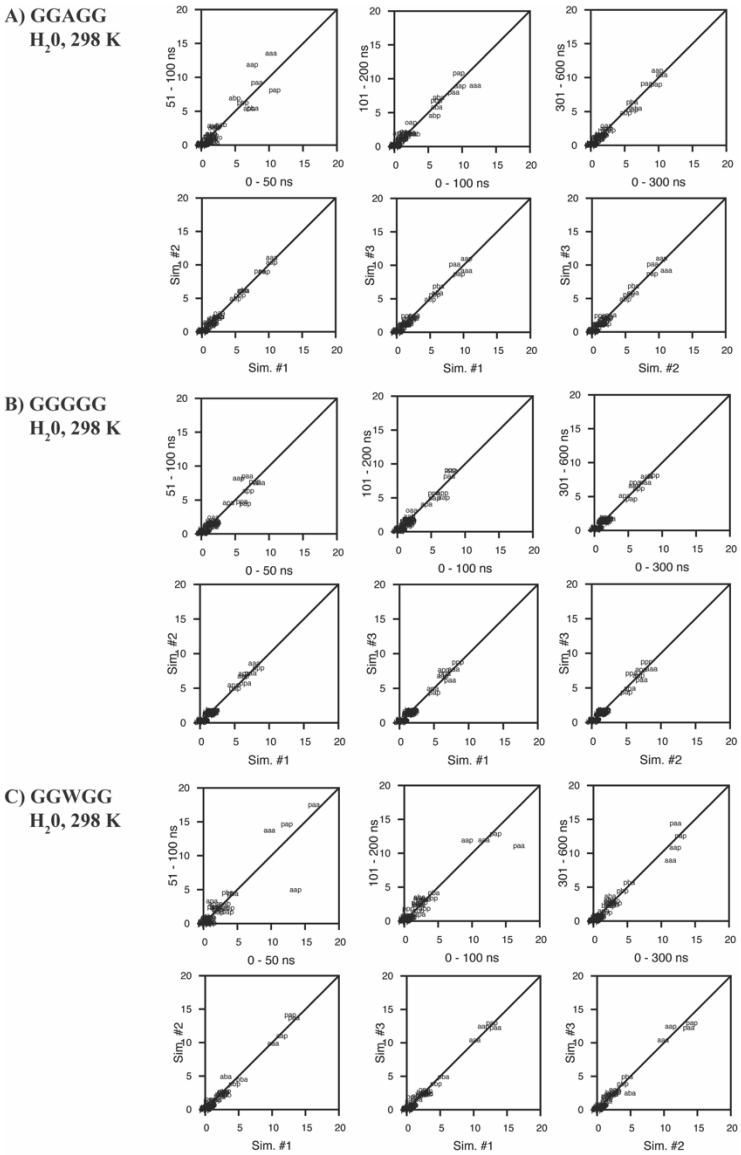

A) GGAGG
H20, 298 K
B) GGGGG
H20, 298 K
C) GGWGG
H20, 298 K
51 - 100 ns
101 - 200 ns
301 - 600 ns
0 - 50 ns
0 - 100 ns
0 - 300 ns
Sim. #2
Sim. #3
Sim. #1

Figure A.6.1

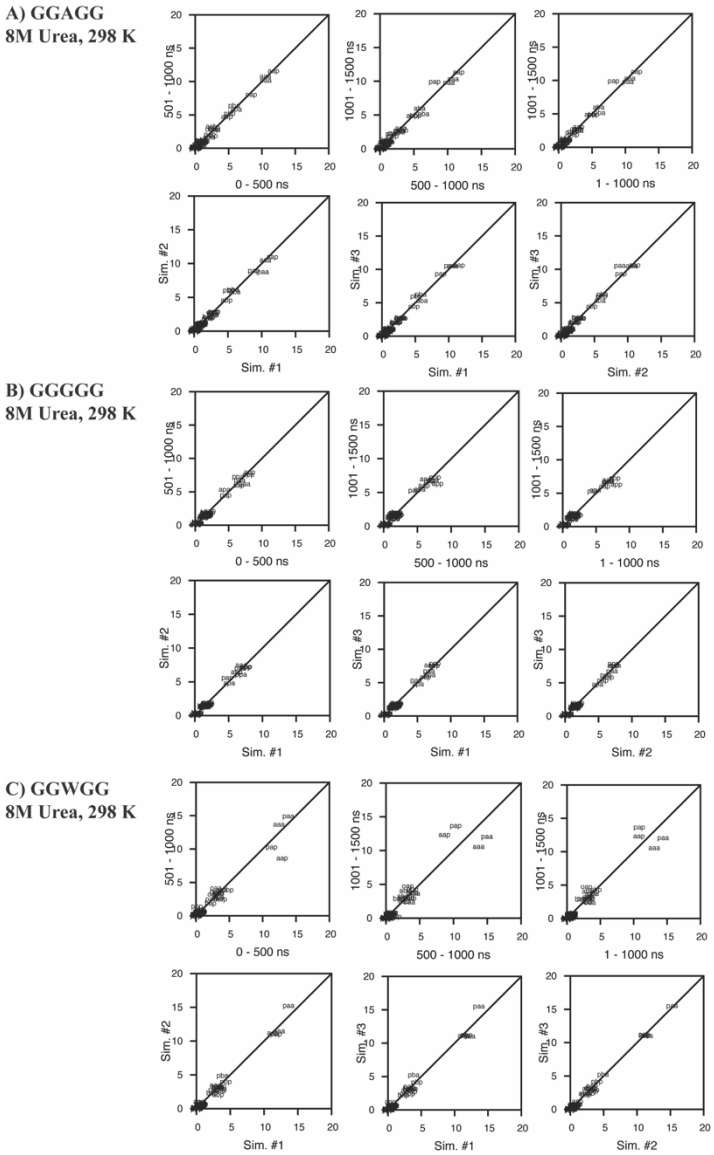

A) GGAGG
8M Urea, 298 K
B) GGGGG
8M Urea, 298 K
C) GGWGG
8M Urea, 298 K
501 - 1000 ns
1001 - 1500 ns
0 - 500 ns
500 - 1000 ns
1 - 1000 ns
Sim. #2
Sim. #3
Sim. #1

Figure A.6.2

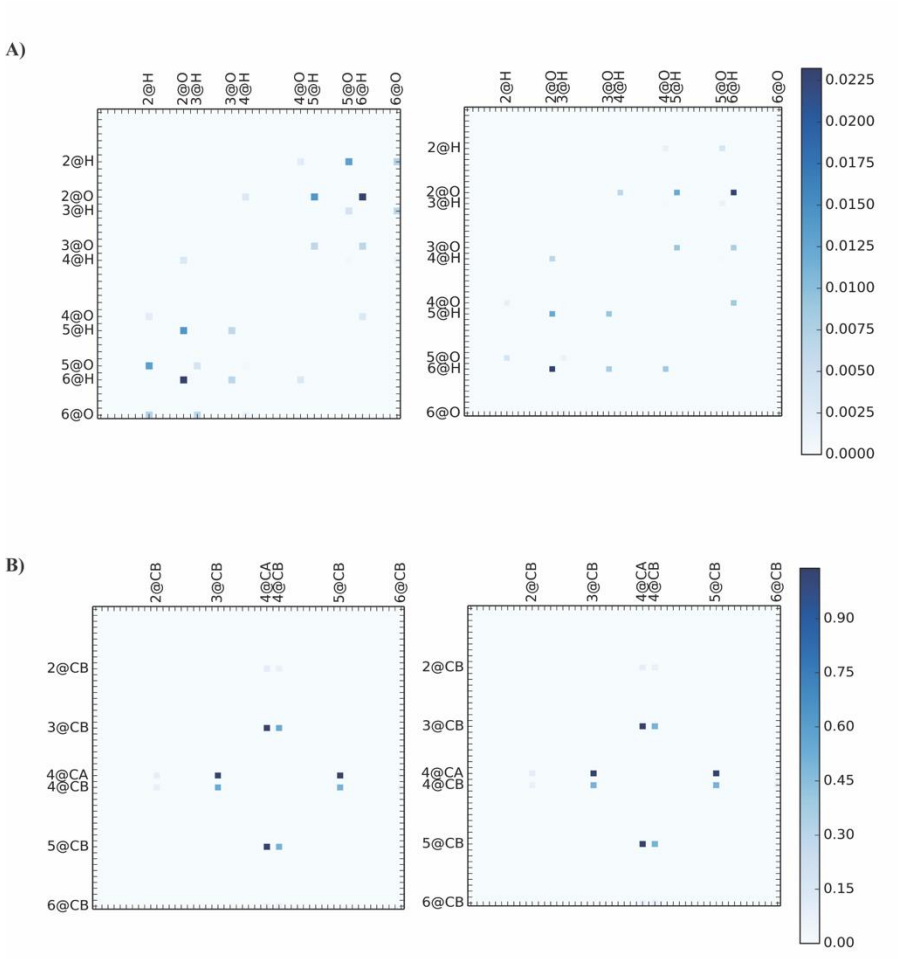

A)
2@H
2@O
3@H
3@O
4@H
4@O
5@H
5@O
6@H
6@O
0.0225
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
B)
2@CB
3@CB
4@CA
4@CB
5@CB
6@CB
0.90
0.75
0.60
0.45
0.30
0.15
0.00

Figure A.6.3

Chapter 7 Supplemental Tables

Table A.7.1. Summary of simulations described in this article.

System	# Guest Residues	Simulation Length (ns)	Aggregate Simulation Length (us)
AAXAA	22	600	13.2
AAxAA	22	600	13.2
aaXaa	22	600	13.2
aaxaa	22	600	13.2
Total	88	600	52.8

Table A.7.2 Summary of residues simulated.

L-Residue	D-Residue	# Chi Angles	# Rotameric Rotamers	Total # Rotamers
Arg	Dar	4	81	81
Ash[7]	Dah	2	3	18
Asn	Dan	2	3	18
Asp	Das	2	3	18
Cyh[1]	Dcp	1	3	3
Gln	Dgn	3	9	54
Glh[8]	Dgh	3	9	27
Glu	Dgl	3	9	27
Hid[2]	Dhd	2	3	18
Hie[3]	Dhe	2	3	18
Hip[4]	Dhp	2	3	18
Ile[5]	Dil	2	9	9
Leu	Dle	2	9	9
Lys	Dly	4	81	81
Met	Med	3	27	27
Phe	Dpn	2	3	12
Pro	Dpr	1	-	2
Ser	Dsn	1	3	3
Thr[6]	Dth	1	3	3
Trp	Dtr	2	9	9
Tyr	Dty	2	3	12
Val	Dva	1	3	3

1. Modeled in the reduced state
2. Protonated on ND
3. Protonated on NE
4. Protonated on ND & NE / acidic pH model
5. Naturally dominant side chain chirality
6. Natrually dominant side chain chirality
7. Protonated Asp / acidic pH model
8. Protonated Glu / acidic pH model

Table A.7.3. Rotamer bin definitions

Group	Conformation	Angular Range	Angles
1	*g+*	$0 \leq \chi < 120$	Arg $\chi_{1\text{-}4}$; Asx χ_1; Cyh χ_1; Glx χ_1; His $\chi_{1;}$ Ile $\chi_{1\text{-}2}$; Leu $\chi_{1\text{-}2}$; Lys $\chi_{1\text{-}4;}$ Met $\chi_{1\text{-}3;}$ Phe $\chi_{1;}$ Ser χ_1; Thr χ_1
1	*t*	$120 \leq \chi < 180$ $-180 \leq \chi < -120$	
1	*g-*	$-120 \leq \chi < 9$	
2	*g+*	$-180 \leq \chi < -60$	Trp χ_2
2	*t*	$-60 \leq \chi < 60$	
2	*g-*	$60 \leq \chi < 180$	
3	*Ng+*	$-150 \leq \chi < -90$	Asn χ_2; Gln χ_3
3	*Og-*	$-90 \leq \chi < -30$	
3	*Nt*	$-30 \leq \chi < 30$	
3	*Og+*	$30 \leq \chi < 90$	
3	*Ng-*	$90 \leq \chi < 150$	
3	*Ot*	$150 \leq \chi < 180$ $-180 \leq \chi < -150$	
4	*g+*	$30 \leq \chi < 90$	Asp χ_2, Glu χ_3
4	*t*	$-30 \leq \chi < 30$	
4	*g-*	$-90 \leq \chi < -30$	
5	*Ng+*	$30 \leq \chi < 90$	His χ_2
5	*Cg-*	$90 \leq \chi < 150$	
5	*Nt*	$150 \leq \chi < 180$ $-180 \leq \chi < -150$	
5	*Cg+*	$-150 \leq \chi < -90$	
5	*Ng-*	$-90 \leq \chi < -30$	
5	*Ct*	$-30 \leq \chi < 30$	
6	*g*	$45 \leq \chi < 135$	Phe χ_2; Tyr χ_2
6	*t*	$135 \leq \chi < 180$ $-180 \leq \chi < -135$	
7	*g-*	$-180 \leq \chi < 0$	Pro χ_1
7	*g+*	$0 \leq \chi < 180$	

Table A.7.4. Residue specific Karplus equation coefficients

J type	Residue	C_0	C_1	C_2
$^3J_{H\alpha,H\beta}$	*fundamental*	7.24		
	Ala	6.63		
	Arg, Asx, Glx, His, Leu, Lys, Met, Phe, Pro, Trp, Tyr	6.01		
	consensus	5.83	-1.37	3.61
	Ile, Val	5.40		
	Cys	5.32		
	Ser	5.03		
	Thr	4.42		
$^3J_{N',H\beta}$	*fundamental*	2.27		
	Ala	2.47		
	Arg, Asx, Glx, His, Leu, Lys, Met, Phe, Trp, Tyr	2.28		
	Pro	2.23		
	consensus	2.22	-0.75	1.15
	Ile, Val	2.08		
	Cys	2.06		
	Ser	1.97		
	Thr	1.77		
$^3J_{C',H\beta}$	*fundamental*	2.21		
	Ala	3.72		
	Arg, Asx, Glx, His, Leu, Lys, Met, Phe, Pro, Trp, Tyr	3.41		
	consensus	3.32	-1.58	2.01
	Ile, Val	3.11		
	Cys	3.07		
	Ser	2.92		
	Thr	2.62		
$^3J_{C',C\gamma}$	Asn	2.52		
	Asp	2.22		
	His	2.14		
	Met	1.76		
	Arg, Glx, Lys, Pro	1.72		
	consensus	1.70	-0.87	1.15
	Leu	1.68		
	Phe, Trp, Tyr	1.64		
	Val	1.61		
	Ile	1.57		
	Thr	1.36		
	fundamental	1.00		
$^3J_{N',C\gamma}$	Asn	1.54		
	Asp	1.35		
	His	1.30		
	fundamental	1.18		

Met	1.06		
Arg, Glx, Lys	1.03		
consensus	1.02	-0.49	0.65
Leu, Pro	1.01		
Phe, Trp, Tyr	0.98		
Val	0.96		
Ile	0.93		
Thr	0.80		

Table A.7.5. L-Valine and D-Valine rotamer distributions in chiral and achiral host pentapeptides.

Residue	Rotamer	GGVGG	GGvGG	Rotamer	Residue
L-Val	*g*+	0.71	1.56	*g*-	D-Val
L-Val	*t*	35.07	38.41	*t*	D-Val
L-Val	*g*-	64.21	60.02	*g*+	D-Val

Chapter 7 Supplemental Figure

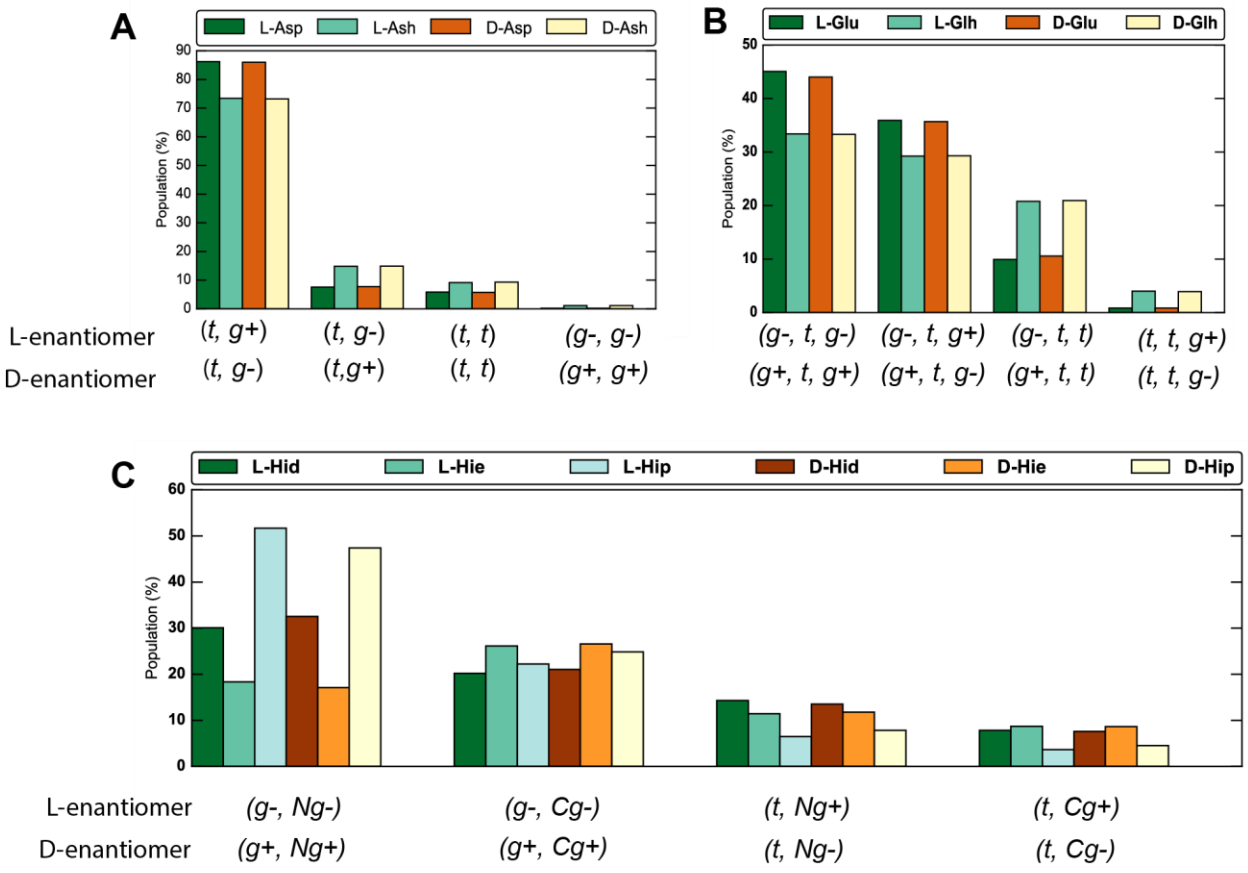


Figure A.7.1. The rotamers of Asp, Glu, and His are sensitive to protonation state. Comparison of the top four most populated rotamers of Asp/Ash (A), Glu/Glh (B), and Hie/Hid/Hip (C) shows that the most populated rotamers for each residue are sensitive to the protonation state of the side chain.

APPENDIX B: ROTAMER LIBRARIES FOR L- AND D-AMINO ACIDS

Table B.7.1. Backbone independent rotamer probabilities for D- and L- amino acids derived from the Dynameomics database, achiral pentapeptides, heterochiral pentapeptides, and homochiral pentapeptides. The ordering of the D-amino acid rotamers reflects the near-equivalent populations of 'mirror-image' rotamers for L- and D- guest residues.

Residue	χ^1	χ^2	χ^3	χ^4	BBIND	AAXAA	aaXaa	AAxAA	aaxaa	χ^1	χ^2	χ^3	χ^4	Residue
ARG	g+	g+	g+	g+	0.02	0.00	0.00	0.00	0.00	g-	g-	g-	g-	DAR
ARG	g+	g+	g+	t	0.03	0.00	0.01	0.07	0.02	g-	g-	g-	t	DAR
ARG	g+	g+	g+	g-	< 0.01	0.00	0.00	0.00	0.00	g-	g-	g-	g+	DAR
ARG	g+	g+	t	g+	0.01	0.00	0.00	0.00	0.00	g-	g-	t	g-	DAR
ARG	g+	g+	t	t	0.08	0.02	0.00	0.00	0.00	g-	g-	t	t	DAR
ARG	g+	g+	t	g-	0.01	0.00	0.00	0.00	0.00	g-	g-	t	g+	DAR
ARG	g+	g+	g-	g+	< 0.01	0.00	0.00	0.00	0.00	g-	g-	g+	g-	DAR
ARG	g+	g+	g-	t	0.02	0.00	0.00	0.00	0.00	g-	g-	g+	t	DAR
ARG	g+	g+	g-	g-	0.01	0.00	0.00	0.00	0.00	g-	g-	g+	g+	DAR
ARG	g+	t	g+	g+	0.48	0.42	0.53	0.63	0.40	g-	t	g-	g-	DAR
ARG	g+	t	g+	t	0.86	0.86	0.56	0.70	1.26	g-	t	g-	t	DAR
ARG	g+	t	g+	g-	0.08	0.01	0.01	0.02	0.02	g-	t	g-	g+	DAR
ARG	g+	t	t	g+	0.68	0.18	0.03	0.05	0.04	g-	t	t	g-	DAR
ARG	g+	t	t	t	0.64	0.64	0.34	0.35	0.51	g-	t	t	t	DAR
ARG	g+	t	t	g-	0.39	0.17	0.04	0.27	0.27	g-	t	t	g+	DAR
ARG	g+	t	g-	g+	0.08	0.02	0.01	0.03	0.01	g-	t	g+	g-	DAR
ARG	g+	t	g-	t	0.70	0.26	0.51	0.67	0.41	g-	t	g+	t	DAR
ARG	g+	t	g-	g-	0.44	0.26	0.48	0.63	0.59	g-	t	g+	g+	DAR
ARG	g+	g-	g+	g+	0.01	0.00	0.00	0.00	0.00	g-	g+	g-	g-	DAR
ARG	g+	g-	g+	t	< 0.01	0.00	0.00	0.00	0.00	g-	g+	g-	t	DAR
ARG	g+	g-	g+	g-	< 0.01	0.00	0.00	0.00	0.00	g-	g+	g-	g+	DAR
ARG	g+	g-	t	g+	< 0.01	0.00	0.00	0.00	0.00	g-	g+	t	g-	DAR
ARG	g+	g-	t	t	0.01	0.00	0.00	0.01	0.01	g-	g+	t	t	DAR
ARG	g+	g-	t	g-	0.01	0.00	0.00	0.00	0.00	g-	g+	t	g+	DAR

Residue	χ^1	χ^2	χ^3	χ^4	BBIND	AAXAA	aaXaa	AAxAA	aaxaa	χ^1	χ^2	χ^3	χ^4	Residue
ARG	g+	g-	g-	g+	< 0.01	0.00	0.00	0.00	0.00	g-	g+	g+	g-	DAR
ARG	g+	g-	g-	t	0.02	0.00	0.02	0.00	0.00	g-	g+	g+	t	DAR
ARG	g+	g-	g-	g-	< 0.01	0.00	0.01	0.00	0.00	g-	g+	g+	g+	DAR
ARG	t	g+	g+	g+	0.31	0.00	0.23	0.10	0.22	t	g-	g-	g-	DAR
ARG	t	g+	g+	t	0.75	0.09	0.69	0.56	0.38	t	g-	g-	t	DAR
ARG	t	g+	g+	g-	0.05	0.00	0.00	0.01	0.01	t	g-	g-	g+	DAR
ARG	t	g+	t	g+	0.58	0.06	0.09	0.34	0.31	t	g-	t	g-	DAR
ARG	t	g+	t	t	1.26	0.43	0.61	0.90	0.69	t	g-	t	t	DAR
ARG	t	g+	t	g-	0.47	0.11	0.07	0.12	0.08	t	g-	t	g+	DAR
ARG	t	g+	g-	g+	0.08	0.00	0.00	0.00	0.00	t	g-	g+	g-	DAR
ARG	t	g+	g-	t	0.11	0.02	0.00	0.02	0.00	t	g-	g+	t	DAR
ARG	t	g+	g-	g-	0.27	0.03	0.00	0.01	0.01	t	g-	g+	g+	DAR
ARG	t	t	g+	g+	2.75	0.55	0.58	0.84	0.67	t	t	g-	g-	DAR
ARG	t	t	g+	t	3.37	0.88	1.31	1.05	1.51	t	t	g-	t	DAR
ARG	t	t	g+	g-	0.62	0.08	0.02	0.03	0.07	t	t	g-	g+	DAR
ARG	t	t	t	g+	0.71	0.29	0.33	0.31	0.29	t	t	t	g-	DAR
ARG	t	t	t	t	1.92	0.75	1.14	1.01	0.78	t	t	t	t	DAR
ARG	t	t	t	g-	1.05	0.32	0.09	0.18	0.22	t	t	t	g+	DAR
ARG	t	t	g-	g+	0.45	0.05	0.04	0.04	0.02	t	t	g+	g-	DAR
ARG	t	t	g-	t	2.22	1.91	2.54	2.10	1.56	t	t	g+	t	DAR
ARG	t	t	g-	g-	1.17	0.44	0.90	0.63	0.40	t	t	g+	g+	DAR
ARG	t	g-	g+	g+	0.03	0.00	0.00	0.00	0.00	t	g+	g-	g-	DAR
ARG	t	g-	g+	t	0.03	0.00	0.00	0.00	0.00	t	g+	g-	t	DAR
ARG	t	g-	g+	g-	< 0.01	0.00	0.00	0.00	0.00	t	g+	g-	g+	DAR
ARG	t	g-	t	g+	0.11	0.01	0.01	0.01	0.02	t	g+	t	g-	DAR
ARG	t	g-	t	t	0.32	0.05	0.05	0.07	0.07	t	g+	t	t	DAR
ARG	t	g-	t	g-	0.05	0.01	0.04	0.00	0.01	t	g+	t	g+	DAR
ARG	t	g-	g-	g+	0.09	0.00	0.00	0.00	0.00	t	g+	g+	g-	DAR
ARG	t	g-	g-	t	0.41	0.29	0.11	0.18	0.24	t	g+	g+	t	DAR
ARG	t	g-	g-	g-	0.06	0.01	0.01	0.02	0.02	t	g+	g+	g+	DAR
ARG	g-	g+	g+	g+	0.17	0.02	0.03	0.05	0.06	g+	g-	g-	g-	DAR

Residue	χ^1	χ^2	χ^3	χ^4	BBIND	AAXAA	aaXaa	AAxAA	aaxaa	χ^1	χ^2	χ^3	χ^4	Residue
ARG	g-	g+	g+	t	0.44	0.35	0.74	0.44	0.43	g+	g-	g-	t	DAR
ARG	g-	g+	g+	g-	0.04	0.00	0.00	0.00	0.00	g+	g-	g-	g+	DAR
ARG	g-	g+	t	g+	0.10	0.05	0.06	0.12	0.17	g+	g-	t	g-	DAR
ARG	g-	g+	t	t	0.39	0.24	0.37	0.27	0.39	g+	g-	t	t	DAR
ARG	g-	g+	t	g-	0.11	0.05	0.05	0.08	0.04	g+	g-	t	g+	DAR
ARG	g-	g+	g-	g+	0.01	0.00	0.00	0.00	0.00	g+	g-	g+	g-	DAR
ARG	g-	g+	g-	t	0.11	0.02	0.02	0.00	0.00	g+	g-	g+	t	DAR
ARG	g-	g+	g-	g-	0.03	0.00	0.00	0.00	0.00	g+	g-	g+	g+	DAR
ARG	g-	t	g+	g+	4.67	5.78	5.82	6.33	6.18	g+	t	g-	g-	DAR
ARG	g-	t	g+	t	11.72	15.45	20.70	19.94	18.15	g+	t	g-	t	DAR
ARG	g-	t	g+	g-	0.75	0.51	0.43	0.37	0.62	g+	t	g-	g+	DAR
ARG	g-	t	t	g+	3.48	3.24	2.90	2.94	3.34	g+	t	t	g-	DAR
ARG	g-	t	t	t	11.68	11.07	14.36	15.24	10.39	g+	t	t	t	DAR
ARG	g-	t	t	g-	4.00	4.05	3.94	3.49	2.77	g+	t	t	g+	DAR
ARG	g-	t	g-	g+	2.17	0.80	0.73	0.82	0.68	g+	t	g+	g-	DAR
ARG	g-	t	g-	t	10.92	20.69	14.17	13.02	19.52	g+	t	g+	t	DAR
ARG	g-	t	g-	g-	11.06	12.17	8.22	9.30	11.08	g+	t	g+	g+	DAR
ARG	g-	g-	g+	g+	0.69	0.17	0.27	0.18	0.07	g+	g+	g-	g-	DAR
ARG	g-	g-	g+	t	0.83	0.14	0.11	0.11	0.14	g+	g+	g-	t	DAR
ARG	g-	g-	g+	g-	0.13	0.02	0.01	0.01	0.01	g+	g+	g-	g+	DAR
ARG	g-	g-	t	g+	1.30	0.97	1.41	1.51	0.87	g+	g+	t	g-	DAR
ARG	g-	g-	t	t	3.76	5.48	5.18	5.63	6.05	g+	g+	t	t	DAR
ARG	g-	g-	t	g-	2.16	1.40	1.44	1.97	1.03	g+	g+	t	g+	DAR
ARG	g-	g-	g-	g+	0.40	0.09	0.13	0.06	0.05	g+	g+	g+	g-	DAR
ARG	g-	g-	g-	t	3.64	6.78	6.15	5.16	5.94	g+	g+	g+	t	DAR
ARG	g-	g-	g-	g-	1.41	1.19	1.34	0.97	0.89	g+	g+	g+	g+	DAR
ASN	g+	Nt	-	-	0.87	0.14	0.09	0.05	0.19	g-	Nt	-	-	DAN
ASN	g+	Og+	-	-	2.38	0.81	0.56	0.49	1.19	g-	Og-	-	-	DAN
ASN	g+	Ng-	-	-	0.25	0.12	0.10	0.11	0.22	g-	Ng+	-	-	DAN
ASN	g+	Ot	-	-	< 0.01	0.00	0.00	0.00	0.00	g-	Ot	-	-	DAN

Residue	χ^1	χ^2	χ^3	χ^4	BBIND	AAXAA	aaXaa	AAxAA	aaxaa	χ^1	χ^2	χ^3	χ^4	Residue
ASN	g+	Ng+	-	-	0.15	0.07	0.05	0.06	0.14	g-	Ng-	-	-	DAN
ASN	g+	Og-	-	-	1.01	0.57	0.47	0.30	0.79	g-	Og+	-	-	DAN
ASN	t	Nt	-	-	11.18	11.00	12.09	12.96	11.87	t	Nt	-	-	DAN
ASN	t	Og+	-	-	25.14	32.48	29.31	30.33	34.06	t	Og-	-	-	DAN
ASN	t	Ng-	-	-	1.05	1.09	1.14	1.05	1.01	t	Ng+	-	-	DAN
ASN	t	Ot	-	-	0.19	0.17	0.19	0.18	0.18	t	Ot	-	-	DAN
ASN	t	Ng+	-	-	25.68	26.48	27.76	26.37	24.47	t	Ng-	-	-	DAN
ASN	t	Og-	-	-	20.39	16.40	18.29	18.64	15.38	t	Og+	-	-	DAN
ASN	g-	Nt	-	-	2.09	1.50	1.73	1.63	1.52	g+	Nt	-	-	DAN
ASN	g-	Og+	-	-	1.85	2.35	2.37	2.29	2.28	g+	Og-	-	-	DAN
ASN	g-	Ng-	-	-	1.83	2.45	2.13	1.95	2.31	g+	Ng+	-	-	DAN
ASN	g-	Ot	-	-	0.05	0.03	0.04	0.04	0.03	g+	Ot	-	-	DAN
ASN	g-	Ng+	-	-	0.62	0.25	0.20	0.20	0.25	g+	Ng-	-	-	DAN
ASN	g-	Og-	-	-	5.27	4.12	3.49	3.36	4.11	g+	Og+	-	-	DAN
ASP	g+	t	-	-	0.30	0.00	0.00	0.00	0.00	g-	t	-	-	DAS
ASP	g+	g+	-	-	1.40	0.03	0.00	0.06	0.07	g-	g-	-	-	DAS
ASP	g+	g-	-	-	2.56	0.05	0.02	0.20	0.17	g-	g+	-	-	DAS
ASP	t	t	-	-	7.38	5.82	4.90	4.57	5.71	t	t	-	-	DAS
ASP	t	g+	-	-	66.45	86.24	85.34	85.63	86.02	t	g-	-	-	DAS
ASP	t	g-	-	-	12.81	7.55	9.41	9.33	7.73	t	g+	-	-	DAS
ASP	g-	t	-	-	1.48	0.02	0.02	0.00	0.02	g+	t	-	-	DAS
ASP	g-	g+	-	-	2.10	0.10	0.10	0.06	0.05	g+	g-	-	-	DAS
ASP	g-	g-	-	-	5.51	0.18	0.21	0.14	0.22	g+	g+	-	-	DAS
ASH	g+	t	-	-	-	0.02	0.01	0.01	0.03	g-	t	-	-	DAH
ASH	g+	g+	-	-	-	0.22	0.15	0.19	0.21	g-	g-	-	-	DAH
ASH	g+	g-	-	-	-	0.56	0.33	0.38	0.51	g-	g+	-	-	DAH
ASH	t	t	-	-	-	9.13	9.26	8.96	9.32	t	t	-	-	DAH
ASH	t	g+	-	-	-	73.44	73.23	73.33	73.23	t	g-	-	-	DAH
ASH	t	g-	-	-	-	14.84	15.79	15.54	14.91	t	g+	-	-	DAH

Residue	χ^1	χ^2	χ^3	χ^4	BBIND	AAXAA	aaXaa	AAxAA	aaxaa	χ^1	χ^2	χ^3	χ^4	Residue
ASH	g-	t	-	-	-	0.21	0.17	0.22	0.20	g+	t	-	-	DAH
ASH	g-	g+	-	-	-	0.47	0.38	0.49	0.48	g+	g-	-	-	DAH
ASH	g-	g-	-	-	-	1.09	0.68	0.89	1.12	g+	g+	-	-	DAH
CYH	g+	-	-	-	14	15.59	12.47	12.80	16.79	g-	-	-	-	DCH
CYH	t	-	-	-	16.97	8.03	8.91	9.41	7.68	t	-	-	-	DCH
CYH	g-	-	-	-	69.03	76.38	78.62	77.79	75.53	g+	-	-	-	DCH
GLN	g+	g+	Nt	-	< 0.01	0.00	0.00	0.00	0.00	g-	g-	Nt	-	DGN
GLN	g+	g+	Og+	-	0.01	0.01	0.00	0.00	0.01	g-	g-	Og-	-	DGN
GLN	g+	g+	Ng-	-	< 0.01	0.00	0.00	0.00	0.00	g-	g-	Ng+	-	DGN
GLN	g+	g+	Ot	-	< 0.01	0.00	0.00	0.00	0.00	g-	g-	Ot	-	DGN
GLN	g+	g+	Ng+	-	0.01	0.01	0.01	0.00	0.00	g-	g-	Ng-	-	DGN
GLN	g+	g+	Og-	-	0.01	0.00	0.01	0.00	0.00	g-	g-	Og+	-	DGN
GLN	g+	t	Nt	-	0.29	0.46	0.37	0.31	0.49	g-	t	Nt	-	DGN
GLN	g+	t	Og+	-	0.58	0.77	0.63	0.52	0.75	g-	t	Og-	-	DGN
GLN	g+	t	Ng-	-	0.42	0.44	0.40	0.38	0.34	g-	t	Ng+	-	DGN
GLN	g+	t	Ot	-	0.08	0.11	0.08	0.08	0.09	g-	t	Ot	-	DGN
GLN	g+	t	Ng+	-	0.52	0.58	0.61	0.50	0.61	g-	t	Ng-	-	DGN
GLN	g+	t	Og-	-	0.51	0.57	0.63	0.52	0.67	g-	t	Og+	-	DGN
GLN	g+	g-	Nt	-	0.01	0.01	0.01	0.00	0.01	g-	g+	Nt	-	DGN
GLN	g+	g-	Og+	-	0.04	0.06	0.05	0.03	0.05	g-	g+	Og-	-	DGN
GLN	g+	g-	Ng-	-	0.03	0.06	0.03	0.02	0.04	g-	g+	Ng+	-	DGN
GLN	g+	g-	Ot	-	< 0.01	0.00	0.00	0.00	0.00	g-	g+	Ot	-	DGN
GLN	g+	g-	Ng+	-	< 0.01	0.00	0.00	0.00	0.00	g-	g+	Ng-	-	DGN
GLN	g+	g-	Og-	-	0.01	0.01	0.01	0.00	0.00	g-	g+	Og+	-	DGN
GLN	t	g+	Nt	-	0.12	0.07	0.12	0.08	0.08	t	g-	Nt	-	DGN
GLN	t	g+	Og+	-	0.65	0.22	0.36	0.31	0.24	t	g-	Og-	-	DGN
GLN	t	g+	Ng-	-	0.04	0.01	0.02	0.02	0.01	t	g-	Ng+	-	DGN
GLN	t	g+	Ot	-	0.01	0.01	0.01	0.01	0.01	t	g-	Ot	-	DGN
GLN	t	g+	Ng+	-	0.45	0.32	0.39	0.30	0.26	t	g-	Ng-	-	DGN

Residue	χ^1	χ^2	χ^3	χ^4	BBIND	AAXAA	aaXaa	AAxAA	aaxaa	χ^1	χ^2	χ^3	χ^4	Residue
GLN	t	g+	Og-	-	0.20	0.14	0.16	0.12	0.14	t	g-	Og+	-	DGN
GLN	t	t	Nt	-	1.10	0.85	1.13	1.25	0.84	t	t	Nt	-	DGN
GLN	t	t	Og+	-	2.09	1.58	1.96	2.10	1.53	t	t	Og-	-	DGN
GLN	t	t	Ng-	-	1.43	1.11	1.33	1.37	1.10	t	t	Ng+	-	DGN
GLN	t	t	Ot	-	0.26	0.20	0.21	0.21	0.19	t	t	Ot	-	DGN
GLN	t	t	Ng+	-	1.45	1.09	1.14	0.99	1.00	t	t	Ng-	-	DGN
GLN	t	t	Og-	-	1.74	1.31	1.66	1.55	1.30	t	t	Og+	-	DGN
GLN	t	g-	Nt	-	0.04	0.01	0.02	0.03	0.01	t	g+	Nt	-	DGN
GLN	t	g-	Og+	-	0.06	0.03	0.05	0.06	0.02	t	g+	Og-	-	DGN
GLN	t	g-	Ng-	-	0.10	0.05	0.07	0.07	0.07	t	g+	Ng+	-	DGN
GLN	t	g-	Ot	-	< 0.01	0.00	0.00	0.00	0.00	t	g+	Ot	-	DGN
GLN	t	g-	Ng+	-	0.01	0.00	0.01	0.01	0.00	t	g+	Ng-	-	DGN
GLN	t	g-	Og-	-	0.14	0.04	0.09	0.11	0.05	t	g+	Og+	-	DGN
GLN	g-	g+	Nt	-	0.34	0.31	0.27	0.29	0.28	g+	g-	Nt	-	DGN
GLN	g-	g+	Og+	-	0.45	0.62	0.66	0.76	0.58	g+	g-	Og-	-	DGN
GLN	g-	g+	Ng-	-	0.01	0.01	0.02	0.01	0.01	g+	g-	Ng+	-	DGN
GLN	g-	g+	Ot	-	< 0.01	0.00	0.00	0.00	0.00	g+	g-	Ot	-	DGN
GLN	g-	g+	Ng+	-	0.43	0.58	0.53	0.55	0.62	g+	g-	Ng-	-	DGN
GLN	g-	g+	Og-	-	0.83	0.59	0.52	0.57	0.51	g+	g-	Og+	-	DGN
GLN	g-	t	Nt	-	10.62	11.06	13.08	13.36	10.84	g+	t	Nt	-	DGN
GLN	g-	t	Og+	-	20.10	16.14	13.74	14.04	16.43	g+	t	Og-	-	DGN
GLN	g-	t	Ng-	-	15.39	15.38	12.75	12.44	15.51	g+	t	Ng+	-	DGN
GLN	g-	t	Ot	-	2.22	2.52	2.70	2.62	2.58	g+	t	Ot	-	DGN
GLN	g-	t	Ng+	-	10.84	12.35	13.15	12.81	12.38	g+	t	Ng-	-	DGN
GLN	g-	t	Og-	-	17.59	20.03	20.71	20.54	20.08	g+	t	Og+	-	DGN
GLN	g-	g-	Nt	-	0.93	0.86	0.87	0.96	0.81	g+	g+	Nt	-	DGN
GLN	g-	g-	Og+	-	1.26	1.35	1.25	1.27	1.45	g+	g+	Og-	-	DGN
GLN	g-	g-	Ng-	-	2.54	3.41	3.52	3.73	3.74	g+	g+	Ng+	-	DGN
GLN	g-	g-	Ot	-	0.04	0.04	0.04	0.05	0.04	g+	g+	Ot	-	DGN
GLN	g-	g-	Ng+	-	0.02	0.26	0.24	0.27	0.25	g+	g+	Ng-	-	DGN
GLN	g-	g-	Og-	-	3.83	4.32	4.40	4.76	3.97	g+	g+	Og+	-	DGN

Residue	χ^1	χ^2	χ^3	χ^4	BBIND	AAXAA	aaXaa	AAxAA	aaxaa	χ^1	χ^2	χ^3	χ^4	Residue
GLU	g+	g+	t	-	0.01	0.00	0.00	0.00	0.00	g-	g-	t	-	DGL
GLU	g+	g+	g+	-	0.09	0.00	0.00	0.00	0.00	g-	g-	g-	-	DGL
GLU	g+	g+	g-	-	0.03	0.00	0.00	0.00	0.00	g-	g-	g+	-	DGL
GLU	g+	t	t	-	0.32	0.23	0.27	0.31	0.25	g-	t	t	-	DGL
GLU	g+	t	g+	-	1.31	0.80	1.26	1.02	0.91	g-	t	g-	-	DGL
GLU	g+	t	g-	-	1.12	0.69	1.06	0.95	0.67	g-	t	g+	-	DGL
GLU	g+	g-	t	-	0.03	0.00	0.00	0.00	0.00	g-	g+	t	-	DGL
GLU	g+	g-	g+	-	0.19	0.06	0.02	0.01	0.00	g-	g+	g-	-	DGL
GLU	g+	g-	g-	-	0.16	0.01	0.01	0.04	0.00	g-	g+	g+	-	DGL
GLU	t	g+	t	-	0.34	0.00	0.01	0.01	0.00	t	g-	t	-	DGL
GLU	t	g+	g+	-	1.69	0.07	0.14	0.11	0.04	t	g-	g-	-	DGL
GLU	t	g+	g-	-	0.35	0.02	0.02	0.02	0.01	t	g-	g+	-	DGL
GLU	t	t	t	-	1.59	0.75	0.96	1.03	0.81	t	t	t	-	DGL
GLU	t	t	g+	-	5.47	2.78	2.93	3.21	3.17	t	t	g-	-	DGL
GLU	t	t	g-	-	6.06	2.42	2.74	2.89	2.50	t	t	g+	-	DGL
GLU	t	g-	t	-	0.05	0.00	0.00	0.00	0.00	t	g+	t	-	DGL
GLU	t	g-	g+	-	0.08	0.00	0.00	0.00	0.00	t	g+	g-	-	DGL
GLU	t	g-	g-	-	0.71	0.00	0.01	0.01	0.00	t	g+	g+	-	DGL
GLU	g-	g+	t	-	0.48	0.02	0.08	0.06	0.07	g+	g-	t	-	DGL
GLU	g-	g+	g+	-	1.22	0.10	0.14	0.07	0.09	g+	g-	g-	-	DGL
GLU	g-	g+	g-	-	1.12	0.02	0.13	0.11	0.06	g+	g-	g+	-	DGL
GLU	g-	t	t	-	12.73	9.94	10.62	10.73	10.58	g+	t	t	-	DGL
GLU	g-	t	g+	-	25.45	35.92	34.15	33.81	35.67	g+	t	g-	-	DGL
GLU	g-	t	g-	-	29.69	45.05	44.51	44.52	44.03	g+	t	g+	-	DGL
GLU	g-	g-	t	-	1.28	0.04	0.09	0.08	0.10	g+	g+	t	-	DGL
GLU	g-	g-	g+	-	1.48	0.21	0.17	0.18	0.20	g+	g+	g-	-	DGL
GLU	g-	g-	g-	-	6.97	0.85	0.70	0.83	0.85	g+	g+	g+	-	DGL
GLH	g+	g+	t	-	-	0.00	0.00	0.00	0.00	g-	g-	t	-	DGH
GLH	g+	g+	g+	-	-	0.01	0.01	0.01	0.01	g-	g-	g-	-	DGH

Residue	χ^1	χ^2	χ^3	χ^4	BBIND	AAXAA	aaXaa	AAxAA	aaxaa	χ^1	χ^2	χ^3	χ^4	Residue
GLH	g+	g+	g-	-	-	0.00	0.00	0.00	0.00	g-	g-	g+	-	DGH
GLH	g+	t	t	-	-	0.73	0.35	0.43	0.68	g-	t	t	-	DGH
GLH	g+	t	g+	-	-	1.39	0.74	0.87	1.24	g-	t	g-	-	DGH
GLH	g+	t	g-	-	-	1.19	0.62	0.76	1.09	g-	t	g+	-	DGH
GLH	g+	g-	t	-	-	0.00	0.00	0.00	0.00	g-	g+	t	-	DGH
GLH	g+	g-	g+	-	-	0.00	0.00	0.00	0.01	g-	g+	g-	-	DGH
GLH	g+	g-	g-	-	-	0.01	0.00	0.00	0.02	g-	g+	g+	-	DGH
GLH	t	g+	t	-	-	0.04	0.04	0.05	0.06	t	g-	t	-	DGH
GLH	t	g+	g+	-	-	0.24	0.25	0.26	0.28	t	g-	g-	-	DGH
GLH	t	g+	g-	-	-	0.07	0.07	0.07	0.07	t	g-	g+	-	DGH
GLH	t	t	t	-	-	1.39	1.73	1.61	1.49	t	t	t	-	DGH
GLH	t	t	g+	-	-	2.50	3.06	2.89	2.55	t	t	g-	-	DGH
GLH	t	t	g-	-	-	2.29	2.61	2.46	2.38	t	t	g+	-	DGH
GLH	t	g-	t	-	-	0.01	0.01	0.01	0.01	t	g+	t	-	DGH
GLH	t	g-	g+	-	-	0.01	0.02	0.02	0.01	t	g+	g-	-	DGH
GLH	t	g-	g-	-	-	0.07	0.12	0.11	0.10	t	g+	g+	-	DGH
GLH	g-	g+	t	-	-	0.11	0.11	0.12	0.11	g+	g-	t	-	DGH
GLH	g-	g+	g+	-	-	0.74	0.68	0.70	0.72	g+	g-	g-	-	DGH
GLH	g-	g+	g-	-	-	0.19	0.19	0.18	0.22	g+	g-	g+	-	DGH
GLH	g-	t	t	-	-	20.79	23.36	23.62	20.94	g+	t	t	-	DGH
GLH	g-	t	g+	-	-	29.24	27.82	27.69	29.29	g+	t	g-	-	DGH
GLH	g-	t	g-	-	-	33.38	32.81	32.77	33.29	g+	t	g+	-	DGH
GLH	g-	g-	t	-	-	0.65	0.61	0.61	0.62	g+	g+	t	-	DGH
GLH	g-	g-	g+	-	-	0.90	0.90	0.86	0.87	g+	g+	g-	-	DGH
GLH	g-	g-	g-	-	-	4.01	3.87	3.91	3.92	g+	g+	g+	-	DGH
HID	g+	Ct	-	-	0.17	0.06	0.03	0.05	0.05	g-	Ct	-	-	DHD
HID	g+	Ng+	-	-	2.31	1.25	0.82	1.10	1.18	g-	Ng-	-	-	DHD
HID	g+	Cg-	-	-	1.46	0.49	0.30	0.37	0.46	g-	Cg+	-	-	DHD
HID	g+	Nt	-	-	0.04	0.01	0.00	0.00	0.01	g-	Nt	-	-	DHD
HID	g+	Cg+	-	-	2.59	1.41	0.76	1.40	1.33	g-	Cg-	-	-	DHD

Residue	χ^1	χ^2	χ^3	χ^4	BBIND	AAXAA	aaXaa	AAxAA	aaxaa	χ^1	χ^2	χ^3	χ^4	Residue
HID	g+	Ng-	-	-	1.78	0.62	0.26	0.68	0.56	g-	Ng+	-	-	DHD
HID	t	Ct	-	-	2.88	2.01	2.35	2.28	1.75	t	Ct	-	-	DHD
HID	t	Ng+	-	-	14.91	14.31	13.44	12.83	13.52	t	Ng-	-	-	DHD
HID	t	Cg-	-	-	3.48	2.83	3.03	2.92	2.69	t	Cg+	-	-	DHD
HID	t	Nt	-	-	0.90	0.56	0.66	0.67	0.52	t	Nt	-	-	DHD
HID	t	Cg+	-	-	9.56	7.85	8.74	8.76	7.59	t	Cg-	-	-	DHD
HID	t	Ng-	-	-	3.50	2.55	3.20	3.12	2.30	t	Ng+	-	-	DHD
HID	g-	Ct	-	-	5.62	5.84	7.04	7.21	4.96	g+	Ct	-	-	DHD
HID	g-	Ng+	-	-	7.93	4.07	3.26	3.45	3.59	g+	Ng-	-	-	DHD
HID	g-	Cg-	-	-	16.32	20.20	20.54	20.99	21.06	g+	Cg+	-	-	DHD
HID	g-	Nt	-	-	2.00	1.69	2.38	2.45	1.53	g+	Nt	-	-	DHD
HID	g-	Cg+	-	-	4.51	4.16	3.98	3.77	4.37	g+	Cg-	-	-	DHD
HID	g-	Ng-	-	-	20.04	30.08	29.20	27.96	32.55	g+	Ng+	-	-	DHD
HIE	g+	Ct	-	-	0.31	0.28	0.23	0.27	0.24	g-	Ct	-	-	DHE
HIE	g+	Ng+	-	-	2.62	2.76	2.65	3.21	2.55	g-	Ng-	-	-	DHE
HIE	g+	Cg-	-	-	0.81	0.73	0.50	0.58	0.59	g-	Cg+	-	-	DHE
HIE	g+	Nt	-	-	0.01	0.00	0.00	0.00	0.01	g-	Nt	-	-	DHE
HIE	g+	Cg+	-	-	2.58	3.77	2.37	2.97	4.20	g-	Cg-	-	-	DHE
HIE	g+	Ng-	-	-	4.68	4.75	3.18	4.45	5.25	g-	Ng+	-	-	DHE
HIE	t	Ct	-	-	4.2	2.86	2.34	2.24	2.82	t	Ct	-	-	DHE
HIE	t	Ng+	-	-	13.17	11.45	8.22	7.82	11.78	t	Ng-	-	-	DHE
HIE	t	Cg-	-	-	3.04	2.73	2.37	2.26	2.71	t	Cg+	-	-	DHE
HIE	t	Nt	-	-	0.78	0.54	0.65	0.64	0.53	t	Nt	-	-	DHE
HIE	t	Cg+	-	-	12.9	8.70	13.13	12.21	8.63	t	Cg-	-	-	DHE
HIE	t	Ng-	-	-	6.73	3.73	4.45	4.23	3.75	t	Ng+	-	-	DHE
HIE	g-	Ct	-	-	4.29	2.50	3.53	3.44	2.55	g+	Ct	-	-	DHE
HIE	g-	Ng+	-	-	5.66	6.02	6.90	7.15	6.17	g+	Ng-	-	-	DHE
HIE	g-	Cg-	-	-	16.2	26.14	30.22	30.94	26.57	g+	Cg+	-	-	DHE
HIE	g-	Nt	-	-	1.98	1.39	2.08	2.16	1.35	g+	Nt	-	-	DHE
HIE	g-	Cg+	-	-	4.02	3.30	2.71	2.48	3.20	g+	Cg-	-	-	DHE

Residue	χ^1	χ^2	χ^3	χ^4	BBIND	AAXAA	aaXaa	AAxAA	aaxaa	χ^1	χ^2	χ^3	χ^4	Residue
HIE	g-	Ng-	-	-	16.03	18.36	14.48	12.94	17.11	g+	Ng+	-	-	DHE
HIP	g+	Ct	-	-	-	0.00	0.00	0.00	0.00	g-	Ct	-	-	DHP
HIP	g+	Ng+	-	-	-	0.61	0.19	0.74	0.07	g-	Ng-	-	-	DHP
HIP	g+	Cg-	-	-	-	0.65	0.20	0.82	0.05	g-	Cg+	-	-	DHP
HIP	g+	Nt	-	-	-	0.00	0.00	0.01	0.00	g-	Nt	-	-	DHP
HIP	g+	Cg+	-	-	-	0.44	0.66	0.77	0.56	g-	Cg-	-	-	DHP
HIP	g+	Ng-	-	-	-	0.11	0.17	0.19	0.10	g-	Ng+	-	-	DHP
HIP	t	Ct	-	-	-	0.35	1.06	0.95	0.42	t	Ct	-	-	DHP
HIP	t	Ng+	-	-	-	6.49	14.20	13.65	7.83	t	Ng-	-	-	DHP
HIP	t	Cg-	-	-	-	1.71	3.60	3.63	2.17	t	Cg+	-	-	DHP
HIP	t	Nt	-	-	-	0.27	0.62	0.56	0.40	t	Nt	-	-	DHP
HIP	t	Cg+	-	-	-	3.66	5.20	4.71	4.53	t	Cg-	-	-	DHP
HIP	t	Ng-	-	-	-	0.59	0.80	0.80	0.69	t	Ng+	-	-	DHP
HIP	g-	Ct	-	-	-	1.87	3.26	3.02	1.49	g+	Ct	-	-	DHP
HIP	g-	Ng+	-	-	-	1.30	0.87	0.80	1.33	g+	Ng-	-	-	DHP
HIP	g-	Cg-	-	-	-	22.23	12.52	15.51	24.87	g+	Cg+	-	-	DHP
HIP	g-	Nt	-	-	-	1.77	3.46	3.96	1.99	g+	Nt	-	-	DHP
HIP	g-	Cg+	-	-	-	6.25	6.02	6.04	6.10	g+	Cg-	-	-	DHP
HIP	g-	Ng-	-	-	-	51.69	47.17	43.82	47.40	g+	Ng+	-	-	DHP
ILE	g+	g+	-	-	0.75	1.43	0.92	1.49	2.62	g-	g-	-	-	DIL
ILE	g+	t	-	-	43.32	75.05	65.80	29.24	38.67	g-	t	-	-	DIL
ILE	g+	g-	-	-	0.10	0.16	0.15	8.77	11.95	g-	g+	-	-	DIL
ILE	t	g+	-	-	0.98	0.54	0.13	6.28	5.50	t	g-	-	-	DIL
ILE	t	t	-	-	2.77	1.10	0.42	2.97	2.97	t	t	-	-	DIL
ILE	t	g-	-	-	0.05	0.00	0.00	8.91	8.43	t	g+	-	-	DIL
ILE	g-	g+	-	-	0.36	0.28	0.28	0.28	0.26	g+	g-	-	-	DIL
ILE	g-	t	-	-	28.55	9.33	11.73	2.47	1.68	g+	t	-	-	DIL
ILE	g-	g-	-	-	23.12	12.12	20.57	39.59	27.93	g+	g+	-	-	DIL

Residue	χ^1	χ^2	χ^3	χ^4	BBIND	AAXAA	aaXaa	AAxAA	aaxaa	χ^1	χ^2	χ^3	χ^4	Residue
LEU	g+	g+	-	-	0.26	0.13	0.05	0.06	0.09	g-	g-	-	-	DLE
LEU	g+	t	-	-	0.19	0.12	0.02	0.06	0.04	g-	t	-	-	DLE
LEU	g+	g-	-	-	< 0.01	0	0	0	0	g-	g+	-	-	DLE
LEU	t	g+	-	-	22.92	14.49	14.67	15.65	14.27	t	g-	-	-	DLE
LEU	t	t	-	-	3.26	1.85	2.08	2.27	1.77	t	t	-	-	DLE
LEU	t	g-	-	-	1.19	0.33	1.69	0.84	1.08	t	g+	-	-	DLE
LEU	g-	g+	-	-	4.18	4.63	4.62	4.45	4.39	g+	g-	-	-	DLE
LEU	g-	t	-	-	66.49	77.66	73.24	74.69	75.96	g+	t	-	-	DLE
LEU	g-	g-	-	-	1.51	0.8	3.62	1.99	2.4	g+	g+	-	-	DLE
LYS	g+	g+	g+	g+	0.01	0.00	0.00	0.00	0.00	g-	g-	g-	g-	DLY
LYS	g+	g+	g+	t	< 0.01	0.00	0.00	0.00	0.00	g-	g-	g-	t	DLY
LYS	g+	g+	g+	g-	< 0.01	0.00	0.00	0.00	0.00	g-	g-	g-	g+	DLY
LYS	g+	g+	t	g+	0.06	0.03	0.00	0.00	0.01	g-	g-	t	g-	DLY
LYS	g+	g+	t	t	0.01	0.00	0.00	0.00	0.00	g-	g-	t	t	DLY
LYS	g+	g+	t	g-	0.08	0.01	0.02	0.01	0.02	g-	g-	t	g+	DLY
LYS	g+	g+	g-	g+	< 0.01	0.00	0.00	0.00	0.00	g-	g-	g+	g-	DLY
LYS	g+	g+	g-	t	< 0.01	0.00	0.00	0.00	0.00	g-	g-	g+	t	DLY
LYS	g+	g+	g-	g-	< 0.01	0.00	0.00	0.00	0.00	g-	g-	g+	g+	DLY
LYS	g+	t	g+	g+	0.22	0.19	0.10	0.14	0.24	g-	t	g-	g-	DLY
LYS	g+	t	g+	t	0.14	0.08	0.07	0.03	0.10	g-	t	g-	t	DLY
LYS	g+	t	g+	g-	0.01	0.00	0.00	0.00	0.00	g-	t	g-	g+	DLY
LYS	g+	t	t	g+	2.25	1.33	0.25	0.46	0.58	g-	t	t	g-	DLY
LYS	g+	t	t	t	0.43	0.29	0.14	0.09	0.21	g-	t	t	t	DLY
LYS	g+	t	t	g-	1.21	0.92	0.42	0.50	0.87	g-	t	t	g+	DLY
LYS	g+	t	g-	g+	0.01	0.00	0.00	0.00	0.00	g-	t	g+	g-	DLY
LYS	g+	t	g-	t	0.13	0.06	0.02	0.02	0.10	g-	t	g+	t	DLY
LYS	g+	t	g-	g-	0.52	0.22	0.13	0.20	0.16	g-	t	g+	g+	DLY
LYS	g+	g-	g+	g+	< 0.01	0.00	0.00	0.00	0.00	g-	g+	g-	g-	DLY
LYS	g+	g-	g+	t	< 0.01	0.00	0.00	0.00	0.00	g-	g+	g-	t	DLY
LYS	g+	g-	g+	g-	< 0.01	0.00	0.00	0.00	0.00	g-	g+	g-	g+	DLY

Residue	χ^1	χ^2	χ^3	χ^4	BBIND	AAXAA	aaXaa	AAxAA	aaxaa	χ^1	χ^2	χ^3	χ^4	Residue
LYS	g+	g-	t	g+	0.02	0.01	0.01	0.00	0.01	g-	g+	t	g-	DLY
LYS	g+	g-	t	t	< 0.01	0.00	0.00	0.00	0.00	g-	g+	t	t	DLY
LYS	g+	g-	t	g-	0.02	0.01	0.01	0.00	0.00	g-	g+	t	g+	DLY
LYS	g+	g-	g-	g+	< 0.01	0.00	0.00	0.00	0.00	g-	g+	g+	g-	DLY
LYS	g+	g-	g-	t	< 0.01	0.00	0.00	0.00	0.00	g-	g+	g+	t	DLY
LYS	g+	g-	g-	g-	< 0.01	0.01	0.00	0.00	0.00	g-	g+	g+	g+	DLY
LYS	t	g+	g+	g+	0.22	0.29	0.16	0.17	0.13	t	g-	g-	g-	DLY
LYS	t	g+	g+	t	0.19	0.07	0.06	0.09	0.06	t	g-	g-	t	DLY
LYS	t	g+	g+	g-	< 0.01	0.00	0.00	0.00	0.00	t	g-	g-	g+	DLY
LYS	t	g+	t	g+	2.02	0.87	0.91	1.72	1.01	t	g-	t	g-	DLY
LYS	t	g+	t	t	0.29	0.13	0.08	0.11	0.13	t	g-	t	t	DLY
LYS	t	g+	t	g-	1.42	0.82	0.80	1.05	0.99	t	g-	t	g+	DLY
LYS	t	g+	g-	g+	< 0.01	0.00	0.00	0.00	0.00	t	g-	g+	g-	DLY
LYS	t	g+	g-	t	0.03	0.00	0.00	0.00	0.00	t	g-	g+	t	DLY
LYS	t	g+	g-	g-	0.04	0.00	0.00	0.01	0.01	t	g-	g+	g+	DLY
LYS	t	t	g+	g+	2.32	0.47	0.39	0.49	0.54	t	t	g-	g-	DLY
LYS	t	t	g+	t	0.87	0.18	0.15	0.29	0.34	t	t	g-	t	DLY
LYS	t	t	g+	g-	0.03	0.01	0.00	0.00	0.00	t	t	g-	g+	DLY
LYS	t	t	t	g+	3.52	2.82	3.33	4.02	3.30	t	t	t	g-	DLY
LYS	t	t	t	t	0.77	0.29	0.29	0.39	0.41	t	t	t	t	DLY
LYS	t	t	t	g-	4.81	1.88	1.98	2.15	2.04	t	t	t	g+	DLY
LYS	t	t	g-	g+	0.02	0.00	0.00	0.00	0.01	t	t	g+	g-	DLY
LYS	t	t	g-	t	0.36	0.17	0.14	0.17	0.18	t	t	g+	t	DLY
LYS	t	t	g-	g-	0.53	0.55	0.71	0.87	0.69	t	t	g+	g+	DLY
LYS	t	g-	g+	g+	0.01	0.00	0.00	0.00	0.00	t	g+	g-	g-	DLY
LYS	t	g-	g+	t	< 0.01	0.00	0.00	0.00	0.00	t	g+	g-	t	DLY
LYS	t	g-	g+	g-	< 0.01	0.00	0.00	0.00	0.00	t	g+	g-	g+	DLY
LYS	t	g-	t	g+	0.3	0.13	0.19	0.13	0.16	t	g+	t	g-	DLY
LYS	t	g-	t	t	0.07	0.01	0.02	0.02	0.01	t	g+	t	t	DLY
LYS	t	g-	t	g-	0.52	0.21	0.27	0.27	0.15	t	g+	t	g+	DLY
LYS	t	g-	g-	g+	< 0.01	0.00	0.00	0.00	0.00	t	g+	g+	g-	DLY

Residue	χ^1	χ^2	χ^3	χ^4	BBIND	AAXAA	aaXaa	AAxAA	aaxaa	χ^1	χ^2	χ^3	χ^4	Residue
LYS	t	g-	g-	t	0.06	0.00	0.00	0.02	0.00	t	g+	g+	t	DLY
LYS	t	g-	g-	g-	0.15	0.04	0.03	0.04	0.04	t	g+	g+	g+	DLY
LYS	g-	g+	g+	g+	0.11	0.29	0.11	0.12	0.14	g+	g-	g-	g-	DLY
LYS	g-	g+	g+	t	0.03	0.04	0.03	0.01	0.02	g+	g-	g-	t	DLY
LYS	g-	g+	g+	g-	0.01	0.00	0.00	0.00	0.00	g+	g-	g-	g+	DLY
LYS	g-	g+	t	g+	0.61	0.73	0.96	1.09	0.79	g+	g-	t	g-	DLY
LYS	g-	g+	t	t	0.06	0.03	0.07	0.05	0.06	g+	g-	t	t	DLY
LYS	g-	g+	t	g-	0.39	0.66	0.69	0.73	0.67	g+	g-	t	g+	DLY
LYS	g-	g+	g-	g+	< 0.01	0.00	0.00	0.00	0.00	g+	g-	g+	g-	DLY
LYS	g-	g+	g-	t	< 0.01	0.00	0.00	0.00	0.00	g+	g-	g+	t	DLY
LYS	g-	g+	g-	g-	0.01	0.00	0.00	0.00	0.00	g+	g-	g+	g+	DLY
LYS	g-	t	g+	g+	4.79	6.98	8.81	7.31	7.65	g+	t	g-	g-	DLY
LYS	g-	t	g+	t	2.01	1.54	1.71	1.25	1.63	g+	t	g-	t	DLY
LYS	g-	t	g+	g-	0.09	0.08	0.06	0.06	0.09	g+	t	g-	g+	DLY
LYS	g-	t	t	g+	14.99	19.17	13.20	13.06	17.95	g+	t	t	g-	DLY
LYS	g-	t	t	t	4.96	3.05	3.32	3.33	2.95	g+	t	t	t	DLY
LYS	g-	t	t	g-	22.34	29.95	37.45	36.37	30.52	g+	t	t	g+	DLY
LYS	g-	t	g-	g+	0.19	0.02	0.02	0.05	0.04	g+	t	g+	g-	DLY
LYS	g-	t	g-	t	3.67	2.34	2.24	2.18	2.34	g+	t	g+	t	DLY
LYS	g-	t	g-	g-	5.13	4.20	2.09	2.10	4.40	g+	t	g+	g+	DLY
LYS	g-	g-	g+	g+	0.26	0.08	0.10	0.08	0.06	g+	g+	g-	g-	DLY
LYS	g-	g-	g+	t	0.09	0.03	0.03	0.04	0.02	g+	g+	g-	t	DLY
LYS	g-	g-	g+	g-	< 0.01	0.00	0.00	0.00	0.00	g+	g+	g-	g+	DLY
LYS	g-	g-	t	g+	6.06	8.26	8.74	8.07	6.82	g+	g+	t	g-	DLY
LYS	g-	g-	t	t	1.04	1.27	1.19	1.23	1.12	g+	g+	t	t	DLY
LYS	g-	g-	t	g-	7.64	6.84	6.74	7.18	7.54	g+	g+	t	g+	DLY
LYS	g-	g-	g-	g+	0.02	0.01	0.01	0.02	0.01	g+	g+	g+	g-	DLY
LYS	g-	g-	g-	t	0.66	0.50	0.43	0.36	0.53	g+	g+	g+	t	DLY
LYS	g-	g-	g-	g-	1.19	1.78	1.32	1.83	2.10	g+	g+	g+	g+	DLY
MET	g+	g+	g+	-	0.07	0.06	0.04	0.01	0.05	g-	g-	g-	-	DMT

Residue	χ^1	χ^2	χ^3	χ^4	BBIND	AAXAA	aaXaa	AAxAA	aaxaa	χ^1	χ^2	χ^3	χ^4	Residue
MET	g+	g+	t	-	0.07	0.07	0.03	0.02	0.04	g-	g-	t	-	DMT
MET	g+	g+	g-	-	< 0.01	0.00	0.00	0.00	0.00	g-	g-	g+	-	DMT
MET	g+	t	g+	-	0.73	0.59	0.66	0.54	0.66	g-	t	g-	-	DMT
MET	g+	t	t	-	0.76	0.65	0.49	0.42	0.73	g-	t	t	-	DMT
MET	g+	t	g-	-	0.79	0.73	0.59	0.33	0.57	g-	t	g+	-	DMT
MET	g+	g-	g+	-	< 0.01	0.00	0.00	0.00	0.00	g-	g+	g-	-	DMT
MET	g+	g-	t	-	0.08	0.04	0.01	0.02	0.03	g-	g+	t	-	DMT
MET	g+	g-	g-	-	0.07	0.03	0.03	0.02	0.04	g-	g+	g+	-	DMT
MET	t	g+	g+	-	1.92	1.79	1.95	1.91	1.81	t	g-	g-	-	DMT
MET	t	g+	t	-	2.17	1.68	1.65	1.62	1.46	t	g-	t	-	DMT
MET	t	g+	g-	-	0.09	0.06	0.06	0.06	0.05	t	g-	g+	-	DMT
MET	t	t	g+	-	4.5	1.73	1.74	1.80	2.11	t	t	g-	-	DMT
MET	t	t	t	-	2.53	1.29	1.55	1.67	1.29	t	t	t	-	DMT
MET	t	t	g-	-	2.55	1.38	1.58	1.61	1.33	t	t	g+	-	DMT
MET	t	g-	g+	-	0.01	0.01	0.01	0.01	0.01	t	g+	g-	-	DMT
MET	t	g-	t	-	0.52	0.25	0.26	0.32	0.29	t	g+	t	-	DMT
MET	t	g-	g-	-	1.16	0.39	0.39	0.43	0.37	t	g+	g+	-	DMT
MET	g-	g+	g+	-	0.72	1.96	1.52	1.88	1.93	g+	g-	g-	-	DMT
MET	g-	g+	t	-	0.83	1.81	1.91	1.95	1.71	g+	g-	t	-	DMT
MET	g-	g+	g-	-	0.03	0.05	0.07	0.07	0.05	g+	g-	g+	-	DMT
MET	g-	t	g+	-	16.4	14.80	15.33	15.42	14.46	g+	t	g-	-	DMT
MET	g-	t	t	-	15.82	14.86	17.37	16.88	14.85	g+	t	t	-	DMT
MET	g-	t	g-	-	20.74	21.90	18.87	18.67	21.11	g+	t	g+	-	DMT
MET	g-	g-	g+	-	0.94	0.77	0.89	0.90	0.81	g+	g+	g-	-	DMT
MET	g-	g-	t	-	11.48	15.14	16.83	16.81	15.53	g+	g+	t	-	DMT
MET	g-	g-	g-	-	15.06	17.95	16.16	16.62	18.71	g+	g+	g+	-	DMT
PHE	g+	t	-	-	0.02	0.01	0.01	0.01	0.01	g-	t	-	-	DPN
PHE	g+	g	-	-	6.47	3.84	3.04	2.79	4.38	g-	g	-	-	DPN
PHE	t	t	-	-	2.79	1.65	2.02	2.03	1.52	t	t	-	-	DPN
PHE	t	g	-	-	39.54	36.28	39.62	39.27	32.76	t	g	-	-	DPN

Residue	χ^1	χ^2	χ^3	χ^4	BBIND	AAXAA	aaXaa	AAxAA	aaxaa	χ^1	χ^2	χ^3	χ^4	Residue
PHE	g-	t	-	-	5.19	3.03	4.36	4.38	3.22	g+	t	-	-	DPN
PHE	g-	g	-	-	45.99	55.20	50.95	51.52	58.11	g+	g	-	-	DPN
PRO	g+	-	-	-	62.93	64.31	67.00	67.69	64.29	g-	-	-	-	DPR
PRO	g-	-	-	-	37.07	35.69	33.00	32.31	35.71	g+	-	-	-	DPR
SER	g+	-	-	-	22.69	18.04	17.91	18.04	17.63	g-	-	-	-	DSN
SER	t	-	-	-	2.05	0.83	0.90	1.17	0.90	t	-	-	-	DSN
SER	g-	-	-	-	75.25	81.13	81.19	80.79	81.47	g+	-	-	-	DSN
THR	g+	-	-	-	45.14	49.21	46.43	25.25	16.18	g-	-	-	-	DTH
THR	t	-	-	-	1.71	0.05	0.09	0.07	12.10	t	-	-	-	DTH
THR	g-	-	-	-	53.15	50.73	53.49	74.68	71.72	g+	-	-	-	DTH
TRP	g+	t	-	-	0.51	0.25	0.19	0.19	0.33	g-	t	-	-	DTR
TRP	g+	g-	-	-	3.94	5.45	2.63	2.71	4.73	g-	g+	-	-	DTR
TRP	g+	g+	-	-	6.57	1.63	2.18	2.12	2.94	g-	g-	-	-	DTR
TRP	t	t	-	-	10.53	9.29	9.46	9.01	7.19	t	t	-	-	DTR
TRP	t	g-	-	-	13.28	13.21	14.07	12.96	10.57	t	g+	-	-	DTR
TRP	t	g+	-	-	16.59	13.75	15.27	16.80	13.59	t	g-	-	-	DTR
TRP	g-	t	-	-	14.98	14.97	20.34	19.79	17.10	g+	t	-	-	DTR
TRP	g-	g-	-	-	26.4	23.70	20.25	21.90	24.46	g+	g+	-	-	DTR
TRP	g-	g+	-	-	7.19	17.74	15.62	14.53	19.10	g+	g-	-	-	DTR
TYR	g+	t	-	-	0.02	0.01	0.00	0.00	0.01	g-	t	-	-	DTY
TYR	g+	g	-	-	8.27	3.58	2.04	1.88	4.76	g-	g	-	-	DTY
TYR	t	t	-	-	2.85	1.68	1.91	1.88	1.69	t	t	-	-	DTY
TYR	t	g	-	-	37.99	35.00	39.09	38.00	35.17	t	g	-	-	DTY
TYR	g-	t	-	-	5.29	3.19	4.24	4.07	3.28	g+	t	-	-	DTY
TYR	g-	g	-	-	45.57	56.54	52.72	54.16	55.10	g+	g	-	-	DTY

Residue	χ^1	χ^2	χ^3	χ^4	BBIND	AAXAA	aaXaa	AAxAA	aaxaa	χ^1	χ^2	χ^3	χ^4	Residue
VAL	g+	-	-	-	1.36	0.74	0.38	0.51	1.86	g-	-	-	-	DVA
VAL	t	-	-	-	55.53	36.04	45.66	48.67	10.26	t	-	-	-	DVA
VAL	g-	-	-	-	43.11	63.23	53.96	50.82	87.88	g+	-	-	-	DVA

Copyright Permissions

Chapter 2 reproduced in part with permission from:

> Childers, M. C. and Daggett, V. "Validating molecular dynamics simulations against experimental observables in light of underlying conformational ensembles." *J. Phys. Chem. B.*, **122** (26), 6673-6689 (2018). Copyright 2019 American Chemical Society.

Chapter 6 reproduced in part with permission from:

> Childers, M. C., Towse, C.L., and Daggett, V. "The effect of chirality and steric hindrance on intrinsic backbone conformational propensities: tools for protein design." *Protein Eng. Des. Sel.*, **29** (7), 271-280 (2016).

Chapter 7 reproduced in part with permission from:

> Childers, M. C. and Daggett, V. "Molecular dynamics-derived rotamer libraries for D-amino acids within homochiral and heterochiral polypeptides." *Protein Eng. Des. Sel.,* **31** (6), 191-204 (2018).