

Molecular Dynamics Simulations of Intrinsically Disordered Proteins and Biomolecular Assemblies

Charles Emile McAnany

Shawnee, Kansas

B.S. Chemistry, Rose-Hulman Institute of Technology, 2012
B.S. Chemical Engineering, Rose-Hulman Institute of Technology, 2012

A Dissertation presented to the Graduate Faculty
Of the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Chemistry

University of Virginia
December, 2017

Copyright:

Except for the chapters on Desmoplakin and LG-ELP, this work is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/4.0/>.

The chapter on Desmoplakin is adapted with permission from C. E. McAnany and C. Mura. Claws, disorder, and conformational dynamics of the C-terminal region of human desmoplakin. *J Phys Chem B*, 120(33):8654–8667, Aug 2016. Copyright 2016 American Chemical Society.

The chapter on LG-ELP is adapted with permission from J. D. Tang, C. E. McAnany, C. Mura, and K. J. Lampe. Toward a designable extracellular matrix: Molecular dynamics simulations of an engineered laminin-mimetic, elastin-like fusion protein. *Biomacromolecules*, 17(10):3222–3233, Oct 2016. Copyright 2016 American Chemical Society.

Abstract

This dissertation describes four projects using computer simulations to study the molecules of life. The first project aims to better understand how adhesive junctions between cells form, and the role of protein dynamics in this process. The second aims to design a new biomaterial that can be used in therapy after a traumatic brain injury. Here, again, protein dynamics is likely to play a pivotal role. The third aims to understand how the distinct members of an evolutionary family of proteins, all with the same basic shape, can assemble into very different complexes. The final project aims to understand how RNA interacts with the surface of an ancient protein. While these problems come from diverse areas of biology, the methodological approach used for all questions is the same: given an initial set of atomic coordinates, a computer program predicts how those atoms will move over time, thereby simulating the molecular dynamics of the system. This technique can give an atomically-detailed, femtosecond-by-femtosecond view of otherwise-murky biological processes.

Dedication

This dissertation is dedicated to everyone in the world, in the misguided hope that people will remark, “Wow, this work is dedicated to me! I should cite the author’s work!”

Contents

1	Introduction	1
1.1	Molecular Dynamics	1
1.2	The Timescales of Motion in Proteins	2
1.3	Molecular simulations do not occur at equilibrium.	3
2	Claws, Disorder, and Conformational Dynamics of the C-terminal Tail of Human Desmoplakin	5
2.1	Abstract	7
2.2	Introduction	7
2.2.1	Desmosomes mediate cellular adhesion	7
2.2.2	Post-translational modifications alter the behavior of desmoplakin	8
2.2.3	Arginine claws can structurally rigidify disordered regions	9
2.2.4	Simulations of disordered structural ensembles	9
2.2.5	Our MD simulations of DP	10
2.3	Methods and Procedure	11
2.3.1	Molecular dynamics simulations	11
2.3.2	Methylarginine parameterization	12
2.3.3	Analysis pipeline	12
2.4	Results	13
2.4.1	Arginine claws occur in the DP _{CTT}	13
2.4.2	RCs typically exclude solvent	17
2.4.3	Methylation and phosphorylation prime DP _{CTT} for GSK3 activity	17
2.4.4	The serine-rich region of DP _{CTT} is not entirely free in solution	18
2.5	Discussion	18
2.6	Conclusion	21
2.7	Acknowledgments	21
3	Toward a Designable Extracellular Matrix: Molecular Dynamics Simulations of an Engineered Laminin-mimetic, Elastin-like Fusion Protein	23
3.1	Abstract	25
3.2	Introduction	25
3.3	Methods of Procedure	27
3.3.1	LG-ELP Fusion Protein Design Methodology	27
3.3.2	MD Simulations of LG-ELP.	28
3.3.3	Analysis of Relative Solvent-Accessible Surface Area.	28
3.3.4	Hydrogen-Bonding Analysis.	28
3.3.5	Statistical Data Analysis.	28
3.4	Results and Discussion	29
3.4.1	Temperature-Dependent Structural Transitions of LG-ELP.	29
3.4.2	Secondary Structure Composition and Temperature Dependence.	30
3.4.3	Relative SASA and Association Interactions.	31
3.4.4	Role of Hydration in Compact Conformations.	33
3.5	Conclusions	35
3.6	Acknowledgements	35
4	The Oligomeric Plasticity of Cyclic Protein Assembly: A Simulation-based Analysis of Sm Rings	37
4.1	Abstract	39

4.2	Introduction	39
4.2.1	Many proteins assemble into oligomers.	39
4.2.2	The Sm family of ancient, structurally-conserved RNA-binding proteins.	40
4.3	Results And Discussion	41
4.3.1	A “Pizza Tensor” quantifies the structural relationship between Sm subunits.	41
4.3.2	The oligomeric state of the ring can be predicted based on the dimer.	42
4.3.3	Sm dimers are structurally stable on the 200-ns timescale.	42
4.3.4	The PT reveals remarkable flexibility in Sm dimers.	44
4.3.5	The POS shows that oligomeric plasticity is inherent to Sm proteins.	46
4.3.6	The flexibility of Sm proteins can be probed experimentally.	47
4.4	Conclusions	49
4.5	Materials and Methods	49
4.5.1	System preparation	49
4.5.2	Simulations	50
4.5.3	Analysis	50
4.6	Acknowledgements	50
5	A Simulation-based Approach to the Dynamical Basis of Hfq-RNA Interactions	51
5.1	Abstract	53
5.2	Introduction	53
5.2.1	Hfq’s Role in Annealing sRNAs and mRNAs	53
5.2.2	MD simulations of RNA	54
5.3	Methods	55
5.3.1	Constrained insertion to generate an Hfq-U ₆ odel	55
5.3.2	Unrestrained dynamics of the Hfq-RNA complex	55
5.4	Results	55
5.4.1	Constrained insertion suggests bound conformations	55
5.4.2	Two nucleotides bind firmly, four others explore the Hfq surface	56
5.4.3	A uracil-specific binding site on the rim	56
5.5	Future Directions	57
	Bibliography	59
	S2 Supplementary Information for Desmoplakin	S2.73
	S2.1 Overview	S2.73
	S3 Supplementary Information for LG-ELP	S3.93
	S4 Supplementary Information for Sm Oligomeric Plasticity	S4.105
	S4.1 PT, POS, and Validation statistics for all simulations	S4.105
	S4.2 POS dependence on atom selection	S4.129
	S4.3 Principal components analysis of PT values	S4.130

List of Figures

2.1	Architecture of the desmosome	8
2.2	A recognizable RC in the DP _{CTT}	10
2.3	Representative results from the analysis pipeline, showing a strong RC	14
2.4	Backbone conformations across all simulations	16
2.5	One-degree-of-freedom docking	19
2.6	S2849 in close proximity to the PRD	19
3.1	Proposed LG-ELP fusion protein	27
3.2	Temperature-dependent conformational states	29
3.3	Representative structures of the LG-ELP fusion protein at different temperatures	30
3.4	Contact maps of the dynamical interactions in our LG-ELP design	31
3.5	Secondary structural content of the ELP region as a function of temperature	32
3.6	Temperature-dependent changes in relative SASA of individual residues	33
3.7	Changes in hydration, hydrogen bonding, and overall structure	34
4.1	Structural similarity of Sm proteins	40
4.2	A pizza displays oligomeric plasticity	41
4.3	Representative PT alignment	42
4.4	Calculating the PT	43
4.5	The POS Algorithm	44
4.6	The POS of a dimer	45
4.7	Buried surface area in the Sm-Sm interface	47
4.8	The PT of 4PNO	47
4.9	Pizza tensor for all systems	48
4.10	POS distributions for all systems	49
5.1	Hfq's role in Class I sRNA-mRNA annealing	54
5.2	Constrained insertion step	55
5.3	Motion in the binding pocket over 300 ns	56
5.4	Unrestrained dynamics and RNA behavior on the protein surface	57
5.5	Average Hfq-RNA distance	58
5.6	Clusters in the U-specific pocket	58
S2.1	WT_PARM99SB	S2.74
S2.2	WT_CHARMM36	S2.75
S2.3	S2849S1P_PARM99SB	S2.76
S2.4	S2849S2P_PARM99SB	S2.77
S2.5	S2849S2P_PARM99SB_cycle2	S2.78
S2.6	S2849S2P_CHARMM36	S2.79
S2.7	S2849S2P_CHARMM36_cycle2	S2.80
S2.8	R2834H_PARM99SB	S2.81
S2.9	R2834H_CHARMM36	S2.82
S2.10	R2834H_S2849S2P_PARM99SB	S2.83
S2.11	R2834H_S2849S2P_PARM99SB_cycle2	S2.84
S2.12	R2834H_S2849S2P_CHARMM36	S2.85
S2.13	R2834H_S2849S2P_CHARMM36_cycle2	S2.86
S2.14	R2834MeMe_PARM99SB	S2.87
S2.15	R2834MeMe_CHARMM36	S2.88
S2.16	R2834MeMe_S2849S2P_PARM99SB	S2.89

S2.17	R2834MeMe_S2849S2P_PARM99SB_cycle2	S2.90
S2.18	R2834MeMe_S2849S2P_CHARMM36	S2.91
S2.19	R2834MeMe_S2849S2P_CHARMM36_cycle2	S2.92
S3.1	Initial starting structure of the LG-ELP protein	S3.94
S3.2	RDF of oxygen atoms around the ELP backbone	S3.95
S3.3	Secondary structural content across a range of temperatures as a function of time	S3.96
S3.4	Secondary structural content in extended simulations	S3.97
S3.5	Frequency of occurrence for secondary structural content	S3.98
S3.6	The time evolution of the secondary structure ensemble	S3.99
S3.7	Secondary structural content of the ELP region	S3.100
S3.8	Protein contact maps of dynamical interactions	S3.101
S3.9	Hydration of the ELP region	S3.102
S3.10	Time evolution of the radius of gyration	S3.103
S3.11	Time evolution of the radius of gyration for extended simulations	S3.103
S3.12	Correlation between the radius of gyration and end-to-end distance of the ELP region	S3.104
S4.1	1I4K	S4.106
S4.2	1I81	S4.107
S4.3	1I8F	S4.108
S4.4	1I8F.ring	S4.109
S4.5	1LJO	S4.110
S4.6	2QTX	S4.111
S4.7	2X4J Cter, Nter	S4.112
S4.8	2X4J Nter Cter	S4.113
S4.9	3BW1	S4.114
S4.10	3BY7	S4.115
S4.11	3HFO	S4.116
S4.12	4PNO	S4.117
S4.13	4PNO ring	S4.118
S4.14	4PNO tetramer	S4.119
S4.15	4WZJ bd1	S4.120
S4.16	4WZJ d1d2	S4.121
S4.17	4WZJ d2f	S4.122
S4.18	4WZJ d3b	S4.123
S4.19	4WZJ eg	S4.124
S4.20	4WZJ fe	S4.125
S4.21	4WZJ gd3	S4.126
S4.22	PAESM2	S4.127
S4.23	PAESM2 ring	S4.128
S4.24	POS calculated for random subsets of atoms	S4.129
S4.25	PCA of PT values	S4.130

List of Tables

2.1	Simulation systems and their Cy_*^R histograms	15
4.1	Simulated Sm protein systems	46
S3.1	Summary of MD simulation systems of the LG-ELP protein	S3.94

Chapter 1

Introduction

In this dissertation, I will present the results of three diverse projects that ultimately bear on skin diseases, biomaterials for neural tissue engineering, and the evolutionary origin of ribonucleoprotein assemblies. While the principal target of my research has been the RNA-associated Sm protein family, two of the chapters in this dissertation describe projects in collaboration with other labs focused on other biological systems. While the biological systems I have studied are only weakly related, my approach to studying them has been the same: molecular dynamics (MD) simulations. For the sake of clarity, I will focus in this introduction on MD and its applications. The biological background is provided in each of the individual chapters as appropriate.

1.1 Molecular Dynamics

MD allows us to probe classical biological systems with literally-unlimited precision[143]. At its core, MD applies classical Newtonian mechanics to a simple model of atomic interactions. Given the positions $\vec{x}(t)$, velocities $\vec{v}(t)$, and accelerations $\vec{a}(t)$ for each degree of freedom in the system (the x , y , or z coordinate for every atom at time t), one can write the equations for the *Verlet integrator*[166]:

$$\vec{v}(t + \frac{1}{2}\Delta t) = \vec{v}(t) + \frac{1}{2}\vec{a}(t)\Delta t \quad (1.1)$$

$$\vec{x}(t + \Delta t) = \vec{x}(t) + \vec{v}(t + \frac{1}{2}\Delta t) \quad (1.2)$$

$$\vec{a}(t + \Delta t) = \vec{f}(\vec{x}(t + \Delta t)) \quad (1.3)$$

$$\vec{v}(t + \Delta t) = \vec{v}(t + \frac{1}{2}\Delta t) + \frac{1}{2}\vec{a}(t + \Delta t)\Delta t \quad (1.4)$$

where $\vec{f}(\vec{x})$ is some function of the position of all the particles in the system. To solve for the unknown acceleration function, $\vec{f}(\vec{x})$, MD uses Newton's law,

$$\vec{F}(\vec{x}) = \vec{M} \odot \vec{f}(\vec{x}) \quad (1.5)$$

where \vec{M} is the vector of masses of each degree of freedom in the system ($[m_1, m_1, m_1, m_2, m_2, m_2, m_3, \dots]$, since the degrees of freedom are the x , y , and z coordinates for each atom) and \odot indicates element-wise multiplication of the two vectors. The force $\vec{F}(\vec{x})$ is in turn calculated from the system's potential energy $U(\vec{x})$:

$$\vec{F}(\vec{x}) = -\nabla U(\vec{x}) \quad (1.6)$$

This potential energy, $U(\vec{x})$, is estimated based on a simple model of interatomic interactions, and the accuracy of this model limits the ultimate accuracy of the simulation. Most MD integrators use the following approximation for $U(\vec{x})$ [143]:

$$U(\vec{x}) = U_{bond} + U_{angle} + U_{dihedral} + U_{vdW} + U_{Coulomb} \quad (1.7)$$

where

$$U_{bond} = \sum_{i=1}^{nbonds} k_{bond,i} (r_i - r_{0,i})^2 \quad (1.8)$$

$$U_{angle} = \sum_{i=1}^{nangles} k_{angle,i} (\theta_i - \theta_{0,i})^2 \quad (1.9)$$

$$U_{dihedral} = \sum_{i=1}^{ndihedrals} k_{dihedral,i} (1 + \cos(n_i \phi_i - \gamma_i)) \quad (1.10)$$

$$U_{vdW} = \sum_{i=1}^{natoms} \sum_{j=i+1}^{natoms} 4\epsilon_{ij} \left(\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right) \quad (1.11)$$

$$U_{Coulomb} = \sum_{i=1}^{natoms} \sum_{j=i+1}^{natoms} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \quad (1.12)$$

In (1.8), $nbonds$ is the total number of bonds in the system, $k_{bond,i}$ is a Hooke's law spring constant, r_i is the current length of that bond (the distance between the relevant atomic nuclei in \vec{x}), and $r_{0,i}$ is the equilibrium length for bond i . In (1.9), the terms are analogous to (1.8): $k_{angle,i}$ is the strength of the angle, modeled as a Hookean spring, and θ_i is the angle between the atoms in angle i . In other words, angular distortions are modelled with a harmonic potential. In (1.10), we model the energetics of rotation about single bonds with the cosine of the dihedral angle. In the dihedral term, n_i is the periodicity of the potential (n would be three for an ethane molecule rotating along its H-C-C-H dihedral, for example), ϕ_i is the current dihedral angle between the atoms contained in dihedral i , and γ_i is the offset for the dihedral potential. The van der Waals term, (1.11), iterates over all pairs of atoms (i, j) (except those atoms which are already linked by a bond, angle, or dihedral term) and estimates the contribution of London dispersion and steric clash forces. ϵ_{ij} is the strength of the interaction between atoms i and j . σ_{ij} is the distance at which dispersion and steric clash exactly cancel out, and r_{ij} is the current distance between the two atoms. (1.12) accounts for electrostatics using Coulomb's law[56]. q_i refers to the partial charge on atom i , ϵ_0 is the vacuum permittivity, and r_{ij} is, again, the distance between atoms i and j . The actual values of the fixed parameters are called the *force field*.

The above discussion is a simplification that glosses over many of the advanced features found in a modern MD engine, three of which are particularly relevant to the following chapters. First, my simulations (except as noted) are carried out in periodic boundary conditions so that the protein molecules can diffuse in an infinite periodic bath. This change significantly complicates the evaluation of electrostatic interactions[52]. Second, the integrator I have just described operates in the microcanonical (NVE) ensemble, while my simulations are actually performed in the isothermal-isobaric (NPT) ensemble. The MD program I use maintains pressure by growing and shrinking the periodic cell in response to pressure fluctuations, and it maintains temperature by perturbing atomic velocities to generate the correct (Boltzmann) distribution of velocities[8]. Third, some of my simulations use an implicit solvent model, which modifies $U_{Coulomb}$ to mimic bulk solvent[208, 64]. Simulations in implicit solvent are significantly faster to run, but are less accurate than simulations including explicit water molecules.

1.2 The Timescales of Motion in Proteins

The literally-unlimited spatial and temporal resolution of MD is also its Achilles' heel. The very best MD engines require time proportional to $N \log(N)$ to perform one step in the simulation on a system containing N particles[143, 166]. To capture motions at the atomic scale, the timestep using the integrator above cannot be much larger than 2 femtoseconds[143]. Thus, it takes 500,000 integration steps to simulate a system for a single nanosecond. The very longest MD simulations, using specialized hardware dedicated to molecular simulation, are on the order of milliseconds, and a typical simulation today is on the order of 500 nanoseconds[114, 185].

Many important biological processes occur on the microsecond timescale and beyond. Local motions in protein backbone and sidechain moieties occur on the picosecond to nanosecond timescale[143]. Correlated motions in the form of standing waves across β -sheets occur "occur on the slower timescale extending from tens of nanoseconds to a few milliseconds"[22]. Recent electron paramagnetic resonance (EPR) measurements have shown that intrinsically disordered proteins (IDPs) have correlation times that "range from 0.1 to 2 ns", indicating that structural rearrangements can occur easily in these systems[28]. The villin headpiece, a well-known fast-folding peptide, folds on the timescale of several microseconds[68]. Interestingly, some systems are notably less dynamic at certain timescales. For example, duplex DNA helices show a gap in dynamics on the microsecond to millisecond timescale[72], with gross structural changes occurring more slowly and fast motion captured at higher frequencies. A 1-millisecond

simulation of a protein showed that fast and slow motions are distinct[194]. Fast motions arose primarily from side chain movement, and occurred over timescales up to 10 ns. Slow motions arose from the backbone hopping between distinct conformational states with lifetimes on the order of 10 μ s.

In Chapters 2 and 3, I start with a peptide in a non-native state so that it can explore some of its conformational space without needing to escape from an initial energy well. While these simulations are far too short to thoroughly sample conformational space, I am able to use local properties (on the scale of a few residues) to account for biological phenomena. In Chapter 4, I observe a relaxation event that is likely to be essentially barrierless, namely the relaxation of a dimer when it is suddenly removed from a larger oligomer. This fast relaxation is sufficient to show the flexibility of Sm proteins, and thus answer the underlying biological question. Chapter 5, reports a new project that is using steered molecular dynamics (SMD) to force an Sm-RNA system out of its energy well in a matter of nanoseconds[160].

1.3 Molecular simulations do not occur at equilibrium.

A typical macroscopic definition of chemical equilibrium emphasizes that the bulk properties of a system at equilibrium do not change in time, and that there is no net flow of matter or energy. For example, Sandler’s classic thermodynamics textbook ties equilibrium to entropy: “ S = maximum at equilibrium in a closed system at constant U and V ”[183]. This definition is inapplicable on the microscale, where the properties of interest in a system are not bulk averages, but specific torsion angles, interatomic distances, and other microscopic quantities that rapidly change as the system explores its conformational space. Importantly, if the microstate of a system is completely specified (all atomic positions and momenta are known exactly), then the entropy is exactly zero, because $S = k \log(\Omega)$ and Ω , the number of microstates, is one[237]. A microscopic definition of equilibrium must therefore account for the time it takes for a system to sample all of the microstates in its equilibrium ensemble. Since no simulation of finite duration can sample the entire ensemble, I adopt the following working definition for equilibrium: A simulation approaches equilibrium for some parameter Θ if and only if the distribution of values that Θ adopts during the simulation approaches the distribution that Θ would have if the simulation were extended to an arbitrarily long time. With enough sampling, an MD trajectory can provide a good approximation of the equilibrium ensemble for a particular property[143, 194]. From this, one can extract the free energies of interesting transitions, or determine the pathway that a reaction of interest takes[166]. Without sufficient sampling, however, techniques like principal component analysis (PCA) can fail to capture the biologically important behavior of the underlying system[11].

There are some remarkable techniques to extract free energies from MD experiments without the need to sample extensively. Many of these techniques require a clearly-defined reaction coordinate, such as the position of a substrate in a transport protein[166]. Other enhanced-sampling techniques are well-suited to sampling systems with few particles, or systems where the states of interest have similar conformations[136]. Unfortunately, these techniques are not applicable to the systems I have studied. For simulations of large intrinsically disordered proteins (IDPs), the computational cost of extensive sampling makes it impossible to capture a complete equilibrium ensemble for the system. By their nature, IDPs have flat potential energy surfaces with small barriers between states, so myriad conformations, including gross structural rearrangements, contribute to the equilibrium ensemble[211].

Even without an equilibrium ensemble, many techniques have been developed to derive useful biological insight from MD trajectories. Perhaps the most common method of analysis in MD is root-mean-square deviation (RMSD)[118, 29, 68, 114, 194].

$$RMSD = \sqrt{\frac{\sum_{i=1}^n (x_{i,t} - x_{i,r})^2}{n}} \quad (1.13)$$

where $x_{i,t}$ is position of atom i at time t , and $x_{i,r}$ is the position of atom i in a reference structure, and n is the total number of atoms. RMSD is useful for determining if two structures are similar, but can introduce subtle errors when used to argue that structures are dissimilar. For example, RMSD would report a change when a phenyl ring is rotated 180°, even though the resulting structure would be identical to the reference. Another popular technique uses a set of geometric criteria to quantitatively classify secondary structure[190, 177]. This measure is particularly attractive for IDPs, because it is experimentally measurable (by circular dichroism spectroscopy) and secondary structure sampled more quickly than global structure[68]. A popular global measure is radius of gyration (R_g),

$$R_g^2 = \frac{\sum_{i=1}^n w_i (x_i - \bar{x})^2}{\sum_{i=1}^n w_i} \quad (1.14)$$

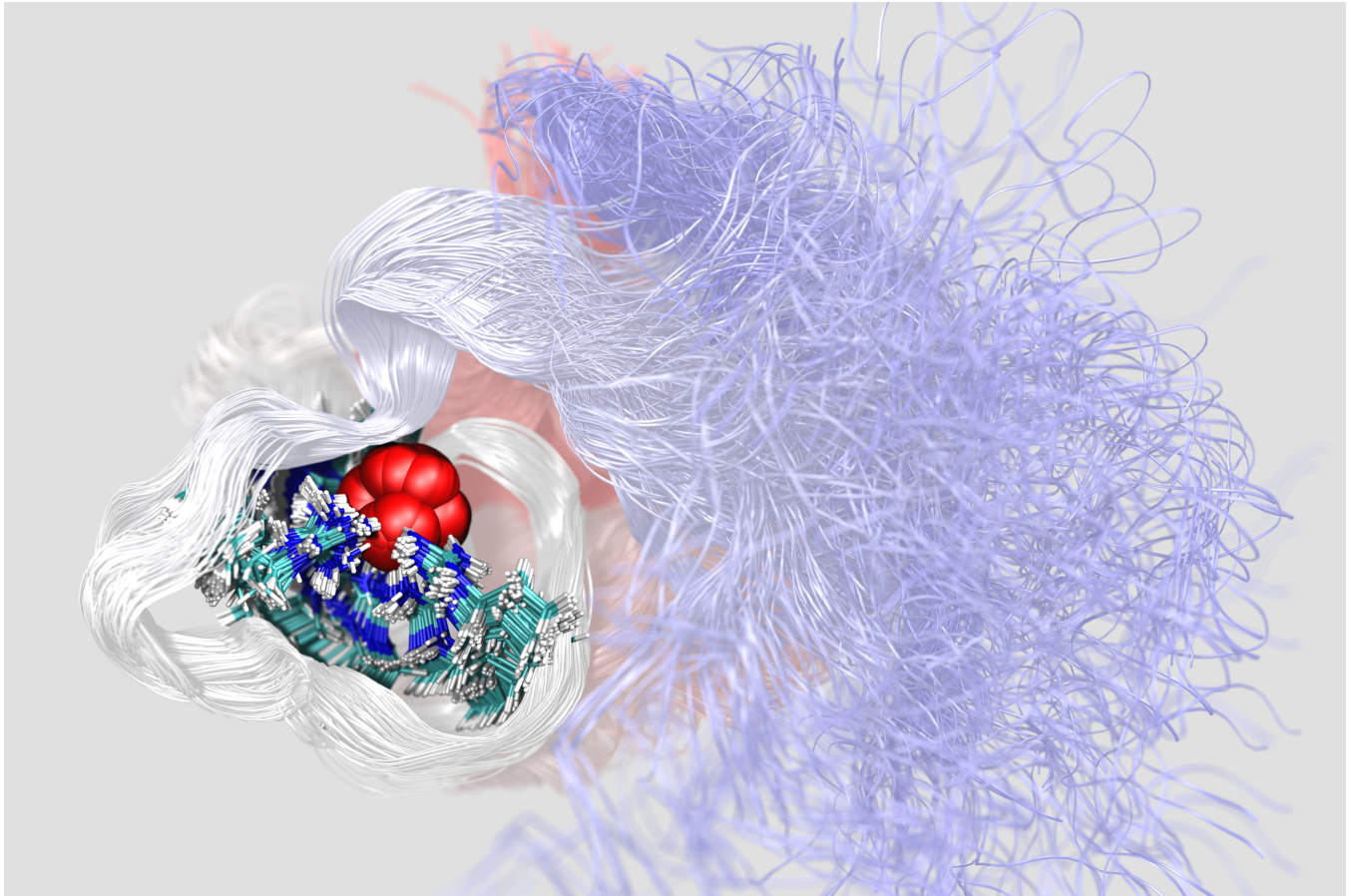
Where w_i is the weight assigned to each atom, x_i is the position of atom i , and \bar{x} is the center of mass of the system (with masses given by w). R_g gives a convenient estimate of the overall size of a system[148, 211], which can be linked to its transport behavior. Other techniques include solvent-accessible surface area (SASA) measurement[68, 29], contact maps[118, 114, 177, 193], clustering of multiple trajectories, analysis of correlated motion[194, 114], and prediction of nuclear magnetic resonance (NMR) observables[114, 117, 177, 185, 193]. For many systems, some

particular parameter of interest leads to the development of a new, system-specific analysis technique. For example, one might measure the unfolding pathway for each residue in a peptide[185], or the diameter of an ion channel as it switches oligomeric states[29].

For my work on IDPs, I have focused on the analysis of *geometrical* properties that lend themselves to experimental testing. In Chapter 2, for example, I show that a particular pair of modifications primes a protein for further action, and that other modifications do not lead to the same priming effect. I cannot use the priming effect as a reaction coordinate, since there are many different conformations that would look equally primed, and the transition between two primed states may go through a non-primed intermediate. In Chapter 3, I show the beginning of a tool to predict the phase behavior of biomaterials based on data from short MD simulations. Since an equilibrium ensemble is inaccessible, I use secondary structure as a reporter for the bulk behavior. Sampling secondary structure, of course, happens more quickly than sampling tertiary structure, so I am able to sample more of the relevant phase space during the short simulation. In Chapter 4, I allow a protein to relax from an initial, constrained state. In this case, an incomplete ensemble is still informative, as it shows structures that the protein is capable of adopting easily. By performing similar computational experiments on many different proteins, I find that these proteins are unexpectedly flexible. In Chapter 5, I describe an experiment using steered molecular dynamics to force an RNA molecule to dissociate from the a binding pocket in a protein. This approach allows me to sample relevant conformational states that would not be observed in an equilibrium simulation.

Chapter 2

Claws, Disorder, and Conformational Dynamics of the C-terminal Tail of Human Desmoplakin



This chapter is adapted with permission from:

C. E. McAnany and C. Mura. Claws, disorder, and conformational dynamics of the C-terminal region of human desmoplakin. *J Phys Chem B*, 120(33):8654–8667, Aug 2016

2.1 Abstract

Multicellular organisms consist of cells that interact via elaborate adhesion complexes. Desmosomes are membrane-associated adhesion complexes that mechanically tether the cytoskeletal intermediate filaments (IFs) between two adjacent cells, creating a network of tough connections in tissues such as skin and heart. Desmoplakin (DP) is the key desmosomal protein that binds IFs, and the DP-IF association poses a quandary: desmoplakin must stably and tightly bind IFs to maintain the structural integrity of the desmosome. Yet, newly synthesized DP must traffick along the cytoskeleton to the site of nascent desmosome assembly without ‘sticking’ to the IF network, implying weak or transient DP-IF contacts. Recent work reveals that these contacts are modulated by post-translational modifications (PTMs) in DP’s C-terminal tail. Using molecular dynamics simulations, we have elucidated the structural basis of these PTM-induced effects. Our simulations, nearing 2 μ s in aggregate, indicate that phosphorylation of S2849 induces an ‘arginine claw’ in desmoplakin’s C-terminal tail (DP_{CTT}). If a key arginine, R2834, is methylated, the DP_{CTT} preferentially samples conformations that are geometrically well-suited as substrates for processive phosphorylation by the cognate kinase GSK3. We suggest that DP_{CTT} is a molecular switch that modulates, via its conformational dynamics, DP’s efficacy as a substrate for GSK3. Finally, we show that the fluctuating DP_{CTT} can contact other parts of DP, suggesting a competitive binding mechanism for the modulation of DP-IF interactions.

2.2 Introduction

2.2.1 Desmosomes mediate cellular adhesion

Desmosomes are inter-cellular junctions found in epithelial and cardiac tissue[203, 89, 2, 73, 57, 59, 102]. By connecting the intermediate filaments (IFs) of neighboring cells, desmosomes create a network of adhesive structural interactions that impart tensile strength and durability to these tissues. The general architecture of the desmosome is shown in Figure 2.1. Desmosomes expose the extracellular regions of two transmembrane cadherins, desmocollin and desmoglein, on the cell surface; these proteins bind the cadherins of neighboring cells via Ca^{2+} -dependent homo- or heterophilic interactions. The desmosomal cadherins traverse the plasma membrane and bind two other key proteins, plakoglobin and plakophilin, which in turn bind a large, essential protein known as desmoplakin (DP; Figure 2.1). DP binds to the cytoskeletal IFs and, because the cytoskeleton spans the cytosol of one cell and binds to other desmosomes (which in turn bind to other, neighboring cells), this extended network of adhesive molecular contacts links together cells into tissues.

The IFs bind to DP’s three plakin repeat domains (PRDs), which correspond to residues 1960-2208 and are denoted PRD A, PRD B, and PRD C (Figure 2.1)[38, 89]. The C-terminal PRDs are connected to the plakin domain, in DP’s N-terminal region, via a fibrous rod (residues 1057-1945, central coiled-coil in Figure 2.1). The coiled-coil region is responsible for DP dimerization and, ultimately, links an electron-dense region known as the *outer dense plaque* (near the cell membrane) to the *inner dense plaque* (proximal to the IF network), across a span of ≈ 10 -20 nm (Figure 2.1)[80, 73]. The plakin domain of spectrin repeats (residues 178-883, leftmost structure in DP in Figure 2.1)[40] provides a relatively rigid N’-terminal connection that binds to the plakophilin (PKP) and plakoglobin (PG) proteins, thereby helping target DP to the desmosome[196]. Plakoglobin binds to the intracellular regions of desmocollin and desmoglein, denoted as the cadherin cytoplasmic regions (CCR) in Figure 2.1. Crystallographic structures of PRDs have revealed a basic groove that can sterically accommodate IFs, suggesting that as a potential mode of DP-IF interactions[38, 40, 96, 66].

Because desmosomes impart structural integrity and mechanical strength to cell-cell junctions, aberrant desmosome function underlies several diseases of the skin and heart[102]. For example, pemphigus is an autoimmune disease caused by antibodies to desmoglein[195], the DP mutation S2594P is linked to Carvajal syndrome[171], and several DP mutations are associated with the lethal heart disease arrhythmogenic right ventricular cardiomyopathy[3, 4, 171]. Down-regulation of DP has been linked to metastasis of tumor cells[157], and desmosome function in cancer remains an active area of research[89]. Several point mutations in the desmoplakin C-terminal tail (DP_{CTT}) have been examined previously[201, 246, 88], providing evidence that the DP_{CTT} region regulates DP-IF adhesion.

Cellular adhesion by desmosomes is regulated by two principal mechanisms: Ca^{2+} -dependent adhesion of extracellular cadherin domains[88] and phosphorylation-dependent adhesion of DP to IFs[102]. During desmosome formation, DP must be translocated to the desmosome along the cytoskeletal network, and therefore must bind only loosely to IFs. Once DP reaches the desmosome and is properly localized, it binds more tightly to the IFs in order to create stable and persistent intercellular connections in epithelial tissues (e.g., skin) and cardiac muscle. The IF-binding site of DP is required for normal desmosome assembly *in vivo*, suggesting that DP transport occurs along IFs[75]; however, the S2849G mutation, which is in the DP_{CTT}, causes DP to associate abnormally strongly with IFs, thereby retarding desmosome assembly[75].

arginine residues in the DP_{CTT} were dimethylated.) In cases where a single arginine is dimethylated, some evidence indicates that both methyls are on the same nitrogen, yielding an asymmetric dimethylarginine residue[14]. In DP_{CTT}, methylation appears to be necessary before GSK3 can initiate processive phosphorylation. Since their initial discovery in the 1960s[156], methylated protein residues often have been found to occur in serine-rich region (SRR) regions; indeed, such sequences serve as a common substrate for methyltransferases[32, 15, 44]. However, unlike phosphorylation, methylation is not known to be metabolically reversible, at least not outside the context of histones[30]. Apart from regulating the phosphorylation cascade of DP_{CTT}, any functional roles of these methylations remain unexplored.

2.2.3 Arginine claws can structurally rigidify disordered regions

The DP_{CTT} contains an SRR which is multiply phosphorylated[5], but any structural and dynamical effects of PTMs in the DP_{CTT} remain unknown. The structural dynamics of heavily-phosphorylated SRRs have been studied in other systems, and phosphorylation of SRRs is a common regulatory mechanism in eukarya[161]. A three-dimensional (3D) structure known as the *arginine claw* provides a rationale for some of these interactions and effects.

The arginine claw (RC), a relatively recently-identified structural element of SRRs, was first characterized in the C-terminal region of ASF/SF2[84], a protein involved in mRNA splicing, spliceosome assembly, and mRNA nuclear trafficking[249]. This protein is phosphorylated in an SRR, and this modification serves as a nuclear import signal. Fundamentally, the compaction of a peptide region into an RC sequesters charged side-chains away from the protein surface (Figure 2.2). Implicit-solvent molecular dynamics (MD) simulations of a fully-phosphorylated (RSPO₃)₈ peptide[84] initially revealed a compact structure, with one phosphate group coordinated by the guanidinium moieties of several arginine residues. An RC such as we find in the DP_{CTT} (see below) is shown in Figure 2.2c, alongside an illustration of the RC originally characterized by Hamelberg et al.[84] in Figure 2.2d. Such structures as shown in Figure 2.2d were found to stably persist over the 200-ns fully-atomistic explicit-solvent MD simulations of the (RS)₈ system[84]. In multiply-phosphorylated SRRs, those phosphate groups not involved in the RC are solvent-exposed, and this dynamically-varying surface exposure has been proposed as the recognition mechanism for nuclear import of a serine/arginine-rich ASF/SF2 (this particular ‘SR protein’ is also known as SRSF1)[84]. NMR studies of the ASF/SF2 system, as well as hPrp28 (another RNA-splicing-related system), have complemented the results of MD simulations, demonstrating that the phosphorylation of SRRs rigidifies the region[242]. Further simulation-based studies of RCs showed that claw formation allows the SRR of the lamin B receptor to bind to histones, despite the large positive charges of both interacting proteins[190]. Crystallographic studies of the RNA splicing factor SF1 have also revealed a partial RC[235]. As a final recent example, simulations have detected a claw-like structure in the long-time dynamics of a small, apoptosis-related intrinsically disordered protein (IDP) known as Noxa[97].

2.2.4 Simulations of disordered structural ensembles

MD simulations[128, 143] have been used to examine SRRs, IDPs, PTMs and, to a lesser extent, the interplay between these[118, 211, 225, 165, 229, 242, 200, 55, 245, 138]. The long timescales of conformational transitions and structure formation in SRRs has often prompted the use of relatively inexpensive implicit-solvent models. However, continuum solvent models likely overestimate the electrostatic effects of salt bridges in determining three-dimensional structure[250], and simulations studying RCs performed with implicit solvent models predict more compact structures than do analogous explicit solvent simulations[84]. Another important consideration is the force-field (FF) used to describe the potential energy landscape of a system. Modern FFs have been used to predict protein structures, albeit with limited success[67]; any FF shortcomings are exacerbated in simulations of IDPs due to the small energy differences between conformations[172]. Recent work has shown that CHARMM36 and ff03* predict substantially different secondary structures in glycosylated IDPs[245]. Simulations of highly-charged systems are also affected by the inadequate representation of electronic polarizability in current FFs. The classical Coulomb model of electrostatic interactions has been extended to include polarizability, though polarizable FF parameters are not yet available for PTMs such as in the systems studied here[117]. FFs are generally parameterized against the physicochemical properties of well-characterized model systems, for which experimental data or high-level quantum mechanical calculations are available[120, 228]. Disordered peptides are often underrepresented in these parameterization processes, as validating a structural ensemble generated by simulations of an IDP may be experimentally challenging (versus non-IDP systems)[138]. Not only are structural parameters difficult to determine experimentally[137], trajectory analysis is seldom straightforward and many complex techniques have been employed in analyzing IDP simulation results[192]. Finally, note that the RC is a somewhat unusual system insofar as it has a highly-charged core, while FFs are parameterized against the more common cases wherein charged residues are solvent-exposed. For these reasons, we note that simulations of systems of this type should be considered more suggestive and predictive rather than conclusive.

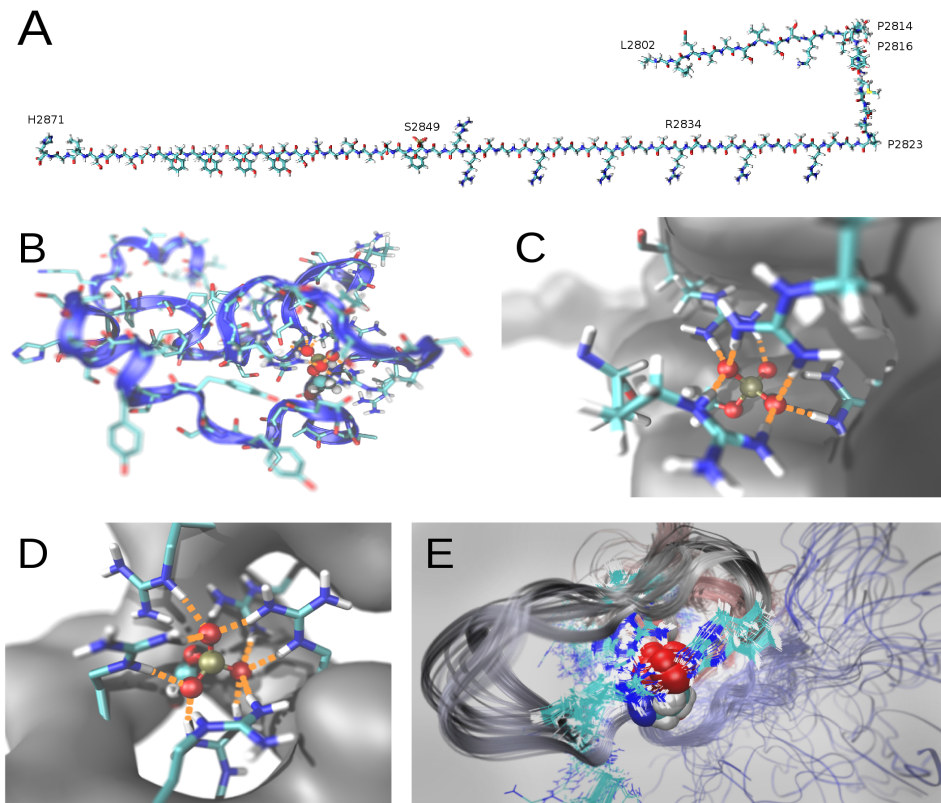


Figure 2.2: **A recognizable RC in the DP_{CTT}.** All simulation systems started in an extended backbone conformation (with bends at each proline residue), as exemplified in (a). After 10 ns of implicit-solvent simulation under CHARMM36, the R2834H S2849S_{PO₃} system can be seen to have collapsed and formed an RC (b). A sample RC, at 95 ns for the R2834H S2849S_{PO₃} system under CHARMM36, is shown in (c). The gray surface surrounds all residues other than arginines (shown as bonds), and the S_{PO₃} 2849 phosphosite is explicitly shown (ball-and-stick). Hydrogen bonds are represented as dashed orange lines. For reference, a previously-identified RC in an unrelated system[84] is shown in (d), rendered in a similar manner as (c); the coordination geometry of arginines and a phosphoserine is similar to that seen for the DP_{CTT} in (c). The structural stability of an RC is demonstrated in (e) by overlaying multiple frames from the simulation trajectory of (c). The regions near the RC remain stable for the duration of the 100-ns simulation, with the chelating arginines (shown as bonds) moving very little relative to S_{PO₃} 2849 (ball-and-stick, only oxygens can be seen). The rest of the DP_{CTT} backbone (thin ribbons) does not adopt a single, stable structure.

For slow processes and rare events, the computational cost of simulating a system such as an IDP for a sufficient length of time may be untenable. Several enhanced sampling methods have been developed[205, 232, 155]. However, the size of the DP_{CTT}, with its extended starting conformation (and requisite number of solvent molecules; Figure 2.2a), necessitates a large number of replicas (>100) for replica-exchange simulations, and correspondingly long trajectories (on the scale of 1 ms) are required for adequate mixing[1] of the replicas (McAnany & Mura, data not shown).

2.2.5 Our MD simulations of DP

We used classical, all-atom MD simulations to examine the structural effects of PTMs in the DP_{CTT}, with a specific aim of elucidating the conformational dynamics of this 70-residue region (Figure 1) and the riddle of strong/weak DP-IF interactions (might DP_{CTT} be a PTM-modulated molecular switch?). To mitigate the effects of FF inaccuracies and limited sampling, each system was simulated under two independent FFs (from the Amber and CHARMM families), and each production trajectory is at least 100-ns long. Simulations were extended to 200 ns for all phosphorylated systems; for consistency in scaling the figures, the 200-ns simulations were split into 100-ns chunks.

When we refer to a simulation without explicitly mentioning a time, we refer exclusively to the first 100 ns; when referring to the second 100 ns, we call this ‘cycle2’.

We begin by proposing a quantitative definition of an RC, and we show that simultaneous methylation and phosphorylation cause DP_{CTT} to assume conformations that are compatible with GSK3-binding. We propose that DP_{CTT}, in its various claw and non-claw states, competes with IF molecules for binding sites on the neighboring PRD elements (Figure 2.1), thus suggesting a straightforward dynamical mechanism for the regulation of DP-IF interactions.

2.3 Methods and Procedure

2.3.1 Molecular dynamics simulations

Classical, all-atom MD simulations were performed using NAMD 2.9[166], using either the PARM99SB[27] or the CHARMM36[17, 121] force-field. Parameters for modified residues, such as diprotic phosphoserine (S_{PO3}), were drawn from [90] and [158] as available (see Section 2.3.2). As no crystallographic or NMR structure of the DP_{CTT} is currently available, we constructed the peptide ²⁸⁰²LLEAASVSS KGLPSYNMS SAPGSRSGSR SGSRSGSRSG SRSGSRGSGF DATGNSSYSY SYFSSSSIG H²⁸⁷¹ using VMD’s `Molefacture` plugin in protein builder mode (VMD v1.9.1)[93]; note that the above sequence numbering matches human DP (UniProt ID P15924), and the simulated DP_{CTT} peptide ends at the very C’-terminus of DP. The peptide was constructed in an extended conformation ($\phi = 180^\circ, \psi = 180^\circ$), as shown in Figure 2.2a. PTMs were applied to specific residues (Figure 2.1) by using either leap (for PARM99SB, LEaP from AmberTools13[27]) or patches in VMD’s `psfgen` tool (for CHARMM36). Each initially-extended peptide system was subjected to a brief conformational relaxation simulation in implicit solvent. These relaxation simulations were performed with rigid hydrogen atoms, a nonbonded cutoff distance of at least 11.0 Å, and a Langevin thermostat set to human physiological temperature (310 K) with a damping constant of 1/ps. NAMD’s generalized Born implicit solvent model[208] was used with an ion concentration of 0.15 M. A 2-fs integration timestep was used in all simulations, as is common in MD simulations including rigid hydrogen[166]. The relaxation simulation consisted of 10,000 steps of conjugate gradient potential energy minimization, followed by 10 ns of unrestrained MD. A representative relaxed structure is shown in Figure 2.2b.

Periodic boundary conditions were set up by solvating the final structures from the relaxation simulations in a truncated octahedral cell of water molecules, of sufficient dimensions such that there would be at least 15 Å of water between the peptide and the envelope of the cell (this worst case scenario being reached if the peptide were to adopt the most extended state found in the last 5 ns of the relaxation simulation). This heuristic was adopted because of the periodic boundary conditions used in the explicit-solvent simulations: the peptide will be flexible during the production runs, and any prolonged violation of a 30 Å distance between periodic images of the DP_{CTT} solute could introduce artifacts. To mitigate computational costs, a “worst case” expanded size for the peptide was estimated based on the last half of the relaxation run; the first half of the relaxation run was not used in our geometry calculations, as the peptide is still collapsing during that time from its initial (extended) state. Even with the relaxation simulation, most of our simulated systems still contained over 200,000 particles (mostly H₂O). Waters were placed about the compactified peptide using the SOLVATE program[82], with custom modifications introduced in-house to enhance its performance. (These modifications do not affect the final positions of water molecules generated by SOLVATE.) Ions were placed by VMD’s `Autoionize` plugin (for CHARMM36) or LEaP (for PARM99SB) to reach 0.15 M NaCl. Because LEaP’s ion placement was observed to be non-random, a 10-ns water equilibration run was performed on those systems simulated using PARM99SB; this run comprised 100 steps of energy minimization, followed by 10 ns of dynamics with the protein atoms harmonically restrained by a force constant of 1 kcal/mol/Å². All other parameters were the same as in the equilibration runs.

For consistency, all PARM99SB and CHARMM36 systems were equilibrated in the same way, using the general approach of Mura & McCammon[144]. Again, a 2-fs timestep, with at least an 11.0 Å nonbonded cutoff and a 310 K Langevin thermostat, were used. Periodic boundary conditions were employed with PME electrostatics and a grid spacing of better than 1/Å per direction. NAMD’s `langevinPiston` feature was used to maintain pressure at 1 atm. Protein atoms were initially harmonically restrained to their initial positions by a 50 kcal/mol/Å² spring. The systems were minimized for 1000 steps, then gradually heated in 10 K increments, with 2 ps of dynamics at each new temperature. Once the system temperature reached 310 K, the restraints were weakened to 0.01 kcal/mol/Å² by repeatedly halving the restraint strength and simulating for 2 ps. Finally, the restraints were completely removed and the system was equilibrated for 10 ns in the NPT ensemble.

Production trajectories were computed using the same simulation parameters as the equilibration runs described above, and were extended to at least 100 ns each (Table 2.1). All production simulations were performed on the Rivanna supercomputer at the University of Virginia. In total, our simulations used nearly 2 million CPU-hours, taking approximately 1,000 CPU-hours per nanosecond with 1000 cores used for each simulation. Analyses were

performed using VMD and custom scripts written in the Python[226] and D[6] languages. All simulation and analysis scripts are available upon request, as are dehydrated trajectories.


2.3.2 Methyllarginine parameterization

Parameters for dimethylarginine, R_{Me_2} , in the CHARMM family of FFs were generously contributed by the Dejaegere laboratory [79]. These parameters lacked a term for the CK1–NH1–CK2 angle, subtended by the carbons of the added methyl groups and the nitrogen to which they are bonded; therefore, the value of this term was estimated using *ab initio* quantum mechanical calculations on a single R_{Me_2} residue. Specifically, the GAMESS[186, 77] program was used to perform geometry optimizations at the RHF/3-21G level in implicit water [214]. First, the optimal equilibrium geometry was determined, then the relevant bond angle was constrained 1° higher than the equilibrium angle and the equilibrium geometry re-calculated subject to this constraint. The derivative of energy with respect to angle provides the necessary value for this new FF parameter. The angle constraint was found to be 95.467 kcal/mol/rad², with an equilibrium angle value of 115.252°.

2.3.3 Analysis pipeline

For the sake of data-processing consistency, comparability, and automation, software tools were developed into a pipeline to analyze each simulation trajectory in a standardized manner. The detailed results of these analyses are shown in Figures S2.1 to S2.19. Figures were prepared using matplotlib and Python 3.3, with some analysis steps performed in VMD and the D programming language. Detailed descriptions of our analysis modules follow in the remaining subsections.

Arginine Clawicity, Cy_*^R (panel A)

Plots of the arginine clawicity (Cy_*^R), show, at each trajectory time-step, the Cy_*^R of the simulated system. For each residue in the sequence, the number of hydrogen bonds (donor-acceptor distance <3 Å, donor-hydrogen-acceptor angle $<20^\circ$) made to arginine are calculated, and the largest of these numbers (the Cy_*^R , by definition, where the ‘*’ wildcard denotes any residue) is plotted as a blue point. A green trace, representing a 1-ns running average, smoothens the noisy behavior of Cy_*^R . On the right of the panel, a vertically-oriented histogram shows the distribution of Cy_*^R values over the entire simulation; an example is given in Figure 2.3a. These histograms (e.g., ) are also used within the text to succinctly convey the Cy_*^R behavior of a given simulation.

Residue-specific arginine clawicity, Cy^R (panel b)

Plots of residue-specific arginine clawicity (Cy^R) show which residues are contained in an RC, as exemplified in Figure 2.3b. For each residue, at each time-step, the number of hydrogen bonds to arginine is calculated. These data are averaged with a 1-ns window before plotting, in order to avoid aliasing. White areas indicate that no hydrogen bonds were made to arginine by a particular residue at a particular time. For clarity, the DP_{CTT} sequence is staggered (up/down) along the horizontal axes of these plots: Residues on the top line align with inward-facing ticks and residues on the bottom line align with the extended outward-facing ticks. The key residues H2834 and S2849 are marked with asterisks.

Solvent accessible surface area, SASA, of residues 2849 and 2834 (panels c and d)

Solvent-accessible surface areas were calculated using VMD’s SASA tool, with a solvent probe radius of 1.4 Å. As with Cy_*^R , the SASA for each frame is shown as a blue point and a green trace shows a 1-ns running average. SASA values were calculated for the entirety of a residue, so comparison between systems with different residue modifications or mutations requires caution, as the residues are of different size. The histogram adjoined to the right axis (200 bins) shows the distribution of SASA over the entire simulation.

The S2849-S2845 distance (panel e)

For each frame in the simulation, the distance between the hydroxyl oxygens of S2849 and S2845 was calculated and plotted as a blue point. The green trace shows a 1-ns running average, and the histogram on the right (200 bins) shows the distribution of distances for the entire simulation.

GSK3 clash scores (panel f)

The GSK3 steric clash scores (as defined below) were evaluated via what effectively became a one-dimensional docking procedure (Figure 2.5). We began with the 3D structure of GSK3, taken as chain A from PDB entry 1I09[210]. The (side-chain) oxygen of residue 338 of chain B (the *recognition site* landmark), and the solvent-facing oxygen of the phosphate docked to chain A (the *active site* landmark), were used as reference points for alignment. These two reference points correspond to the recognition site and active site of GSK3. Note that only those chain A protein atoms built into the crystal structure were considered in the evaluation of clash scores. The corresponding pair of atoms from DP are the side-chain oxygens of S2849 (phosphorylated prior to GSK3 interaction) and S2845 (destined for phosphorylation by GSK3). In phosphorylated systems, the oxygen attached to the carbon was used. For each frame of each trajectory, DP and GSK3 were aligned based on the two pairs of atoms described above. GSK3 was then rotated, in 1° increments, about the axis defined from these four reference points. For each configuration, the number of clashes was taken as the number of contacts between atoms in GSK3 and atoms in DP (sans hydrogens for computational efficiency), with a 2 Å sweep radius. The minimum number of contacts, considered among all rotated positions for the trajectory frame in question, is defined as the *clash score* for that frame; it is this quantity which is plotted in the panels f.

Contact maps (panel g)

The contact map shows the pairwise contacts within a protein 3D structure, measured as a symmetric matrix of interatomic distances, $d_{i,j}$, for all pairs of residues i and j . The distance is defined so as to account for side-chain interactions: for a given residue pair, all pairs of atoms within each of the two residues (i_x, j_y) are considered, where atom x (i_x) is from residue i and atom y (j_y) is from residue j . The contact map distance for (i, j) is then taken as the distance between the closest pair of atoms for all of those pairs within the residue pair. In our illustrations, the lower-left triangle of the contact map shows the *average inter-residue distance* for the duration of the simulation, while the upper triangle gives the *minimum distance* considered over the entire trajectory. The horizontal axis is identical to that used for Cy^{R} , and the vertical axis is marked every ten residues and at the residues that were PTM sites in this study (asterisks).

Ramachandran plots (panel h)

Ramachandran plots show the distribution of peptide backbone torsion angles, (ϕ, ψ) , for each system, along the entire trajectory. Colors are graded by the logarithm of the probability density of a given (ϕ, ψ) configuration. Regions corresponding to canonical secondary structures are demarcated by guidelines, with the boundaries drawn from the MolProbity source code[33]. The percent of observations in each region is given at the top of the panel, and these regions roughly correspond to secondary structures: ‘La’ = left-handed α -helix; ‘La+’ = generously-allowed left-handed α -helix; ‘e’ = ϵ -turn regions, often found ahead of a helix or strand; ‘ α ’ = standard (right-handed) α -helix; ‘ β ’ = β -strand; ‘g+’ = generously-allowed helix or strand; ‘o’ = other structures.

2.4 Results

2.4.1 Arginine claws occur in the DP_{CTT}

A claw can be quantitatively defined, and occurs in the DP_{CTT}

Past efforts have qualitatively detected RCs based on visual analysis of trajectories, such as the one shown in Figure 2.2d[84, 190]. These past claws (i) were characterized as multiple arginines interacting with a phosphate, (ii) were found to be stable on the timescale of a 100-ns simulation, and (iii) had estimated free energies of formation of ≈ -5 kcal/mol[84]. While those attributes describe the behavior of a claw, they are not suitable metrics for determining the claw-forming propensity across a number of trajectories, which is a goal in our current study. First, the above set of descriptors does not, in and of itself, provide an algorithmic solution to the decision problem of whether a particular structure is or is not an RC. Second, the above description involves kinetic and thermodynamic information, both of which require more computationally expensive calculations than would a straightforward geometric definition of an RC. Finally, the above description of a claw does not work well for a trajectory that transiently adopts a claw or claw-like conformation. Therefore, we propose a definition of a claw that is akin to that of a protein secondary structural element.

Our definition is purely geometric, based only on a definition of the hydrogen bond[107, 9], and our parameter is easily evaluated for an arbitrary 3D structure. We define the *clawicity*, $\text{Cy}_{\text{B}}^{\text{A}}$, as the maximum number of hydrogen bonds made by any residue in B to all residues in A. For example, $\text{Cy}_{\text{S51}}^{\text{R}}$ refers to the number of hydrogen bonds made by S51 to all arginine (R) residues. $\text{Cy}_{\text{S}}^{\text{R}}$ refers to the number of hydrogen bonds made to an arginine by the

serine (*any* serine) with the greatest number of hydrogen bonds to arginine. We define the *arginine clawicity* (Cy_*^R) as the number of hydrogen bonds made to arginine residues by the residue with the most hydrogen bonds to arginine (here, the “*” wildcard means *any residue*). The *residue-specific arginine clawicity* (Cy_i^R) is defined as the number of hydrogen bonds made to arginine by each residue, i . Thus, for a peptide containing n residues, Cy^R would contain n values: $Cy_1^R, Cy_2^R, Cy_3^R, \dots, Cy_{n-1}^R, Cy_n^R$. In the current work, we consider a hydrogen bond to have a donor-acceptor distance below 3 Å and a donor-hydrogen-acceptor angle less than 20°[93]. This definition is trivially extended to other residues and may be made smoother by incorporating a definition of hydrogen bonds with non-integer order. For example, the order of a hydrogen bond might smoothly decrease from 1 to 0 as the donor-acceptor distance varies from 3 to 4 Å.

As an initial observation, note that representative plots of Cy_*^R (Figure 2.3a) and the site-specific Cy^R (Figure 2.3b) reveal a rather strong RC when the R2834H S_{PO₃}2849 system is simulated under the CHARMM36 force-field.

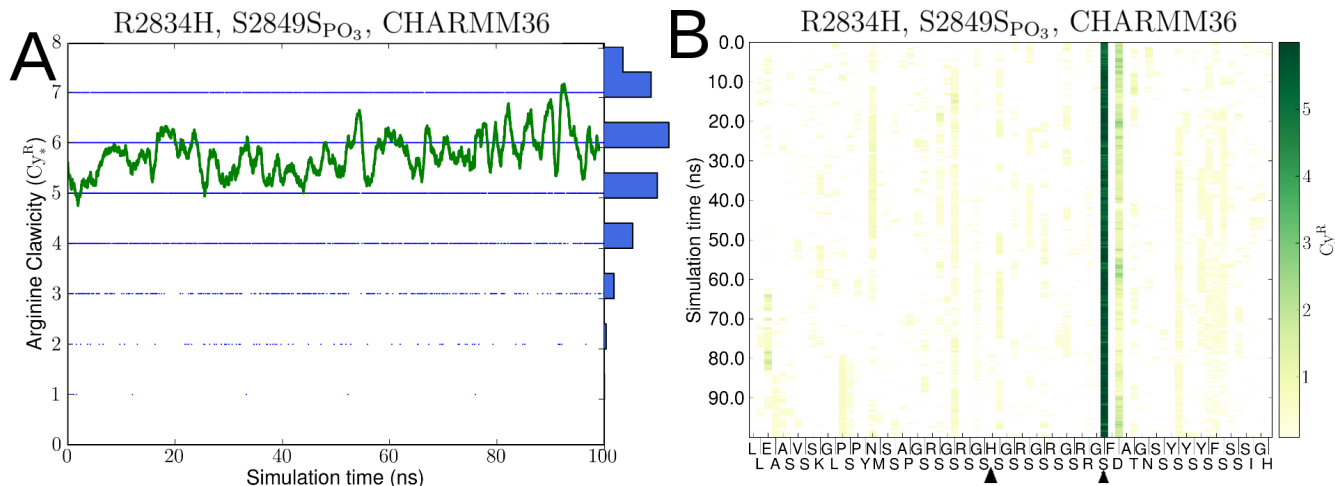
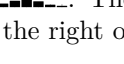


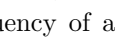
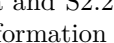



Figure 2.3: **Representative results from the analysis pipeline, showing a strong RC.** The Cy_*^R for the first 100 ns of the R2834H S_{2849SPO₃} simulation under CHARMM36 is shown in (a), demonstrating the appearance of a strong claw (large, persistent clawicity value). Blue points show the Cy_*^R , defined as the number of hydrogen bonds made to arginines by the residue with the most hydrogen bonds to arginine; the green line is a 1-ns running average. The marginal distribution on the right is a histogram of piled-up Cy_*^R values, ranging from 0 to 8: . The residue-specific Cy^R (b) shows that Cy_{S2849}^R frequently exceeds 6, and that only D2851 (immediately to the right of the dark strip) makes any other substantial contribution to this system’s arginine clawicity (Cy_*^R).

To facilitate communication in this text, we represent Cy_*^R values using histograms as in-line strip charts, e.g. . Each bar denotes the frequency of a particular Cy_*^R value across a trajectory, with the leftmost bar representing an Cy_*^R of zero. For example,  tends to adopt structures of Cy_*^R equal to 1, 2 or 3. Conversely,  shows a system with a particularly strong RC. Distributions of Cy_*^R values for the last 100 ns of each simulation system are shown in Table 2.1.

We discovered an RC in the conformational states sampled by the DP_{CTT}, as shown in Figure 2.2. Several arginine residues in DP_{CTT} surround S_{PO₃}2849 and form numerous hydrogen bonds and ion-pairs. Notably, some of the RCs found in the DP_{CTT} are long-lived structures, such as were those identified by Hamelberg et al.[84]. To our knowledge, DP_{CTT} is the largest unstructured peptide wherein an RC has been found.

Non-phosphorylated DP_{CTT} systems do not form strong claws

The unmodified (non-phosphorylated) wild-type DP_{CTT} peptide does not adopt a strong RC, as shown in Figures S2.1a and S2.2a for the Amber and CHARMM force-fields, respectively. Simulations under PARM99SB show little RC formation () and Figure S2.1b shows that no residue consistently hydrogen-bonds with any arginine with an Cy^R exceeding unity. The simulations using CHARMM36 predict an average Cy_*^R about 1 higher than do those using PARM99SB: . Amino acids D2851 and H2871 (the final C'-terminal residue) account for most of the Cy_*^R , as shown in Figure S2.2b.

As mentioned above, a newly-discovered PTM in the DP_{CTT} is asymmetric dimethylation of R2834, yielding R_{Me₂}2834[5]. For this modified peptide system, we find a slight increase in the average Cy_*^R when PARM99SB is

Table 2.1: **Simulation systems and their Cy_*^R histograms.** Cy_*^R values from the last 100 ns of each simulation are presented as histograms, where the intensity in a particular bin represents the frequency that the system had the corresponding Cy_*^R value. As an example, the bin numbers are explicitly shown in 012345678, which represents a simulation that frequently displayed Cy_*^R values of 2, 3, and 4 (highest peaks in the histogram). CHARMM36 was consistently found to predict higher Cy_*^R values than PARM99SB; in terms of clawicity, CHARMM36 also predicts a stronger response to phosphorylation.

Simulation system	Duration (per FF)	Force-field	
		PARM99SB	CHARMM36
Wild-type, unmodified	100 ns		
S2849S _{PO3}	200 ns		
R2834H	100 ns		
R2834H and S2849S _{PO3}	200 ns		
R2834R _{Me2}	100 ns		
R2834R _{Me2} and S2849S _{PO3}	200 ns		
S2849S _{HPO3} (PARM99SB only)	100 ns		—

used, . D2851 is the primary contributor to this weak RC (Figure S2.14b). CHARMM36 predicts a slightly higher Cy_*^R than that seen in the unmodified peptide: . Consistent with the PARM99SB simulation of this system, D2851 is the primary residue creating the RC in the CHARMM36 trajectories (Figure S2.15b). This particular RC structure does not appear to be dynamically stable: it briefly dissociates 40 ns into the trajectory, and then re-forms at ≈ 60 ns. This observation suggests that, although a claw can form in this system, the DP_{CTT} would be unlikely to adopt a collapsed RC conformation as a stable, long-lived structure.

The R2834H mutant exhibits low Cy_*^R values under PARM99SB (), with Figure S2.8b showing E2804 forming the center of a weak RC. Under CHARMM36, D2851 forms no RC and the overall Cy_*^R is low: . For R2834H simulations under both FFs, the clawicity behavior is similar to that in the unmodified system.

The behavior of DP_{CTT} is sensitive to force-field

The backbone dihedral angle distributions for PARM99SB and CHARMM36 are shown in Figure 2.4. A recent methodological study of an arginine/serine (RS)-rich peptide (unrelated to DP), using several FFs, found that CHARMM36 tends to favor the formation of left-handed helices[172]. We found that DP_{CTT}, which also contains an SRR, does not show this trend, at least not on the timescales of our present simulations. Instead, CHARMM36 frequently predicts more β -strand character (54.5%) than does PARM99SB (40.9%), as indicated in Figure 2.4. The total helical content (including left-handed helices) is somewhat higher under PARM99SB (23.0%) than it is under CHARMM36 (18.4%).

Site R2834H provides a striking example of the differential structural effects of various FFs. In the phosphorylated R2834H system, PARM99SB predicts that H2834 will be essentially entirely buried in the protein, or at least occluded from solvent (Figures S2.10d and S2.11d). In contrast, CHARMM36 predicts that this residue will be solvent-exposed, perhaps as a result of the constraints imposed by the strong RC that forms in this system (Figures S2.12d and S2.13d). Similarly, in the methylated system, PARM99SB predicts a more buried R_{Me2} (Figures S2.16d and S2.17d) than that predicted by CHARMM36 (Figures S2.18d and S2.19d).

Phosphorylation of S2849 leads to claw formation in the wild-type system

Trajectories computed under both CHARMM36 and PARM99SB are consistent, inasmuch as the S_{PO3}2849 system (with no other modifications) frequently forms an RC. PARM99SB predicts a substantial shift from the unmodified Cy_*^R profile, . The S_{PO3}2849 system adopts a high Cy_*^R , , in the first 100 ns of the production run, followed by in the next 100 ns (cycle2). Figures S2.4a and S2.5a indicate that this system's DP_{CTT}'s claw is less stable than that reported for the (RS)₈ peptide[84], and Figure S2.4a also shows a dramatic re-structuring at ≈ 70 ns in the production run. CHARMM36 shows a similar trend, moving from the unmodified Cy_*^R (to

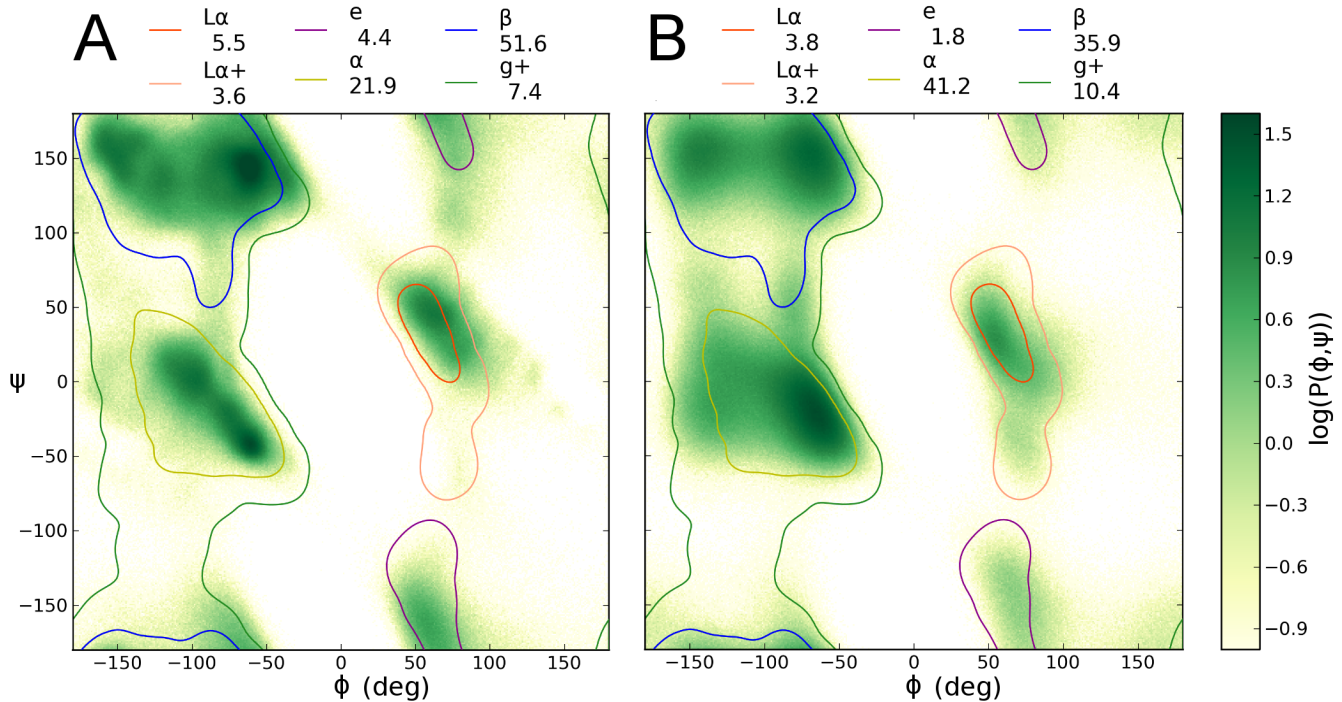


Figure 2.4: **Backbone conformations across all simulations.** To gauge the frequency of any unusual (non-canonical) secondary structures, Ramachandran plots are shown for all of our (a) CHARMM36 and (b) PARM99SB data, compiled across all trajectories for all simulation systems. Regions corresponding to canonical structural elements are indicated as contour lines (see also the Methods section), and the color-scale is graded by the \log_{10} -likelihood of a particular (ϕ, ψ) conformation. In contrast to a recent report[172], we find that DP_{CTT} does not show a preference for left-handed α -helices under CHARMM36; those recent simulations of a non-phosphorylated SRR, unrelated to our DP systems, found that CHARMM36 predicts that over 40% of the residues in the simulated peptide are in left-handed helices, even though only $\approx 6\%$ of the residues in a reference set of known protein structures exhibit such a structure. Our CHARMM36 trajectory data do not indicate that left-handed helices are a problem, at least in our simulation systems. Combining all frames of every simulation, CHARMM36 predicts 9% left-handed helix, while PARM99SB predicts 7%.

..... in the first 100 ns, and then in the second 100 ns (cycle2). Figures S2.6a and S2.7a show a more stable RC, akin to that seen previously[84]. For both FFs, the RC that forms is centered around position S_{PO₃}2849 (panels (b) in Figures S2.4 to S2.7).

For comparative purposes, an additional simulation was performed with the monoprotic phosphate modification, S_{HPO₃}2849 (versus the diprotic S_{PO₃}), in the Amber FF. This system, under PARM99SB, exhibited essentially no Cy_{*}^R (.....). Figure S2.3b shows that the RC does not form around S_{HPO₃}2849 to any appreciable extent; instead, an aspartate residue (D2851) makes occasional structures with Cy_{D2851}^R values of 2, and this tendency diminishes after ≈ 40 ns.

Methylation of R2834 weakens the RC

Simulations of the phosphorylated peptide with PARM99SB show that methylation of R2834, in conjunction with the phosphorylation at S2849, significantly weakens the RC, with in the first 100 ns followed by in the second 100 ns. The second cycle even has slightly lower Cy_{*}^R than the non-phosphorylated R2834H system. Figures S2.16b and S2.17b show that the principal residue involved in the RC is still S_{PO₃}2849. The effect predicted by CHARMM36 is more subtle: the Cy_{*}^R values remain similar to the non-methylated system in terms of their distribution, but Figures S2.18a and S2.19a show that the RC is more labile in this system. The sliding-window average (green trace) shows an increased variability compared to the nearly-constant behavior seen in Figures S2.6a and S2.7a (compare also the panels (c) [SASA of S2849] in Figures S2.6, S2.7, S2.18 and S2.19). Again, the CHARMM36 RC is centered on S_{PO₃}2849 (Figures S2.18b and S2.19b).

Mutation R2834H may disrupt the RC structure

For simulation systems containing the R2834H point-mutant as well as phosphorylation at S2849 (i.e., $\text{S}_{\text{PO}_3}\text{2849}$), the two FFs give differing results. Specifically, PARM99SB predicts essentially no change from the non-phosphorylated system in terms of Cy_*^{R} , with $\text{---}\blacksquare\text{---}\text{---}$ for the first 100 ns and $\text{---}\blacksquare\text{---}\text{---}$ for the next 100 ns; Figures S2.10b and S2.11b show that the RC stably settles at $\text{S}_{\text{PO}_3}\text{2849}$ by the second half of the 200-ns trajectory. In contrast, the CHARMM36 simulation of this system gives the strongest RCs observed in any of our trajectories, with $\text{---}\blacksquare\text{---}\text{---}$ for the first 100 ns, followed by $\text{---}\blacksquare\text{---}\text{---}$ for the remaining 100 ns, with the RC forming essentially near the start of the trajectory. The running averages for CHARMM36 (Figures S2.12a and S2.13a) reach clawicity values of 6, while no other simulation system ever reaches 5.

2.4.2 RCs typically exclude solvent

Analysis of the SASA of $\text{S}_{\text{PO}_3}\text{2849}$ can be used to reveal the general solvation features (buried, partially exposed, or fully exposed) of the RC in a 3D structure. The negatively-charged S_{PO_3} residue will be electrostatically attracted to arginine side-chains and, as expected, this is borne out in our observations of Cy_*^{R} values. In those simulation systems containing $\text{S}_{\text{PO}_3}\text{2849}$ but not exhibiting an RC, one might expect that the phosphate would be solvent-exposed and engaged in hydrogen bonds with water. We find that, for the non-methylated systems, this model works well. In systems containing a strong RC, a plot of the SASA of $\text{S}_{\text{PO}_3}\text{2849}$ shows that the phosphoserine is buried within the protein, as seen by comparing the (a) (RCy) and (c) (SASA) pairs of panels in Figures S2.4 to S2.7, S2.12 and S2.13, and note the anticorrelation between RCy values and the SASA. In the PARM99SB simulation of the S2849-phosphorylated R2834H mutant, the Cy_*^{R} was relatively low in the first (Figure S2.10a) and second (Figure S2.11a) 100-ns bins, and this agrees with the higher SASA observed for $\text{S}_{\text{PO}_3}\text{2849}$ in Figures S2.10c and S2.11c; the negative correlation can be seen here, again, most clearly by comparing the trend in Figure S2.10a (increasing values) and Figure S2.10c (decreasing values). The monoprotic system, $\text{S}_{\text{HPO}_3}\text{2849}$ under the Amber FF, similarly shows low Cy_*^{R} and high SASA values for $\text{S}_{\text{HPO}_3}\text{2849}$ (Figure S2.3c); the SASA values at this site are quite broadly distributed (Figure S2.3c, marginal histogram), implying a structurally heterogeneous ensemble of conformational states.

The methylated systems present two deviations from this inverse trend between Cy_*^{R} and SASA values. Under PARM99SB, $\text{S}_{\text{PO}_3}\text{2849}$ interacts with non-arginine residues on the protein surface, in the methylated system. In Figure S2.17c, the SASA of $\text{S}_{\text{PO}_3}\text{2849}$ can be seen to jump from a buried state to an exposed state after 40 ns, with no concomitant change in the Cy_*^{R} (Figure S2.17b). $\text{S}_{\text{PO}_3}\text{2849}$ interacts with S2861 and S2835 until 140 ns in the production trajectory, at which point it disengages from these residues while remaining attached to R2838 (until 197 ns). When the methylated system is simulated using CHARMM36, a solvent-exposed RC forms. The phosphate is clearly solvent-exposed, as shown in Figures S2.18c and S2.19c, but this system still forms an RC ($\text{---}\blacksquare\text{---}\text{---}$).

2.4.3 Methylation and phosphorylation prime DP_{CTT} for GSK3 activity

The DP_{CTT} sequence (Figure 2.1) contains several potential phosphorylation sites, including consensus sites for the GSK3 kinase. Recent experiments have revealed that DP is phosphorylated in its CTR by GSK3[5]. Thus, we used two simple metrics to assess the ability (not necessarily the propensity) of the DP_{CTT} to interact with GSK3 throughout the entire MD trajectory: (i) the $\text{S}_{\text{PO}_3}\text{2849}$ -S2845 distance, and (ii) the extent of steric clash between the DP_{CTT} and GSK3 molecules. First, the simple geometric distance between $\text{S}_{\text{PO}_3}\text{2849}$ and S2845 was measured and compared to the distance between the recognition site and active site in GSK3. This distance was used because $\text{S}_{\text{PO}_3}\text{2849}$ maps to GSK3's recognition site and S2845 corresponds to the kinase's active site. In the GSK3 crystal structure[210], this distance is ≈ 12 Å (some variability in this value is expected, as the active site was not occupied by a substrate in this GSK3 crystal structure). As a rudimentary gauge of DP_{CTT} 's ability to bind to GSK3, we suggest that DP_{CTT} conformations wherein the $\text{S}_{\text{PO}_3}\text{2849}$ -S2845 distance is ≈ 12 Å will be more favored to bind to GSK3 as a result of simple geometric matching, without requiring substantial structural rearrangement of the DP_{CTT} .

While non-phosphorylated DP_{CTT} systems would not be expected (biologically) to interact with GSK3, it is nevertheless informative to consider, as a background distribution, how these distances compare for the non-phosphorylated and phosphorylated systems. We find that the distances in the non-phosphorylated systems show a strong dependence on FF. PARM99SB yields distances that are substantially less than 12 Å for the completely unmodified wild-type system (Figure S2.1e) and the methylated, non-phosphorylated wild-type system (Figure S2.14e). The non-phosphorylated R2834H mutant system starts with GSK3-compatible distances, but collapses at ≈ 70 ns to incompatible distances (Figure S2.8e). In general, the CHARMM36 simulations predict longer distances than PARM99SB, and tend to predict distances that are more compatible with GSK3 binding (see Figures S2.2e, S2.9e and S2.15e).

Simulations of the phosphorylated DP_{CTT} systems exhibit good agreement between the distance distributions for PARM99SB and CHARMM36. Our distance parameter consistently lies between $\approx 12\text{--}15$ Å, which is compatible with GSK3 binding. The R2834H mutation in the phosphorylated system leads to a slight decrease in the distance under both PARM99SB and CHARMM36 (panels (e) in Figures S2.10 to S2.13), compared to that seen in the other two phosphorylated wild-type systems—namely, (i) the phosphorylated (S_{PO₃}2849) system in panels (e) of Figures S2.4 to S2.7, and (ii) the phosphorylated & dimethylated systems (S_{PO₃}2849 & R_{Me₂}2834) in panels (e) of Figures S2.16 to S2.19.

Our second GSK3-compatibility criterion, described in Figure 2.5 and the Methods section, assesses the ability of GSK3 to sterically accommodate various structural states of DP_{CTT}. Specifically, we align (i) the active site of GSK3 with S2845 of DP_{CTT} (as this is where the next phosphorylation event will occur), and (ii) the recognition site of GSK3 to S_{PO₃}2849 (as this is the landmark in DP_{CTT} that is recognized). These spatial transformations and geometric constraints effectively reduce the problem to a one-dimensional protein-protein docking exercise, the one degree-of-freedom being rotation about the line defined by constraints (i) and (ii); this construction is schematized in Figure 2.5. If there exists a rotation wherein GSK3 and DP_{CTT} can be brought together without substantial steric clash (literally, overlap of atomic van der Waals envelopes), then this suggests that GSK3 can readily bind to that conformation of DP_{CTT} (or at least that there is no enthalpic barrier to doing so). By this measure, we find that the only phosphorylated DP_{CTT} systems which exhibit steric compatibility along the trajectory frames are the methylated systems (Figures S2.17f and S2.19f). This accommodation is seen with both the PARM99SB and CHARMM36 FFs in the last 50 ns of the production run. Therefore, based on these data we suggest that methylation at R2834, yielding R_{Me₂}2834, ‘primes’ DP_{CTT} for processive phosphorylation by biasing its structural ensemble towards conformations that are amenable to GSK3 phosphorylation.

2.4.4 The serine-rich region of DP_{CTT} is not entirely free in solution

Potential interactions between the SRR of the DP_{CTT} and the rest of the large DP protein (Figure 2.1) were explored by analyzing pairwise inter-residue distances. As detailed in the Methods section, the full suite of contact maps, shown in panels (g) of Figures S2.1 to S2.19, show the mean inter-residue distances (lower triangle), averaged over entire trajectories, while the upper-right triangle gives the minimum inter-residue distance across an entire trajectory.

One may be tempted to view the DP_{CTT} as a disordered string that thermally fluctuates in solution, but this is not entirely accurate: the dynamical DP_{CTT} may in fact double back on the plakin repeat domains (as a reminder, see the PRDs in Figure 2.1). While simulations of larger DP systems, including an entire PRD in addition to the DP_{CTT}, are beyond the scope of this work, the first few residues of our DP_{CTT} simulation system are from a PRD (the third PRD in Figure 2.1). Therefore, if the phosphorylated S_{PO₃}2849 samples conformations that bring it near the first few residues of DP_{CTT}, then that suggests that regions within the DP_{CTT} may interact directly with the PRDs to regulate IF binding (and also that simulations limited to only the SRR might not account for all the factors that govern the structure and dynamics of this region). In all of our simulations, the SRR comes into close spatial proximity to other regions of DP_{CTT}, including the more N'-terminal residues that are part of PRD-C. A possible mechanism by which the DP_{CTT} can attenuate the overall strength of DP-IF binding may involve a simple binding competition between IFs and DP_{CTT} for the IF-binding site of the plakin repeat domain; in this model, the precise pattern of PTMs, and therefore the clasticity and dynamics of the DP_{CTT}, would modulate the competitive binding events. When DP_{CTT} is fully phosphorylated, its strongly negative charge could compete with the (negatively-charged) IFs for the binding groove on the PRD, as suggested by crystallographic studies[38, 40, 96].

2.5 Discussion

Arginine claws can form in partially phosphorylated systems—Past work on the arginine claw considered only fully-phosphorylated (RS)_n repeats[84, 190]. To our knowledge, our present study provides the first evidence that RCs can form in other systems too. Experimental and computational studies of a 36-residue peptide from myelin basic protein suggested that phosphorylated threonines can confer structure to disordered regions, via electrostatic interactions with basic residues. However, unlike an RC, the interactions in that system did not result in burial of the phosphate group within the protein[229]. Our present simulations, focused on the DP_{CTT}, predict that a strong RC can form in protein segments with only half the arginine density of RS-repeat peptides, and even when only a single serine is phosphorylated. Therefore, the RC may be a common, or at least underappreciated, structural element in phosphorylation-based regulation of protein function via molecular switches, even for protein sequences that lack canonical (RS)_n repeat regions.

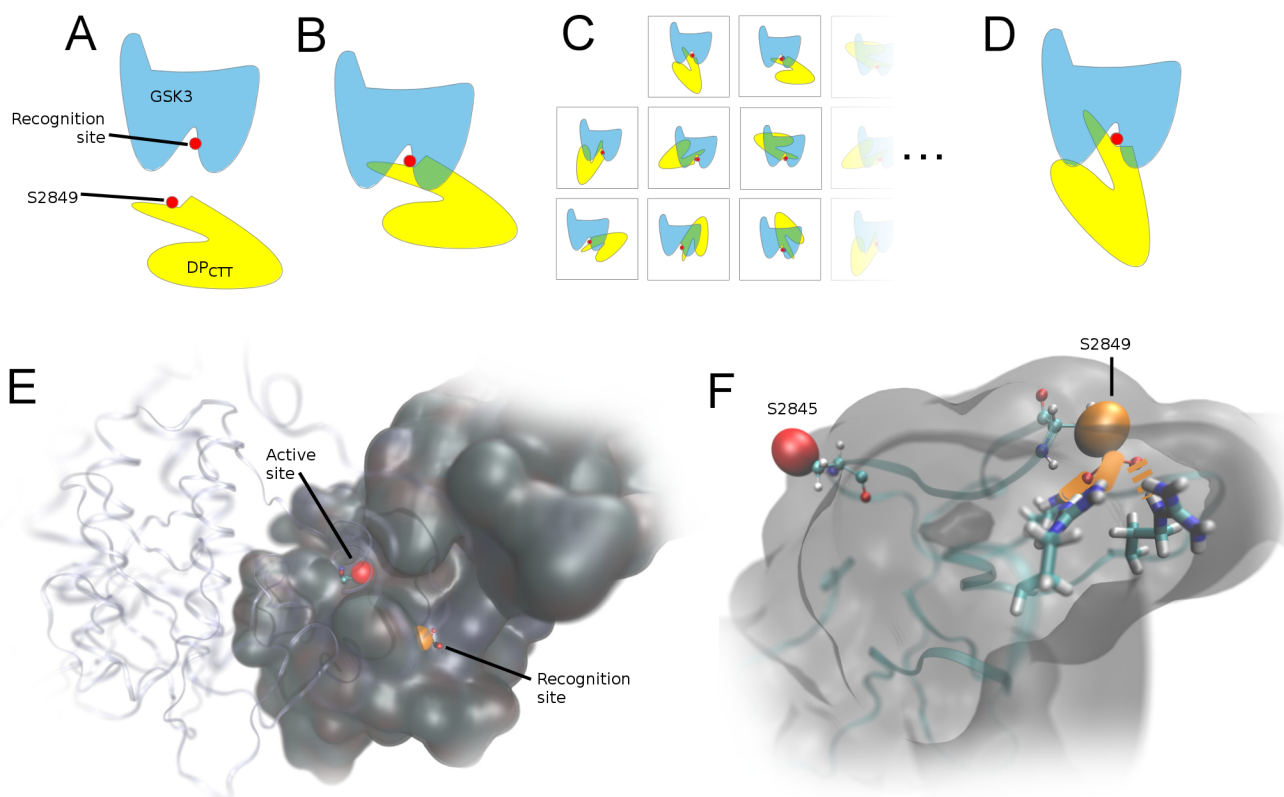


Figure 2.5: A method to evaluate potential DP-GSK3 geometric complementarity via one-degree-of-freedom docking. We begin with two molecules to be docked, as schematized in (a), and we know that the red atom on the yellow molecule aligns with the red atom on the blue molecule when the molecules interact. (In this two-dimensional case, only one atom is needed per molecule. For the three dimensional case, two atoms per molecule are needed to define an axis of rotation.) In step (b), the molecules have been aligned, in arbitrary angular orientation, based only on the positions of their red atoms. Next, the yellow molecule is rotated about the axis defined by the red atoms (c). At each step in the full rotational sweep (1° increments in our implementation), the *clash score* is computed as the number of pairs of atoms (one from DP and one from GSK3) with an interatomic distance less than 2 Å. The best pose (d) is taken as the one with the minimal clash score. This procedure is repeated for each frame along the trajectory. In a realistic example (e), two atoms (red, orange) from GSK3 are used to perform the alignment. The outward-facing oxygen of the phosphate (orange sphere) defines the recognition site, while the active site is defined by the side-chain oxygen of S261 (red sphere) of chain B (gray ribbon). The protein atoms from chain A (dark surface) were used to calculate the clash. The atoms in DP that were used to perform the alignment are highlighted in (f). This frame, from 173.48-ns in the production trajectory of $R_{Me_2}2834\ S_{PO_3}2849$ (CHARMM36), has a very low clash score of 18; all non-hydrogen atoms from DP (dark surface) were used to calculate the clash. The side-chain oxygen of S2845 (red) will be phosphorylated by GSK3 only if S2849 (orange) is phosphorylated. Note that the three terminal oxygens of the phosphate are still engaged in an RC (hydrogen bonds in orange).

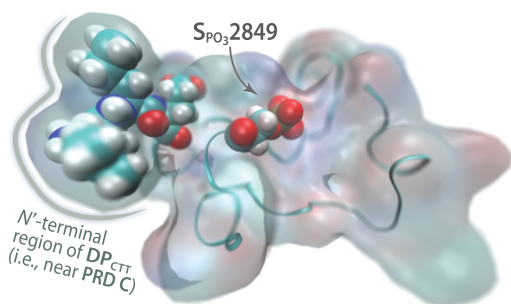


Figure 2.6: S2849 in close proximity to the PRD. This frame, from 71-ns in the $R2834H, S2849S_{PO_3}$ simulation under PARM99SB, exemplifies the contacts made between S2849 and residues that are part of the last plakin repeat domain (PRD C) in DP. Residues 2802–2805 are shown as van der Waals spheres on the left, and $S2849S_{PO_3}$ is shown as vdW spheres in the center. These close contacts suggest that DP_{CTT} can directly interact with the PRDs.

Claws are predicted by several force-fields

The original RC was described as “very stable and, once formed, persist[ing] for the rest of the simulation”[84]; that initial study employed only the Amber FF03 parameter set. A subsequent study of another RS-rich peptide found that RCs form under the Amber PARM99SB-ILDN FF[190]. Our simulations of DP_{CTT} show that RCs can form under both PARM99SB and CHARMM36. Nevertheless, the fine details of RC dynamics are sensitive to the FF; for instance, for many of our systems CHARMM36 frequently predicts higher Cy_*^R values than does PARM99SB.

The FF-dependence of our Cy_*^R parameter is substantial, and this may reflect the somewhat unusual chemical nature of RC sequences, versus most protein sequences. In addition to charged moieties buried in a proteinaceous core, arginine-phosphate interactions are characterized by a “covalent-like” stability[240] that may be inadequately described as point charges interacting via simple Coulombic electrostatics. An RC was not detected in recent NMR experiments with another RNA splicing-related, serine/arginine-rich system[242]; however, the structural ensembles reported in that work were derived via an approach (a ‘sub-ensemble selection procedure’ against the NMR data) differing from the simple, naive equilibrium MD simulations reported here and elsewhere[84, 190], and the trajectories in that work sampled shorter (≈ 50 -ns) timescales. In short, it remains to be established if RCs occur in solution, and under what conditions. A recent crystal structure has shown that (solvent-exposed) RCs can form in the RNA splicing factor SF1[235]. In that system, the RC acts as a secondary structural element in an otherwise disordered region; notably, electron density could be detected for residues immediately upstream of the phosphoserine, but only in the phosphorylated, not the non-phosphorylated, system[235].

Methylation in the DP_{CTT} may promote GSK3 binding

From the simulations presented here, we suggest that the R_{Me2}2834 and S_{PO3}2849 PTMs are required for productive DP-GSK3 interactions. This claim is based upon three lines of evidence. First, our modified DP_{CTT} systems were found to present the phosphate group on the surface, rather than buried within the protein. This surface exposure did not occur in phosphorylated systems with unmodified R2834, suggesting that methylation is coupled to the dynamics of S_{PO3}2849 accessibility. As the processive kinase GSK3 recognizes proteins already containing a phosphate, exposure of S_{PO3}2849 may facilitate GSK3 binding. Second, we find that in some trajectories the S_{PO3}2849-S2845 distance closely matches the distance between the active site and substrate recognition site of GSK3. Upon GSK3 binding to S_{PO3}2849, S2845 can reach the active site of GSK3 without DP having to undergo conformational changes. Third, the steric clash (Figure 2.5) between DP and GSK3, computed along entire trajectories, is far lower in systems containing R_{Me2}2845 than in those without this PTM. The degree to which DP must deform to bind to GSK3 is therefore much lower, increasing the probability that contact between DP and GSK3 leads to the addition of a phosphate at S2845; that is, PTMs may help ‘pre-structure’ the DP substrate in a binding-competent state, thereby decreasing the entropic cost associated with forming a DP-GSK3 complex. In our mechanistic model for GSK3 regulation, DP_{CTT} essentially self-regulates its processive phosphorylation by GSK3; DP_{CTT} achieves this by sampling conformational states that vary in their suitability as substrates for GSK3.

The serine-rich region of DP_{CTT} contacts other parts of DP

Past studies of RCs have examined short, (RS)_n-containing peptides in isolation. The serine-rich region of DP_{CTT} is not well-described by these past models, as we have shown that the SRR can interact with other regions of DP. In particular, the SRR can contact residues that have been resolved in a crystal structure of a plakin repeat domain[38]. The charge-complementarity between a fully-phosphorylated SRR in DP_{CTT} and the positively-charged IF-binding groove on a PRD[38], combined with the tendency for DP_{CTT} to explore the surface of DP, suggests that a simple competition for PRD binding sites may account for the cellular effects of DP_{CTT} phosphorylation. That the DP_{CTT} is covalently linked to the upstream PRDs (Figure 2.1) implies a high local density of negative charge, and this could compete with the negatively-charged IFs to cause DP to detach from the IF network; examination of the ionic strength-dependence of this process would be telling. Finally, note that in our mechanistic model any structural role for arginine claw conformational dynamics (apart from its role in GSK3 processive phosphorylation) would require a further series of simulations, ideally including as many structured PRD regions as possible.

Our computational results can be experimentally tested.

Given the difficulty in simulating disordered proteins, it is essential to consider the results in this paper as hypotheses for further experimental work. We propose four experiments to probe the conclusions we have drawn. First, the existence of an arginine claw can be quantified spectroscopically. By synthesizing the peptide, one can isotopically label one arginine and compare its chemical shift when the post-translational modifications described here are included in the peptide. By seeing which arginine residues change when a phosphate is added, it will be possible to quantify the strength of the arginine claw in a way analogous to Cy_*^R . Second, we propose inhibiting methylases to prevent

methylation of R2834. We predict that phosphorylation of DP_{CTT} will be slowed in this case. A non-methylated peptide with phosphorylation at S2849 (created by peptide synthesis) can be assayed for GSK3 activity in vitro, and this data can be compared to the rate for a methylated peptide. Third, an in vivo experiment can be performed to further test the role of methylation. If our conclusions are correct, DP will bind tightly to IFs when methylases are inhibited since GSK3 will not be able to activate the processive phosphorylation cascade. Fourth, we suggest that PRD C and DP_{CTT} can be spin-labelled, and the distance between the PRD and the tail can be measured using EPR or FRET. We expect that the distance between these regions will be larger in the non-phosphorylated state than in the phosphorylated state, though in both cases the DP_{CTT} will be within 30 Å or so of the PRD’s surface.

2.6 Conclusion

Recent experiments have revealed that desmoplakin’s activity is regulated by PTMs in its presumably-disordered C-terminal tail. Using MD simulations, we have elucidated the structural effects of three modifications in the 70-residue DP_{CTT} region: phosphorylation of S2849, methylation of R2834, and mutagenesis of R2834 to histidine. Our simulations indicate that an RC can form in some of the phosphorylated systems, sequestering the phosphate within the protein. To our knowledge, DP_{CTT} is the largest system that has been shown to form an RC by MD simulation. Our findings build on past studies of RC formation in SR repeats, and are corroborated by recent crystallographic results for other SR systems. Upon methylation of R2834, S_{PO₃}2849 becomes solvent-exposed, which may enhance its detection by the cognate kinase GSK3. Methylation of R2834 has the further effect of biasing the structural ensemble towards conformations that are sterically compatible as substrates for GSK3.

We also find that DP_{CTT}’s SRR is not isolated from the rest of DP, suggesting that studies of short peptides excised from larger systems may miss some of the interactions that define the conformational ensemble of such regions. This point is illustrated by the effects of R2834 methylation: The position of the RC, and the overall conformation of the DP_{CTT}, are affected by this seemingly minor chemical modification, many residues away from the site of phosphorylation. The common self-contact in DP_{CTT}, seen in contact maps for all our simulated systems, suggests that a regulatory mechanism of DP-IF adhesion may be a simple binding competition between DP_{CTT} and IFs for the positively-charged groove on plakin repeat domains.

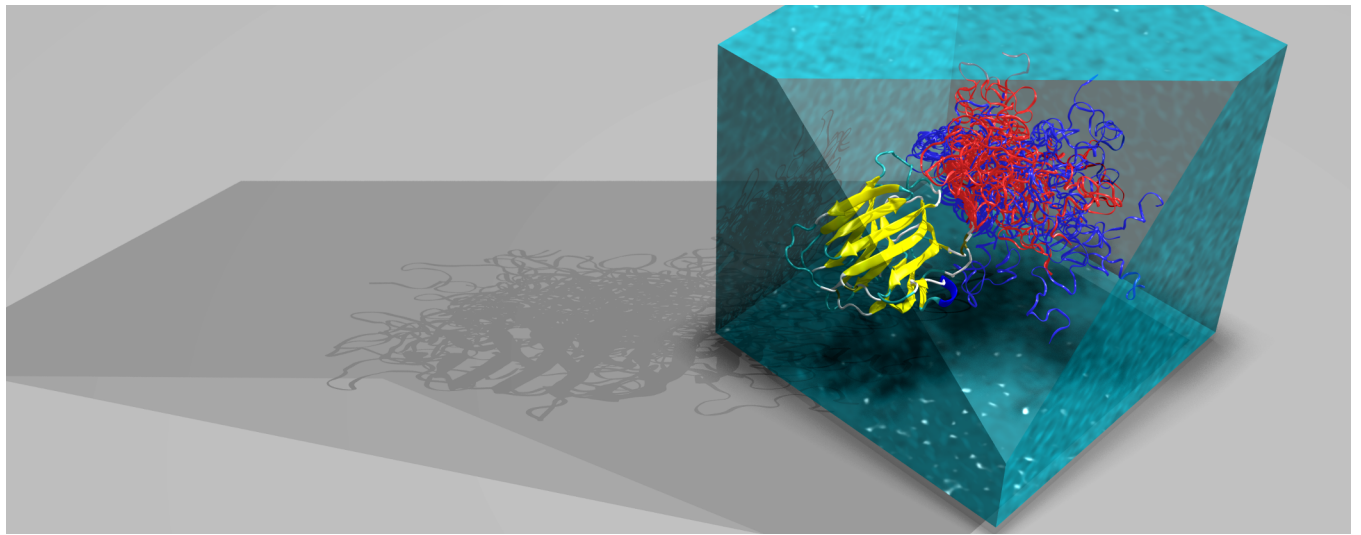
By elucidating the roles and linkages between protein conformational dynamics, PTMs, and claw-like structural elements, our simulations of the C-terminal region of human desmoplakin synthesize several strands of evidence and shed light on the underlying molecular mechanism of DP-IF interactions, including the riddle of strong/weak interactions with the IF network. We predict that RCs can form when S2849 is phosphorylated, and that methylation of the disease-associated site R2834 promotes processive phosphorylation by GSK3. Our data also suggest that DP_{CTT} may bind to a PRD, thus providing a simple, atomically-detailed competition mechanism for the regulation of DP-IF adhesion.

2.7 Acknowledgments

We thank D. Hunt & L. Zhang (UVA), as well as K. Green & L. Albrecht (Northwestern), for helpful discussions. We thank A. Dejaegere (IGBMC) for providing force-field parameters for R_{Me₂}, and D. Hamelberg (Georgia State) for providing the coordinates used to generate Figure 2.2d. K. Holcomb & A. Munro (UVA) are thanked for exceptional computer support; UVA’s Advanced Research Computing Services and Information Technology Services provided computational resources and technical support that contributed to the results reported herein. Portions of this work were supported by UVA, the Jeffress Memorial Trust (J-971), and NSF CAREER award MCB-1350957.

Chapter 3

Toward a Designable Extracellular Matrix: Molecular Dynamics Simulations of an Engineered Laminin-mimetic, Elastin-like Fusion Protein



This chapter is adapted with permission from:

J. D. Tang, C. E. McAnany, C. Mura, and K. J. Lampe. Toward a designable extracellular matrix: Molecular dynamics simulations of an engineered laminin-mimetic, elastin-like fusion protein. *Biomacromolecules*, 17(10):3222–3233, Oct 2016

3.1 Abstract

Native extracellular matrices (ECMs) exhibit networks of molecular interactions between specific matrix proteins and other tissue components. Guided by these naturally self-assembling supramolecular systems, we have designed a matrix-derived protein chimera that contains a laminin globular (LG) domain fused to an elastin-like polypeptide (ELP). This bipartite design offers a flexible protein engineering platform: (i) laminin is a key multifunctional component of the ECM in human brains and other neural tissues, making it an ideal bioactive component of our fusion, and (ii) ELPs, known to be well-tolerated *in vivo*, provide a self-assembly scaffold with tunable physicochemical (viscoelastic and thermoresponsive) properties. Experimental characterization of novel proteins is resource-intensive, and examining many conceivable designs would be a formidable challenge in the laboratory. Computational approaches provide a way forward: molecular dynamics (MD) simulations can be used to analyze the structural/physical behavior of candidate LG-ELP fusion proteins, particularly in terms of the conformational properties salient to our design goals, such as assembly propensity in a temperature range spanning the inverse temperature transition seen in ELPs. As a first step in examining the physical characteristics of a model LG-ELP fusion protein, including its temperature-dependent structural behavior, we simulated the protein over a range of physiologically-relevant temperatures (290-320 K). We find that the ELP region, built upon the archetypal (VPGXG)₅ scaffold, is quite flexible and has a propensity for β -rich secondary structures near physiological (310-315 K) temperatures. Our trajectories indicate that the temperature-dependent burial of hydrophobic patches in the ELP region, coupled to the local water structure dynamics and mediated by intramolecular contacts between aliphatic side-chains, correlates with the temperature-dependent structural transitions in known ELP polymers. Because of the link between compaction of ELP segments into β -rich structures and differential solvation properties of this region, we posit that future variation of ELP sequence and composition can be used to systematically alter the phase transition profiles and, thus, general functionality of our LG-ELP fusion protein system.

3.2 Introduction

A major challenge in neural tissue engineering and regenerative medicine is one of tissue construction: what bio-material, in terms of chemical composition and physical properties, might best mimic the native ECM that houses neural stem cells (NSCs), neurons, glia, and other cells? Engineered proteins afford an opportunity to systematically control both biological functionality and the structural/mechanical properties of the resulting ECM mimetic, thus enabling one to guide the behavior of encapsulated cells[106, 204]. For instance, neural cells encapsulated in engineered protein or peptide materials extend neurites hundreds of microns into the surrounding three-dimensional (3D) matrix[105]. These materials permit cellular remodeling and bioresorption via cell-controlled proteolytic degradation and inherently behave in a more physiologically native manner than other biomimetics (e.g. commonly-used synthetic hydrogels). Tissue engineering can benefit immensely from artificial ECMs designed from naturally occurring protein sequences: such polymers promote native cellular interactions and elicit desired regenerative behaviors *in vivo*[236, 63] while enabling control over bioactive and structural properties (porosity, proteolytic remodeling, cellular adhesion, stiffness, etc.). In short, biologically-based ECM mimetics provide a suitable matrix for the controlled organization of viable cells into physiologically relevant tissues[25, 99].

The ECM in neural tissue is a hierarchically structured composite material, consisting of proteoglycans and large (typically >400 kDa) structural proteins collagen, fibronectin, and laminin. In the central nervous system (CNS), laminin is a particularly vital component of the ECM[83, 13]. Following a neural tissue injury, temporal regulation of laminin expression is critical in the production of potential neurotrophic and neurite-promoting factors by reactive astrocytes[94]. Laminin also plays an important role in axonal growth in the developing mammalian CNS and in concurrent mechanotransduction events, such as in astrocyte cell adhesion and spreading[83, 129].

Laminins are glycoproteins that provide a key linkage between cells and the broader ECM scaffold. Human laminin is an immense (900 kDa), disulfide-linked heterotrimer that consists of many globular domains and α -, β -, and γ -rod-like chains; together, these entities assemble into a four-armed cruciform shape[243]. Several adhesion peptides have been identified within the laminin amino acid sequence; in particular, the ¹¹²⁴RGD, ⁹²⁵YIGSR and ²¹⁰¹IKVAV segments are known recognition sites for as many as 20 integrins[167], the 67 kDa laminin-1 receptor[78], and the 110-kDa laminin-binding protein[209], respectively. These recognition sequences have been used to functionalize non-adhesive polymeric scaffolds, such as in hydrogels based on polyethylene glycol and hyaluronic acid[135, 233, 182]. However, these short ECM-derived peptide fragments are often imperfect in mediating cell-signaling events in neural tissue (cell attachment, axonal growth, etc.), likely because of (i) insufficient binding with cell-surface receptors and (ii) failure to initiate anchoring for assembly of basement membrane scaffolds[212, 213, 239, 218, 113].

The fifth globular domain from the C-terminal region of the laminin $\alpha 2$ chain, denoted LG5, plays a key role as a binding site for integrins, heparin, and α -dystroglycan (α -DG)[46, 224, 95, 206]. Heparin is a highly anionic, polysulfated glycosaminoglycan (GAG) that binds exogenous growth factors and thereby helps regulate and maintain

NSC differentiation[71, 150]. In neural cells, the α -DG glycoprotein complex plays a fundamental role in facilitating new laminin polymerization at the cell surface and supporting cellular adhesion[134, 141]. LG5 also contains a region that binds integrin β 1[224, 206], which is part of an integrin adhesive complex that links the cytoskeleton and the ECM. Past work has focused on engineering hydrogels that contain only the short integrin-binding peptides from LG modules. A more effective biomimicry strategy might incorporate longer laminin sequences, enabling multifunctional biomaterials with native-like cell-binding capacities and targeted selectivity for growth factors (which, in turn, initiate stem cell self-renewal and differentiation programs). There is a precedent for engineering proteins functionalized with the LG5 domain to mediate cellular behavior[49, 91]. A further design criterion for ECM-mimetic fusion proteins is that they contain regions that enable assembly into higher-order structures, via either noncovalent (self-assembly) or covalent (chemical crosslinking) mechanisms. ELPs have generated much interest in the tissue engineering field, as the hierarchical self-assembly of these relatively ordered (via local interactions) peptides provides structural support in ECM materials, as well as the ability to control viscoelastic properties. The ability to tune the physical properties of ELP-containing regions offers a versatile way to modulate protein-mediated interactions between cells and the ECM that are critical in cellular adhesion, spreading, and migration.

ELPs undergo thermally-triggered first-order phase transitions[221] characterized by a system-specific transition temperature known as the lower critical solution temperature (LCST). This behavior is also termed an inverse temperature transition as the polymer becomes more structured upon reaching the LCST, separating into polymer-rich and water-rich phases. Interestingly, the latent heat of these phase transitions are so small that they “challenge the sensitivity and stability of instrumentation”[221]. The solution behavior at/near the LCST depends on both (i) intrinsic factors, such as the amino acid composition[74, 175] and the number of (VPGXG)_n pentapeptide repeats (X denotes a guest residue, which can vary from one repeat to another), as well as (ii) extrinsic parameters, such as the concentration, pH, ionic strength, and other bulk solution properties[132, 119, 36, 219, 133, 217]. Both sets of factors are useful in the context of protein design and engineering, as they are entirely manipulable: various ELP regions can be fused to a target protein and combined with systematic perturbation of experimental conditions to modulate protein/solution properties at and near the LCST. The assembly behavior at the LCST has been introduced into otherwise soluble polypeptides by fusing them to ELP regions[42, 169]. The thermoresponsive behavior of recombinant ELP fusions then allows simple purification via inverse transition cycling[132], thus, obviating expensive chromatographic resins and enabling large-scale production. Also, biocompatibility of ELP fusion proteins with biomechano-responsive properties has recently been demonstrated in animals[116].

Fundamental progress in biomaterials discovery has been limited by a lack of high-resolution data about the structural dynamics of the underlying polymeric network. The properties of any material ultimately stem from the 3D structures and dynamics of its molecular constituents, from the level of individual proteins to their higher-order assembly into matrices. These structural and dynamical properties, in turn, are deeply linked to the patterns of intra- and intermolecular interactions that are thermodynamically accessible (and substantially populated) under a given set of experimental conditions. The structural and thermodynamic properties of a fusion protein design can be quantitatively characterized via experimental means (e.g., X-ray scattering), but systematically doing so on the scale of many dozens or even hundreds of designs would be prohibitively laborious and resource-intensive. Moreover, such approaches do not, in general, provide the atomic-resolution information on structure and dynamics that we need in order to iteratively refine and systematically improve our designs.

The thermodynamic properties and structural dynamics of various ELPs, above and below their LCSTs, have been studied by experimental and computational means[112, 247, 34, 111, 223, 221]. However, a universally accepted, atomically detailed description of the physicochemical and structural basis of this phase transition remains elusive[34, 7, 248]; also, past studies have generally examined short ELP regions in isolation, not fused to other protein domains. Deeper knowledge of the phase behavior and interfacial properties of ELPs would expand their general utility in biomaterial applications and would mitigate the costs of producing and characterizing what end up being poorly structured (or otherwise undesirable) ECM candidates. Here, we have designed and simulated a multifunctional fusion construct, with the ultimate goal of driving neural differentiation via an engineered ECM that assembles under cyto-compatible conditions. We use the LG5 domain to supply crucial cell-protein matrix interactions, while the ELP component of our modular design provides control over desired micro- and nanostructures. Being able to control the properties of our fusion goes in tandem with the architecture and physical properties of these matrices being stimuli-responsive, so environmental parameters such as temperature must be able to modulate the individual protein structures that compose such a matrix.

Using classical, all-atom MD simulations[143], we have examined the behavior of our LG-ELP design near its putative phase transition, as well as the temperature-dependent conformational and structural dynamics leading up to the LCST. These simulations supply picosecond-resolved, atomically detailed information on discrete structural and functional states for our protein, on the overall time scale of about 100 ns. Thus, we can both analyze the molecular events near the presumed LCST transition of our fusion protein and also obtain an a priori view of the structural properties of our design, before dedicating experimental resources to the synthesis and characterization of a novel biopolymer with unknown (and otherwise unpredictable) LCST behavior.

3.3 Methods of Procedure

3.3.1 LG-ELP Fusion Protein Design Methodology

We designed an LG5-ELP fusion protein with the intention that it be able to undergo a temperature-induced structural transition, leading to formation of a functional ECM suitable for CNS tissue regeneration. Four design criteria were applied: (i) The fusion protein should be thermodynamically stable (i.e., retain native structure) under physiological conditions (temperature, pH, ionic strength). (ii) The fusion protein should feature bioactive sites along the LG portion of the peptide chain, and the ELP must not interact with the LG portion in a manner that occludes these bioactive sites (proteolytic sites, cell-binding domains, binding sites for other ECM molecules or growth factors, etc.). (iii) The fusion protein should be capable of self-assembly via noncovalent interactions. (iv) The self-assembly properties should be readily controllable by altering the assembly driving sequence element (ELP, in our case).

ELPs consist of a pentapeptide repeat, $(VPGXG)_n$, where X is any guest residue other than proline. ELPs are described using the notation $ELP[W_iY_jZ_k]_n$, where W, Y, and Z are the single-letter codes for the amino acids at X, the subscripts i , j , and k indicate the number of pentamers featuring that guest residue, and n is the total number of repeats. From our estimates using the T_t -based hydrophobicity scale of amino acids[223, 220, 222] and the LCST behavior of various other engineered ELP fusions[169, 85], we designed an ELP with the sequence $ELP[K_2L_2I_2K_2]_1$. We predict that this motif will satisfy the aforementioned design criteria. The repeated Gly-Leu and Gly-Ile dipeptides serve as cleavage sites for type IV collagenase (gelatinase)[191], rendering the hydrogel susceptible to enzymatic cleavage and thereby allowing cell spreading and migration. In addition, the primary amine functionality of the lysine side chain ($\epsilon\text{-NH}_3^+$) enables site-specific coupling or cross-linking reactions[43].

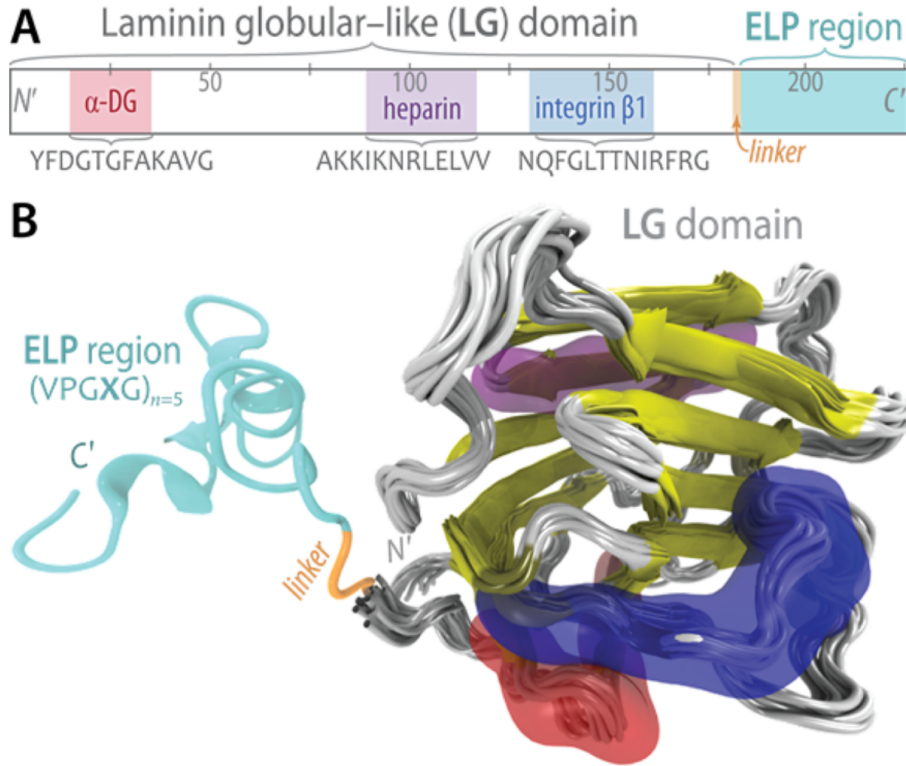


Figure 3.1: Proposed LG-ELP fusion protein. This schematic of our protein engineering design shows a laminin globular (LG) domain fused to a C-terminal elastin-like polypeptide (ELP). (A) Biologically active segments[224, 206] in the LG domain function as recognition/binding sites for α -dystroglycan (α -DG) (red), heparin (purple), and integrin- β 1 (blue). Our ELP repeat region (cyan), consisting of 42 residues of the ELP pentapeptide repeat motif and a three-residue linker (orange), comprises the C-terminal tail of our fusion construct; this ELP region is intended to act as a self-assembly module. (B) A three-dimensional (3D) structural rendition of the fusion protein (ribbon representation) shows the LG domain as an overlay of multiple snapshots from the 100 ns simulation. The LG domain folds as a β -sandwich, with two sheets (one with six strands and the other with seven strands) stacked atop one another; the colored regions correspond to the recognition sequences in (A). The ELP tail is indicated (cyan), with the specific structure shown here drawn from the 315 K trajectory at $t = 1$ ns (i.e., after energy minimization and initial trajectory equilibration).

3.3.2 MD Simulations of LG-ELP.

Our LG-ELP design fuses the LG5 domain, known to adopt an antiparallel β -sandwich fold, to a C-terminal ELP tail (Figure 3.1). Our starting 3D model for the LG5 domain was drawn from the crystal structure of the mouse homologue of the laminin $\alpha 2$ chain (PDB 1DYK)[212], which contains residues 2934-3117 of that particular laminin chain. An initial 3D structure for the 42-residue $ELP-[K_2L_2I_2K_2]_1$ sequence, GVG-VPGKG-VPGKG-VPGLG-VPGLG-VPGIG-VPGIG-VPGKG-VPKG (hyphens are used to visually highlight the pentapeptide repeat motif), was built using the peptide builder tool in the program Avogadro[112] the N-terminal GVG in the above sequence is a linker from the C-terminus of the LG5 domain. The ELP starting structure was modeled as a canonical α -helix, with backbone torsion angles of $\phi = -60^\circ$, $\psi = -40^\circ$ (Figure S3.1). ELPs are likely only loosely structured in solution[152], so the helical starting structure was not anticipated to bias the equilibrium structural ensemble (at least not if given sufficient sampling). Atomistic MD simulations were performed in NAMD, under the CHARMM36 force-field for proteins[58, 121].

To prepare for simulations under periodic boundary conditions, the initial 3D model of LG-ELP was solvated in a cube of explicit TIP3P water molecules, using the “solvation box” extension in VMD[93]; a 15-Å padding of solvent, between the solute and nearest box face, was used to mitigate interactions between the protein and its periodic images. Physiological concentrations (150 mM) of Na^+ ions, including sufficient Cl^- ions to neutralize the solute’s charge, were then added to the solvated system using VMD’s “ionize” plugin. The final simulation cell contained 166,137 atoms, with a cubic box of water measuring 120 Å/edge. The internal energy was minimized for 10000 steps, and the system was then equilibrated for 10 ns (with a 2 fs integration step) in the NPT ensemble (Figure 3.2, initial pose). Simulations were conducted over a range of seven temperatures: 290, 295, 300, 305, 310, 315, and 320 K. In each case, temperature and pressure (1 atm) were maintained using a Langevin thermostat and piston. NAMD 2.9 was used for all simulations[166], with each trajectory extended to a final production time of at least 100 ns. To assess whether trajectory-derived quantities were consistent across our various final (production) runs, and not merely consequences of insufficient/limited sampling, we performed extended simulations. Using the final structure (trajectory frame) from the 310 K simulation (effectively providing a negative control), we computed the corresponding structural quantities of 100-140 ns trajectories at 290, 300, and 320 K. Moreover, we extended the 310 K simulation to 200 ns, as interesting transitions occur near this temperature.

Trajectories were processed and further analyzed using in-house scripts written in the Python[226] and D[6] programming languages, as well as VMD. Root-mean-square deviations (RMSD) for C^α atoms were computed with VMD’s RMSD extension toolbox. Secondary structures in the ELP region were assigned using STRIDE[70, 86]. Table S3.1 summarizes all of our LG-ELP-related simulations. All simulation configuration files and analysis scripts are available upon request.

3.3.3 Analysis of Relative Solvent-Accessible Surface Area.

We calculated solvent-accessible surface areas (SASAs) with the SASA tool in VMD, using a standard water probe radius of 1.4 Å. Rost and Sander’s method[178] was used to determine the relative solvent accessibility, $RelAcc_i$, of each residue i in the ELP region; this relative accessibility is simply the ordinary accessibility of a residue in a 3D structure (Acc_i) normalized by total surface area for that residue type ($RelAcc_i = \frac{Acc_i}{\max Acc_i}$). In our analyses, $RelAcc_i$ values were computed over the entirety of the production trajectories for each simulation temperature.

3.3.4 Hydrogen-Bonding Analysis.

Hydrogen bonds were computed using VMD’s geometric criteria: namely, a distance cutoff of 3.5 Å and a D-H-A angle cutoff of 30° . Hydrogen bonds between two water molecules were excluded from our calculations. The number of water molecules surrounding the ELP backbone was determined by counting the number of waters within 3.15 Å of the peptide, as previously described[248]. This distance corresponds to the first minimum in the radial distribution function between the oxygen atoms of water molecules and atoms in the peptide backbone (Figure S3.2).

3.3.5 Statistical Data Analysis.

Output data from our Tcl/Tk scripts (used with VMD’s Tcl API) were analyzed using tools from the NumPy and SciPy Python packages. Note that all simulations, and subsequent trajectory and data analyses, were of the full-length (225 amino acid) LG-ELP protein. In many cases, we show only the ELP region in certain sections of our analyses; this is purely for the sake of clarity and simplicity. Spearman rank-order correlation coefficients, and associated p -values, for trajectory-derived data (taken from the beginning to the end of the trajectory), such as intramolecular hydrogen bonding statistics, the number of neighboring water molecules, and so on, were calculated using SciPy’s statistical modules.

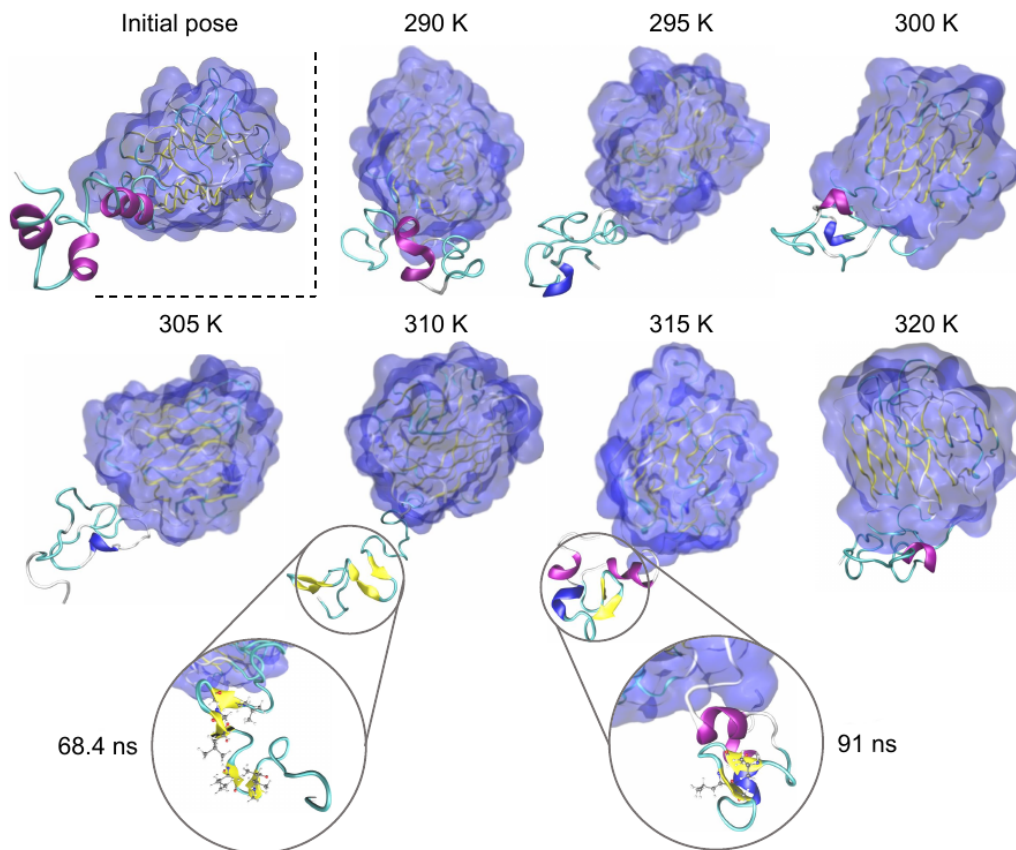


Figure 3.2: **Representative structures, illustrating temperature-dependent conformational states of the LG-ELP fusion protein.** In the initial pose, the LG-ELP protein is shown immediately after minimization and equilibration of the simulation system, with the LG domain (ribbon diagrams) enclosed by a semitransparent blue surface. This initial pose was the starting model for simulations at each temperature. The ELP region (ribbons) in this starting state can be seen to be a mixture of helices and coils; the C-terminus is labeled in this view with α -helices colored purple, 310 helices blue, β -strands yellow, the β -turn motif cyan and irregular coil regions white. LG-ELP structures are shown from each of the 290-320 K trajectories, with each temperature indicated and each structural snapshot taken at 100.0 ns. Insets are representative snapshots at 310 and 315 K, taken from the 68.4 and 91 ns time points, respectively; the side chains that contact one another to mediate β -sheet formation are depicted as ball-and-stick representations (gray carbons, blue nitrogens, red oxygens, and silver hydrogens). These trajectory frames illustrate the formation of β -sheet regions within the ELP tail.

3.4 Results and Discussion

3.4.1 Temperature-Dependent Structural Transitions of LG-ELP.

To explore the structural properties and conformational dynamics of our model LG-ELP fusion protein (Figure 3.1) at various temperatures and illuminate its phase transition behavior, we performed all-atom MD simulations of the protein immersed in a bath of explicit solvent. This system was simulated at temperatures ranging from 290 to 320 K, with each trajectory extended to at least 100 ns duration. Representative structures from the trajectories show that the ELP region in the initial pose is a mixture of helices and coils, and this region forms more structured β -strands near 310-315 K (Figure 3.2). This finding agrees with other studies of the assembly propensity of similar ELP segments (albeit in isolation, not as a fusion partner)[181, 65]. We find that the ELP does not associate with the LG domain, and thus, the LG domain remains accessible in solution for binding of bioactive agents such as integrins, heparin, and α -DG. The structural stability and general rigidity of the N-terminal LG domain is largely maintained throughout each simulation, with the root-mean-square deviation (RMSD) never exceeding 5 Å (data not shown), as opposed to the far more flexible ELP region (Figure 3.3).

As shown by the overlaid structural snapshots in Figure 3.3, the LG domain’s initial structure is largely preserved throughout each simulation. The “frayed” appearance of the ELP region highlights the structural disorder/flexibility inherent to native elastin-based sequences. At temperatures below 305 K, we see a collapse of the ELP from its initial conformation. A hydrophobic cluster within ELP, toward the end of the 100 ns trajectory, is present for all

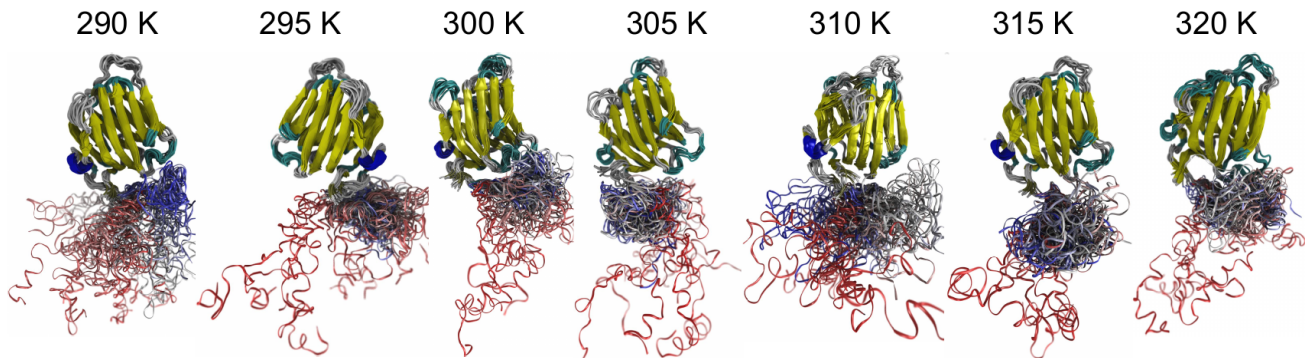


Figure 3.3: **Representative structures of the LG-ELP fusion protein simulated at different temperatures.** Spatiotemporal evolution of the LG-ELP fusion protein is demonstrated by superimposing frames, taken at 10 ns intervals, from the simulation of the entire fusion protein. The ELP region is colored so as to convey the simulation time, graded from early (red) to middle (gray) to late (blue) timesteps along the MD trajectory. Note the structural rigidity of the LG domain and the conformational flexibility of the ELP region.

temperatures except 310 K, where the ELP region becomes extended; this point can also be seen in each contact map (Figure 3.4). Contact maps are matrices that show, for each residue in a 3D structure, the pairwise distance to all other residues. These symmetric matrices compactly represent the pattern of intramolecular contacts, and in our case reveal a lack of interatomic contacts between the LG and ELP regions (Figure 3.4). At 310 K, a transient, but noticeable, extension of the ELP occurred, starting at 75 ns and highlighted by the loss of intrastrand hydrogen bonds (data not shown). This thermally induced rearrangement of the ELP region may well correspond to the sampling of conformations that would favor higher-order (intermolecular) assembly, and we do not see this structural extension at 315 K (though, as for any simulation, absence of an observation could reflect limited sampling).

3.4.2 Secondary Structure Composition and Temperature Dependence.

We examined the structural transitions from the initial starting peptide structure to the final conformational ensemble, focusing on the ELP region of the LG-ELP fusion. At all temperatures, the ELP region exhibits a significant amount of unstructured character (β -turn and “other” in STRIDE), with these two classes accounting for most of the secondary structures in the ELP (Figures 3.5, S3.3 and S3.4). These findings are consistent with solid-state nuclear magnetic resonance (NMR) data[152, 104] and circular dichroism (CD) spectroscopy[174, 140], of similar ELP sequences, where residues within the pentapeptide repeat preferentially adopt β -turn structures. We found that the ELP region accrues β -strand character over the course of a 100 ns trajectory at physiologically relevant temperatures (Figure 3.5), and we posit that this β -strand enrichment can serve as a useful structural property for achieving temperature-triggered LG-ELP assembly; such assembly can occur via intermolecular β -strand- β -strand contacts, for example, by the domain swapping mode of β -rich protein association[179, 54].

In simulations at 305 K, there is a sharp reduction of α -helicity, followed by a complete loss of helical structure after 74 ns (Figures S3.3 and S3.5). The secondary structure distribution at 305 K also shows a bimodal distribution in β -turn and “other” motifs (Figure 3.5), indicating the preferential sampling of these two discrete conformational states. At 310 and 315 K, there is an increase in β -sheet character. The occurrence of β -sheet-like structures at temperatures above the phase transition has been experimentally detected in similar, single-molecule ELP systems[112, 248, 140, 151, 51]. The drastic change in secondary structural content found in our trajectories suggests that heating the system potentially destabilizes polyproline-induced α -helix conformations, perhaps by selectively decreasing the stability of water solvation effects[47, 198]. Such a disruption in helical propensity is consistent with the findings of Li et al.[112] and Ohgo et al.[152], where, at higher temperatures, the proline in (VPGXG) adopts torsion angles similar to type-I and type-II β -turns. This shift in secondary structure in our LG-ELP system is especially prominent at 320 K, where there is a complete loss of β -sheet character, and the reduction of β -bridges with respect to 310 and 315 K is associated with the increase in β -turns within the system. At lower temperatures, the composition favors more α -helical and “other” secondary structures (3₁₀ helices, π -helices, random coils, etc.). The pattern of sampling that we find in secondary structure formation, as a function of temperature, suggests that 310 K is near the target temperature at which macromolecular ordering of the LG-ELP fusion may occur.

This phenomenon associated with the structural changes accompanied by the phase transition is further demonstrated by the distribution of secondary structural content in the 100-200 ns trajectory. The time evolution of the

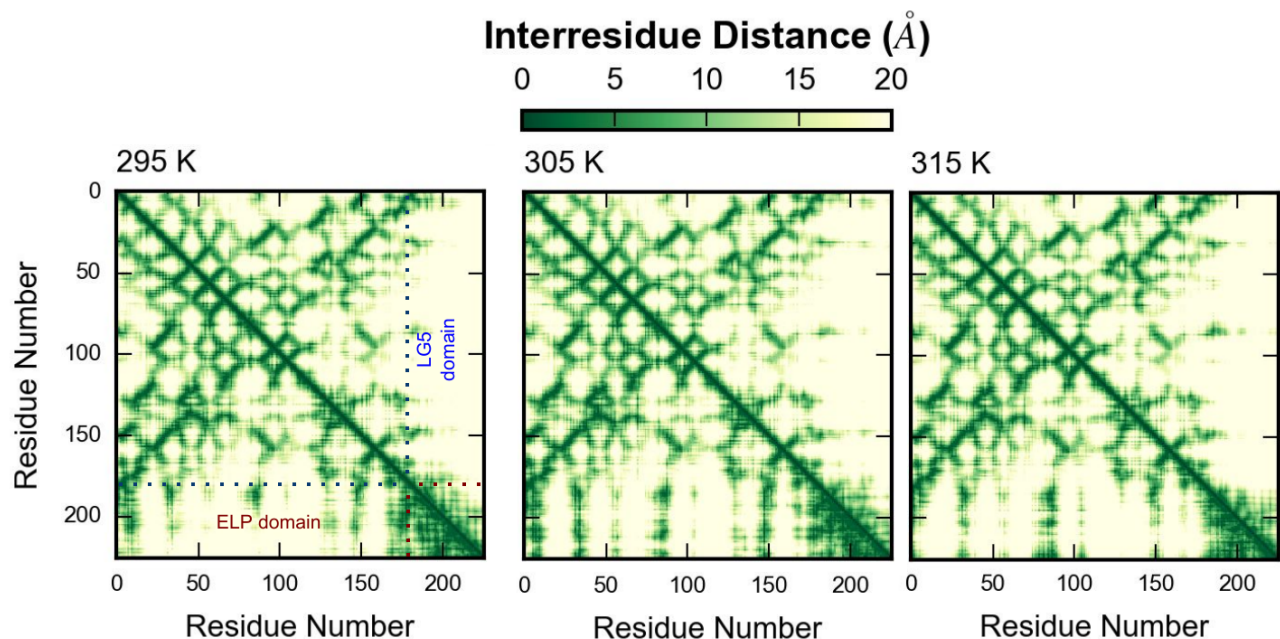


Figure 3.4: **Contact maps of the dynamical interactions in our LG-ELP design** reveal a lack of persistent interactions between the ELP region and the LG domain, independent of temperature. Contact maps are shown for the full length LG-ELP fusion at the indicated simulation temperatures, with colors graded by the pairwise distance (scale bar) between the two side chains under consideration. The LG domain and ELP region, are demarcated by blue and red lines, respectively (for clarity, this is drawn only in the 295 K map). The classic crisscross patterns, highlighted by stripes of contacts perpendicular to the main diagonal, are indicative of the β -sheet core of the LG domain. Because an ordinary (symmetric) contact map contains 2-fold redundant information, here we show (i) the minimum inter-residue distance in the lower triangular matrix, and (ii) the mean inter-residue distance, averaged over an entire trajectory, in the upper triangle. At all simulation temperatures, no stably persistent intramolecular contacts (short distances) are found between the LG and ELP regions, as illustrated by (i) the high-intensity (short-distance) square submatrices at the lower-right of each map, indicating that most ELP residues interact with other ELP residues (not LG residues), and (ii) the vertical white stripes toward the right of each matrix, indicating a dearth contacts between the ELP region and the LG domain. Thus, the ELP region does not engage in spurious/unwanted interactions with the LG domain in solution, at least not on the 100 ns time scale of these simulations. (Contact maps for all simulated temperatures can be found in Figure S3.8.)

secondary structure profile in the extended simulation at 310 K showed four distinct regions of persistent β -sheet like conformations, Leu4 – Gly5200–201 - Leu4 – Gly5205–206 and Ile4 – Gly5210–211 - Ile4 – Gly5214–215 (Figure S3.6) with reduced conformational flexibility. Using the final trajectory frame of 310 K as a starting structure, we extended the simulation from 100 to 140 ns at 290 K, to assess the potential artifacts of limited sampling of structural classes. Reassuringly, we found that the β -sheet state does not persist, and in fact it disappears within 5 ns (Figures S3.4 and S3.7). Similarly, a transition from the 310 K trajectory to 320 K corresponds to a decrease of β -sheet content. At 300 K, however, the temperature shift resulted in a seemingly stable, extended β conformation of the peptide backbone in the Gly5201 – Leu4205 region from 100 to 140 ns (Figures S3.4 and S3.6). This result indicates that the intramolecular contacts between these nonpolar side-chains might be attributable to a population of pre-existing conformations from the previous structural ensemble at 310 K, as these are precisely the same β -sheet forming residues from the initial 100 ns trajectory.

3.4.3 Relative SASA and Association Interactions.

Conformational transitions can be analyzed via dynamical correlation functions, which provide information on how a molecule can interact with the surrounding solvent. We evaluated the SASA of the ELP region in order to characterize the local ordering and solvation dynamics of the system. The SASA can help quantify protein surface-water contacts, and it is a parameter that has long been associated with the thermodynamics of protein structure, as related to the hydrophobic effect and folding[41].

We find no strong trend in solvent exposure properties for residues in the ELP region (Figure 3.6). For all

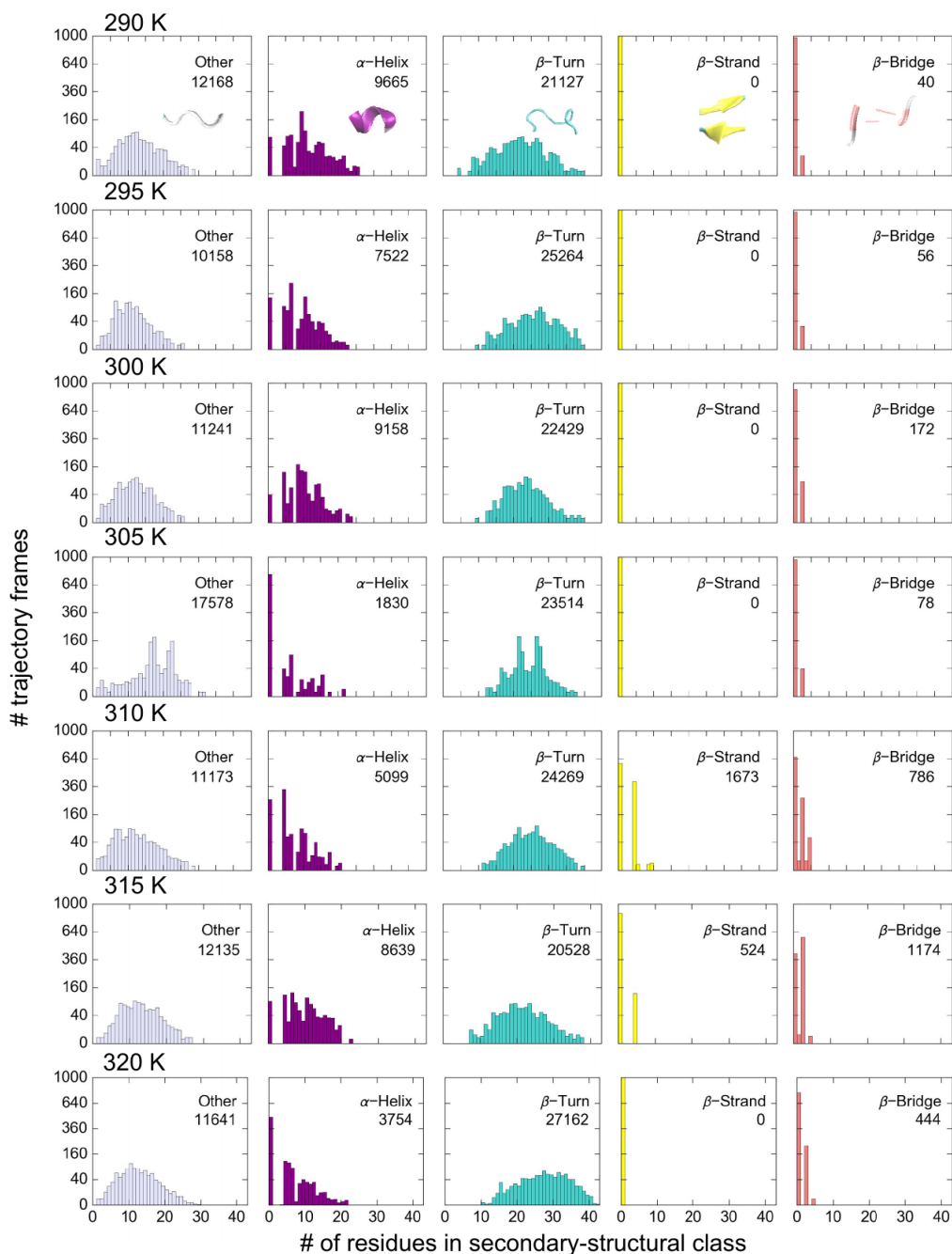


Figure 3.5: **Secondary structural content of the ELP region as a function of temperature** across the 290-320 K series. For simplicity, these trajectory analysis results are shown only for the ELP region, instead of the full-length LG-ELP fusion; there are no noticeable changes in the structural content of the LG domain in all of our simulations. These secondary structure analyses show that the average conformational behavior of a single ELP monomer strongly depends on simulation temperature. Numbers written as insets within each panel give the total number of times that the secondary structure was detected in the simulation. Cartoon representations, shown as secondary structure thumbnail schematics in the first row, match the colors in the histogram. The predominant conformations exhibited by the ELP are β -turns and “other” structures. At low temperatures, α -helical and β -turn structures are prevalent, with minimal β -strand and bridge structures. However, states with greater β -sheet structural content occur as the temperature goes from 305 to 310 K, indicating a possible order/disorder phase transition. Additionally, a significant shift in the character of the β structure, from strand to bridge, occurs at 315 K. The complete lack of β -strand structure at 320 K and subsequent rise in β -turns corresponds to an increased flexibility of the ELP backbone at higher temperatures.

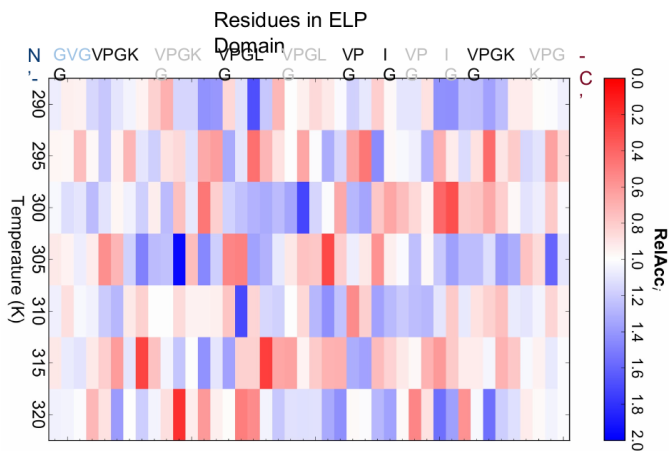


Figure 3.6: **Temperature-dependent changes in relative SASAs of individual residues in the ELP region.** The relative solvent accessibility, $RelAcc_i$, represents the accessible surface area of a residue in the context of a (potentially folded) polypeptide. Blue colors indicate that a residue is more solvent-exposed than average, while red indicates that a residue is more buried than it otherwise would be (outside the context of the peptide).

residues, a linear regression of SASA against temperature yields fits with R^2 values less than 0.5 (data not shown). This result suggests that the structural transitions of ELP regions do not strongly correlate with the SASA of any specific residue, representing a notable departure from previous models of ELP phase transitions[178, 92]. There is also a striking lack of correlation between $RelAcc_i$ and temperature. Linear regression gives R^2 values less than 0.35 for each residue, again suggesting that any ELP phase transition in this temperature regime is not accompanied by gross structural rearrangements. While the hydrophobic regions of ELP have been thought to become more exposed at elevated temperatures (at least for isolated ELP segments, unfused to other proteins)[98], our simulations do not reveal any such correlation. Though the $RelAcc$ of our ELP residues is uncorrelated with temperature, the values do fluctuate (Figure 3.6), and no single residue is consistently buried or consistently exposed. The ELP phase transition, therefore, seems to be marked most strongly by the formation of β -sheet secondary structures, without any concomitant gross structural rearrangements (at least in terms of SASA).

A close examination of the intramolecular contacts, that is, within the fusion protein, reveals that the formation of β -sheets by ELP residues is not occluded or otherwise hindered by interatomic contacts between the ELP region and the LG domain (Figures 3.4, 3.5 and S3.8). From a protein design perspective, this is most reassuring: our simulations suggest that the ELP region will be accessible in solution, free of significant interactions with the nearby LG domain. Similarly, the LG domain’s function should not be abrogated by the presence of ELP, and we expect putative ELP-ELP interactions to mirror those found in previous studies of ELP aggregation[181, 140].

3.4.4 Role of Hydration in Compact Conformations.

We investigated the time-dependent hydration properties of our fusion’s ELP region by examining the intramolecular hydrogen bonding (within ELP) and the number of water molecules hydrogen-bonded with the ELP. The number of surrounding water molecules decreases and the number of intrapeptide hydrogen bonds increases, with increasing temperature from 305 to 315 K, and then a dip occurs at 320 K (Figure 3.7). There is a slow decrease, with time, in the number of solvating water molecules at all simulated temperatures (Figure S3.9a). The 310 K trajectory features an intriguingly abrupt dip in the number of water molecules at 64 and 82 ns. The displacement of water molecules with higher temperatures is consistent with a model, wherein desolvation (e.g., of nonpolar side-chains) biases specific (e.g., polar) segments of the amphipathic ELP chain into more compact conformations, such as β -turns and strand-like conformations. Helical structures are often unfavorable at elevated temperatures for entropic reasons, such as a greater loss, upon folding, of orientational and other conformational degrees of freedom[101, 48]. Thus, higher temperatures may indirectly, via effects on solvation structure, enhance the stability of β -sheet formation in relatively disordered conformational ensembles, such as that of ELP. Changes in hydration density exhibit a correlation with β -sheet propensity along all trajectories (Figures S3.5 and S3.9b). A Spearman’s rank-order correlation coefficient of -0.83 ($p = 0.04$ for 100 ns) indicates a moderately strong negative correlation between intramolecular hydrogen bonding and surrounding water molecules with increasing temperature. This quantity captures the fact that, at elevated temperatures (>310 K), the ELP region preferentially contacts itself rather than water, indicative of a phase transition[152]. The increased number of hydrogen bonds above 305 K suggests a coil-to-globule transition[151]. A possible model is that, at high temperatures, insufficient conformational order exists to allow for formation of a single, well-defined structural state. As such, at lower temperatures the increased rigidity of the system would not facilitate the formation of intramolecular peptide-peptide hydrogen bonds, which would, instead, be replaced by intermolecular hydrogen bonds with the surrounding water structure.

Coupled protein-solvent interactions are a key element of a system’s structural properties and dynamical behavior in any order/disorder transition (e.g., protein folding), but time-resolved experimental data on such interactions are

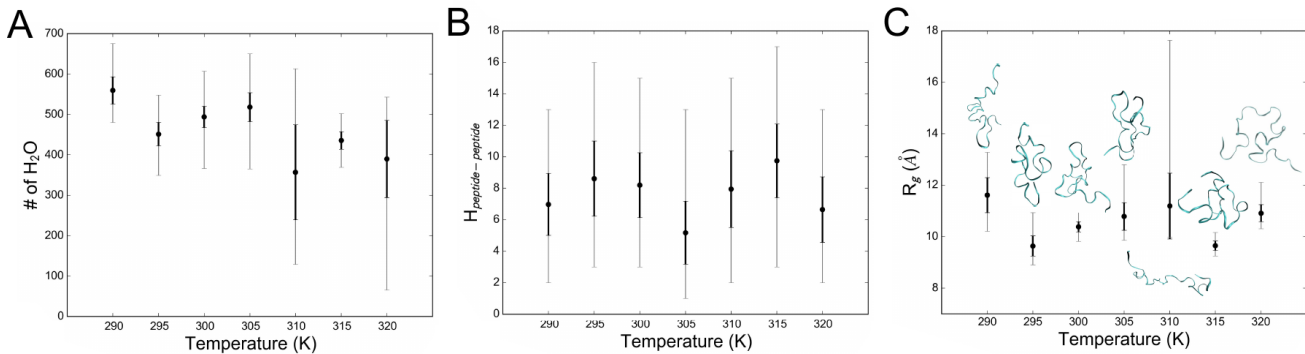


Figure 3.7: Changes in degree of hydration, hydrogen-bonding, and overall structure of the ELP region. (A) Number of water molecules is counted within 3.15 Å of the ELP, with varying temperature. An abrupt decrease in the number of surrounding water molecules suggests that this change is associated with the formation of β -sheets at 310 K in the ELP region. (B) Number of intramolecular peptide-peptide hydrogen bonds ($H_{\text{peptide-peptide}}$). The formation of intramolecular hydrogen bonding has been observed for many peptide aggregates that exhibit LCST behavior; however, the large number of disordered conformational states of our ELP hinders us from discerning any trend as regards a temperature that might be indicative of a phase transition. (C) Temperature dependence of the radius of gyration (R_g). Proteins in all simulated temperatures exhibit temperature-induced collapse (relative to the initial starting structure). Only 310 and 320 K show a slight expansion of the polypeptide chain, while all other proteins exhibit compaction, reminiscent of the “hydrophobic collapse” in typical (water-soluble) globular proteins. In all panels, black error bars represent standard deviations and gray error bars show min/max values. Only the last 40 ns of the trajectories at each temperature are included in the analysis shown here.

not easily obtained, at least at high spatial resolution. Atomistic simulations can provide information about literally each interatomic contact, including the dynamical networks of (i) apolar interactions within a protein, (ii) protein-solvent contacts, and (iii) solvent-solvent contacts, all of which are important factors in macromolecular folding and binding. The compactness of a biomolecular 3D structure, and, by inference, the degree of formation of a hydrophobic “core”, can be measured as the radius of gyration, R_g . The time-evolution of R_g for the ELP region alone (Figures 3.7c, S3.10 and S3.11) does not clearly reveal a sharp phase transition, unlike many biopolymers that exhibit LCST behavior[42, 51, 103]. Though R_g data are, in principle, experimentally accessible via solution-state measurements, for example, Guinier analysis of small-angle X-ray scattering data[173], such approaches to extracting R_g values are confounded by phase changes in going from a soluble to insoluble state, as is common with many polymers that demonstrate LCST behavior[221]. Our simulations reveal that the ELP portion of our fusion protein adopts β -strand secondary structures at high temperatures, implying that this region can undergo structural changes, akin to order/disorder phase transitions, and form ordered complexes. Intriguingly, the drastic solute restructuring that is often associated with LCST behavior[98] does not appear to be a feature in our system’s transition. At higher temperatures, the unfolding or “elongation” of the polypeptide (Figures S3.10 and S3.11) is primarily entropically driven, but at a critical temperature (near 315 K in our system), the chain collapses because the loss of configurational entropy of the side-chains and backbone is counterbalanced by entropic changes in the network of solvent-(solvent, protein) interactions[241, 176]. To assess whether our findings were consistent with our results from the first 100 ns trajectories, we performed additional simulations at 290, 300, 310, and 320 K using the final (100 ns) frame from the 310 K simulation as the starting structure for each different temperature. These extended trajectory data support the argument of a structural transition near 310 K, where it is represented by a gross structural rearrangement of the polypeptide backbone. This transient state is characterized by a “two-state” equilibrium between the collapsed and extended conformation (Figure S3.11) within the ELP region. At low temperatures, that is, 290 and 300 K, we continue to observe a collapsed state, which is stabilized by the relatively strong peptide-peptide and peptide-water interactions, compared to the extended conformation at 310 K.

As a final step of analysis, we considered the “end-to-end” distance, taken as the simple Euclidean distance between the N- and C-termini of a given polypeptide segment, as another geometric measure of peptide compactness. Monitoring the dynamics of the end-to-end distance for the ELP region (Figure S3.12) revealed that this part of our fusion design can explore a substantial region of conformational space without altering its global shape (as indicated by a relatively constant R_g value). Note that this behavior differs from that of larger, “ordinary” globular proteins, where the detailed 3D structural changes that correspond to transitions between nearby local minima on the energy landscape effectively act as barriers to the rapid sampling of conformational space, thereby decreasing kinetic rates of transitions[87, 115, 159].

3.5 Conclusions

Classical, all-atom MD simulations were used to examine the structural properties and conformational dynamics of an engineered, laminin-mimetic elastin-like fusion protein, referred to here as LG-ELP. Analyses of the temperature-dependent conformational changes in full-length LG-ELP, in terms of secondary structural content, solvent accessible surface area, hydrogen bonding, and hydration properties, illuminate the phase transition behavior of this fusion protein. The increased structuring of the protein, and the opportunity that that presents for engineering noncovalent interactions, provide a platform for the rational design of macroscopic material properties[202]. The secondary structural elements in a peptide are known to correlate with the compliance, stiffness, density, and other mechanical properties of hydrogels built upon the given peptide[50, 170]. In this work, we computed atomically detailed MD trajectories of an engineered LG-ELP protein design at several temperatures, thereby providing us with an a priori view of the phase behavior of our design as a function of temperature in the physiological range; reassuringly, we found that the ELP region of our fusion protein did not engage in interactions with the LG domain. This type of information is invaluable in guiding the design of new fusion protein sequences and motifs with desired biological functionalities. Ultimately, our strategy can be used to simulate multiple fusion protein designs, rank-order them, and synthesize those candidates that exhibit the desired phase transition behavior. Because our strategy of using simulations is physics-based, our approach also illuminates the secondary and tertiary structural properties of our LG-ELP fusion, as well as physicochemical properties such as the coupled dynamics of the solvation environment and its influence on the phase transition behavior of our design.

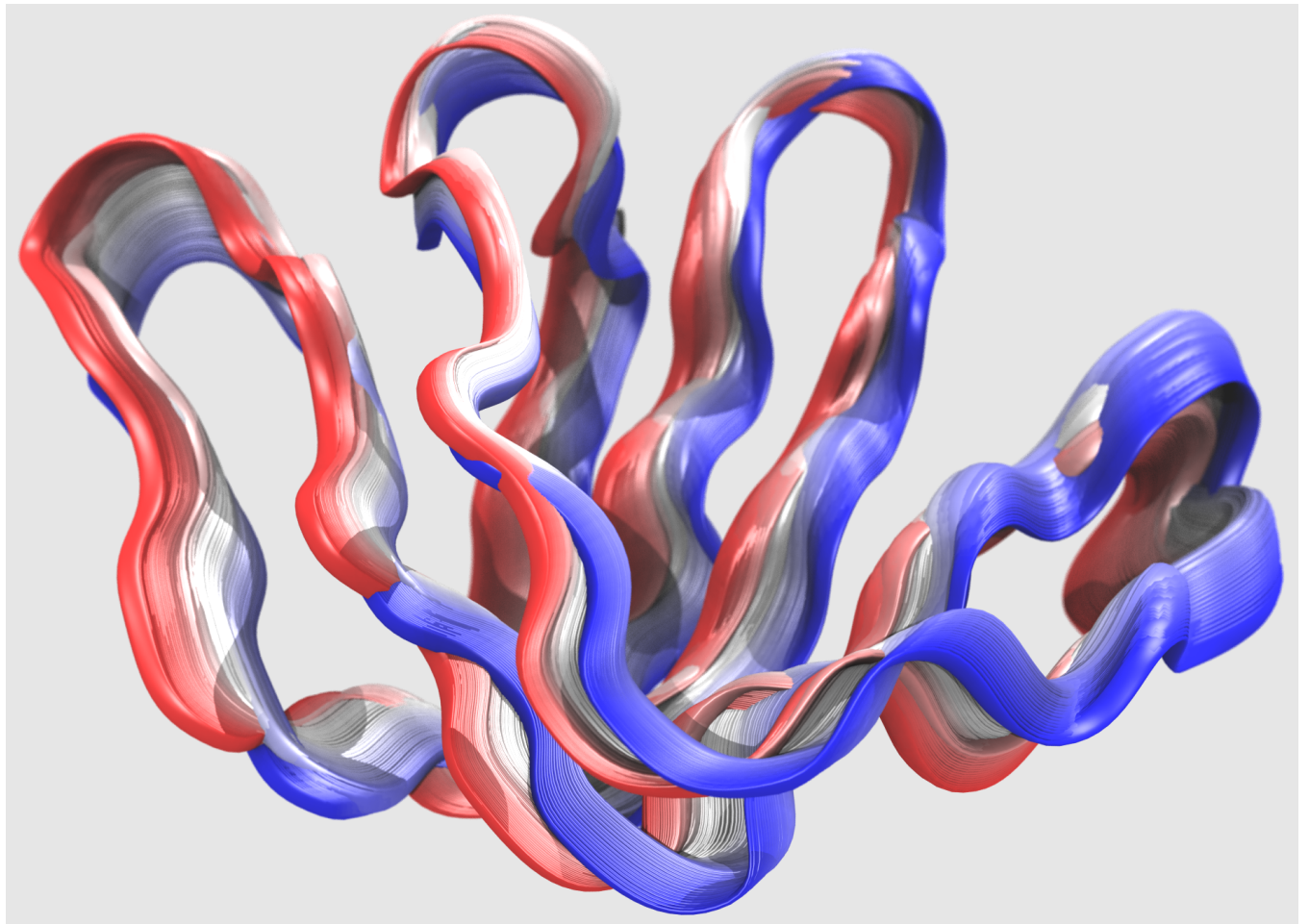
Simulations are enjoying increased use in the analysis of protein structure and function, but to our knowledge an MD-based simulation methodology has not been used in the manner reported here, namely, to help guide the design and iterative refinement of novel fusion proteins that can act as stimuli-responsive cellular matrix materials. The simulations reported here elucidate the relationships between solvation, hydrophobicity, structural dynamics, and other atomically detailed properties, for a novel biomolecular system, and our strategy offers a robust and extensible platform to guide future design and syntheses of protein biomaterials. In particular, our general computational approach can be readily applied in the rational design of engineered extracellular matrix proteins for constructing stimuli-responsive and biocompatible materials for applications in drug delivery, tissue engineering, and regenerative medicine.

3.6 Acknowledgements

We thank K. Holcomb and A. Munro (UVa) of UVa’s Advanced Research Computing Services for exceptional computer support. Computations were performed on UVa’s high-performance cluster, Rivanna, with financial support provided by the School of Engineering and Applied Sciences, Data Science Institute, and College of Arts and Sciences. Portions of this work were supported by UVa (K.J.L.), the Jeffress Memorial Trust Grant J-971 (C.M.), Jeffress Memorial and Carman Trust Grant 2016.Jeffress.Carman.7644 (K.J.L.), and NSF Career Award MCB-1350957 (C.M.).

Chapter 4

The Oligomeric Plasticity of Cyclic Protein Assembly: A Simulation-based Analysis of Sm Rings



4.1 Abstract

The RNA-associated Sm proteins can be found in all three domains of life: archaea, bacteria, and eukarya. In eukarya, Sm proteins are well-studied in connection with their roles in pre-mRNA splicing. In bacteria, the Sm protein Hfq acts as an RNA chaperone, playing vital roles in mRNA-sRNA annealing and RNA-based regulatory networks. In archaea, the functional roles of Sm proteins remains an open question. Sm proteins assemble into cyclic oligomers of 5, 6, 7, or 8 subunits, and the assemblies can be either homo- or heteromeric. Bacterial Sm proteins have only been found as homo-hexamers, while eukaryotic Sm proteins typically assemble into hetero-heptamers. Archaeal Sm proteins have been found as homomeric hexamers, heptamers, and octamers. Despite this variation in quaternary structure, all Sm monomers exhibit nearly identical tertiary structures. How can this be? What is the origin of this oligomeric plasticity, if not encoded in the monomer? We have used a systematic array of molecular dynamics simulations to examine the interfaces between Sm subunits, and have developed several quantitative relations that link the results of dimer simulations to the behavior of complete rings. The simulations reveal that Sm oligomers are remarkably flexible. Sm dimers can adopt multiple conformations, and Sm rings are distinctly asymmetric. For a dimer of the *E. coli* Sm protein, our simulations show one monomer twisting nearly 15° from its crystallographic position. The surprising flexibility of Sm oligomers may be related to the dynamical effects of RNA binding and we are currently investigating these effects in a variety of Sm systems.

4.2 Introduction

4.2.1 Many proteins assemble into oligomers.

Large protein assemblies offer functional and evolutionary advantages. Such oligomeric structures can use allostery to communicate between multiple active sites, and they are more stable against denaturation. Further, for structural proteins such as intermediate filaments and viral capsids, a protein’s large size is essential to its function. Cells can create large protein assemblies either by using one long peptide chain, or by assembling several smaller proteins[163]. The latter option is commonly preferred by evolution; indeed the *majority* of proteins assemble into oligomers[60]. The formation of oligomers provides several advantages to the cell; these advantages have been reviewed in detail in [76]. Oligomers are more resistant to coding errors since only one subunit needs to be remade if an error occurs. They offer better coding efficiency in the genome, by using identical components multiple times (in the case of homooligomers), or by creating modular assemblies where swapping elements of the oligomer changes its function[163]. The favorability of small proteins is shown by their prevalence in the genome: eukaryotic proteins have a median length of 361 amino acids, while bacterial proteins have a much shorter median length of 267 amino acids. Archaeal proteins are shorter still, with a median length of 247 amino acids[24].

In this work, we focus on the geometry of symmetrical, cyclic oligomers. Such a ring-shaped system can form one more intersubunit interface than an open oligomer with, say, helical symmetry[76]. Further, since none of the oligomer-forming interfaces are exposed in a ring, there is a reduced propensity for aggregation. Though the benefits of forming rings are multiple, it remains unclear what unifying principles govern the number of subunits in cyclic oligomers.

In some cases, basic principles of sterics and dynamics dictate, or at least guide, a protein’s oligomeric state. For example, the propensity for β -propellers to adopt seven-fold symmetry has been shown to arise from fundamental steric effects[146]. Even number theory plays a role in oligomerization: Matsunaga et al. suggest that rings with prime numbers of subunits are in general more rigid than those containing highly-composite numbers of subunits[126]. This rule arises from the observation that a vibrational mode of a ring can have all of its nodes at subunit interfaces only if the symmetry of the mode divides the symmetry of the ring ($N_{ring} = 0(\text{mod } N_{mode})$). If the ring has a prime number of subunits, then any vibrational mode must have a node inside a subunit of the ring[126]. In general, though, predicting the oligomeric state of a protein given only the structure of the monomer is a difficult task that must be addressed using search methods such as protein-protein docking[16] and phylogenetic analysis of interfaces[62].

Some proteins simply have more than one allowable oligomeric state. These proteins include ion channels from hepatitis C virus with four to seven subunits[29], the Lo18 protein chaperone that exists as dimers, dodecamers, and 16-mers[123], the TRAP RNA-binding protein from *Bacillus stearothermophilus* with 11 or 12 subunits[126], an archaeal Sm protein that assembles along the edges of β -sheets[100], and leucine-zipper-based coiled-coils of α -helices[154].

While many proteins assemble into cyclic oligomers, they are frequently not perfectly symmetric[20]. The degree of symmetry for a particular assembly can be quantified with the continuous symmetry measure (CSM)[244]. The CSM is 0 for a perfectly symmetric structure and increases up to a maximum value of 1[20]. The CSM is directly comparable between two structures; it does not depend on the number of atoms or the overall system size. These measures have recently been applied to large assemblies including whole proteins[61, 168]. These measures are very informative for nearly-symmetric structures, but are not directly applicable to the task of determining if a

given substructure is a component of a larger, nearly-symmetric structure. We have developed a new geometric measurement that, in a sense, combines CSMs and oligomeric state prediction.

4.2.2 The Sm family of ancient, structurally-conserved RNA-binding proteins.

The Sm family of RNA-binding proteins is found in all three domains of life. Bacterial Sm proteins, known as Hfq, are posttranscriptional regulators that act by binding RNA[139, 184]. Frequently, Hfq oligomers will bind to regulatory sRNAs on one face of the ring and mRNAs on the other, thereby mediating base-pairing interactions between the two RNA strands[145]. This places Hfq at the center of a number of regulatory pathways where a particular sRNA interacts with an mRNA transcript to either promote or inhibit expression. These physiological pathways include quorum sensing[109], virulence,[31], and iron metabolism[125].

Hfq proteins assemble into homohexameric rings, with no known exceptions. The face of the ring containing the N'-terminal α -helix is termed the *proximal* face and the other face of the ring is the *distal* face. Hfq typically binds to mRNA on its distal face, while sRNA typically binds to the proximal face[184]. A third, *lateral*, binding site has recently been discovered and it may facilitate the actual base pairing between mRNA and sRNA[199].

In eukarya, Sm proteins are best-known for their role in mRNA splicing[21]. Eukaryotic Sm proteins form the core of the spliceosome, a humongous machine in which snRNA threads through the pores of several Sm oligomers[110]. While Hfq is believed to operate exclusively as complete rings, the Sm proteins in the eukaryotic spliceosome assemble around the cognate snRNA strand as part of the snRNP biogenesis pathway[145, 81]. There are seven eukaryotic Sm paralogs that are found in the spliceosome: SmD1, SmD2, SmD3, SmB, SmE, SmF, and SmG. While the entire Sm ring is not stable in the absence of RNA, several of the Sm proteins do form stable sub-heptamers: SmD1/SmD2, SmD3/SmB, and SmE/SmF/SmG[81]. The evolutionary origin of this complex set of Sm proteins is reviewed in [230].

Sm-like archaeal proteins (SmAPs) have been found in several archaeal species, but their function remains largely unknown[145]. Archaeal species encode between one and three Sm proteins, and the evolutionary relationship of SmAPs and eukaryotic Sm proteins has been carefully studied in [189].

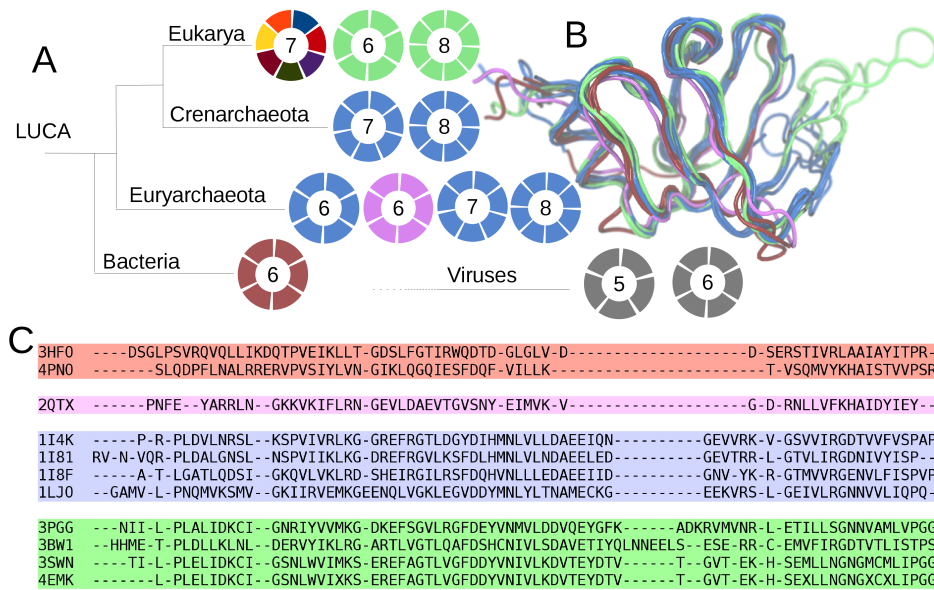


Figure 4.1: **Structural similarity of Sm proteins.** In (A), a phylogenetic tree showing the oligomeric states adopted by Sm proteins in different domains. With the exception of eukaryotic heptamers (multicolored), all Sm proteins form homomeric rings. Aligning several Sm protein structures (B) shows the degree of structural conservation in Sm proteins from bacteria (burgundy traces), eukarya (green traces), and archaea (blue traces). The archaeon *Methanococcus jannaschii* contains an Sm protein (purple trace) that is more similar to Hfq than to other SmAPs. In (C), a sequence alignment of the structures in (B) shows the large variation in sequence among Sm proteins. The two rightmost β -strands in (B) are present only in archaeal and eukaryotic Sm proteins, not Hfq; the lack of this β 3- β 4 hairpin extension in bacteria accounts for the large gap in Hfq-like sequences in (C).

Though Sm proteins show immense sequence diversity, considered across the phylogenetic tree, they are remarkably similar at the level of monomer 3D structure, as illustrated in Figure 4.1. Despite the shared monomer structure, different Sm proteins are able to assemble into different oligomeric forms. Bacterial Hfq is only known to assemble

into homo-hexamers[145]. Eukaryotic Sm proteins assemble typically as hetero-heptamers, though a homo-octamer of the Sm-like protein LSm3 has been observed in a crystal structure[147], though the biological role of these octamers, if any, is unknown. (LSm3 is known to assemble with other LSm paralogs to form hetero-heptamers in vivo.) SmAPs have been found to assemble into homo-hexamers, homo-heptamers, and homo-octamers[215]. While most Sm proteins have one preferred oligomeric state, a SmAP has been identified that forms either a hexamer or heptamer depending on solution pH and the presence of RNA[100]. A particularly interesting Sm-like structure of putative cyanophage origin adopts the Sm fold (though the α -helix is C-terminal, not N-terminal) and assembles into a pentamer[53].

What enables this plasticity in quaternary structure, given that Sm proteins have such similar tertiary structure? Figure 4.2 shows the difficulty with this plasticity: it is analogous to cutting a pizza in six slices, removing a slice, and rearranging the remaining slices without gaps between them. Given that the tertiary structure of the Sm monomer is strongly conserved, what is the source of flexibility that allows different Sm proteins to adopt so many oligomeric states?

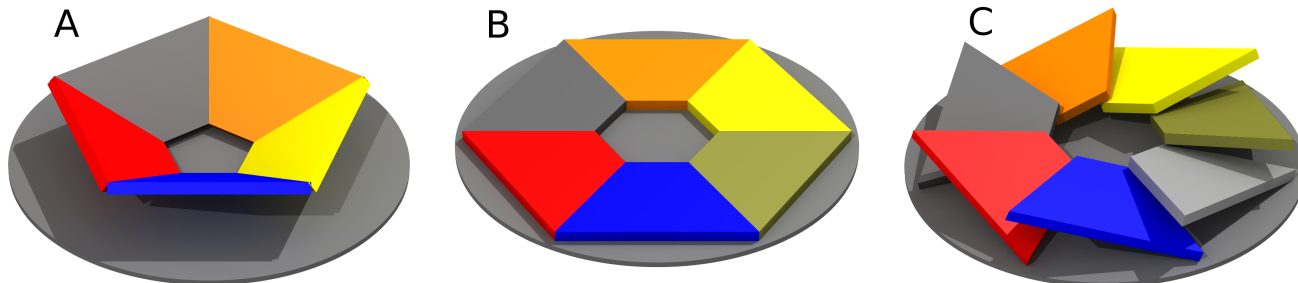


Figure 4.2: **A pizza displays oligomeric plasticity.** Sm proteins, despite having nearly-identical tertiary structures, can assemble into a variety of oligomeric forms. By analogy, this is like cutting a pizza in six slices, removing one slice, and then moving the other pieces to close the created gap. In three dimensions, this is possible. A pentamer (A) can be created by removing a piece from a hexamer (B) and applying a negative pitch to the monomers. Similarly, a positive roll on each monomer creates enough room to add an an extra monomer, creating a heptamer (C). These motions are quantified using the Pizza Tensor.

To explore these questions, we have performed extensive molecular dynamics (MD) simulations of 19 different Sm dimer systems, two Sm rings, and one tetramer. In total, these simulations represent nearly 4 μ s of simulation time in explicit solvent. Where not explicitly stated otherwise, we will consider only dimers drawn from complete rings in this work. When referring to “a dimer from the tetramer simulation”, we are using the center two monomers from the tetramer simulation. When referring to “a dimer from the complete ring”, we have used the data for two adjacent monomers from a simulation of the complete ring.

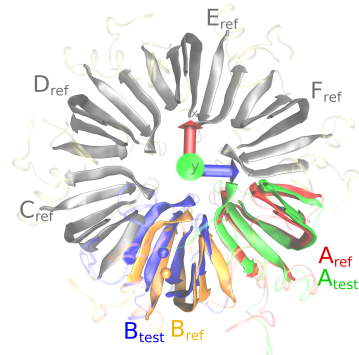
4.3 Results And Discussion

4.3.1 A “Pizza Tensor” quantifies the structural relationship between Sm subunits.

We first define a geometric measurement that allows us to relate all Sm dimer structures to a single, common reference model. The pizza tensor (PT) quantifies how much a given Sm dimer deviates from a reference structure. First, chain A of the dimer of interest is aligned to chain A of the reference dimer. Then, the transformation matrix that best maps chain B of the reference dimer to chain B of the dimer of interest is calculated. This matrix is expressed in terms of dx , dy , dz , $roll$, $pitch$, and yaw to provide an easy-to-visualize description of the motion, and these six values constitute the PT. For the reference structure, we have used the structure of *E. coli* Hfq solved at 0.97 Å resolution by Schulz et al. (PDB ID 4PNO)[188]. Using the hexameric ring as generated by crystallographic symmetry, we define chains A and B by looking at the proximal face of the ring, with chain B at the 6 o’clock position, and chain A at the 4 o’clock position. We have positioned the reference dimer such that the center of mass of chain B is at the origin, and the center of mass of the whole hexamer lies on the positive x-axis. The proximal face of the Hfq ring faces the positive y-axis, and thus chains A, B, C, and so on proceed in a left-handed direction about the y-axis (The global coordinate system is right-handed). Figures 4.3 and 4.4 show how the PT is calculated.

```
void calcPizzaTensor(Dcd testDimer, Molecule refDimer) {
    //testDimer and refDimer both contain only the C-alpha atoms from
    //residues common to the two proteins (as determined by structure alignment)
    Molecule chainRefB = refDimer.select("chain_B");
```

Figure 4.3: **Representative PT alignment.** For the reference structure, chains A, B, and C-F are colored red, orange, and gray, respectively. The dimer of interest shows chains A and B in green and blue, respectively. To calculate the PT, chain A of the dimer of interest (green) is aligned to chain A of the reference structure (red). The least-squares transformation matrix that maps chain B of the reference (orange) to chain B of the dimer of interest (blue) is calculated. In this figure, the centers of mass of chains B are shown as small spheres. In this figure, chain B of the dimer of interest has moved substantially in the positive x direction. For clarity, regions other than the β -sheet are shown as transparent ribbons.



```

Molecule chainRefA = refDimer.select("chain_A");
foreach(Molecule frame; testDimer.frames){
  Molecule chainTestA = frame.select("chain_A");
  float [4,4] fitTestARefA = measureFit(chainTestA, chainRefA);
  //move the whole test dimer by fitTARA.
  frame.move(fitTestARefA);
  Molecule chainTestB = frame.select("chain_B");
  //Find the transformation between chains B.
  float [4,4] pizzaMat = measureFit(chainTestB, chainRefB);
  //Convert from a 4x4 matrix to (dx, dy, dz, roll, pitch, yaw)
  write(toPizzaTensor(pizzaMat));
}
}

```

4.3.2 The oligomeric state of the ring can be predicted based on the dimer.

We next present a method to predict the oligomeric state of a cyclic homooligomer given only the dimer geometry. We present the algorithm first for cyclic oligomers containing an integral number of subunits, and then we present a continuous version. Given a homodimer of chains A and B, calculate the least-squares transformation matrix that maps the atoms in chain A to those in chain B: $B \approx MA$, where B is a $4 \times n$ matrix of homogeneous coordinates of atoms in chain B (the fourth component of each coordinate is always 1), A is a $4 \times n$ matrix of atomic coordinates in chain A, and M is a 4×4 affine transformation matrix. n is the number of atoms in each chain. For a dimer extracted from a k -fold symmetric oligomer, we know that $A = M^k A$, since after k operations the coordinates return to their original position. To determine the predicted oligomeric state (POS) given only a dimer, evaluate $\|A - M^k A\|$ for integer values of k until the second minimum is found or some upper cut-off value is reached. (The second minimum is needed because there will always be a trivial minimum at $k = 0$.) This value of k that minimizes $\|A - M^k A\|$ is the POS.

A dimer structure will not necessarily predict an oligomer with an integer number of subunits. In this case, we extend the above to allow for non-integer k . If M is diagonalizable (which it almost always will be), then $M = SDS^{-1}$, where S is a square matrix whose i^{th} column is the i^{th} eigenvector of M , and D is a diagonal matrix and $D_{i,i} = \lambda_i$, with λ_i the i^{th} eigenvalue. Then $M^k = SD^k S^{-1}$ with $D_{i,i}^k = \lambda_i^k$. The search for the second minimum of k is performed by evaluating $\|A - M^k A\|$ for $k = 0, 0.5, 1, \dots, 10, 10.5, 11$. (The half-integer values are necessary to ensure that the algorithm does not miss the second minimum.) Given that the second minimum was found at k_{min} , a binary search is performed on the range $(k_{min} - 0.5, k_{min} + 0.5)$ to determine the non-integer value of k that minimizes $\|A - M^k A\|$.

In the figures in Section S4.1, the POS is presented as a green trace, and the value of $\|A - M^k A\|$ is presented as a blue trace. Both traces are smoothed with a 1-ns sliding median filter.

4.3.3 Sm dimers are structurally stable on the 200-ns timescale.

We have performed extensive MD simulations of the 22 different Sm protein systems summarized in Table 4.1. We have assessed the stability of each simulation in seven ways. First, we have assessed six root-mean-square deviation (RMSD) values for each simulation. RMSD values were calculated according to Equation (1.13), using the

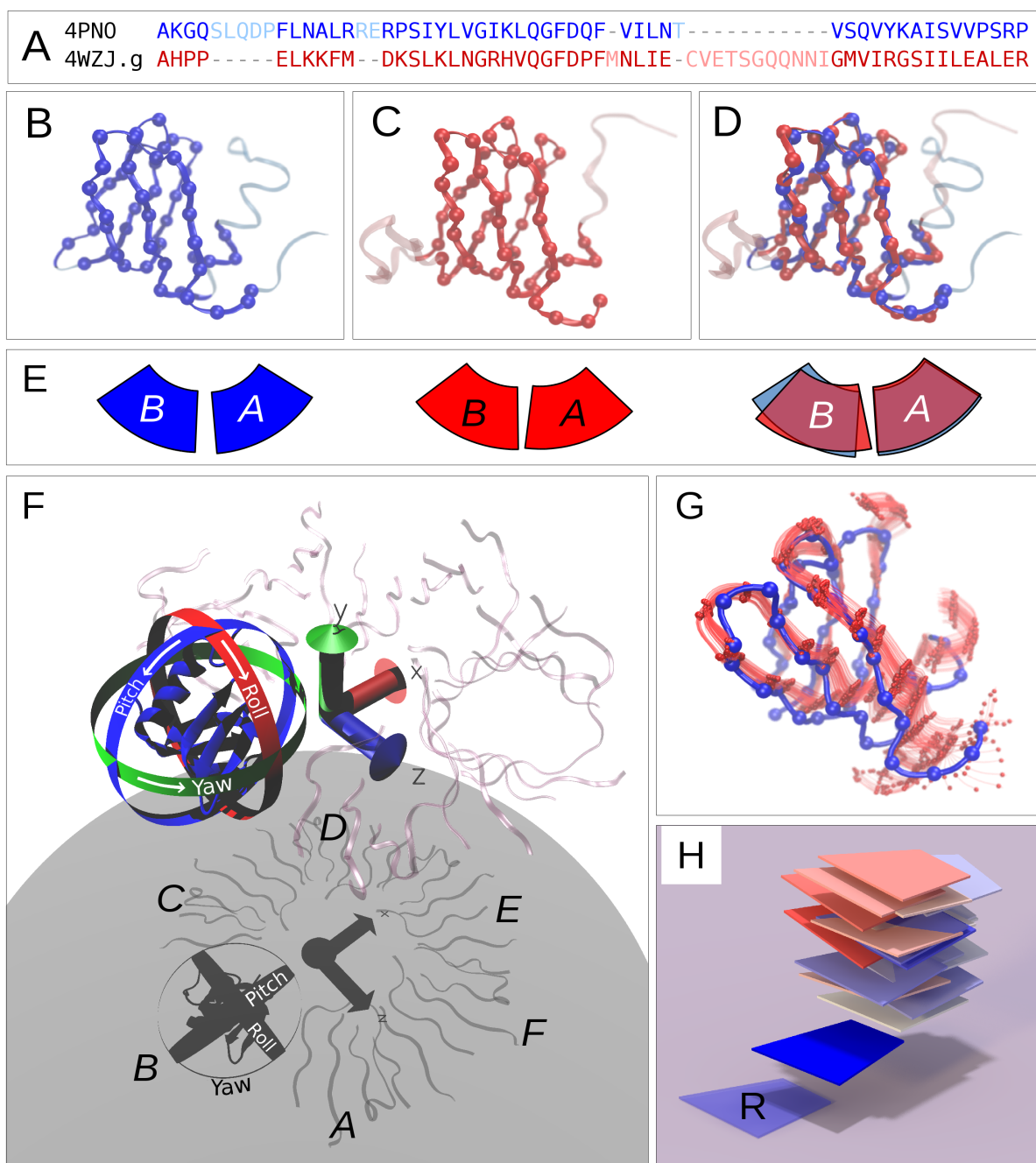


Figure 4.4: **Calculating the PT.** Two Sm dimers are to be aligned. First, a structure alignment of monomers is performed, yielding the sequence alignment shown in (A). (For the sake of text size, many residues have been omitted.) The residues of chain A that matched in the alignment are shown for the reference structure (B) and the target structure (C). In (D), the two structures have been aligned. (E) shows a schematized version of the alignment, where the dimers of each structure (left and middle) have been aligned based on chain A (right). The change between the two chains B is the PT. (F) shows the six degrees of freedom present in the PT. Note that the origin of the x , y , and z axes is actually in the center of the test monomer, not in the center of the ring. (G) shows the results of a simulation, where the target monomer has sampled a number of positions. These positions are represented in (H), using blades that are roughly the same size as Sm monomers. (The orientation in (H) is the same as the blue monomer in (F).) At each 10 ns in the simulation, the reference blade (R) is transformed by the PT and a new blade is drawn. Blue blades occur earlier in the simulation, and red blades occur later. (The translational motion has been exaggerated tenfold to better separate the blades for the sake of this visualization.)

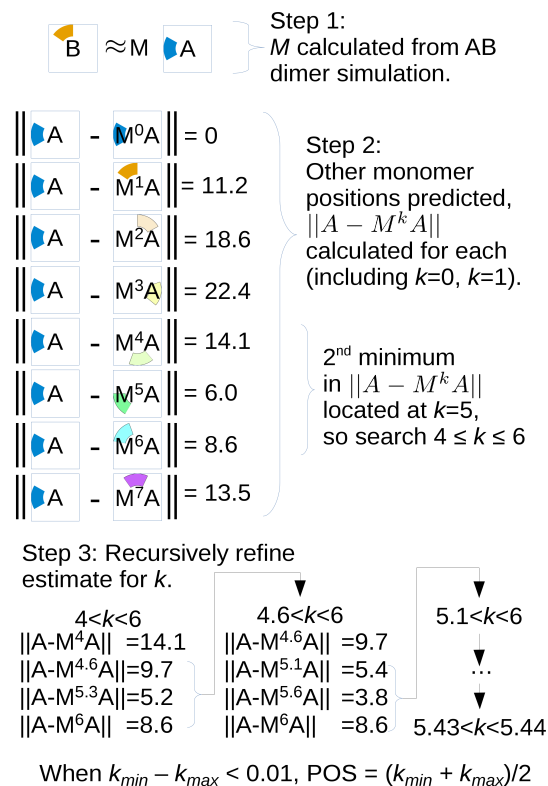


Figure 4.5: **Calculating the predicted oligomeric state (POS).** The POS is calculated in three steps. First, the transformation matrix M from chain A to chain B is calculated based on simulation data. Next, M is applied repeatedly to chain A, and at each iteration k the value of $\|A - M^k A\|$ is calculated. When a second minimum in this quantity has been located, the algorithm then recursively refines the value of k so that $\|A - M^k A\|$ is minimized. The value of k where this minimum occurs is, by definition, the POS. To account for translation, M is a 4x4 affine matrix and A is a 4xn matrix of homogeneous coordinates $[x, y, z, w]$ with $w=1$ for every atom.

initial structure in the production run as the reference. Each system was aligned to the reference to minimize RMSD at each frame. We show RMSD data for each monomer individually and the dimer as a whole. These RMSD values are calculated both for the all-atom system as well as for only backbone atoms that form the Sm core (see Methods for how this is defined). We also show the number of intermonomer interactions by calculating the number of atoms that are within 3 Å of the other monomer. The full data are provided in Section S4.1.

Briefly, the RMSD values for the entire monomers are occasionally surprisingly high (11 Å for chain A of 2QTX (*Mja* Hfq), for example). However, all systems which contain such high values include regions that were not resolved in the crystal structure and were modelled essentially as straight chains extending from the protein. It is not surprising, therefore, that such systems would display great flexibility. This effect is particularly pronounced in the monomers from 4WZJ (*Hsa* Sm hetero-heptamer), many of which contain long extensions to the Sm core. By restricting the RMSD calculations to those atoms which form the canonical Sm core (those which align to the *Eco* Hfq structure), the values are reduced to under 4 Å.

Tracking the number of intersubunit contacts shows that none of the systems dissociate. For each system, we have calculated the average number of atoms within 3 Å of the other monomer during the trajectory. With the exception of 3BY7, the pentamer of putative cyanophage origin, every simulation has an average contact number of at least 13. The plots of the number of contacts over time (Section S4.1) show that no system exhibits a clear, systematic decrease in the number of contacts over time. Systems containing Sm proteins with N- and C-terminal extensions to the Sm core, such as the eukaryotic homologs SmD3, SmD1, and SmD2, show an increase in inter-monomer contacts as the initially-disordered extensions collapse during the simulation. Figure 4.7 quantifies the nature of the actual interface in terms of buried surface area.

4.3.4 The PT reveals remarkable flexibility in Sm dimers.

As an informative starting point, we first consider the simulations of 4PNO (*Eco* Hfq), as that structure was used as the reference in calculations of the PT. Figure 4.8 shows the behavior of the dimer.

The translational components of the PT reveal that the dimer is built upon a flexible Sm-Sm interface, where chain B can move by over 1 Å in the x , y , and z directions. The average values of dx , dy , and dz are all under 1 Å, suggesting that the equilibrium structural ensemble for this dimer contains the crystal structure, but that the dimer is not tightly constrained to this geometry. Examination of the rotational degrees of freedom shows a striking *pitch* of -14.8°. Simulations of a tetramer of 4PNO also reveal a flexible protein; the central dimers shear by 2.4 Å in the x direction, and 1.7 Å in the y direction. In the rotational degrees of freedom, the tetramer is not as flexible as the dimer but still displays *roll* and *yaw* variations above 5°. As might be intuitively expected for structural and steric reasons of steric packing, a simulation of the complete 4PNO ring reveals a more rigid structure than the dimer or

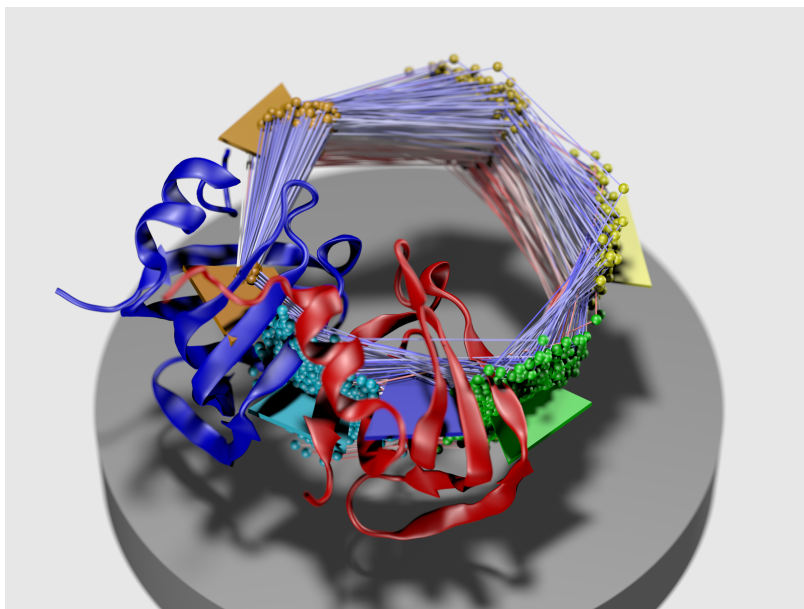


Figure 4.6: **The predicted oligomeric state (POS) of a dimer** The POS of 1LJO (*Afu* SmAP2) is shown at every 50 ps of the first 10 ns of that simulation. First, the transformation matrix M that maps chain A (red) to chain B (blue) is calculated. Then, chain A is moved according to M and a sphere is drawn at the new center of mass of chain A (orange spheres inside chain B). Chain A is moved again according to M and spheres are drawn at its new center of mass. This is repeated four more times to generate the image. Line colors indicate the trajectory timestep, with blue lines occurring earlier in the trajectory and red lines occurring toward the end. When M is applied 5 times (green spheres), the center of mass has not quite returned to chain A’s original center of mass. However, when M is applied 6 times (blue spheres), the center of mass has overshoot the original center of mass. This means that the POS of this system lies somewhere between 5 and 6. The blades (drawn at 6 ns) show that, in addition to the POS being between 5 and 6, the projected monomer has also drifted in the $-y$ direction (toward the distal face of the ring).

tetramer. The complete ring shows a translation in the z direction of nearly 1 Å, and *roll* and *yaw* values near 3.6° and -4.3° , respectively.

Plots of the PT at every frame for each simulation are supplied in Section S4.1 and Figure 4.9. In the following, we note only a few of the most salient features.

Methanococcus jannaschii is a hyperthermophilic archaeon that contains an Sm ortholog that is more similar to Hfq than to other SmAPs[149]. This particular structure (PDB ID 2QTX), is marked by a high *roll* value of 17° . The Sm-like structure of putative cyanophage origin (PDB ID 3BY7) departs drastically from the reference structure, with a *roll* of -45° and a dz of -10.8 Å! Despite such deviation, the characteristic β -sheet between the two monomers is preserved for the entire simulation. Further, the strongly-bent β -sheet is preserved in both monomers, though somewhat distorted.

ORF-137 from *Pyrobaculum* spherical virus contains two tandem Sm domains separated by a short linker; three of these proteins assemble into a ring. Since there are Sm-Sm contacts within the monomers as well as between the monomers, we have simulated both interfaces. First, we simulated a single monomer and calculated the PT by aligning the C-terminal Sm domain to the reference structure and monitoring the motion of the N-terminal Sm domain. (This is the same convention used in Figure 4.4; with the C-terminal domain serving as chain A and the N-terminal domain serving as chain B) Second, we have used the N-terminal domain from one monomer and the C-terminal domain from its neighbor in the ring to form a dimer with no linker. These two systems display very different dynamics. The single-monomer structure (identified as 2X4J.cternter in the Supporting Information) shows small perturbations from its crystal structure. In particular, the dz component of the PT is $+2.3$ Å, showing that the β -sheet is compressed, with the individual β -strands closer together than in 4PNO. The second structure (2X4J.ntercter) shows little dz relative to 4PNO, but every component of the PT shows more variation than 2X4J.cternter. This result suggests that the two distinct Sm-Sm interfaces in this tandem structure have different dynamics: one interface is relatively rigid, while the other is much more flexible.

In general, there is no clear trend in PT values that separates SmAPs from eukaryotic Sm or bacterial Hfq, that separates hexamers from heptamers from octamers, or that distinguishes homomeric from heteromeric assemblies. It would appear, therefore, that there is no single parameter, degree of freedom, or structural determinant that governs the preferred oligomeric state of a given Sm protein. Notably, these results are corroborated by attempts to

Table 4.1: **Simulated Sm protein systems**

Protein	Nat. state	Sim. state	PDB	Sim. time	POS	SI figure
<i>Eco</i> Hfq	6	2	4PNO[188]	200 ns	5.506	S4.12
<i>Eco</i> Hfq	6	6	4PNO[188]	50 ns	5.578	S4.13
<i>Eco</i> Hfq	6	4	4PNO[188]	200 ns	6.196	S4.14
<i>Ssp</i> Hfq	6	2	3HFO[19]	200 ns	6.071	S4.11
<i>Mja</i> Hfq	6	2	2QTX[149]	200 ns	6.603	S4.6
<i>Mth</i> SmAP	7	2	1I81[45]	200 ns	7.177	S4.2
<i>Pae</i> SmAP1	7	2	1I8F[142]	200 ns	6.006	S4.3
<i>Pae</i> SmAP1	7	7	1I8F[142]	50 ns	6.232	S4.4
<i>Pae</i> SmAP2	8	2	N/A	200 ns	7.107	S4.22
<i>Pae</i> SmAP2	8	8	N/A	50 ns	8.576	S4.23
<i>Afu</i> SmAP2	6	2	1LJO[215]	200 ns	5.858	S4.5
<i>Afu</i> SmAP1	7	2	1I4K[216]	200 ns	6.741	S4.1
<i>Sce</i> Lsm3	8	2	3BW1[147]	200 ns	7.490	S4.9
<i>Hsa</i> SmB, SmD1	7	2	4WZJ[110]	200 ns	N/A	S4.15
<i>Hsa</i> SmD1, SmD2	7	2	4WZJ[110]	200 ns	N/A	S4.16
<i>Hsa</i> SmD2, SmF	7	2	4WZJ[110]	200 ns	N/A	S4.17
<i>Hsa</i> SmF, SmE	7	2	4WZJ[110]	200 ns	N/A	S4.20
<i>Hsa</i> SmE, SmG	7	2	4WZJ[110]	200 ns	N/A	S4.19
<i>Hsa</i> SmG, SmD3	7	2	4WZJ[110]	200 ns	N/A	S4.21
<i>Hsa</i> SmD3, SmB	7	2	4WZJ[110]	200 ns	N/A	S4.18
Unknown	5	2	3BY7[53]	200 ns	4.688	S4.10
PSV ORF-137	3	1	2X4J[153]	200 ns	N/A	S4.8
PSV ORF-137	3	2/2	2X4J[153]	200 ns	N/A	S4.7

Abbreviations: *Eco*, *Escherichia coli*; *Ssp*, *Synechocystis sp.*; *Mja*, *Methanococcus jannaschii*; *Mth*, *Methanobacterium thermoautotrophicus*; *Pae*, *Pyrobaculum aerophilum*; *Afu*, *Archaeoglobus fulgidus*; *Sce*, *Saccharomyces cerevisiae*; *Hsa*, *Homo sapiens*; PSV, *Pyrobaculum spherical virus*. PSV ORF-137 contains three monomers, each of which contains two Sm domains. The first simulation of PSV ORF-137 just included one of these monomers, while the second used the C-terminal Sm region from one monomer with the N-terminal region of another to make a dimer. For homomeric systems, the average POS value is shown in the rightmost column.

perform principal component analysis (PCA) on the components of the PT. Figure S4.25 shows that the principal components of the PT are not consistent between different systems, and therefore no simple combined motion (such as a subduction motion comprised of positive *roll* combined with negative *dx*) can describe the movement between the Sm proteins. Instead, this information seems to be encoded more globally, at a finer structural or dynamical level.

4.3.5 The POS shows that oligomeric plasticity is inherent to Sm proteins.

In most cases, the POS computed from a dimer trajectory is similar to that seen in the crystal structure. Data for all simulations are presented in S4.1 and are summarized in Figure 4.10. Several systems are notably unusual. First, *Pae* SmAP1 (PDB 1I8F) has a POS value of 6.0 in the dimer, though it crystallized as a heptamer. More striking is the POS found by extracting a single dimer from a complete ring simulation. (The particular dimer used is irrelevant as asymmetry in any dimer implies asymmetry in the complete ring.) This dimer has a POS of 6.2, even though it was simulated inside a heptameric ring! This result strongly suggests that the *Pae* SmAP1 heptamer ring is highly asymmetric in solution. (Note that the *Pae* SmAP1 structure was refined without the imposition of NCS restraints.) This may reflect one monomer being “special” in some way, as is seen in the ϕ 29 packaging motor[35]; perhaps this asymmetry plays some role in RNA cycling. Or perhaps *Pae* SmAP1 can adopt several oligomeric states, with the heptamer being most stable in the crystallization conditions used to solve the structure, but the hexamer being preferred in our simulations. This mirrors the behavior known for *Afu* SmAP2, which is able to adopt either a hexameric or heptameric state in response to its environment[100].

Pae SmAP2 shows a similar asymmetry. This SmAP was crystallized as an octamer and the octameric state is supported by biophysical data (Randolph & Mura, unpublished data). Our simulation shows a POS of 7.1, suggesting that this protein could also form a stable heptameric ring. Further, a dimer from the complete ring shows a POS of

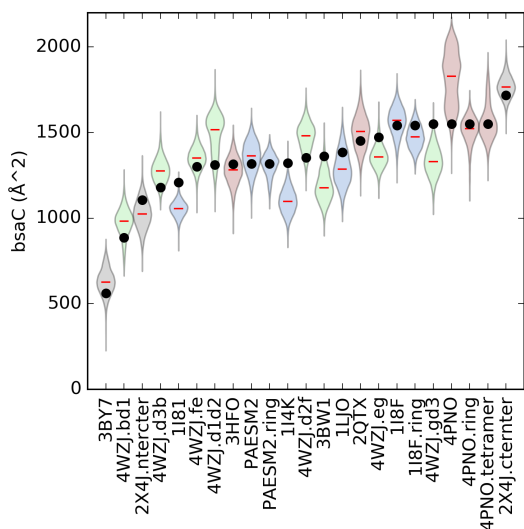


Figure 4.7: **BSA in the Sm-Sm interface.** For each system, we have calculated the amount of buried surface area between the two monomers contributed by residues in the Sm core (based on alignment with 4PNO). Black circles indicate the value in the crystal structure, red lines correspond to the median values, and the colors of each distribution correspond to the phylogenetic tree in Figure 4.1. Most structures remain compatible with their crystallographic values, indicating that the interface does not drastically change during the simulation. 3BY7, a pentamer of putative cyanophage origin, displays a much weaker interface than all other Sm proteins, likely due to the lack of an N-terminal α -helix in that system. 2X4J, containing two tandem Sm-Sm domains, shows that the two Sm interfaces are quite different. The intramonomer Sm-Sm interface has a much lower BSA than the intermonomer interface. 4PNO, from *Eco*, displays a strengthening of the interface over the course of our simulations, in agreement with our PT and POS values for that system.

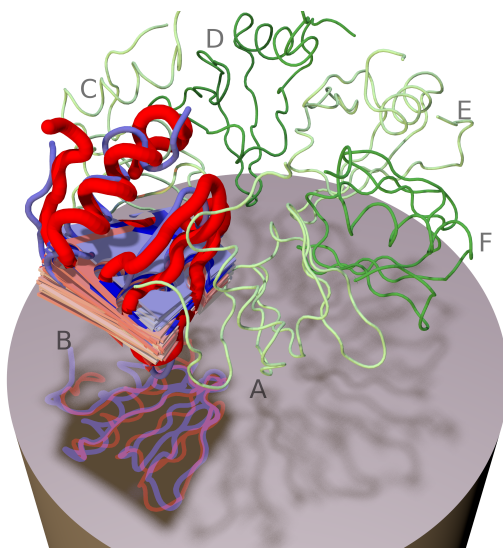


Figure 4.8: **The PT of 4PNO.** At every 2 ns of the simulation of 4PNO, a blade is drawn, translated and rotated according to the PT value at that frame. Green and blue backbone traces show the position of atoms in the 4PNO crystal structure, and the red trace shows the position of chain B (after chain A has been aligned to the reference) at the beginning of the production run. A reference blade, based on the crystal structure, is transformed by the PT at each time step, and the blades are colored according to simulation timestep (blue blades are early in the simulation, red blades occur later). The reference blade, corresponding to *Eco*, is unfortunately obscured by all of the other blades.

8.5, again implying that the complete ring is asymmetric in solution.

The asymmetric ring is also seen in *Eco* Hfq, where a dimer from our simulation of a complete ring shows POS values of 5.72. While not as extreme as the asymmetry seen in *Pae* SmAP1, this asymmetry arises from a perfectly-symmetric starting structure (4PNO was determined in a space group with a 6-fold symmetry axis).

Mja Hfq, an Hfq in a non-bacterial organism, straddles the hexamer-heptamer divide, with a POS of 6.6. This is, by far, the highest POS for an Hfq, and is closer to *Afu* SmAP1 (POS = 6.7) than it is to any other Hfq. The interesting octamer of *Sce* LSm3 adopts a POS of 7.5, far higher than any other Sm protein. An *in vivo* role of an octameric LSm3 assembly is yet unknown, but the dimer at least is accommodating of this state.

Several of the Sm proteins studied here are derived from thermophiles, and it has been found that thermophilic proteins exhibit more small-scale fluctuations around their average structure but reduced large-scale motions of protein domains[131, 238]. Sm proteins, by our analysis, do not clearly demonstrate this trend. The components of the PT do not track with the thermophilicity of the organism containing the protein. For instance, we consider the standard deviations of dx , dy , and dz values. 3HFO, from the mesophilic bacterium *Ssp*, has values that are more similar to those in 1I8F, from the thermophilic archaeon *Pae* than they are to the values from 4PNO, the mesophilic *Eco*. Further, the RMSD data in Section S4.1 show no trend related to thermophilicity. This may be a result of the general stability of Sm proteins against thermal and chemical denaturation.

4.3.6 The flexibility of Sm proteins can be probed experimentally.

While we are confident in our findings regarding the oligomeric plasticity of Sm proteins, all computational studies must be seen as hypothesis-generating. We propose three experiments to validate our findings. First, we propose creating two mutants of *Eco* Hfq: one with mutations such as V62R and F11H, and the other with mutations such

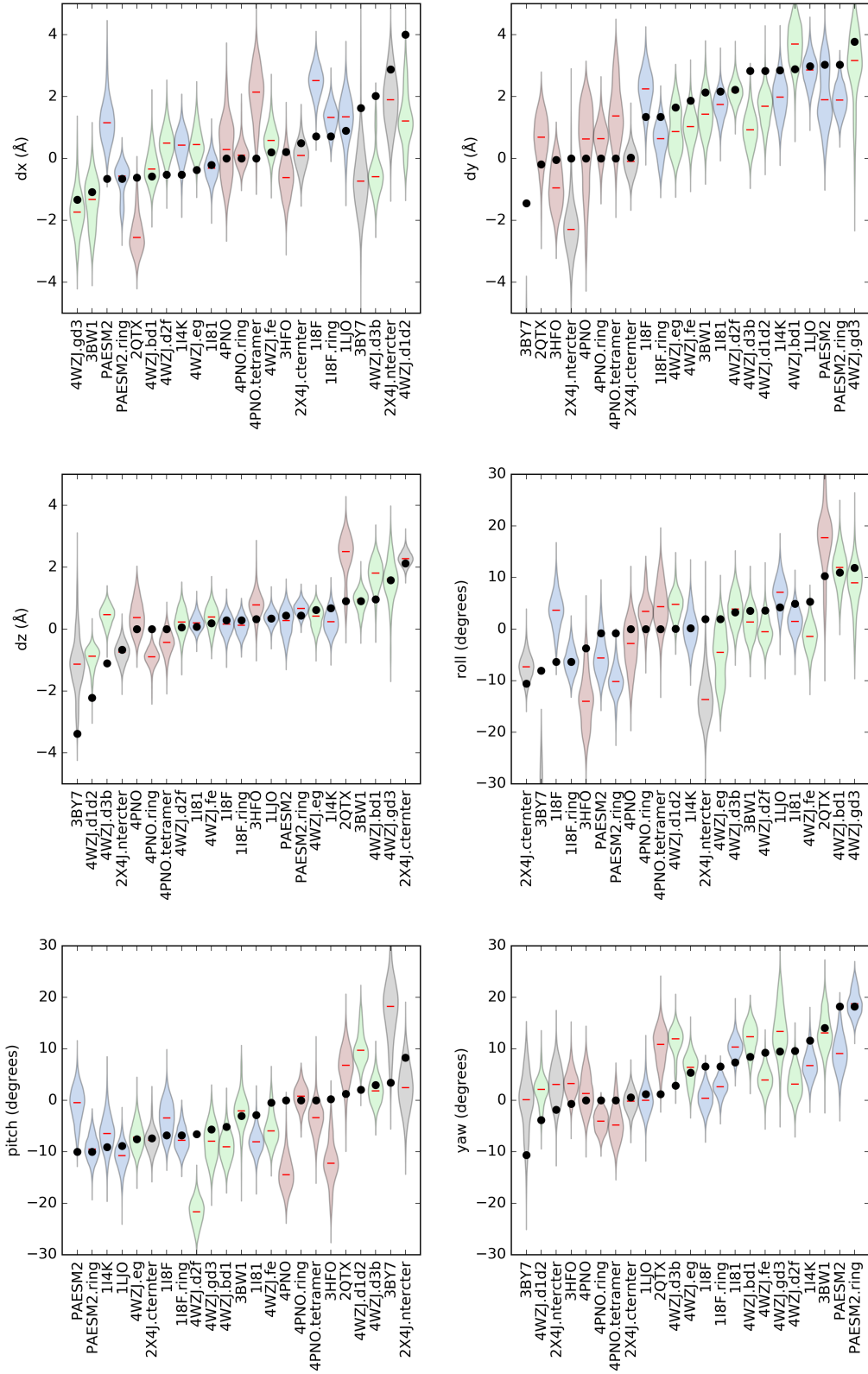


Figure 4.9: **Distribution of PT values for all systems.** For each component of the PT, we show the distribution of values adopted by each system. Colors correspond to the colors used in the phylogenetic tree in Figure 4.1. Black circles indicate the value of that particular component in the crystal structure for that system, and red lines indicate the median value.

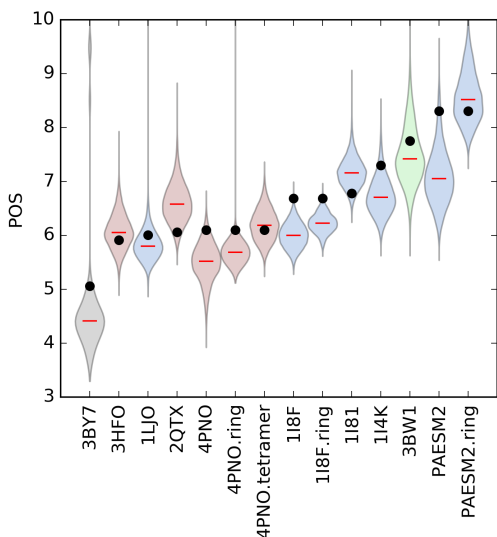


Figure 4.10: **POS distributions for all simulated systems** Colors correspond to the phylogenetic tree in Figure 4.1. In most cases, the POS distribution does not deviate more than one unit from its crystallographic value. Black circles indicate the value of the POS in each crystal structure, and red lines indicate the median value of the POS. On occasion, the POS algorithm is unable to identify the first minimum and instead reports an “overtone” POS, corresponding to $M^{2k}A$. As is clear upon visual inspection, the median value of the POS is not significantly affected by the presence of these overtone POS numbers.

as L45Q and M53K. The first mutant will be unable to oligomerize on the side nearest its α -helix, while the second will be unable to oligomerize on its other edge. As a result, when these proteins are combined, they will assemble into dimers but no more. This system can then be crystallized and solved using X-ray diffraction. We expect to see the system adopt a structure compatible with our simulation data of an *Eco* Hfq dimer. Second, we propose creating chimeric Sm proteins to determine which regions of the protein are responsible for determining the oligomeric state. By creating a chimera with, for example, residues 1 to 40 from *Eco* Hfq and residues 48 to 90 of *Mth* SmAP, we will create a new Sm protein which could assemble is either a hexamer or heptamer. By using different sections of different Sm proteins, we will gain a better understanding of the role of each part of the Sm fold in oligomerization. Third, we propose a crystallographic approach to searching for ring asymmetry. While we cannot currently probe the asymmetry of an Sm ring in solution, we can crystallize an Hfq with a four-nucleotide RNA molecule, U_4 . If this forms a crystal in which the RNA density is consistently located in one place (rather than spread around the ring), it will be possible to quantify the asymmetry of the ring in this situation using the tools we have presented here.

4.4 Conclusions

Taken as a whole, our simulation-based data suggest that Sm proteins are far more flexible than is typically assumed[145]. The behavior of the PT suggests that Sm dimers deform in different ways in different dimers, rather than simply having one particular type of flexibility. This would imply, then, that Sm rings can be flexible in solution, as there are many permissible states that can be sampled by the monomers in a ring, at even just ambient temperatures. Our POS data show that Sm dimers typically adopt conformations that are compatible with experimentally-observed oligomeric states, but there is substantial flexibility that supports the notion that some Sm proteins are able to adopt multiple oligomeric states.

4.5 Materials and Methods

4.5.1 System preparation

Crystal structures were downloaded from the PDB; structures containing less than a complete ring in the asymmetric unit used the biological assembly file from the PDB. In crystal structures containing more than one complete Sm ring, the ring with the most atoms was used. All missing residues in crystal structures were placed with Modeller 9.15[180], using the refine.fast MD level, which performs conjugate gradients optimization followed by simulated annealing from 150 to 1000 K, then down to 300 K. The total simulation time in Modeller was 2 ps for each model, and only those residues added by Modeller were allowed to move during the optimization. In cases where the output from Modeller clashed with other areas of the protein, the torsion angles of the terminal residues in the crystal structures were slightly perturbed until the clash was relieved. Several structures had long extensions that made the system impractically large for explicit-solvent MD simulations. These systems were the SmB-SmD1, SmD1-SmD2, SmD3-SmB, and SmG-SmD3 dimers of 4WZJ. These extended systems were minimized for 10,000 steps then simulated for 10 ns in implicit solvent (GBIS in NAMD[166] with 0.15M ion concentration) with the CHARMM36 force field[17]. During these implicit-solvent simulations, atoms that were present in the initial crystal structure (that

is, those not added by Modeller) were restrained with harmonic restraints of $50 \text{ kcal/mol/\AA}^2$. All other simulation parameters were as used in the production runs. The final frame of these implicit-solvent runs was used as the starting structure for the explicit-solvent simulations described below.

The complete structures were solvated in TIP3P water using the Solvate program[82] and the resulting solvated systems were truncated into a truncated octahedron of sufficient size that the nearest image-image distance is 30 Å. The systems were ionized to 0.15 M Na⁺ and the charge was neutralized with Cl⁻ using the LeAP program from AmberTools2015[26]. All simulations were performed with the ff14SB force field[122], which has been shown to describe interprotein contacts well.

4.5.2 Simulations

All simulations were performed using NAMD 2.9[166]. During the simulations, all bonds involving hydrogen were kept rigid, the vdW cutoff distance was at least 11 Å, and periodic boundary conditions based on the truncated octahedron mentioned above were used. PME electrostatics were used with a grid spacing of at least 1/Å. Pressure and temperature were maintained at 300 K and 1 atm using the Langevin thermostat and piston in NAMD. The integration timestep for all simulations was 2.0 fs.

For all systems, protein backbone atoms were initially restrained with $50 \text{ kcal/mol/\AA}^2$. The systems were minimized for at least 500 steps, then simulated for 1 ns to equilibrate the water and ions. The system was then re-minimized for 1000 steps, then gradually heated from 0 K to 300 K in increments of 10 K, with 1 ps of dynamics at each temperature. The restraints were then repeatedly cut in half with 1 ps of dynamics after every halving. Once the restraints were under $0.01 \text{ kcal/mol/\AA}^2$, they were released completely.

The minimized and thermalized systems were then equilibrated for 10 ns of unrestrained dynamics, and production runs (typically 200 ns) were conducted starting with the final frame of the 10 ns equilibration phase.

All scripts used in this work, as well as dehydrated trajectories, are available upon request.

4.5.3 Analysis

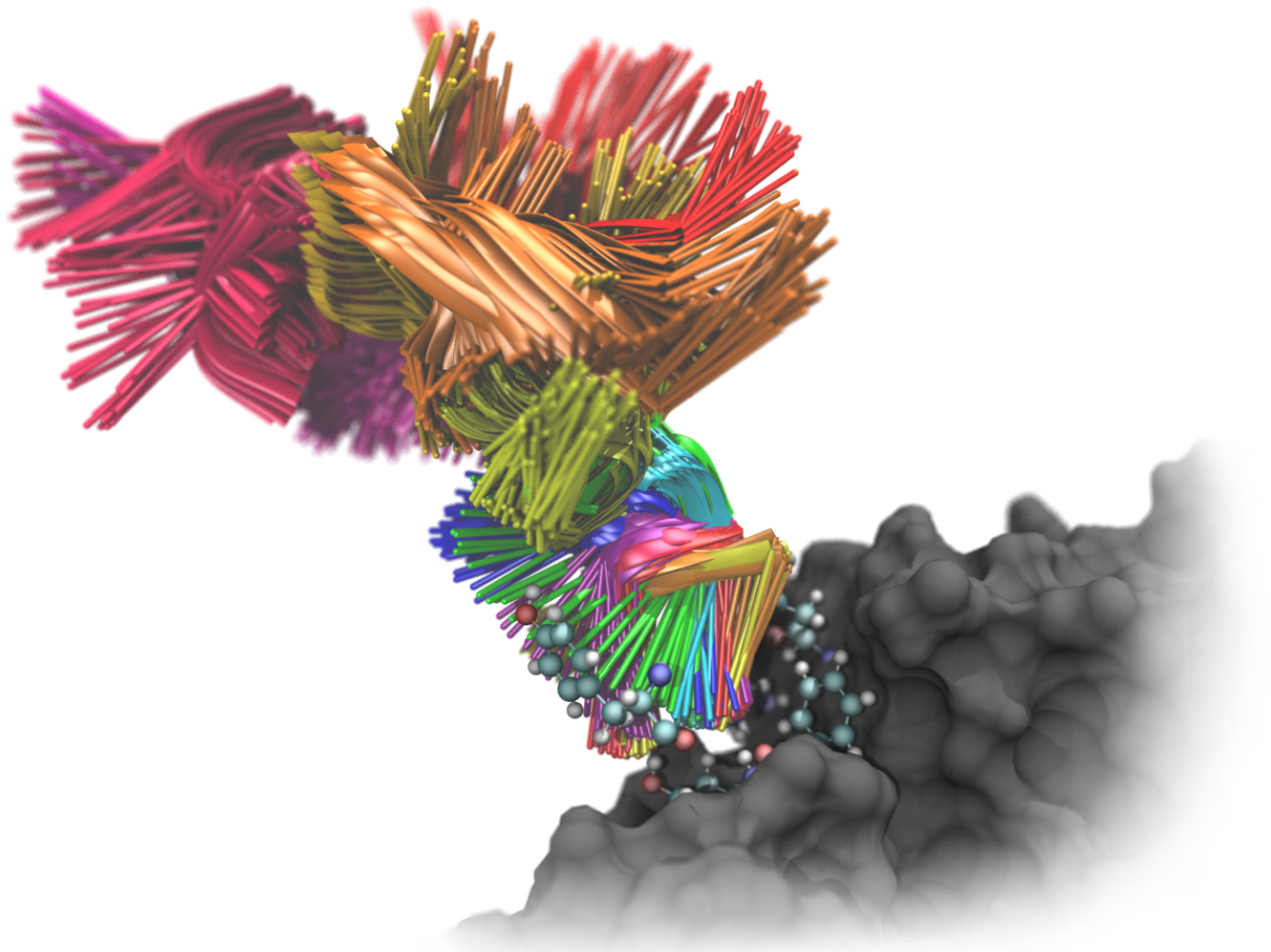
The definitions of the various structural and geometric quantities that we used are described in the Results section. RMSD and contact values were calculated using VMD[93]. PT and POS values were calculated using in-house code written in Python[226] and the D programming language[6]. Figures were created using geometry from VMD and POV-ray 3.7[164]. To define the Sm 'core' domain, we performed a structural alignment against 4PNO using DeepAlign[234]. Full scripts used for the analysis are available upon request.

4.6 Acknowledgements

We thank P. Randolph, K. Stanek, and S. Coupe for helpful discussions, and K. Holcomb and A. Munroe from ARCS at UVa for computing resources. We also thank the members of the vmd-l, namd-l, and digitalmars.d communities for technical advice. This work was supported by the University of Virginia, the National Science Foundation (MCB-1350957), and the Jefferson Scholars Foundation.

Chapter 5

A Simulation-based Approach to the Dynamical Basis of Hfq-RNA Interactions



5.1 Abstract

The bacterial Sm protein, known as Hfq, acts as a generic RNA chaperone that facilitates interactions between two RNA strands, typically a noncoding small RNA (sRNA) and a regulatory target (e.g., a messenger RNA (mRNA)). Many sRNAs play key roles in post-transcriptional regulation, including protein translational control and RNA decay pathways. While many crystal structures have provided static snapshots of Hfq and, more recently, Hfq-RNA complexes, and biochemical studies have supplied valuable information about Hfq function, the physicochemical behavior of the interaction between RNA and Hfq remains unexplored. Hfq self-assembles into hexameric rings in the absence of RNA, with two distinct RNA-binding regions. One side of the ring (the *distal* face) binds U-rich RNAs, while the other (*proximal*) face binds A-rich RNAs. Our recent crystal structures of an *Aquifex aeolicus* Hfq, with and without RNA, reveal, in addition to the proximal and distal sites, a conserved lateral RNA binding site on the periphery of the ring. This lateral site binds U-rich RNA with lower affinity than the proximal site. To see how RNA interacts with Hfq, we are pursuing an extensive suite of MD simulations. By using positional restraints to drive two nucleotides of RNA toward the lateral site, our simulations start with a physically plausible, partially-bound state. We then simulated the unconstrained system to examine RNA interactions with the neighboring protein surface. To gain insight into the dynamic role of the lateral site in RNA annealing, we are now pursuing a battery of steered molecular dynamics (SMD) simulations, guided by specific questions such as: Can RNA simultaneously bind both the lateral and distal (or lateral and proximal) sites? How stable and persistent (thermodynamically and kinetically) are RNA interactions with the lateral site? These simulations will illuminate, in atomic detail, a key mechanistic step in Hfq-mediated RNA annealing.

5.2 Introduction

The roles of Hfq in RNA processing are many, but a unifying feature of all of Hfq’s known functionality stems from Hfq’s affinity to bind U-rich RNA sequences on one face of the protein, and A-rich RNA sequences on the other face[184, 187]. To give several examples, an RNA that is susceptible to degradation could be protected by binding to Hfq[231]. Alternatively, an RNA with a secondary structure that prevents degradation, such as a stem-loop hairpin, could bind to Hfq and adopt a new secondary structure that could be attacked by an RNase[231]. In a more involved example, Hfq could bind to two RNA molecules simultaneously and facilitate their interaction; such a chaperoning function is particularly crucial if the two RNAs have only partial base-pair complementarity[145]. The work in this chapter is driven by a desire to understand this third case, called annealing. For the sake of clarity, this introduction will briefly review the modes of Hfq-RNA interaction, but will focus on only a subset of this third case called “class I” RNA annealing. The results and discussion will implicitly assume class I annealing throughout, though the results are anticipated to be applicable to any other case of Hfq-RNA interaction.

5.2.1 Hfq’s Role in Annealing sRNAs and mRNAs

Recently, Schu et al. have shown that, *in vivo*, there are two separate classes of Hfq-mediated sRNA-mRNA interactions[187]. The process of class I annealing is shown in Figure 5.1. Class I annealing requires three components. First, an mRNA that is a target of regulation that has adopted a secondary structure that occludes the ribosome binding site (RBS). This mRNA must contain an AAN repeat in its sequence, where N is any nucleotide[23]. Second, an sRNA with a complementary sequence to some region of the mRNA. The sRNA must contain a U₆ sequence and a UA-rich region. Third, Hfq. The proximal face of Hfq binds strongly to U-rich RNA regions, so the sRNA will bind at this site. The lateral surface of Hfq binds to UA-rich regions, so the UA-rich region of the sRNA will bind there. The distal face of Hfq binds AAN repeat motifs, so the mRNA will bind there. Hfq serves a dual purpose in this scheme. First, Hfq brings the two RNA molecules close together, thereby increasing the chance of contact between the two RNA molecules and decreasing the entropic cost associated with the formation of the sRNA-mRNA complex. Second, by binding to nucleotides in RNA, Hfq allows the RNAs to adopt secondary structures that would be unfavorable in solution[162]. These new RNA secondary structures may be better-suited for annealing. Class II annealing is similar, except that the sRNA contains a U₆ sequence and an AAN motif, while the mRNA contains a UA-rich region. In class II, the sRNA binds to the proximal and distal sites, while the mRNA binds to the rim[187].

sRNA-mRNA annealing is not limited to enhancing translation of the mRNA. Lenz et al. showed that an Hfq-sRNA complex serves as a silencer for quorum sensing mRNAs; the sRNA-mRNA complex in this case is more susceptible to degradation[109]. The benefit of using Hfq to silence certain genes is its sensitivity, as explained by Lenz et al.:

“As base pairing of an sRNA with its target message is known to promote degradation of both the sRNA and the message, this “mutual destruction” provides an elegant mechanism for ultrasensitivity. Specifically [...] if the rate of synthesis of a particular sRNA exceeds the rate of synthesis of its target message, even if

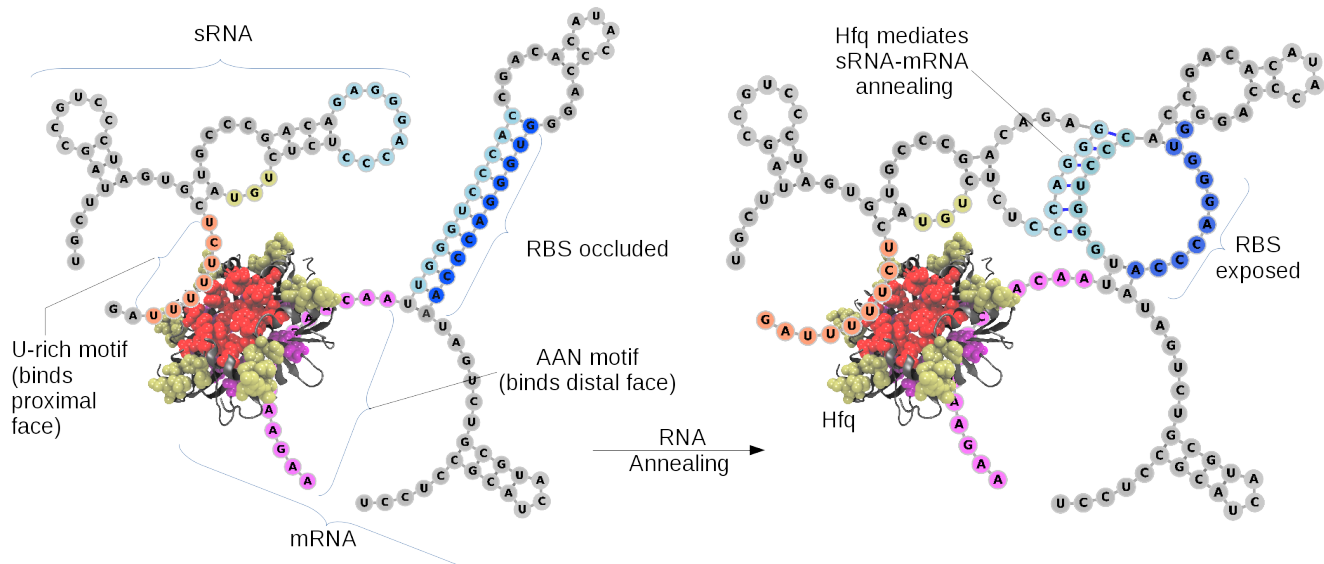


Figure 5.1: Hfq’s role in Class I sRNA-mRNA annealing. An mRNA may adopt a secondary structure that occludes the RBS (left). An sRNA with some sequence complementarity to the RBS would disrupt the base-pairing that occludes the RBS, thus allowing the mRNA to be translated. Hfq facilitates such RNA interactions. At left, a U-rich region (orange) of an sRNA is bound to the proximal face of Hfq (red). A nearby U-rich patch in the sRNA (yellow) interacts with the lateral site of the Hfq (tan). On the distal face of Hfq (pink), an mRNA has bound in a region containing an AAN repeat (pink). The RBS (dark blue) is paired within an internal hairpin in the mRNA (light blue), hindering translation on the mRNA. At right, the mRNA and sRNA have partially annealed, releasing the RBS for binding by the ribosome and initiation of translation. (All sequences in this figure are fictional.)

only slightly, then the sRNA can accumulate in the cell, and target message levels can be reduced to very low levels. In contrast, if the rate of synthesis of a particular target message exceeds that of its regulatory sRNA, then the message can accumulate“[109].

Since a great deal of work is necessary before the complete mechanism of Hfq-mediated RNA annealing will be available, we are focusing in this work on just the role of the lateral site of Hfq and its interaction with RNA. While crystal structures have provided static snapshots of this interaction, we currently know very little about the dynamics of the interaction. We are using molecular dynamics (MD) simulations to fill this gap in our understanding.

5.2.2 MD simulations of RNA

RNA molecules are highly charged, often highly flexible, and interact with polyvalent metal atoms such as Mg^{2+} . The high charge densities in these systems necessitate an accurate model of electrostatic interactions; the usual Coulombic approximation (Equation (1.12)) struggles to accurately predict this[197]. Additionally, the intrinsic flexibility of RNA creates challenges that are similar to those found in intrinsically disordered protein (IDP) studies. Long simulations, on the order of 0.5 ms, are needed in order to adequately sample conformational space for even small tetraloops[227]. Polyvalent metal atoms are critically important to RNA structure and dynamics, but current force fields struggle to reproduce the behavior of these metals[108]. In short, “it is universally acknowledged that current RNA force fields are inadequate”[227]. In this study, we are primarily interested in the dynamical interaction between a short RNA molecule and a protein surface. While we cannot currently measure the accuracy of our simulation, we have identified several phenomena that can be explained in terms of fundamental physicochemical principles and what is known from the literature about the structure and properties of Hfq-RNA systems. As with all MD simulations, the data presented here should be compared with data from experiments such as those we suggest at the end of this chapter.

This chapter presents the first two stages of our three-pronged approach to investigate the interaction of U-rich RNA with the lateral site of *Aquifex aeolicus* (*Aae*) Hfq. In the first part, we use constraints to gently force a free U_6 strand to bind to the Hfq rim in a physically plausible way. The second part consists of simulating this bound structure long enough to observe the behavior of RNA as it explores the surface of Hfq. The third part, which will be discussed briefly at the end of this chapter, is to pull RNA off of the Hfq ring in a variety of different (systematically-sampled) directions, to understand the possible mechanisms for RNA dissociation. We hope that the detailed interactions described herein will prove useful in developing a complete structural and dynamical model of

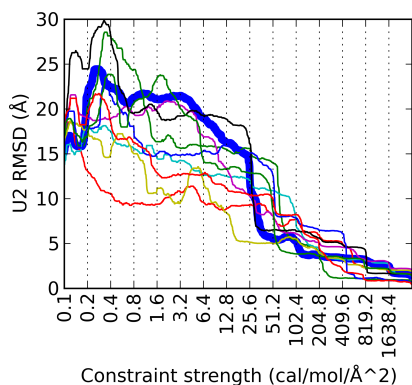


Figure 5.2: **During the constrained insertion step**, RNA entered the binding site and adopted a conformation that was stable over a four-fold change in constraint strength. This plot shows the RMSD of the U_2 that was forced into the binding pocket, relative to the crystal structure for those nucleotides. Dotted vertical lines show the moments when constraint strength was doubled. The ten traces correspond to the ten starting structures generated by Barnacle. The thick blue trace highlights the simulation that was used for the unrestrained dynamics; it ended with the highest buried surface area. Note that the constraint strengths are indicated in $\text{cal/mol}/\text{\AA}^2$ rather than the usual $\text{kcal/mol}/\text{\AA}^2$ since they are particularly weak.

Hfq-mediated RNA annealing.

5.3 Methods

The starting structure of the Hfq hexamer was taken from [199], and the starting structure of the RNA was generated by Barnacle[69] by taking the most likely (according to the energy reported by Barnacle) of 50 randomly-generated structures. All simulations were performed using NAMD 2.9[166]. The ff14SB force field[26] was used for protein atoms, RNA.ROC[10] for RNA, and TIP3P for water. Sodium and chloride ions were included at 150 mM (for implicit-solvent simulations, a 300 mM ion concentration was used). The system was maintained at 310 K using a Langevin thermostat. A nonbonded cutoff of 12 Å was used with an 11 Å switching distance.

5.3.1 Constrained insertion to generate an Hfq- U_6 model

Our crystal structure of Hfq contains only U_2 bound to the lateral site. To obtain a structure including U_6 that is physically plausible, we placed U_6 about 10 Å away from the surface of Hfq. We constrained the position of the last two nucleotides (in 5' to 3' order) of the U_6 molecule to the crystallographic positions. These constraints started very weak ($0.00001 \text{ kcal/mol}/\text{\AA}^2$). Additionally, we constrained all protein heavy atoms to their crystallographic positions with restraints that were 100 times stronger than the constraints on the U_2 . The constraint strength was doubled 15 times over a 1.5 ns simulation in implicit solvent using an integrator step size of 1 fs. We performed ten constrained insertion simulations, using ten randomly-generated RNA structures from Barnacle.

5.3.2 Unrestrained dynamics of the Hfq-RNA complex

For each of the constrained insertion runs, we evaluated the buried surface area between Hfq and RNA and chose the system with the greatest such value for further simulations. This system was solvated in a truncated octahedron of TIP3P water molecules, with a 10 Å padding between the solute and the nearest face of the truncated octahedron. Constraints of $5 \text{ kcal/mol}/\text{\AA}^2$ were applied to all protein and RNA atoms. These constraints were halved every 50 ps over a 350-ps simulation, and the system was then simulated for 10 ns without any constraints. This equilibration run was followed by a 365-ns production, with coordinates saved every picosecond. The integration timestep was 2 fs, and all bonds to hydrogen were kept rigid. Long-range electrostatics were evaluated using particle mesh Ewald summation with a grid spacing of at most $1/\text{\AA}$ [52].

5.4 Results

5.4.1 Constrained insertion suggests bound conformations

The constrained insertion step was necessary because the crystal structure of *Aae* Hfq contains only two (complete) nucleotides and we wished to simulate a longer RNA strand. While it is possible to graft nucleotides directly onto the crystal structure, there is no way to know if the result is physically plausible - the nucleotides could be inserted in a conformation that is not easily reachable, or the nucleotides could even intersect protein atoms. Our approach was to model the insertion process itself, albeit in a somewhat heavy-handed fashion in order to ensure that the final structure of our complex was physically plausible.

In a representative example of such trajectories, after bringing the RNA near the binding pocket using weak constraints ($0.001 \text{ kcal/mol}/\text{\AA}^2$), the 3' nucleotide “jumps” into a bound conformation that is similar to the crystal

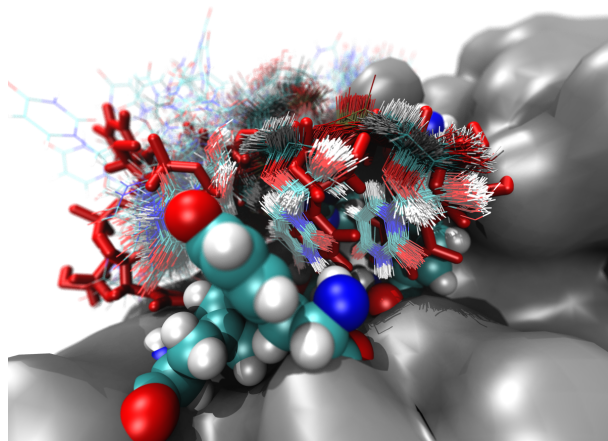


Figure 5.3: **Motion in the binding pocket over 300 ns.** Thin sticks show the conformations adopted by the RNA during 300 ns of unrestrained dynamics. (The trajectory has been aligned using the binding site as a reference). The thick red sticks show the final pose from the constrained simulation. Previous simulations on this system had suggested that the RNA might exit the binding pocket and explore the surface of the Hfq, but the rigidity of this pocket suggests that RNA detachment is a rare event.

structure. This would suggest that the insertion is driven by the underlying structural dynamics of the system, rather than the constraints. After the constraints are made stronger ($0.01 \text{ kcal/mol/\AA}^2$), the second nucleotide jumps into its binding pocket. The bound conformation persists until the constraints are made strong enough to compete with the forces generated by the force field, at which point the conformation distorts slightly to better match the crystallographic positions (Figure 5.2).

5.4.2 Two nucleotides bind firmly, four others explore the Hfq surface

A key goal of this work was to determine if RNA can dissociate from the lateral site in the course of our simulated trajectories. Previous simulations (data not shown) had suggested that U_2 could escape the binding pocket on the sub-100-ns timescale and explore the surface, but we see no evidence of this tendency in our 365-ns production simulation. Instead, the two nucleotides in the binding pocket quickly move from the crystallographic position (which was enforced by the constraints in the previous simulation) to the pose that the nucleotides “jumped” to at low constraint strength. The pattern of residue contacts adopted by the two 3' nucleotides in the course of our simulation mirrors that found crystallographically: N11, R14, S36, D37, and Y3 interact with RNA as seen in the crystal structure, though K15 frequently detaches from the RNA. The two nucleotides then stay in that pose for the duration of the simulation. This structural relaxation of our Hfq-RNA complex indicates two things. First, it indicates that our physical model nearly, but not exactly, reproduces the bound conformation identified crystallographically. Second, since the constrained insertion was able to reach this conformation with weak restraints, the nucleotides outside the binding pocket did not experience strongly unfavorable conformations en route to the bound state. Figure 5.3 shows the limited movement of the bound nucleotides.

The remaining nucleotides beyond the 3'-terminal U_2 explore the surface of the Hfq rim. Figure 5.4a shows the position of each phosphorus atom in U_6 throughout the simulation. Figure 5.4b shows the extent to which the surface area of RNA is buried by contact with Hfq. Unsurprisingly, the two bases in the binding pocket (B5 and B6) are buried for the entire simulation. The remaining RNA nucleotides make numerous transient contacts during the simulation, as shown in Figure 5.4b. The surface of Hfq interacts with all parts of the nucleotides; various bases, sugars, and phosphates all show intermittent burial against Hfq.

The nucleotides exhibit a strong tendency to associate with the arginines on the rim. An analysis of the average distance between each nucleotide and each Hfq residue reveals that the weakly-conserved basic patch in Hfq plays a key role in orienting RNA on the lateral face[199]. Figure 5.5 shows that, with the exception of the final (3') nucleotide, every nucleotide is closest, on average, to an arginine residue. The residues on this binding face (R14, K15, R17, R33, R35, E45, R66 are the residues comprising the five most frequently-contacted residues for the three 3' bases) are, with the exception of R14 and K15, poorly conserved. This poor conservation is seen even in other extremophiles such as *Thermotoga maritima*, which features the amino acids RVKFREY at the corresponding peptide region. *Escherichia coli* has residues RRRQKEP at the corresponding positions.

5.4.3 A uracil-specific binding site on the rim

At 270 ns in the simulation, the second uracil base enters a binding pocket on the lateral surface (Figure 5.4, cyan mark). This base hydrogen-bonds to three residues: R66* (from another Hfq monomer), E45, and G47, and it stacks with the guanidinium moiety of R33. This binding pocket is not conserved in other bacterial species[199]. Since the pocket forms three hydrogen bonds to the uracil base, it is sequence-specific. R33 is relatively quite mobile during the simulation, and occasionally occludes the binding site. Figure 5.6 shows the motion in the pocket. We clustered

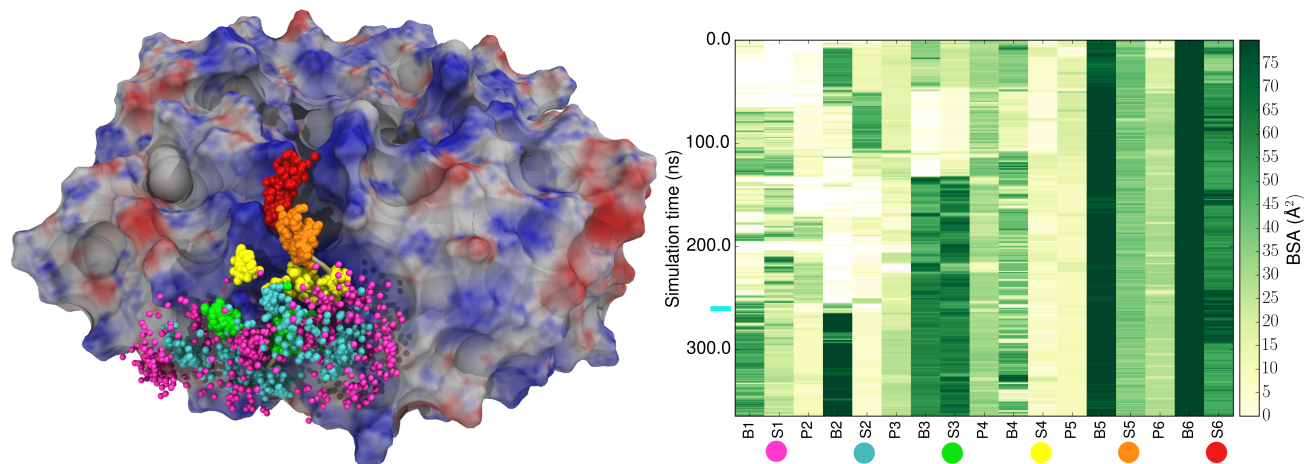


Figure 5.4: **Unrestrained dynamics and RNA behavior on the protein surface.** In the left panel, the protein is colored by its surface electrostatic potential, with blue being more positive. The colored spheres (purple, blue, green, yellow, orange, red in 5' \rightarrow 3' order) show the positions of the 1' carbon atoms in the RNA during the simulation. The right panel shows that nucleotides outside the binding pocket are solvent-exposed. This plot shows buried surface area (BSA = total SA - SASA) for each base (B), sugar (S, colored dot corresponding to colors in the left panel), and phosphate (P) in the system. The two 3' bases are in the binding pocket and therefore completely buried. The behavior of the nucleotides outside this region is consistent with a mechanism where the positively-charged rim attracts the phosphates in the RNA, and this leaves some of the nucleosides solvent-exposed and able to base-pair with mRNA. The burial of base 2 at 270 ns (indicated by a cyan mark) is caused by hydrogen bonds made to R66 (from the next subunit), E45, and G47.

the conformations of the residues in this pocket with a 1 Å root-mean-square deviation (RMSD) cutoff in RMSD using the cluster measurement tool in VMD[93]. The most common state had R33 pointing out of the pocket (blue). An intermediate state (cyan) shows R33 flipping over to occlude the binding pocket (red). From this occluded state, R33 can tilt slightly (green) to accommodate the uracil.

5.5 Future Directions

It is not known if RNA annealing occurs directly on the lateral site, nor is it known how the mRNA and sRNA first come into contact. We also do not know the precise mechanism of RNA attachment and detachment from the lateral site. What we do know is that at least one U-rich RNA strand can both bind to and leave this site. Starting from the final structure of our unconstrained simulation, we are performing a systematic array of SMD simulations, each pulling in a different direction. Our design for these simulations is to constrain the center of mass of the 5' nucleotide to a plane that slowly moves away from the protein surface. This allows lateral movement of the RNA during detachment, which reflects the types of forces present in single-molecule pulling experiments. From this, we hope to unveil the most likely mechanisms of RNA dissociation. For example, do the bases detach independently in a stepwise fashion, or is detachment a more concerted process? Must Hfq deform to allow bases in the pocket to escape? Once the RNA leaves the pocket, does it continue to explore the surface of Hfq, or does it detach completely with little force? Our SMD simulations will provide information that will guide future experimental work on this system, both by suggesting key residues involved in the process and by providing a detailed map of the interactions of RNA with the Hfq surface.

The data from this project suggest several new hypotheses that can be experimentally tested. The role of the basic patch on the *Aae* Hfq rim could be analyzed by mutating *Eco* Hfq to include a similar patch, which we predict would lead to stronger interactions between Hfq and U-rich RNA. Biophysical techniques to quantify the effect of mutations on the function of Hfq have been firmly established[187]. The new U-specific binding site could be probed by titrating Hfq with uridine and measuring, for example, the chemical shift of a key glycine, G47[124]. (G47 binds to the uracil ring through its backbone nitrogen.) For studying the force required to detach RNA from the Hfq surface, single-molecule pulling experiments provide a practical method for assessing the mechanisms predicted by our steered simulations, with the elegant benefit of being the experimental analog of the SMD approach[160]. It is our hope that these results will serve as the basis for an atomically-detailed mechanism for Hfq-mediated RNA annealing processes.

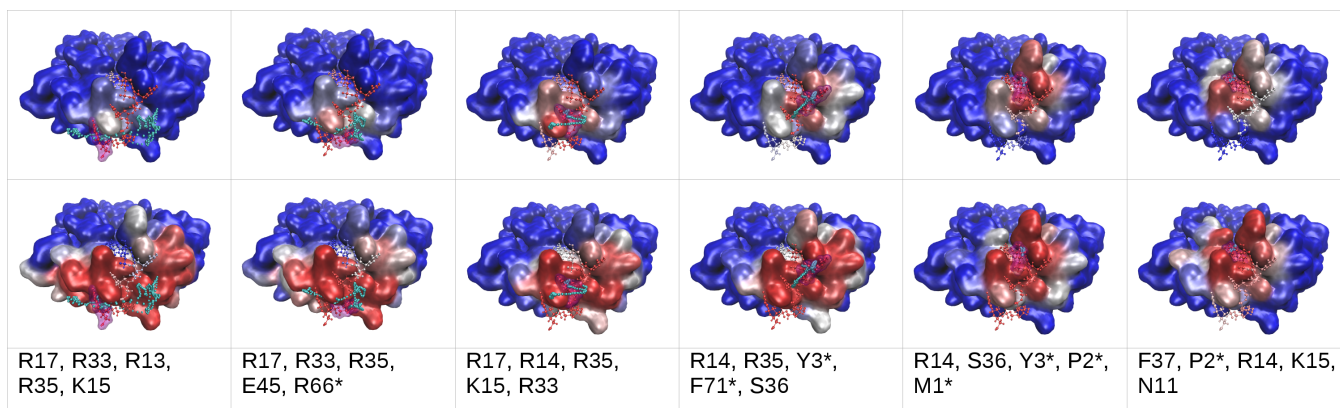


Figure 5.5: **Average distance (top) and nearest distance (bottom) to each residue on the protein surface**, for each nucleotide (left to right, 5' to 3'). Red colors indicate close proximity between the nucleotide and the protein surface, while blue areas indicate that the nucleotide does not contact that area of the Hfq surface. Contact distances are measured between the closest atoms in each nucleotide-residue pair. Cyan spheres show the movement of the carbon atom at the 1' position on each ribose ring throughout the trajectory. The bottom row lists, for each nucleotide, the five residues that are closest, on average, to that nucleotide. An asterisk after a residue indicates that it is from an adjacent Hfq subunit.

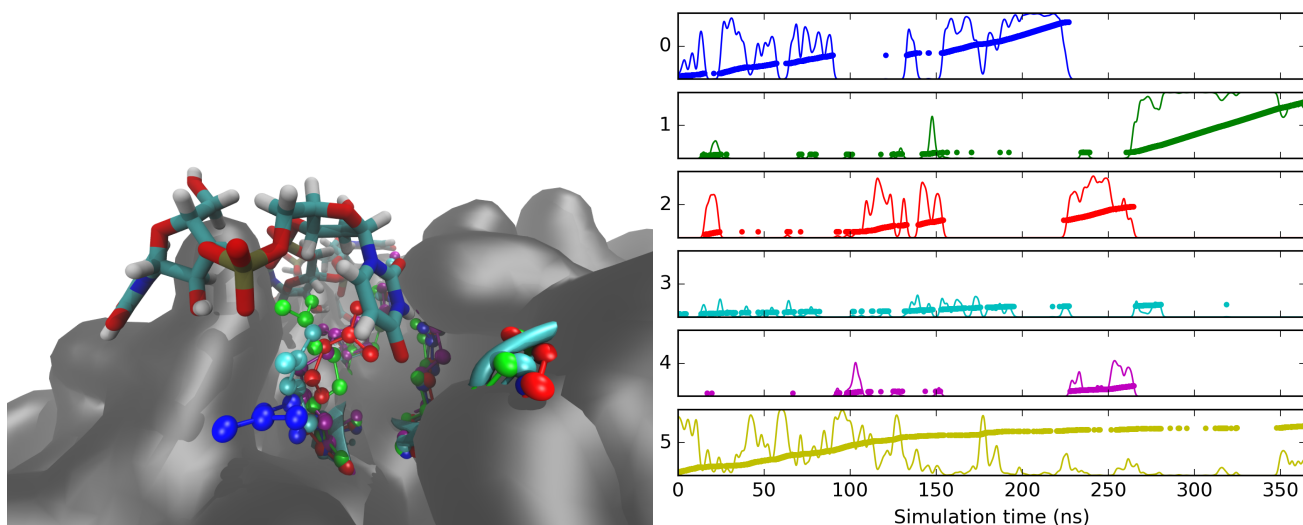


Figure 5.6: **Structural clusters in the uracil-binding pocket.** In the left panel, representative structures of the five structural clusters identified by our analysis are shown, with colors corresponding to the plot on the right. The right panel shows the cluster occupied by the system at every time point (thick line) and the fractional occupancy of each cluster with a small smoothing window (thin line). Each thick line increases in height when the corresponding cluster is occupied; therefore the height of the thick line indicates how frequently the system has been in each cluster. The largest cluster, cluster 0, corresponds to R33 pointing out of the pocket, while cluster 2 is occupied when R33 occludes the binding pocket. Cluster 1, in green, is occupied when the base of U₂ enters the binding pocket. Cluster 5 corresponds to system states that do not fall within any of the other clusters.

Bibliography

- [1] M. J. Abraham and J. E. Gready. Ensuring mixing efficiency of replica-exchange molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 4(7):1119–1128, 2008.
- [2] A. Al-Amoudi and A. S. Frangakis. Structural studies on desmosomes. *Biochemical Society Transactions*, 36(2):181–187, 2008.
- [3] C. Al-Jassar, H. Bikker, M. Overduin, and M. Chidgey. Mechanistic basis of desmosome-targeted diseases. *J. Mol. Biol.*, 425(21):4006–4022, Nov 2013.
- [4] C. Al-Jassar, T. Knowles, M. Jeeves, K. Kami, E. Behr, H. Bikker, M. Overduin, and M. Chidgey. The Nonlinear Structure of the Desmoplakin Plakin Domain and the Effects of Cardiomyopathy-linked Mutations. *J. Mol. Biol.*, 411(5):1049–1061, Sep 2011.
- [5] L. V. Albrecht, L. Zhang, J. Shabanowitz, E. Purevjav, J. A. Towbin, D. F. Hunt, and K. J. Green. GSK3- and PRMT-1-dependent Modifications of Desmoplakin Control Desmoplakin-cytoskeleton Dynamics. *J. Cell Biol.*, 208(5):597–612, Mar 2015.
- [6] A. Alexandrescu. *The D Programming Language*. Pearson Education, 2010.
- [7] M. Amiram, F. G. Quiroz, D. J. Callahan, and A. Chilkoti. A highly parallel method for synthesizing DNA repeats enables the discovery of ‘smart’ protein polymers. *Nature Materials*, 10(2):141–148, 2011.
- [8] H. C. Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of Chemical Physics*, 72(4):2384–2393, 1980.
- [9] E. Arunan, G. R. Desiraju, R. A. Klein, J. Sadlej, S. Scheiner, I. Alkorta, D. C. Clary, R. H. Crabtree, J. J. Dannenberg, P. Hobza, H. G. Kjaergaard, A. C. Legon, B. Mennucci, and D. J. Nesbitt. Definition of the hydrogen bond (IUPAC recommendations 2011). *Pure and Applied Chemistry*, 83(8):1637–1641, 2011.
- [10] A. H. Aytenfisu, A. Spasic, A. Grossfield, H. A. Stern, and D. H. Mathews. Revised RNA dihedral parameters for the Amber force field improve RNA molecular dynamics. *J Chem Theory Comput*, 13(2):900–915, Feb 2017.
- [11] M. A. Balsera, W. Wriggers, Y. Oono, and K. Schulten. Principal component analysis and long time protein dynamics. *The Journal of Physical Chemistry*, 100(7):2567–2572, 1996.
- [12] A. E. Bass-Zubek, R. P. Hobbs, E. V. Amargo, N. J. Garcia, S. N. Hsieh, X. Chen, J. K. Wahl, M. F. Denning, and K. J. Green. Plakophilin 2: a critical scaffold for PKC α that regulates intercellular junction assembly. *J. Cell Biol.*, 181(4):605–613, May 2008.
- [13] K. Beck, I. Hunter, and J. Engel. Structure and function of laminin: anatomy of a multidomain glycoprotein. *FASEB journal*, 4(2):148–60, 1990.
- [14] M. Bedford and S. Clarke. Protein arginine methylation in mammals: Who, what, and why. *Mol. Cell*, 33(1):1–13, Jan 2009.
- [15] M. T. Bedford. Arginine methylation at a glance. *J. Cell. Sci.*, 120(Pt 24):4243–4246, Dec 2007.
- [16] A. Berchanski, D. Segal, and M. Eisenstein. Modeling oligomers with C n or D n symmetry: application to CAPRI target 10. *Proteins*, 60(2):202–206, Aug 2005.
- [17] R. B. Best, X. Zhu, J. Shim, P. E. Lopes, J. Mittal, M. Feig, and A. D. Mackerell. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ , ψ and side-chain χ (1) and χ (2) dihedral angles. *J Chem Theory Comput*, 8(9):3257–3273, Sep 2012.
- [18] E. Beurel, S. F. Grieco, and R. S. Jope. Glycogen synthase kinase-3 (GSK3): Regulation, actions, and diseases. *Pharmacol. Ther.*, 148:114–131, Apr 2015.
- [19] A. Boggild, M. Overgaard, P. Valentin-Hansen, and D. E. Brodersen. Cyanobacteria contain a structural homologue of the Hfq protein with altered RNA-binding properties. *FEBS J.*, 276(14):3904–3915, Jul 2009.
- [20] M. Bonjack-Shterengartz and D. Avnir. The near-symmetry of proteins. *Proteins*, 83(4):722–734, Apr 2015.

- [21] S. Bonnal and J. Valcarcel. Rnatomy of the spliceosome’s heart. *EMBO J.*, 32(21):2785–2787, Oct 2013.
- [22] G. Bouvignies, P. Bernado, S. Meier, K. Cho, S. Grzesiek, R. Bruschweiler, and M. Blackledge. Identification of slow correlated motions in proteins using residual dipolar and hydrogen-bond scalar couplings. *Proc. Natl. Acad. Sci. U.S.A.*, 102(39):13885–13890, Sep 2005.
- [23] R. G. Brennan and T. M. Link. Hfq structure, function and ligand binding. *Curr. Opin. Microbiol.*, 10(2):125–133, Apr 2007.
- [24] L. Brocchieri and S. Karlin. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res.*, 33(10):3390–3400, 2005.
- [25] L. Cai and S. C. Heilshorn. Designing ECM-mimetic materials using protein engineering. *Acta Biomaterialia*, 10(4):1751–1760, 2014.
- [26] D. Case, J. Berryman, R. Betz, D. Cerutti, T. C. III, T. Darden, R. Duke, T. Giese, H. Gohlke, A. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T. Lee, S. LeGrand, P. Li, T. Luchko, R. Luo, B. Madej, K. Merz, G. Monard, P. Needham, H. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, D. Roe, A. Roitberg, R. Salomon-Ferrer, C. Simmerling, W. Smith, J. Swails, R. Walker, J. Wang, R. Wolf, X. Wu, D. York, and P. Kollma. AMBER 15, 2015.
- [27] D. A. Case, T. A. Darden, T. E. Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, R. C. Walker, W. Zhang, K. M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A. W. Goetz, I. Kolossváry, K. F. Wong, F. Paesani, J. Vanicek, R. M. Wolf, J. Liu, X. Wu, S. R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M. J. Hsieh, G. Cui, D. R. Roe, D. H. Mathews, M. G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P. A. Kollman. AMBER 12, 2012. Accessed 2016-05-10.
- [28] T. M. Casey, Z. Liu, J. M. Esquiaqui, N. L. Pirman, E. Milshteyn, and G. E. Fanucci. Continuous wave W- and D-band EPR spectroscopy offer "sweet-spots" for characterizing conformational changes and dynamics in intrinsically disordered proteins. *Biochem. Biophys. Res. Commun.*, 450(1):723–728, Jul 2014.
- [29] D. E. Chandler, F. Penin, K. Schulten, and C. Chipot. The p7 protein of hepatitis C virus forms structurally plastic, minimalist ion channels. *PLoS Comput. Biol.*, 8(9):e1002702, 2012.
- [30] B. Chang, Y. Chen, Y. Zhao, and R. K. Bruick. JMJD6 is a histone arginine demethylase. *Science*, 318(5849):444–447, Oct 2007.
- [31] Y. Chao and J. Vogel. The role of Hfq in bacterial pathogens. *Curr. Opin. Microbiol.*, 13(1):24–33, Feb 2010.
- [32] C. Chen, T. J. Nott, J. Jin, and T. Pawson. Deciphering arginine methylation: Tudor tells the tale. *Nat. Rev. Mol. Cell Biol.*, 12(10):629–642, Oct 2011.
- [33] V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson. MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.*, 66(Pt 1):12–21, Jan 2010.
- [34] Y.-L. Chen and Z.-B. Guan. Bioinspired modular synthesis of elastin-mimic polymers to probe the mechanism of elastin elasticity. *Journal of the American Chemical Society*, 132(13):4577–4579, 2010.
- [35] G. Chistol, S. Liu, C. L. Hetherington, J. R. Moffitt, S. Grimes, P. J. Jardine, and C. Bustamante. High degree of coordination and division of labor among subunits in a homomeric ring ATPase. *Cell*, 151(5):1017–1028, Nov 2012.
- [36] Y. Cho, Y. Zhang, T. Christensen, L. B. Sagle, A. Chilkoti, and P. S. Cremer. Effects of hofmeister anions on the phase transition temperature of elastin-like polypeptides. *Journal of Physical Chemistry B*, 112(44):13765–13771, 2008.
- [37] H. J. Choi, J. C. Gross, S. Pokutta, and W. I. Weis. Interactions of plakoglobin and β -catenin with desmosomal cadherins: Basis of selective exclusion of α - and β -catenin from desmosomes. *J. Biol. Chem.*, 284(46):31776–31788, Nov 2009.
- [38] H. J. Choi, S. Park-Snyder, L. T. Pascoe, K. J. Green, and W. I. Weis. Structures of two intermediate filament-binding fragments of desmoplakin reveal a unique repeat motif structure. *Nat. Struct. Biol.*, 9(8):612–620, Aug 2002.

- [39] H. J. Choi and W. I. Weis. Structure of the armadillo repeat domain of plakophilin 1. *J. Mol. Biol.*, 346(1):367–376, Feb 2005.
- [40] H. J. Choi and W. I. Weis. Crystal structure of a rigid four-spectrin-repeat fragment of the human desmoplakin plakin domain. *J. Mol. Biol.*, 409(5):800–812, Jun 2011.
- [41] C. Chothia. Hydrophobic bonding and accessible surface area in proteins. *Nature (London, United Kingdom)*, 248(5446):338–9, 1974.
- [42] T. Christensen, W. Hassouneh, K. Trabbic-Carlson, and A. Chilkoti. Predicting transition temperatures of elastin-like polypeptide fusion proteins. *Biomacromolecules*, 14(5):1514–1519, 2013.
- [43] C. Chung, K. J. Lampe, and S. C. Heilshorn. Tetrakis(hydroxymethyl) phosphonium chloride as a covalent cross-linking agent for cell encapsulation within protein-based hydrogels. *Biomacromolecules*, 13(12):3912–3916, 2012.
- [44] S. G. Clarke. Protein methylation at the surface and buried deep: Thinking outside the histone box. *Trends Biochem. Sci.*, 38(5):243–252, May 2013.
- [45] B. M. Collins, S. J. Harrop, G. D. Kornfeld, I. W. Dawes, P. M. Curmi, and B. C. Mabbutt. Crystal structure of a heptameric Sm-like protein complex from archaea: implications for the structure and evolution of snRNPs. *J. Mol. Biol.*, 309(4):915–923, Jun 2001.
- [46] H. Colognato, D. A. Winkelmann, and P. D. Yurchenco. Laminin polymerization induces a receptor-cytoskeleton network. *The Journal of cell biology*, 145(3):619–31, 1999.
- [47] T. P. Creamer and M. N. Campbell. Determinants of the polyproline ii helix from modeling studies. *Advances in Protein Chemistry*, 62(Unfolded Proteins):263–282, 2002.
- [48] T. P. Creamer and G. D. Rose. Alpha-helix-forming propensities in peptides and proteins. *Proteins*, 19(2):85–97, 1994.
- [49] S. Crosson, K. Mckee, M. Ruegg, and P. D. Yurchenco. Restoring laminin polymerization by transgenic expression of α LNNd in skeletal muscle improves muscle integrity of laminin- α 2-deficient mice. *Glycobiology*, 22(11):1536–1536, 2012.
- [50] Y. S. Dagdas, A. Tombuloglu, A. B. Tekinay, A. Dana, and M. O. Guler. Interfiber interactions alter the stiffness of gels formed by supramolecular self-assembled nanofibers. *Soft Matter*, 7(7):3524–3532, 2011.
- [51] J. Dandurand, V. Samouillan, C. Lacabanne, A. Pepe, and B. Bochicchio. Water structure and elastin-like peptide aggregation - a differential calorimetric approach. *Journal of Thermal Analysis and Calorimetry*, 120(1):419–426, 2015.
- [52] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *Journal of Chemical Physics*, 98(12):10089, 1993.
- [53] D. Das, P. Kozbial, H. L. Axelrod, M. D. Miller, D. McMullan, S. S. Krishna, P. Abdubek, C. Acosta, T. Astakhova, P. Burra, D. Carlton, C. Chen, H.-J. Chiu, T. Clayton, M. C. Deller, L. Duan, Y. Elias, M.-A. Elsliger, D. Ernst, C. Farr, J. Feuerhelm, A. Grzechnik, S. K. Grzechnik, J. Hale, G. W. Han, L. Jaroszewski, K. K. Jin, H. A. Johnson, H. E. Klock, M. W. Knuth, A. Kumar, D. Marciano, A. T. Morse, K. D. Murphy, E. Nigoghossian, A. Nopakun, L. Okach, S. Oommachen, J. Paulsen, C. Puckett, R. Reyes, C. L. Rife, N. Sefcovic, S. Sudek, H. Tien, C. Trame, C. V. Trout, H. van den Bedem, D. Weekes, A. White, Q. Xu, K. O. Hodgson, J. Wooley, A. M. Deacon, A. Godzik, S. A. Lesley, and I. A. Wilson. Crystal structure of a novel Sm-like protein of putative cyanophage origin at 2.60 Å resolution. *Proteins*, 75(2):296–307, may 2009.
- [54] P. Das, J. A. King, and R. Zhou. Aggregation of γ -crystallins associated with human cataracts via domain swapping at the C-terminal beta-strands. *Proc. Natl. Acad. Sci. U.S.A.*, 108(26):10514–10519, Jun 2011.
- [55] R. K. Das, K. M. Ruff, and R. V. Pappu. Relating sequence encoded information to form and function of intrinsically disordered proteins. *Current Opinion in Structural Biology*, 32:102–112, 2015.
- [56] C.-A. de Coulomb. Premier mémoire sur l’électricité et le magnétisme. In *Histoire de l’Académie Royale des Sciences*. Imprimerie Royale, 1785.
- [57] E. Delva, D. K. Tucker, and A. P. Kowalczyk. The desmosome. *Cold Spring Harbor Perspectives in Biology*, 1(2), 2009.

- [58] E. J. Denning, U. D. Priyakumar, L. Nilsson, and J. Mackerell, Alexander D. Impact of 2'-hydroxyl sampling on the conformational properties of RNA: Update of the CHARMM all-atom additive force field for RNA. *Journal of Computational Chemistry*, 32(9):1929–1943, 2011.
- [59] B. V. Desai, R. M. Harmon, and K. J. Green. Desmosomes at a glance. *J. Cell. Sci.*, 122(Pt 24):4401–4407, Dec 2009.
- [60] C. M. Doyle, J. A. Rumfeldt, H. R. Broom, A. Broom, P. B. Stathopoulos, K. A. Vassall, J. J. Almey, and E. M. Meiering. Energetics of oligomeric protein folding and association. *Arch. Biochem. Biophys.*, 531(1-2):44–64, Mar 2013.
- [61] C. Dryzun, A. Zait, and D. Avnir. Quantitative symmetry and chirality—a fast computational algorithm for large structures: proteins, macromolecules, nanotubes, and unit cells. *J Comput Chem*, 32(12):2526–2538, Sep 2011.
- [62] A. H. Elcock and J. A. McCammon. Identification of protein oligomerization states by analysis of interface conservation. *Proc. Natl. Acad. Sci. U.S.A.*, 98(6):2990–2994, Mar 2001.
- [63] P. Fagerholm, N. S. Lagali, D. J. Carlsson, K. Merrett, and M. Griffith. Corneal regeneration following implantation of a biomimetic tissue-engineered substitute. *Clinical and Translational Science*, 2(2):162–164, 2009.
- [64] M. Feig and C. L. B. III. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Current Opinion in Structural Biology*, 14(2):217 – 224, 2004.
- [65] R. Flamia, G. Lanza, A. M. Salvi, J. E. Castle, and A. M. Tamburro. Conformational study and hydrogen bonds detection on elastin-related polypeptides using X-ray photoelectron spectroscopy. *Biomacromolecules*, 6(3):1299–1309, 2005.
- [66] C. Fogl, F. Mohammed, C. Al-Jassar, M. Jeeves, T. Knowles, P. Rodriguez-Zamora, S. White, E. Odintsova, M. Overduin, and M. Chidgey. Mechanism of intermediate filament recognition by plakin repeat domains revealed by envoplakin targeting of vimentin. *Nature Communications*, 7(5):10827, 2016.
- [67] P. L. Freddolino, C. B. Harrison, Y. Liu, and K. Schulten. Challenges in protein folding simulations: Timescale, representation, and analysis. *Nat Phys*, 6(10):751–758, Oct 2010.
- [68] P. L. Freddolino and K. Schulten. Common structural transitions in explicit-solvent simulations of villin headpiece folding. *Biophys. J.*, 97(8):2338–2347, Oct 2009.
- [69] J. Frellsen, I. Moltke, M. Thiim, K. V. Mardia, J. Ferkinghoff-Borg, and T. Hamelryck. A probabilistic model of RNA conformational space. *PLoS Comput. Biol.*, 5(6):e1000406, Jun 2009.
- [70] D. Frishman and P. Argos. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Genetics*, 23(4):566–79, 1995.
- [71] J. R. Fromm, R. E. Hileman, J. M. Weiler, and R. J. Linhardt. Interaction of fibroblast growth factor-1 and related peptides with heparan sulfate and its oligosaccharides. *Archives of biochemistry and biophysics*, 346(2):252–62, 1997.
- [72] R. Galindo-Murillo, D. R. Roe, and T. E. Cheatham. On the absence of intrahelical DNA dynamics on the us to ms timescale. *Nat Commun*, 5:5152, Oct 2014.
- [73] D. Garrod and M. Chidgey. Desmosome structure, composition and function. *Biochimica et Biophysica Acta (BBA)–Biomembranes*, 1778(3):572 – 587, 2008.
- [74] A. Girotti, J. Reguera, F. J. Arias, M. Alonso, A. M. Testera, and J. C. Rodriguez-Cabello. Influence of the molecular weight on the inverse temperature transition of a model genetically engineered elastin-like pH-responsive polymer. *Macromolecules*, 37(9):3396–3400, 2004.
- [75] L. M. Godsel, S. N. Hsieh, E. V. Amargo, A. E. Bass, L. T. Pascoe-McGillicuddy, A. C. Huen, M. E. Thorne, C. A. Gaudry, J. K. Park, K. Myung, R. D. Goldman, T. L. Chew, and K. J. Green. Desmoplakin assembly dynamics in four dimensions: Multiple phases differentially regulated by intermediate filaments and actin. *J. Cell Biol.*, 171(6):1045–1059, Dec 2005.
- [76] D. S. Goodsell and A. J. Olson. Structural symmetry and protein function. *Annu Rev Biophys Biomol Struct*, 29:105–153, 2000.

- [77] M. S. Gordon and M. W. Schmidt. *Advances in electronic structure theory: GAMESS a decade later*, pages 1167–1189. Elsevier, Amsterdam, 2005.
- [78] J. Graf, Y. Iwamoto, M. Sasaki, G. R. Martin, H. K. Kieinman, F. A. Robey, and Y. Yamada. Identification of an amino acid sequence in laminin mediating cell attachment, chemotaxis, and receptor binding. *Cell (Cambridge, MA, United States)*, 48(6):989–96, 1987.
- [79] C. Grauffel, R. H. Stote, and A. Dejaegere. Force field parameters for the simulation of modified histone tails. *J Comput Chem*, 31(13):2434–2451, Oct 2010.
- [80] K. J. Green, D. A. Parry, P. M. Steinert, M. L. Virata, R. M. Wagner, B. D. Angst, and L. A. Nilles. Structure of the human desmoplakins. Implications for function in the desmosomal plaque. *J. Biol. Chem.*, 265(5):2603–2612, Feb 1990.
- [81] C. Grimm, A. Chari, J. P. Pelz, J. Kuper, C. Kisker, K. Diederichs, H. Stark, H. Schindelin, and U. Fischer. Structural basis of assembly chaperone- mediated snRNP formation. *Mol. Cell*, 49(4):692–703, Feb 2013.
- [82] H. Grubmuller, B. Heymann, and P. Tavan. Ligand binding: Molecular mechanics calculation of the streptavidin-biotin rupture force. *Science*, 271(5251):997–999, Feb 1996.
- [83] R. Hallmann, N. Horn, M. Selg, O. Wendler, F. Pausch, and L. M. Sorokin. Expression and function of laminins in the embryonic and mature vasculature. *Physiological Reviews*, 85(3):979–1000, 2005.
- [84] D. Hamelberg, T. Shen, and J. A. McCammon. A proposed signaling motif for nuclear import in mRNA processing via the formation of arginine claw. *Proc. Natl. Acad. Sci. U.S.A.*, 104(38):14947–14951, Sep 2007.
- [85] W. Hassouneh, E. B. Zhulina, A. Chilkoti, and M. Rubinstein. Elastin-like polypeptide diblock copolymers self-assemble into weak micelles. *Macromolecules (Washington, DC, United States)*, 48(12):4183–4195, 2015.
- [86] M. Heinig and D. Frishman. STRIDE: A web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Research*, 32(Web Server):W500–W502–, 2004.
- [87] J. Henriques, C. Cragnell, and M. Skepoe. Molecular dynamics simulations of intrinsically disordered proteins: Force field evaluation and comparison with experiment. *Journal of Chemical Theory and Computation*, 11(7):3420–3431, 2015.
- [88] R. P. Hobbs and K. J. Green. Desmoplakin regulates desmosome hyperadhesion. *J Invest Dermatol*, 132(2):482–485, Feb. 2012.
- [89] B. Holthofer, R. Windoffer, S. Troyanovsky, and R. E. Leube. Structure and function of desmosomes. *Int. Rev. Cytol.*, 264:65–163, 2007.
- [90] N. Homeyer, A. H. Horn, H. Lanig, and H. Sticht. AMBER force-field parameters for phosphorylated amino acids in different protonation states: Phosphoserine, phosphothreonine, phosphotyrosine, and phosphohistidine. *J Mol Model*, 12(3):281–289, Feb 2006.
- [91] K. Hozumi, D. Otagiri, Y. Yamada, A. Sasaki, C. Fujimori, Y. Wakai, T. Uchida, F. Katagiri, Y. Kikkawa, and M. Nomizu. Cell surface receptor-specific scaffold requirements for adhesion to laminin-derived peptide-chitosan membranes. *Biomaterials*, 31(12):3237–3243, 2010.
- [92] J. Huang, C. Sun, O. Mitchell, N. Ng, Z. N. Wang, and G. S. Boutis. On the inverse temperature transition and development of an entropic elastomeric force of the elastin mimetic peptide [LGGVG]₃, 7. *Journal of Chemical Physics*, 136(8):085101/1–085101/9, 2012.
- [93] W. Humphrey, A. Dalke, and K. Schulten. VMD: visual molecular dynamics. *J Mol Graph*, 14(1):33–38, Feb 1996.
- [94] M. Jucker, P. Bialobok, H. K. Kleinman, L. C. Walker, T. Hagg, and D. K. Ingram. Laminin-like and laminin-binding protein-like immunoreactive astrocytes in rat hippocampus after transient ischemia: antibody to laminin-binding protein is a sensitive marker of neural injury and degeneration. *Annals of the New York Academy of Sciences*, 679(Markers of Neuronal Injury and Degeneration):245–52, 1993.
- [95] S. Y. Jung, J.-M. Kim, H. K. Kang, D. H. Jang, and B.-M. Min. A biologically active sequence of the laminin α large globular 1 domain promotes cell adhesion through syndecan-1 by inducing phosphorylation and membrane localization of protein kinase C δ . *Journal of Biological Chemistry*, 284(46):31764–31775, 2009.

- [96] H. Kang, T. M. Weiss, I. Bang, W. I. Weis, and H. J. Choi. Structure of the intermediate filament-binding region of desmoplakin. *PLoS ONE*, 11(1):e0147641, 2016.
- [97] C. B. Karim, L. M. Espinoza-Fonseca, Z. M. James, E. A. Hanse, J. S. Gaynes, D. D. Thomas, and A. Kelekar. Structural mechanism for regulation of Bcl-2 protein Noxa by phosphorylation. *Sci Rep*, 5:14557, 2015.
- [98] I. L. Karle and D. W. Urry. Crystal structure of cyclic (APGVGV)₂, an analog of elastin, and a suggested mechanism for elongation/contraction of the molecule. *Biopolymers*, 77(4):198–204, 2005.
- [99] Z. Z. Khaing, R. C. Thomas, S. A. Geissler, and C. E. Schmidt. Advanced biomaterials for repairing the nervous system: what can hydrogels do for the brain?. *Materials Today (Oxford, United Kingdom)*, 17(7):332–340, 2014.
- [100] T. Kilic, S. Sanglier, A. Van Dorsselaer, and D. Suck. Oligomerization behavior of the archaeal Sm2-type protein from *Archaeoglobus fulgidus*. *Protein Science*, 15(10):2310–2317, 2006.
- [101] M. Kjaergaard, A.-B. Noerholm, R. Hendus-Altenburger, S. F. Pedersen, F. M. Poulsen, and B. B. Kragelund. Temperature-dependent structural changes in intrinsically disordered proteins: formation of α -helices or loss of polyproline ii?. *Protein Science*, 19(8):1555–1564, 2010.
- [102] A. P. Kowalczyk and K. J. Green. Structure, function, and regulation of desmosomes. *Prog Mol Biol Transl Sci*, 116:95–118, 2013.
- [103] A. Krukau, I. Brovchenko, and A. Geiger. Temperature-induced conformational transition of a model elastin-like peptide GVG(VPGVG)₃ in water. *Biomacromolecules*, 8(7):2196–2202, 2007.
- [104] K. K. Kumashiro, T. L. Kurano, W. P. Niemczura, M. Martino, and A. M. Tamburro. ¹³C CPMAS NMR studies of the elastin-like polypeptide (LGGVG)_n. *Biopolymers*, 70(2):221–226, 2003.
- [105] K. J. Lampe, A. L. Antaris, and S. C. Heilshorn. Design of three-dimensional engineered protein hydrogels for tailored control of neurite growth. *Acta Biomaterialia*, 9(3):5590–5599, 2013.
- [106] K. J. Lampe and S. C. Heilshorn. Building stem cell niches from the molecule up through engineered peptide materials. *Neuroscience Letters*, 519(2):138–146, 2012.
- [107] W. M. Latimer and W. H. Rodebush. Polarity and ionization from the standpoint of the Lewis theory of valence. *Journal of the American Chemical Society*, 42(7):1419–1433, 1920.
- [108] J. A. Lemkul and A. D. MacKerell. Balancing the interactions of Mg(2+) in aqueous solution and with nucleic acid moieties for a polarizable force field based on the classical Drude oscillator model. *J Phys Chem B*, 120(44):11436–11448, Nov 2016.
- [109] D. H. Lenz, K. C. Mok, B. N. Lilley, R. V. Kulkarni, N. S. Wingreen, and B. L. Bassler. The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in *Vibrio harveyi* and *Vibrio cholerae*. *Cell*, 118(1):69 – 82, 2004.
- [110] A. K. W. Leung, K. Nagai, and J. Li. Structure of the spliceosomal U4 snRNP core domain and its implication for snRNP biogenesis. *Nature*, 473(7348):536–539, May 2011.
- [111] B. Li, D. O. Alonso, and V. Daggett. The molecular basis for the inverse temperature transition of elastin. *Journal of molecular biology*, 305(3):581–92, 2001.
- [112] N. K. Li, F. G. Quiroz, C. K. Hall, A. Chilkoti, and Y. G. Yingling. Molecular description of the LCST behavior of an elastin-like polypeptide. *Biomacromolecules*, 15(10):3522–3530, 2014.
- [113] S. Li, P. Liquari, K. K. McKee, D. Harrison, R. Patel, S. Lee, and P. D. Yurchenco. Laminin-sulfatide binding initiates basement membrane assembly and enables receptor signaling in Schwann cells and fibroblasts. *Journal of Cell Biology*, 169(1):179–189, 2005.
- [114] K. Lindorff-Larsen, P. Maragakis, S. Piana, and D. E. Shaw. Picosecond to millisecond structural dynamics in human ubiquitin. *J Phys Chem B*, 120(33):8313–8320, Aug 2016.
- [115] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. How fast-folding proteins fold. *Science (Washington, DC, United States)*, 334(6055):517–520, 2011.
- [116] H. Liu, S. G. Wise, J. Rnjak-Kovacina, D. L. Kaplan, M. M. M. Bilek, A. S. Weiss, J. Fei, and S. Bao. Biocompatibility of silk-tropoelastin protein polymers. *Biomaterials*, 35(19):5138–5147, 2014.

- [117] P. E. Lopes, J. Huang, J. Shim, Y. Luo, H. Li, B. Roux, and A. D. Mackerell. Force field for peptides and proteins based on the classical Drude oscillator. *J Chem Theory Comput*, 9(12):5430–5449, Dec 2013.
- [118] S. Y. Lu, Y. J. Jiang, J. W. Zou, and T. X. Wu. Molecular modeling and molecular dynamics simulation studies of the GSK3/ATP/substrate complex: Understanding the unique P+4 primed phosphorylation specificity for GSK3 substrates. *J Chem Inf Model*, 51(5):1025–1036, May 2011.
- [119] J. A. MacKay, D. J. Callahan, K. N. FitzGerald, and A. Chilkoti. Quantitative model of the phase behavior of recombinant ph-responsive elastin-like polypeptides. *Biomacromolecules*, 11(11):2873–2879, 2010.
- [120] A. D. Mackerell. Empirical force fields for biological macromolecules: Overview and issues. *J Comput Chem*, 25(13):1584–1604, Oct 2004.
- [121] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry. B*, 102(18):3586–616, 1998.
- [122] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling. ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J Chem Theory Comput*, 11(8):3696–3713, Aug 2015.
- [123] M. Maitre, S. Weidmann, A. Rieu, D. Fenel, G. Schoehn, C. Ebel, J. Coves, and J. Guzzo. The oligomer plasticity of the small heat-shock protein Lo18 from *Oenococcus oeni* influences its role in both membrane stabilization and protein protection. *Biochem. J.*, 444(1):97–104, May 2012.
- [124] D. Marion. An introduction to biological NMR spectroscopy. *Mol. Cell Proteomics*, 12(11):3006–3025, Nov 2013.
- [125] E. Masse and S. Gottesman. A small RNA regulates the expression of genes involved in iron metabolism in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.*, 99(7):4620–4625, Apr 2002.
- [126] Y. Matsunaga, R. Koike, M. Ota, J. R. Tame, and A. Kidera. Influence of structural symmetry on protein dynamics. *PLoS ONE*, 7(11):e50011, 2012.
- [127] C. E. McAnany and C. Mura. Claws, disorder, and conformational dynamics of the C-terminal region of human desmoplakin. *J Phys Chem B*, 120(33):8654–8667, Aug 2016.
- [128] J. A. McCammon and S. C. Harvey. *Dynamics of Proteins and Nucleic Acids*. Cambridge University Press, 1988.
- [129] K. K. McKee, S. Capizzi, and P. D. Yurchenco. Scaffold-forming and adhesive contributions of synthetic laminin-binding proteins to basement membrane assembly. *Journal of Biological Chemistry*, 284(13):8984–8994, 2009.
- [130] M. Medina and F. Wandosell. Deconstructing GSK-3: The fine regulation of its activity. *Int J Alzheimers Dis*, 2011:479249, 2011.
- [131] L. Meinhold, D. Clement, M. Tehei, R. Daniel, J. L. Finney, and J. C. Smith. Protein dynamics and stability: the distribution of atomic fluctuations in thermophilic and mesophilic dihydrofolate reductase derived using elastic incoherent neutron scattering. *Biophys. J.*, 94(12):4812–4818, Jun 2008.
- [132] D. E. Meyer and A. Chilkoti. Purification of recombinant proteins by fusion with thermally-responsive polypeptides. *Nature biotechnology*, 17(11):1112–5, 1999.
- [133] D. E. Meyer and A. Chilkoti. Quantification of the effects of chain length and concentration on the thermal behavior of elastin-like polypeptides. *Biomacromolecules*, 5(3):846–851, 2004.
- [134] D. E. Michele and K. P. Campbell. Dystrophin-glycoprotein complex: post-translational processing and dystroglycan function. *Journal of Biological Chemistry*, 278(18):15457–15460, 2003.
- [135] J. S. Miller, C. J. Shen, W. R. Legant, J. D. Baranski, B. L. Blakely, and C. S. Chen. Bioactive hydrogels made from step-growth derived PEG-peptide macromers. *Biomaterials*, 31(13):3736–3743, 2010.

- [136] A. Mitsutake, Y. Mori, and Y. Okamoto. Enhanced sampling algorithms. *Methods Mol. Biol.*, 924:153–195, 2013.
- [137] T. Mittag and J. D. Forman-Kay. Atomic-level characterization of disordered protein ensembles. *Curr. Opin. Struct. Biol.*, 17(1):3–14, Feb 2007.
- [138] J. Mittal, T. H. Yoo, G. Georgiou, and T. M. Truskett. Structural ensemble of an intrinsically disordered polypeptide. *J Phys Chem B*, 117(1):118–124, Jan 2013.
- [139] T. Moller, T. Franch, P. Hojrup, D. R. Keene, H. P. Bachinger, R. G. Brennan, and P. Valentin-Hansen. Hfq: A bacterial Sm-like protein that mediates RNA-RNA interaction. *Molecular Cell*, 9(1):23 – 30, 2002.
- [140] P. Moscarelli, F. Boraldi, B. Bochicchio, A. Pepe, A. M. Salvi, and D. Quaglino. Structural characterization and biological properties of the amyloidogenic elastin-like peptide (VGGVG)₃. *Matrix Biology*, 36:15–27, 2014.
- [141] F. Muntoni, S. Torelli, and M. Brockington. Muscular dystrophies due to glycosylation defects. *Neurotherapeutics*, 5(4):627–632, 2008.
- [142] C. Mura, D. Cascio, M. R. Sawaya, and D. S. Eisenberg. The crystal structure of a heptameric archaeal Sm protein: Implications for the eukaryotic snRNP core. *Proceedings of the National Academy of Sciences*, 98(10):5532–5537, 2001.
- [143] C. Mura and C. E. McAnany. An introduction to biomolecular simulations and docking. *Molecular Simulation*, 40(10-11):732–764, 2014.
- [144] C. Mura and J. A. McCammon. Molecular dynamics of a kappaB DNA element: Base flipping via cross-strand intercalative stacking in a microsecond-scale simulation. *Nucleic Acids Res.*, 36(15):4941–4955, Sep 2008.
- [145] C. Mura, P. S. Randolph, J. Patterson, and A. E. Cozen. Archaeal and eukaryotic homologs of Hfq: A structural and evolutionary perspective on Sm function. *RNA Biol*, 10(4):636–651, Apr 2013.
- [146] A. Murzin. Structural principles for the propeller assembly of β -sheets: the preference for seven-fold symmetry. *Proteins*, 14(2):191–201, Oct 1992.
- [147] N. Naidoo, S. J. Harrop, M. Sobti, P. A. Haynes, B. R. Szymczyna, J. R. Williamson, P. M. Curmi, and B. C. Mabbutt. Crystal structure of Lsm3 octamer from *Saccharomyces cerevisiae*: Implications for Lsm ring organisation and recruitment. *Journal of Molecular Biology*, 377(5):1357–1371, apr 2008.
- [148] C. Narayanan, D. S. Weinstock, K. P. Wu, J. Baum, and R. M. Levy. Investigation of the polymeric properties of α -synuclein and comparison with NMR experiments: A replica exchange molecular dynamics study. *J Chem Theory Comput*, 8(10):3929–3942, Oct 2012.
- [149] J. S. Nielsen, A. Bøggild, C. B. Andersen, G. Nielsen, A. Boysen, D. E. Brodersen, and P. Valentin-Hansen. An Hfq-like protein in archaea: Crystal structure and functional characterization of the Sm protein from *Methanococcus jannaschii*. *RNA*, 13(12):2213–2223, 2007.
- [150] M. A. Nugent and R. V. Iozzo. Fibroblast growth factor-2. *The international journal of biochemistry & cell biology*, 32(2):115–20, 2000.
- [151] H. Nuhn and H.-A. Klok. Secondary structure formation and LCST behavior of short elastin-like peptides. *Biomacromolecules*, 9(10):2755–2763, 2008.
- [152] K. Ohgo, J. Ashida, K. K. Kumashiro, and T. Asakura. Structural determination of an elastin-mimetic model peptide, (Val-Pro-Gly-Val-Gly)₆, studied by ¹³C CP/MAS NMR chemical shifts, two-dimensional off magic angle spinning spin-diffusion NMR, rotational echo double resonance, and statistical distribution of torsion angles from Protein Data Bank. *Macromolecules*, 38(14):6038–6047, 2005.
- [153] M. Oke, L. G. Carter, K. A. Johnson, H. Liu, S. A. McMahon, X. Yan, M. Kerou, N. D. Weikart, N. Kadi, M. A. Sheikh, S. Schmelz, M. Dorward, M. Zawadzki, C. Cozens, H. Falconer, H. Powers, I. M. Overton, C. A. van Niekerk, X. Peng, P. Patel, R. A. Garrett, D. Prangishvili, C. H. Botting, P. J. Coote, D. T. Dryden, G. J. Barton, U. Schwarz-Linek, G. L. Challis, G. L. Taylor, M. F. White, and J. H. Naismith. The scottish structural proteomics facility: targets, methods and outputs. *J. Struct. Funct. Genomics*, 11(2):167–180, Jun 2010.
- [154] K. M. Oshaben, R. Salari, D. R. McCaslin, L. T. Chong, and W. S. Horne. The native GCN4 leucine-zipper domain does not uniquely specify a dimeric oligomerization state. *Biochemistry*, 51(47):9581–9591, Nov 2012.

- [155] K. Ostermeir and M. Zacharias. Advanced replica-exchange sampling to study the flexibility and plasticity of peptides and proteins. *Biochim. Biophys. Acta*, 1834(5):847–853, May 2013.
- [156] W. K. Paik, D. C. Paik, and S. Kim. Historical review: the field of protein methylation. *Trends Biochem. Sci.*, 32(3):146–152, Mar 2007.
- [157] S. Papagerakis, A. H. Shabana, B. H. Pollock, P. Papagerakis, J. Depondt, and A. Berdal. Altered desmoplakin expression at transcriptional and protein levels provides prognostic information in human oropharyngeal cancer. *Hum. Pathol.*, 40(9):1320–1329, Sep 2009.
- [158] G. V. Papamokos, G. Tziatzos, D. G. Papageorgiou, S. D. Georgatos, A. S. Politou, and E. Kaxiras. Structural role of RKS motifs in chromatin interactions: a molecular dynamics study of HP1 bound to a variably modified histone tail. *Biophys. J.*, 102(8):1926–1933, Apr 2012.
- [159] G. A. Papoian and P. G. Wolynes. The physics and bioinformatics of binding and folding - an energy landscape perspective. *Biopolymers*, 68(3):333–349, 2003.
- [160] S. Park, F. Khalili-Araghi, E. Tajkhorshid, and K. Schulten. Free energy calculation from steered molecular dynamics simulations using jarzynski’s equality. *The Journal of Chemical Physics*, 119(6):3559–3566, 2003.
- [161] P. Patwardhan and W. T. Miller. Processive phosphorylation: Mechanism and biological importance. *Cell. Signal.*, 19(11):2218–2226, Nov 2007.
- [162] Y. Peng, J. E. Curtis, X. Fang, and S. A. Woodson. Structural model of an mRNA in complex with the bacterial chaperone Hfq. *Proc. Natl. Acad. Sci. U.S.A.*, 111(48):17134–17139, Dec 2014.
- [163] J. B. Pereira-Leal, E. D. Levy, C. Kamp, and S. A. Teichmann. Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol*, 8(4):R51, 2007.
- [164] Persistence of Vision Pty. Ltd. Persistence of vision raytracer (version 3.6), 2004.
- [165] D. Petrov, C. Margreitter, M. Grandits, C. Oostenbrink, and B. Zagrovic. A systematic framework for molecular dynamics simulations of protein post-translational modifications. *PLoS Comput Biol*, 9(7):1–9, 07 2013.
- [166] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.
- [167] M. D. Pierschbacher and E. Ruoslahti. Cell attachment activity of fibronectin can be duplicated by small synthetic fragments of the molecule. *Nature (London, United Kingdom)*, 309(5963):30–3, 1984.
- [168] M. Pinsky, A. Zait, M. Bonjack, and D. Avnir. Continuous symmetry analyses: C(nv) and D(n) measures of molecules, complexes, and proteins. *J Comput Chem*, 34(1):2–9, Jan 2013.
- [169] G. Qin, M. J. Glassman, C. N. Lam, D. Chang, E. Schaible, A. Hexemer, and B. D. Olsen. Topological effects on globular protein-ELP fusion block copolymer self-assembly. *Advanced Functional Materials*, 25(5):729–738, 2015.
- [170] K. Rajagopal, B. Ozbas, D. J. Pochan, and J. P. Schneider. Self-assembly of β -hairpin peptides to hydrogels and tuning of material properties by design of turn sequence. In *Abstracts Papers American Chemical Society*, pages 288, U68, 2004.
- [171] T. B. Rasmussen, J. Hansen, P. H. Nissen, J. Palmfeldt, S. Dalager, U. B. Jensen, W. Y. Kim, L. Heickendorff, H. Mølgaard, H. K. Jensen, K. E. Sørensen, U. T. Baandrup, P. Bross, and J. Mogensen. Protein expression studies of desmoplakin mutations in cardiomyopathy patients reveal different molecular disease mechanisms. *Clin. Genet.*, 84(1):20–30, Jul 2013.
- [172] S. Rauscher, V. Gapsys, M. J. Gajda, M. Zweckstetter, B. L. de Groot, and H. Grubmüller. Structural ensembles of intrinsically disordered proteins depend strongly on force field: A comparison to experiment. *J Chem Theory Comput*, 11(11):5513–5524, Nov 2015.
- [173] V. Receveur-Brechot and D. Durand. How random are intrinsically disordered proteins? a small angle scattering perspective. *Current Protein and Peptide Science*, 13(1):55–75, 2012.

- [174] H. Reiersen, A. R. Clarke, and A. R. Rees. Short elastin-like peptides exhibit the same temperature-induced structural transitions as elastin polymers: implications for protein engineering. *Journal of molecular biology*, 283(1):255–64, 1998.
- [175] A. Ribeiro, F. J. Arias, J. Reguera, M. Alonso, and J. C. Rodriguez-Cabello. Influence of the amino-acid sequence on the inverse temperature transition of elastin-like polymers. *Biophysical Journal*, 97(1):312–320, 2009.
- [176] A. D. Robertson and K. P. Murphy. Protein structure and the energetics of protein stability. *Chemical Reviews (Washington, D. C.)*, 97(5):1251–1267, 1997.
- [177] D. J. Rosenman, C. R. Connors, W. Chen, C. Wang, and A. E. Garcia. A β monomers transiently sample oligomer and fibril-like configurations: ensemble characterization using a combined MD/NMR approach. *J. Mol. Biol.*, 425(18):3338–3359, Sep 2013.
- [178] B. Rost and C. Sander. Conservation and prediction of solvent accessibility in protein families. *Proteins*, 20(3):216–26, 1994.
- [179] F. Rousseau, J. Schymkowitz, and L. S. Itzhaki. Implications of 3D domain swapping for protein folding, misfolding and function. *Advances in Experimental Medicine and Biology*, 747(Protein Dimerization and Oligomerization in Biology):137–152, 2012.
- [180] A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, 234(3):779–815, Dec 1993.
- [181] A. M. Salvi, P. Moscarelli, B. Bochicchio, G. Lanza, and J. E. Castle. Combined effects of solvation and aggregation propensity on the final supramolecular structures adopted by hydrophobic, glycine-rich, elastin-like polypeptides. *Biopolymers*, 99(5):292–313, 2013.
- [182] T. J. Sanborn, P. B. Messersmith, and A. E. Barron. In situ crosslinking of a biomimetic peptide-PEG hydrogel via thermally triggered activation of factor XIII. *Biomaterials*, 23(13):2703–2710, 2002.
- [183] S. Sandler. *Chemical, Biochemical, and Engineering Thermodynamics*. Wiley, 4 edition, 2006.
- [184] E. Sauer and O. Weichenrieder. Structural basis for RNA 3'-end recognition by Hfq. *Proceedings of the National Academy of Sciences*, 108(32):13065–13070, 2011.
- [185] L. Sborgi, A. Verma, S. Piana, K. Lindorff-Larsen, M. Cerminara, C. M. Santiveri, D. E. Shaw, E. de Alba, and V. Munoz. Interaction networks in protein folding via atomic-resolution experiments and long-time-scale molecular dynamics simulations. *J. Am. Chem. Soc.*, 137(20):6506–6516, May 2015.
- [186] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, T. L. Windus, M. Dupuis, and J. A. Montgomery. General atomic and molecular electronic structure system. *Journal of Computational Chemistry*, 14(11):1347–1363, 1993.
- [187] D. J. Schu, A. Zhang, S. Gottesman, and G. Storz. Alternative Hfq-sRNA interaction modes dictate alternative mRNA recognition. *EMBO J.*, 34(20):2557–2573, Oct 2015.
- [188] E. C. Schulz and O. Barabas. Structure of an *Escherichia coli* Hfq:RNA complex at 0.97 Å resolution. *Acta Crystallogr F Struct Biol Commun*, 70(Pt 11):1492–1497, Nov 2014.
- [189] D. G. Scofield and M. Lynch. Evolutionary diversification of the Sm family of RNA-associated proteins. *Molecular Biology and Evolution*, 25(11):2255–2267, 2008.
- [190] D. Sellis, V. Drosou, D. Vlachakis, N. Voukkalis, T. Giannakouros, and M. Vlassi. Phosphorylation of the arginine/serine repeats of lamin B receptor by SRPK1-insights from molecular dynamics simulations. *Biochim. Biophys. Acta*, 1820(1):44–55, Jan 2012.
- [191] J. L. Seltzer, H. Weingarten, K. T. Akers, M. L. Eschbach, G. A. Grant, and A. Z. Eisen. Cleavage specificity of type IV collagenase (gelatinase) from human skin. use of synthetic peptides as model substrates. *The Journal of biological chemistry*, 264(33):19583–6, 1989.
- [192] N. G. Sgourakis, M. Merced-Serrano, C. Boutsidis, P. Drineas, Z. Du, C. Wang, and A. E. Garcia. Atomic-level characterization of the ensemble of the A β (1-42) monomer in water using unbiased molecular dynamics simulations and spectral algorithms. *J. Mol. Biol.*, 405(2):570–583, Jan 2011.

- [193] N. G. Sgourakis, Y. Yan, S. A. McCallum, C. Wang, and A. E. Garcia. The Alzheimer’s peptides A β 40 and 42 adopt distinct conformations in water: a combined MD / NMR study. *J. Mol. Biol.*, 368(5):1448–1457, May 2007.
- [194] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346, Oct 2010.
- [195] A. Shimizu, A. Ishiko, T. Ota, K. Tsunoda, M. Amagai, and T. Nishikawa. IgG binds to desmoglein 3 in desmosomes and causes a desmosomal split without keratin retraction in a pemphigus mouse model. *Journal of Investigative Dermatology*, 122(5):1145 – 1153, 2004.
- [196] E. A. Smith and E. Fuchs. Defining the interactions between intermediate filaments and desmosomes. *J. Cell Biol.*, 141(5):1229–1241, Jun 1998.
- [197] L. G. Smith, J. Zhao, D. H. Mathews, and D. H. Turner. Physics-based all-atom modeling of RNA energetics and structure. *Wiley Interdiscip Rev RNA*, 8(5), Sep 2017.
- [198] N. Sreerama and R. W. Woody. Molecular dynamics simulations of polypeptide conformations in water: A comparison of α , β , and poly(pro)II conformations. *Proteins*, 36(4):400–6, 1999.
- [199] K. A. Stanek, J. Patterson-West, P. S. Randolph, and C. Mura. Crystal structure and RNA-binding properties of an Hfq homolog from the deep-branching *Aquificae*: conservation of the lateral RNA-binding mode. *Acta Crystallogr D Struct Biol*, 73(Pt 4):294–315, Apr 2017.
- [200] N. Stanley, S. Esteban-Martín, and G. De Fabritiis. Kinetic modulation of a disordered protein domain by phosphorylation. *Nature Communications*, 5, 2014.
- [201] T. S. Stappenbeck, J. A. Lamb, C. M. Corcoran, and K. J. Green. Phosphorylation of the desmoplakin COOH terminus negatively regulates its interaction with keratin intermediate filament networks. *J. Biol. Chem.*, 269(47):29351–29354, Nov 1994.
- [202] N. Stephanopoulos, J. H. Ortony, and S. I. Stupp. Self-assembly for the synthesis of functional biomaterials. *Acta Materialia*, 61(3):912–930, 2013.
- [203] D. L. Stokes. Desmosomes from a structural perspective. *Current Opinion in Cell Biology*, 19(5):565 – 571, 2007.
- [204] K. S. Straley and S. C. Heilshorn. Independent tuning of multiple biomaterial properties using protein engineering. *Soft Matter*, 5(1):114–124, 2009.
- [205] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(12):141 – 151, 1999.
- [206] N. Suzuki, K. Hozumi, S. Urushibata, T. Yoshimura, Y. Kikkawa, J. D. Gumerson, D. E. Michele, M. P. Hoffman, Y. Yamada, and M. Nomizu. Identification of α -dystroglycan binding sequences in the laminin α 2 chain LG4-5 module. *Matrix Biology*, 29(2):143–151, 2010.
- [207] J. D. Tang, C. E. McAnany, C. Mura, and K. J. Lampe. Toward a designable extracellular matrix: Molecular dynamics simulations of an engineered laminin-mimetic, elastin-like fusion protein. *Biomacromolecules*, 17(10):3222–3233, Oct 2016.
- [208] D. E. Tanner, K. Y. Chan, J. C. Phillips, and K. Schulten. Parallel generalized Born implicit solvent calculations with NAMD. *J Chem Theory Comput*, 7(11):3635–3642, Nov 2011.
- [209] K. Tashiro, G. C. Sephel, B. Weeks, M. Sasaki, G. R. Martin, H. K. Kleinman, and Y. Yamada. A synthetic peptide containing the IKVAV sequence from the A chain of laminin mediates cell attachment, migration, and neurite outgrowth. *The Journal of biological chemistry*, 264(27):16174–82, 1989.
- [210] E. ter Haar, J. T. Coll, D. A. Austen, H. M. Hsiao, L. Swenson, and J. Jain. Structure of GSK3 β reveals a primed phosphorylation mechanism. *Nat. Struct. Biol.*, 8(7):593–596, Jul 2001.
- [211] T. Terakawa and S. Takada. Multiscale ensemble modeling of intrinsically disordered proteins: p53 N-terminal domain. *Biophys. J.*, 101(6):1450–1458, Sep 2011.

- [212] R. Timpl, D. Tisi, J. F. Talts, Z. Andac, T. Sasaki, and E. Hohenester. Structure and function of laminin LG modules. *Matrix biology : journal of the International Society for Matrix Biology*, 19(4):309–17, 2000.
- [213] D. Tisi, J. F. Talts, R. Timpl, and E. Hohenester. Structure of the C-terminal laminin G-like domain pair of the laminin $\alpha 2$ chain harbouring binding sites for α -dystroglycan and heparin. *The EMBO journal*, 19(7):1432–40, 2000.
- [214] J. Tomasi, B. Mennucci, and R. Cammi. Quantum mechanical continuum solvation models. *Chem. Rev.*, 105(8):2999–3093, Aug 2005.
- [215] I. Toro, J. Basquin, H. Teo-Dreher, and D. Suck. Archaeal Sm proteins form heptameric and hexameric complexes: crystal structures of the Sm1 and Sm2 proteins from the hyperthermophile *Archaeoglobus fulgidus*. *J. Mol. Biol.*, 320(1):129–142, Jun 2002.
- [216] I. Toro, S. Thore, C. Mayer, J. Basquin, B. Seraphin, and D. Suck. RNA binding in an Sm core domain: X-ray structure and functional analysis of an archaeal Sm protein complex. *EMBO J.*, 20(9):2293–2303, May 2001.
- [217] K. Trabbic-Carlson, D. E. Meyer, L. Liu, R. Piervincenzi, N. Nath, T. LaBean, and A. Chilkoti. Effect of protein fusion on the transition temperature of an environmentally responsive elastin-like polypeptide: a role for surface hydrophobicity?. *Protein Engineering, Design & Selection*, 17(1):57–66, 2004.
- [218] M. V. Tsiper and P. D. Yurchenco. Laminin assembles into separate basement membrane and fibrillar matrices in Schwann cells. *Journal of Cell Science*, 115(5):1005–1015, 2002.
- [219] D. W. Urry. On the molecular mechanisms of elastin coacervation and coacervate calcification. *Faraday Discuss Chem Soc*, (61):205–212, 1976.
- [220] D. W. Urry. A new hydrophobicity scale for proteins and its relevance to interactions at interfaces. In *Abstracts Papers American Chemical Society*, pages 207, 220–Coll, 1994.
- [221] D. W. Urry. Physical chemistry of biological free energy transduction as demonstrated by elastic protein-based polymers. *Journal of Physical Chemistry B*, 101(51):11007–11028, 1997.
- [222] D. W. Urry, D. C. Gowda, T. M. Parker, C. H. Luan, M. C. Reid, C. M. Harris, A. Pattanaik, and R. D. Harris. Hydrophobicity scale for proteins based on inverse temperature transitions. *Biopolymers*, 32(9):1243–50, 1992.
- [223] D. W. Urry, T. L. Trapane, and K. U. Prasad. Phase-structure transitions of the elastin polypentapeptide-water system within the framework of composition-temperature studies. *Biopolymers*, 24(12):2345–56, 1985.
- [224] S. Urushibata, K. Hozumi, M. Ishikawa, F. Katagiri, Y. Kikkawa, and M. Nomizu. Identification of biologically active sequences in the laminin $\alpha 2$ chain G domain. *Archives of Biochemistry and Biophysics*, 497(1-2):43–54, 2010.
- [225] V. Vacic, P. R. Markwick, C. J. Oldfield, X. Zhao, C. Haynes, V. N. Uversky, and L. M. Iakoucheva. Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder. *PLoS Comput. Biol.*, 8(10):e1002709, 2012.
- [226] G. van Rossum. The Python language reference, 2014. Accessed 2016-05-10.
- [227] S. Vangaveti, S. V. Ranganathan, and A. A. Chen. Advances in RNA molecular dynamics: a simulator’s guide to RNA force fields. *Wiley Interdiscip Rev RNA*, 8(2), Mar 2017.
- [228] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J Comput Chem*, 31(4):671–690, Mar 2010.
- [229] K. A. Vassall, K. Bessonov, M. De Avila, E. Polverini, and G. Harauz. The effects of threonine phosphorylation on the stability and dynamics of the central molecular switch region of 18.5-kDa myelin basic protein. *PLoS ONE*, 8(7):e68175, 2013.
- [230] S. Veretnik, C. Wills, P. Youkharibache, R. E. Valas, and P. E. Bourne. Sm/Lsm genes provide a glimpse into the early evolution of the spliceosome. *PLoS Computational Biology*, 5(3):1 – 14, 2009.
- [231] J. Vogel and B. F. Luisi. Hfq and its constellation of RNA. *Nature Reviews Microbiology*, 9(8):578 – 589, 2011.

- [232] L. Wang, R. A. Friesner, and B. J. Berne. Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (REST2). *J Phys Chem B*, 115(30):9431–9438, Aug 2011.
- [233] L. S. Wang, F. Lee, J. Lim, C. Du, A. C. Wan, S. S. Lee, and M. Kurisawa. Enzymatic conjugation of a bioactive peptide into an injectable hyaluronic acid-tyramine hydrogel system to promote the formation of functional vasculature. *Acta Biomater*, 10(6):2539–2550, Jun 2014.
- [234] S. Wang, J. Ma, J. Peng, and J. Xu. Protein structure alignment beyond spatial proximity. *Sci Rep*, 3:1448, 2013.
- [235] W. Wang, A. Maucuer, A. Gupta, V. Manceau, K. R. Thickman, W. J. Bauer, S. D. Kennedy, J. E. Wedekind, M. R. Green, and C. L. Kielkopf. Structure of phosphorylated SF1 bound to U2AF65 in an essential splicing factor complex. *Structure*, 21(2):197–208, Feb 2013.
- [236] M. J. Webber, J. Tongers, C. J. Newcomb, K. T. Marquardt, J. Bauersachs, D. W. Losordo, and S. I. Stupp. Supramolecular nanostructures that mimic VEGF as a strategy for ischemic tissue repair. *Proc. Natl. Acad. Sci. U.S.A.*, 108(33):13438–13443, Aug 2011.
- [237] B. Widom. *Statistical Mechanics: A Concise Introduction for Chemists*. Cambridge University Press, 1 edition, May 2002.
- [238] P. L. Wintrode, D. Zhang, N. Vaidehi, F. H. Arnold, and W. A. Goddard. Protein dynamics in a family of laboratory evolved thermophilic enzymes. *J. Mol. Biol.*, 327(3):745–757, Mar 2003.
- [239] H. Wizemann, J. H. O. Garbe, M. V. K. Friedrich, R. Timpl, T. Sasaki, and E. Hohenester. Distinct requirements for heparin and α -dystroglycan binding revealed by structure-based mutagenesis of the laminin α 2 LG4-LG5 domain. *Journal of Molecular Biology*, 332(3):635–642, 2003.
- [240] A. S. Woods and S. Ferre. Amazing stability of the arginine-phosphate electrostatic interaction. *J. Proteome Res.*, 4(4):1397–1402, 2005.
- [241] R. Wuttke, H. Hofmann, D. Nettels, M. B. Borgia, J. Mittal, R. B. Best, and B. Schuler. Temperature-dependent solvation modulates the dimensions of disordered proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 111(14):5213–5218, 2014.
- [242] S. Xiang, V. Gapsys, H.-Y. Kim, S. Bessonov, H.-H. Hsiao, S. Möhlmann, V. Klaukien, R. Ficner, S. Becker, H. Urlaub, R. Lührmann, B. de Groot, and M. Zweckstetter. Phosphorylation drives a dynamic switch in serine/arginine-rich proteins. *Structure*, 21(12):2162 – 2174, 2013.
- [243] P. D. Yurchenco, Y. Quan, H. Colognato, T. Mathus, D. Harrison, Y. Yamada, and J. J. O’Rear. The α chain of laminin-1 is independently secreted and drives secretion of its β - and γ -chain partners. *Proceedings of the National Academy of Sciences of the United States of America*, 94(19):10189–94, 1997.
- [244] H. Zabrodsky, S. Peleg, and D. Avnir. Continuous symmetry measures. *Journal of the American Chemical Society*, 114(20):7843–7851, 1992.
- [245] G. H. Zerze and J. Mittal. Effect of O-linked glycosylation on the equilibrium structural ensemble of intrinsically disordered polypeptides. *J Phys Chem B*, 119(51):15583–15592, Dec 2015.
- [246] L. Zhang. *Identification of Post-translational Modifications of Human Desmoplakins in Regulating Interactions with Intermediate Filaments*. PhD thesis, University of Virginia, 2014.
- [247] Y. Zhang, K. Trabbic-Carlson, F. Albertorio, A. Chilkoti, and P. S. Cremer. Aqueous two-phase system formation kinetics for elastin-like polypeptides of varying chain length. *Biomacromolecules*, 7(7):2192–2199, 2006.
- [248] B. Zhao, N. K. Li, Y. G. Yingling, and C. K. Hall. LCST behavior is manifested in a single molecule: Elastin-like polypeptide (VPGVG)n. *Biomacromolecules*, 17(1):111–118, 2016.
- [249] X.-Y. Zhong, P. Wang, J. Han, M. G. Rosenfeld, and X.-D. Fu. SR proteins in vertical integration of gene expression from transcription to RNA processing to translation. *Molecular Cell*, 35(1):1 – 10, 2009.
- [250] R. Zhou and B. J. Berne. Can a continuum solvent model reproduce the free energy landscape of a β -hairpin folding in water? *Proc. Natl. Acad. Sci. U.S.A.*, 99(20):12777–12782, Oct 2002.

Chapter S2

Supplementary Information for Desmoplakin

S2.1 Overview

This document provides the detailed results of the analysis suite described in the Methods section of the main text, as applied to each of our simulation systems. Each trajectory has been analyzed in terms of (a) Cy_*^{R} , (b) Cy^{R} , (c) solvent-accessible surface area (SASA) of S2849, (d) SASA of R2834, (e) the S2849-R2834 distance, (f) glycogen synthase kinase 3 (GSK3) steric clash scores, (g) inter-residue contact maps, and (h) the distribution of peptide backbone torsion angles (ϕ, ψ) .

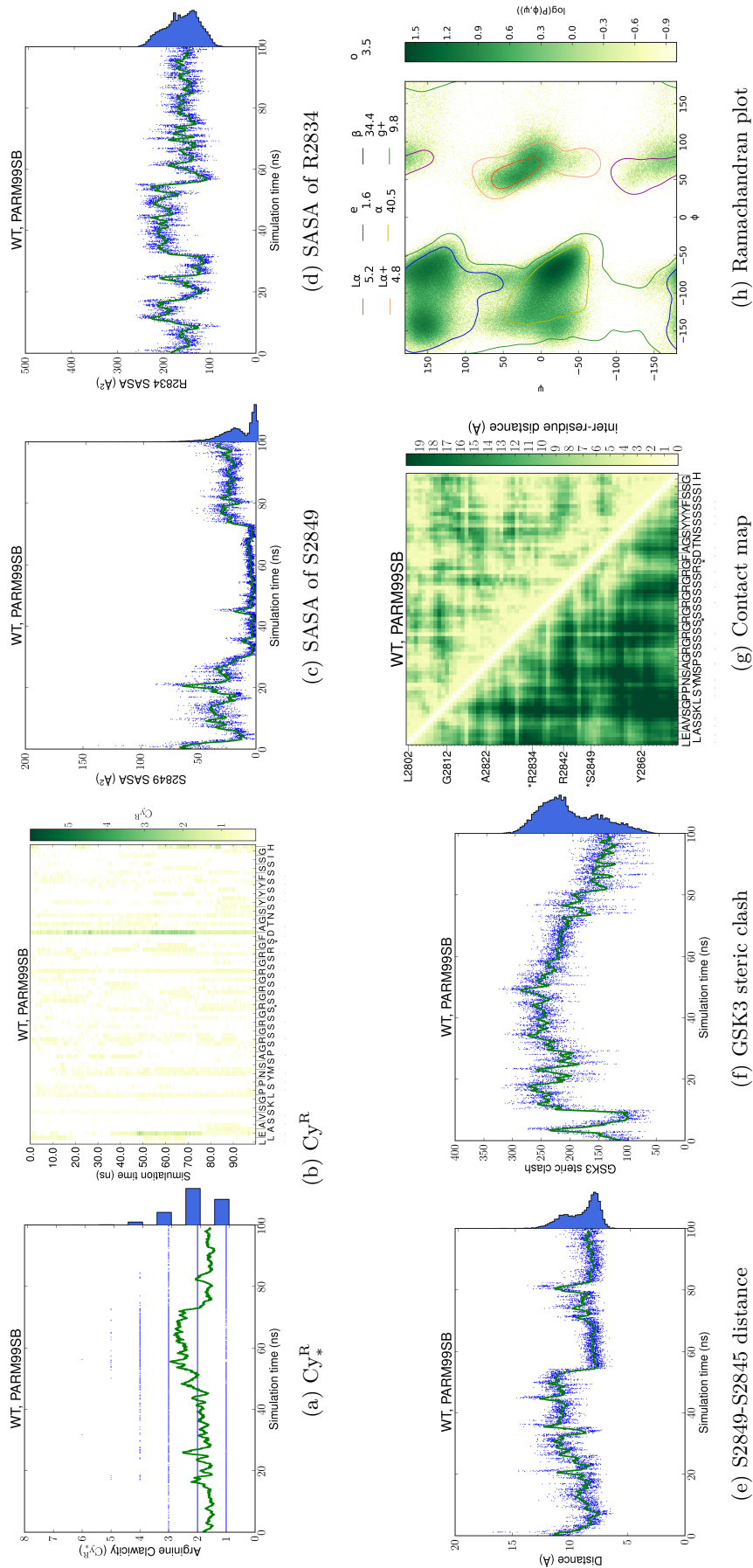


Figure S2.1: Behavior of **WT_PARM99SB**: Time-series plots mark each observation as a blue point and contain a 1-ns running average as a green trace, and the marginal distributions are shown on the right axis, in each of panels a, c, d, e, and f.

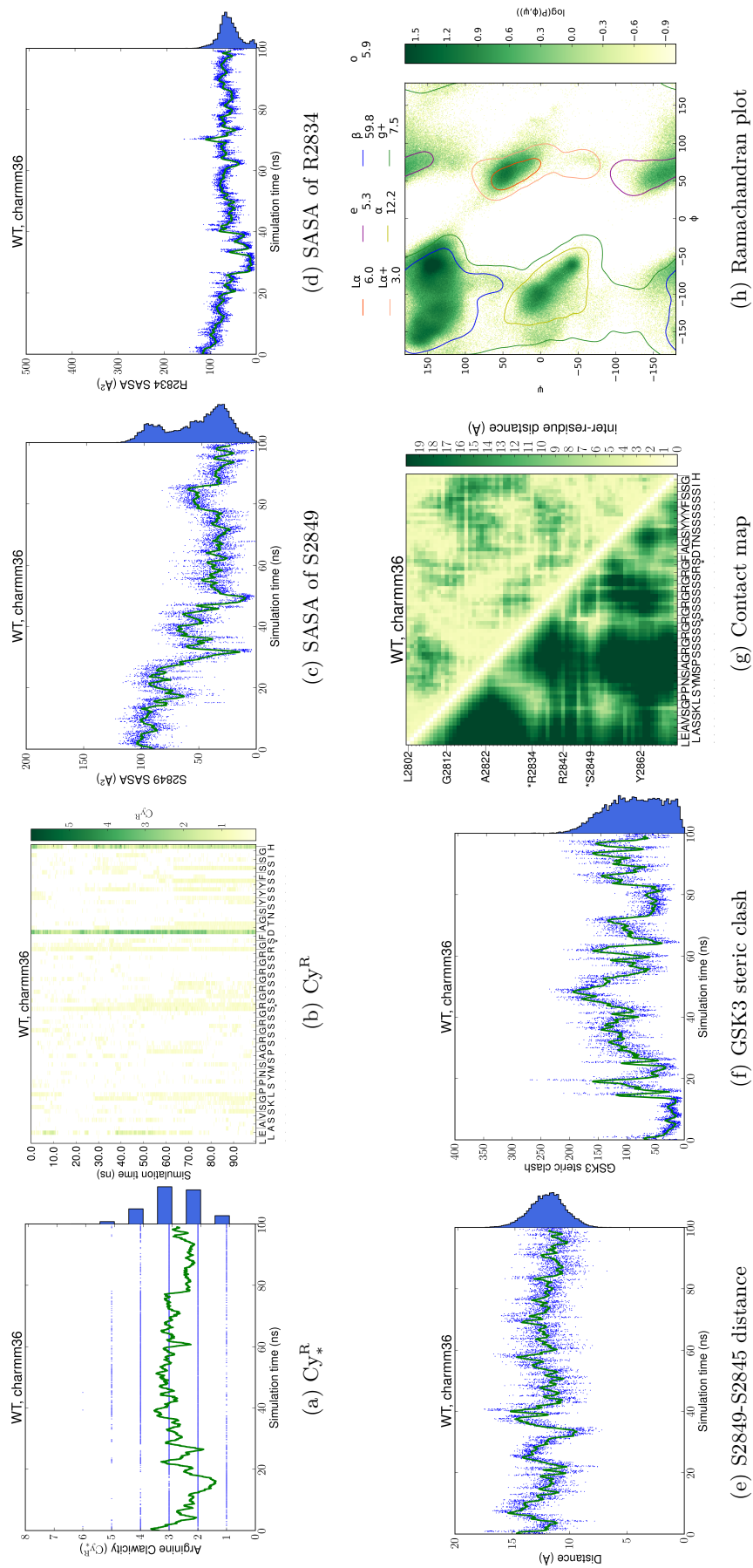


Figure S2.2: Behavior of **WT_CHARMM36**: Time-series plots mark each observation as a blue point and contain a 1-ns running average as a green trace, and the marginal distributions are shown on the right axis, in each of panels a, c, d, e, and f.

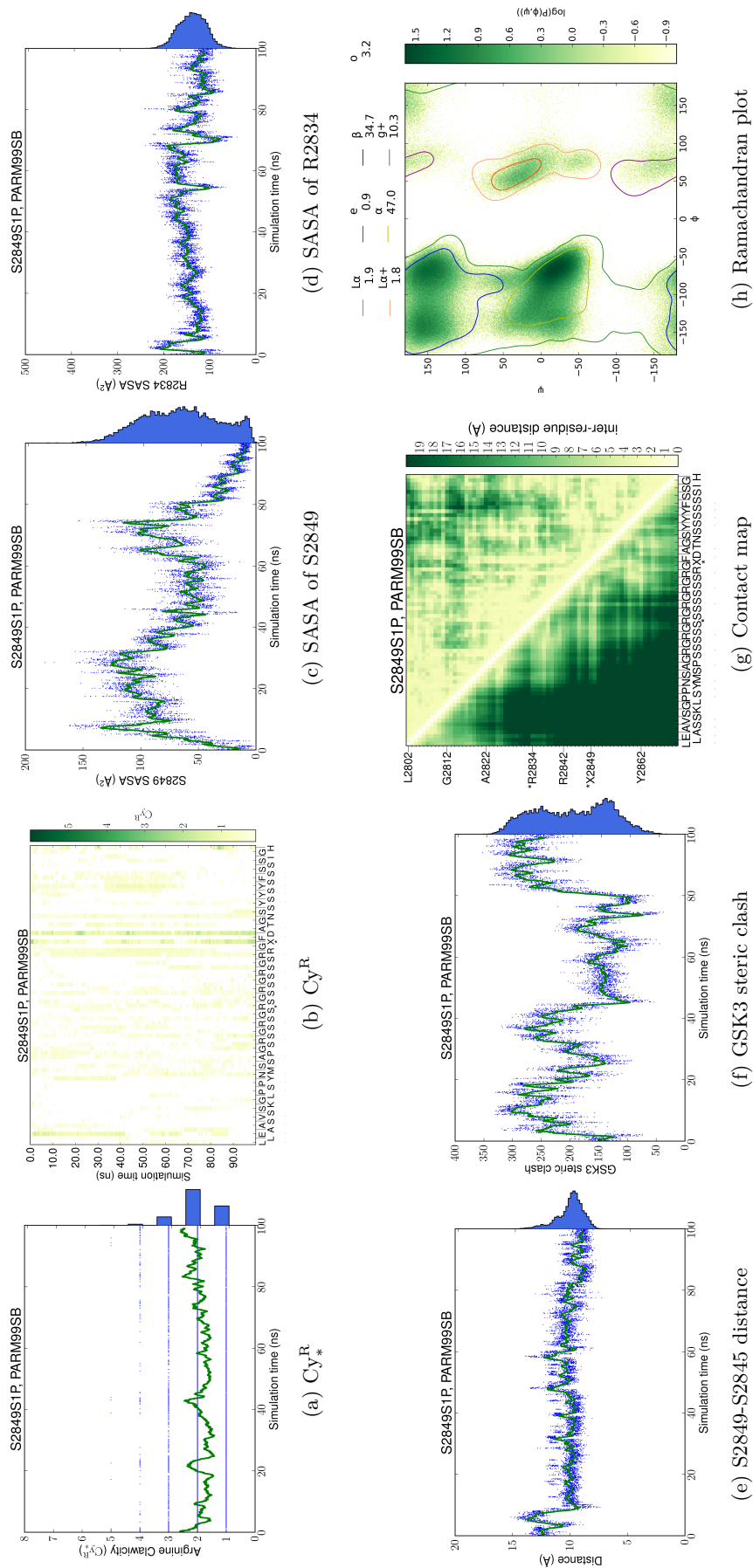


Figure S2.3: Behavior of **S2849S1P_PARM99SB**: Time-series plots mark each observation as a blue point and contain a 1-ns running average as a green trace, and the marginal distributions are shown on the right axis, in each of panels a, c, d, e, and f.

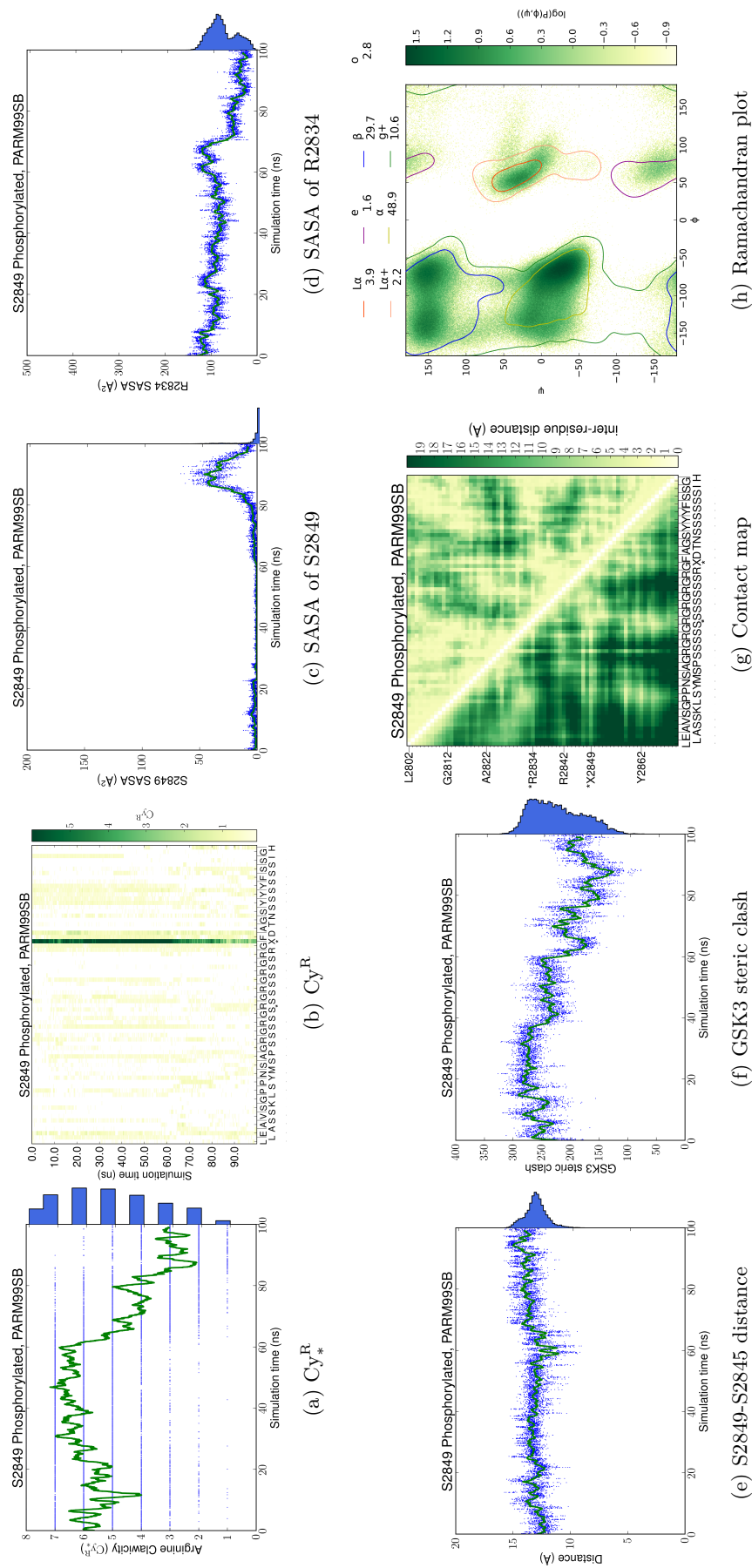


Figure S2.4: Behavior of **S2849S2P_PARM99SB**: Time-series plots mark each observation as a blue point and contain a 1-ns running average as a green trace, and the marginal distributions are shown on the right axis, in each of panels a, c, d, e, and f.

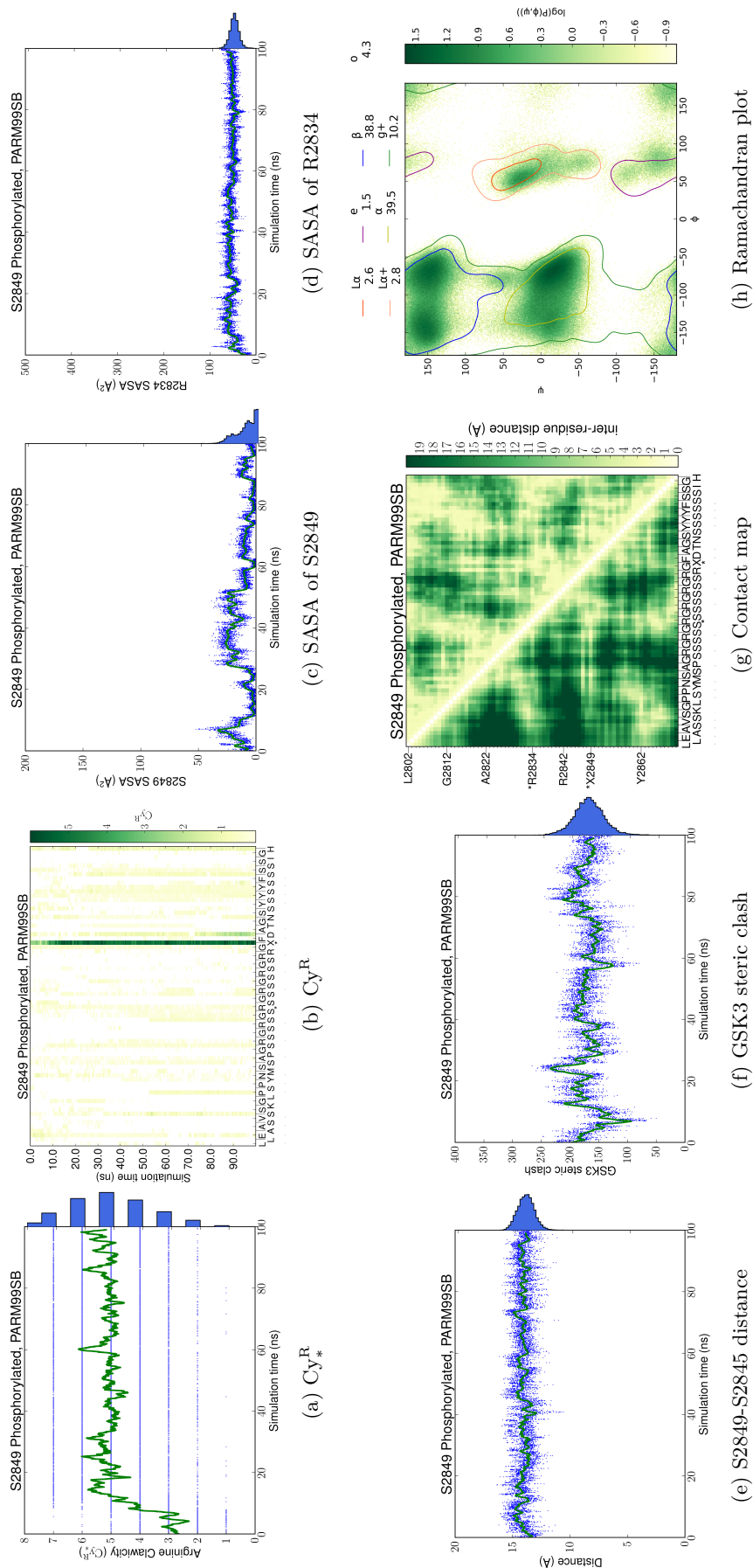


Figure S2.5: Behavior of **S2849S2P_PARM99SB_cycle2**: Time-series plots mark each observation as a blue point and contain a 1-ns running average as a green trace, and the marginal distributions are shown on the right axis, in each of panels a, c, d, e, and f.

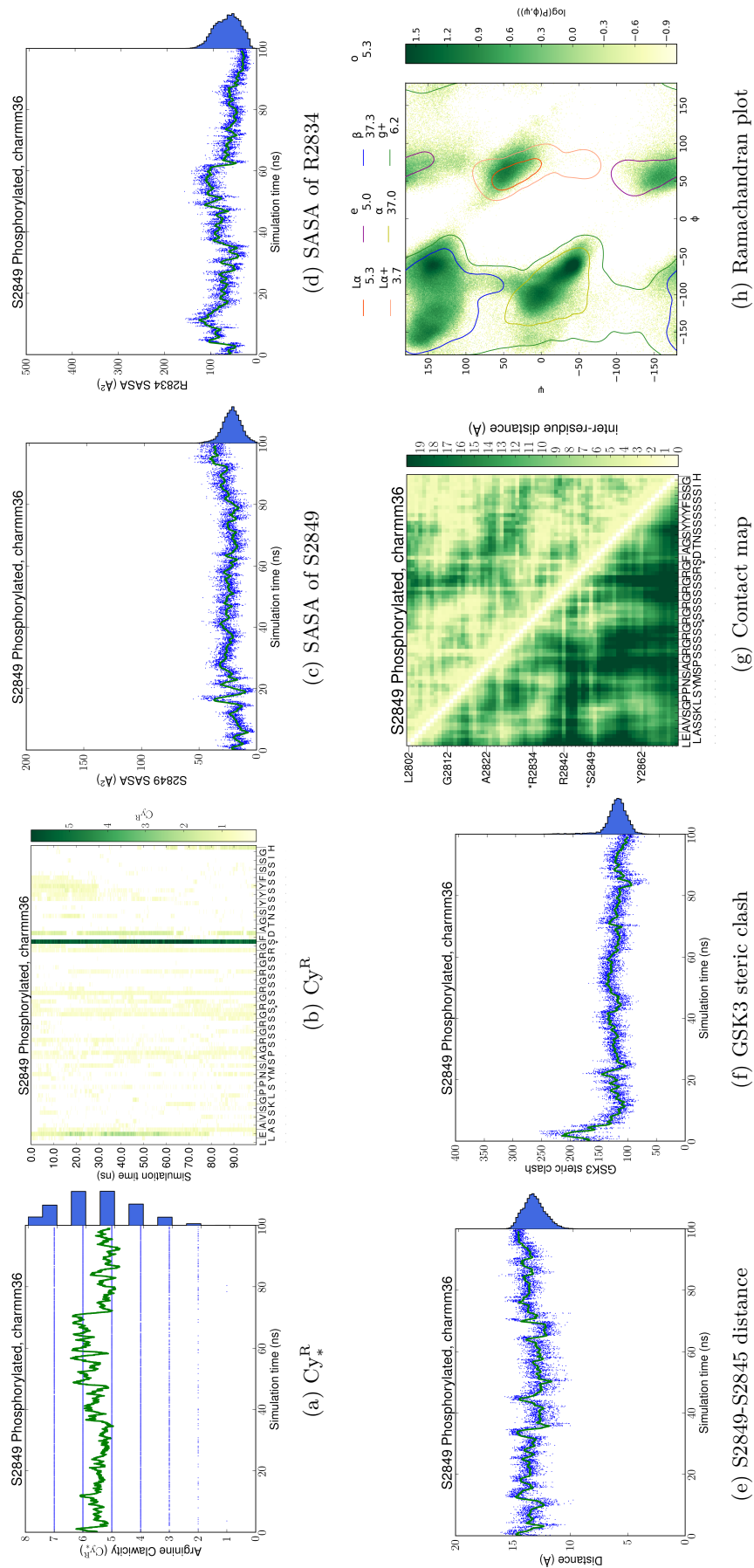


Figure S2.6: Behavior of **S2849S2P_CHARM36**: Time-series plots mark each observation as a blue point and contain a 1-ns running average as a green trace, and the marginal distributions are shown on the right axis, in each of panels a, c, d, e, and f.

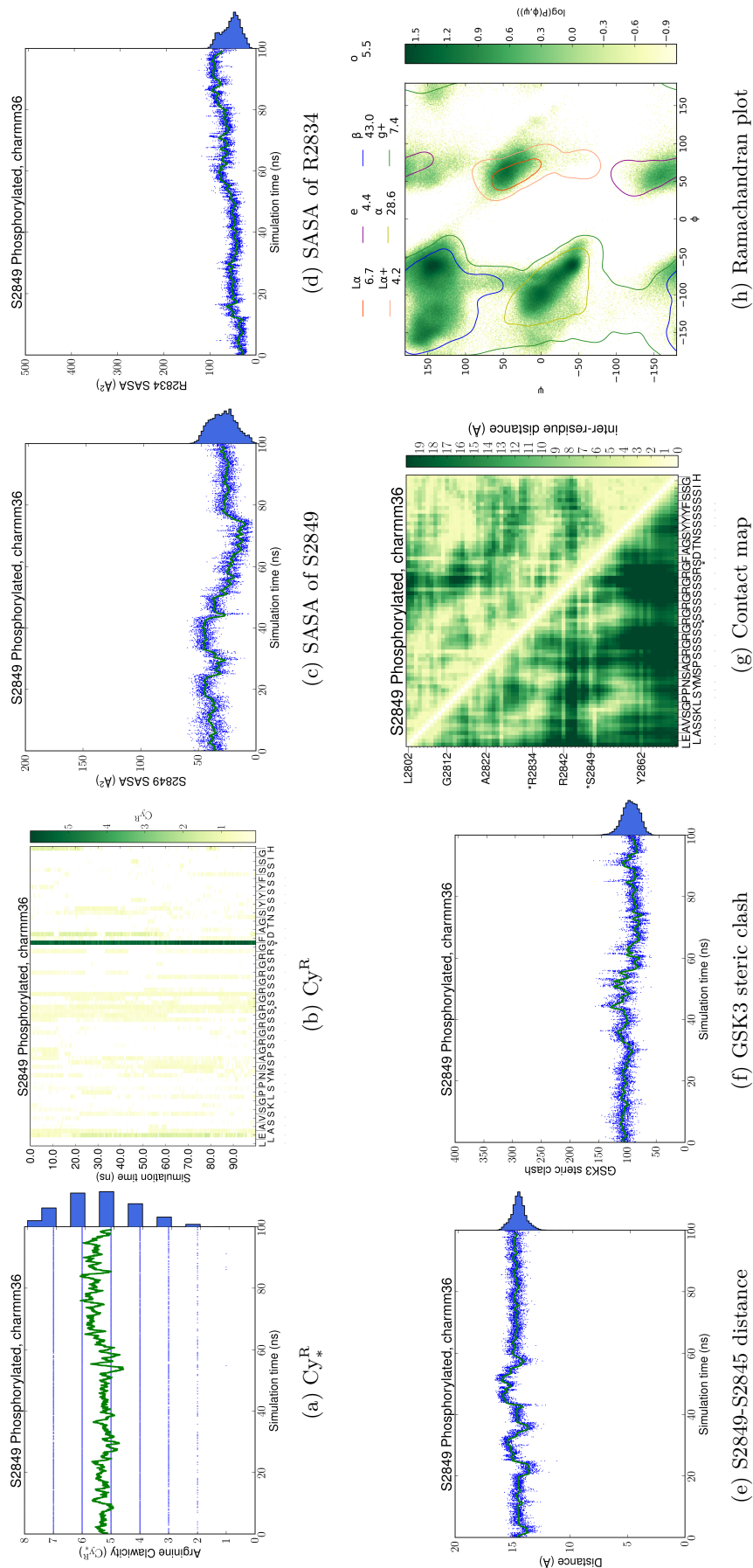


Figure S2.7: Behavior of **S2849S2P_CHARMM36_cycle2**: Time-series plots mark each observation as a blue point and contain a 1-ns running average as a green trace, and the marginal distributions are shown on the right axis, in each of panels a, c, d, e, and f.

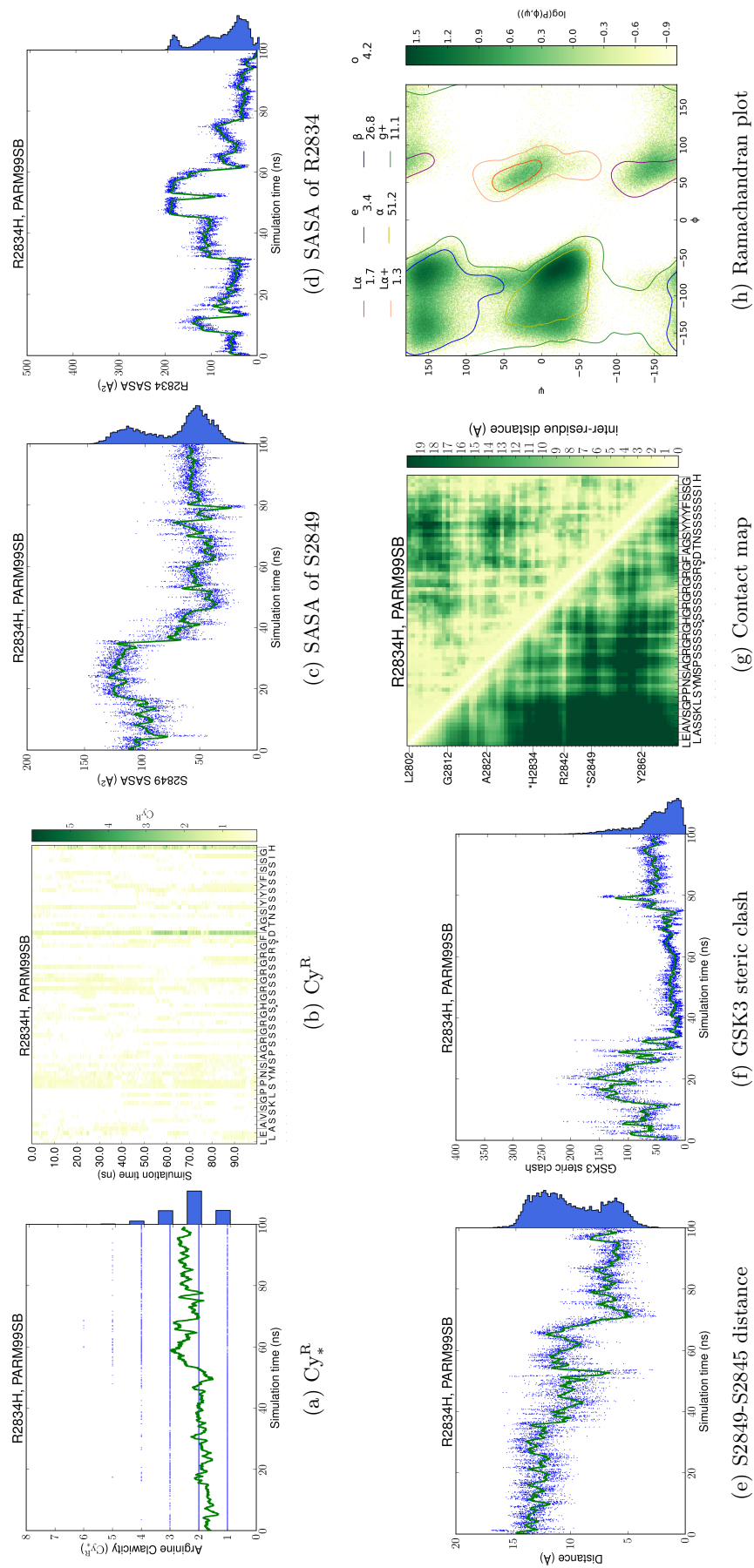


Figure S2.8: Behavior of **R2834H_PARM99SB**: Time-series plots mark each observation as a blue point and contain a 1-ns running average as a green trace, and the marginal distributions are shown on the right axis, in each of panels a, c, d, e, and f.

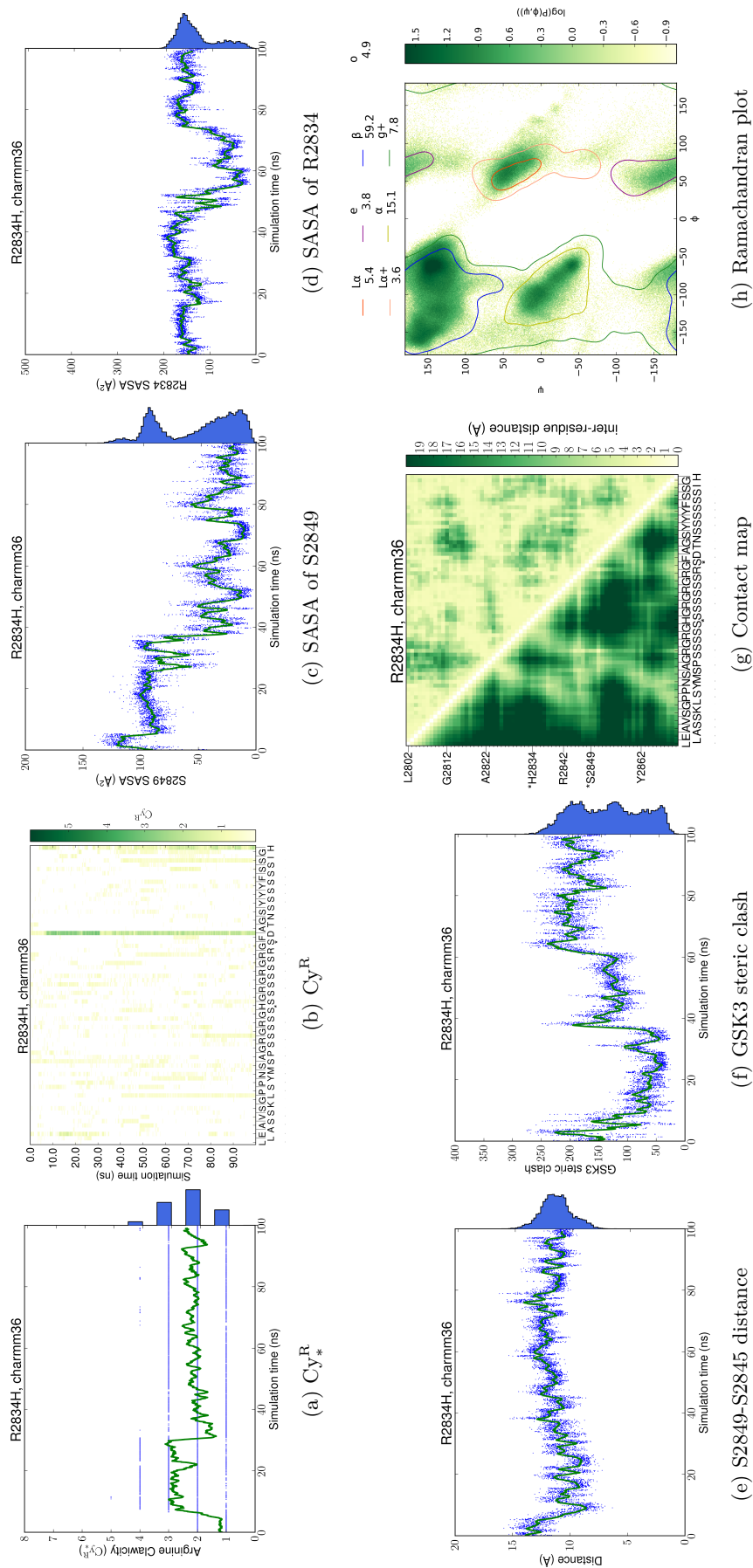


Figure S2.9: Behavior of **R2834H-CHARMM36**: Time-series plots mark each observation as a blue point and contain a 1-ns running average as a green trace, and the marginal distributions are shown on the right axis, in each of panels a, c, d, e, and f.

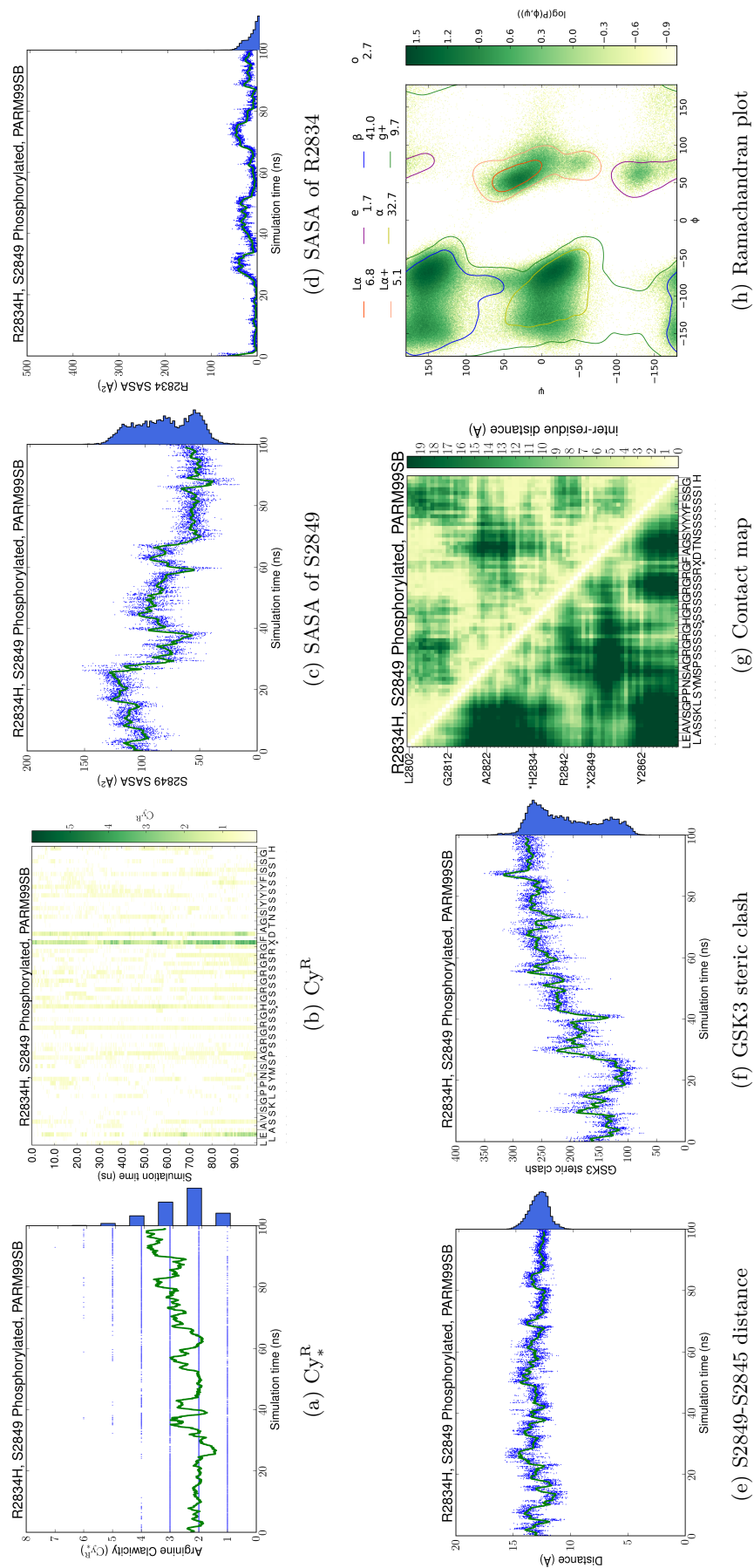


Figure S2.10: Behavior of **R2834H_S2849S2P_PARM99SB**: Time-series plots mark each observation as a blue point and contain a 1-ns running average as a green trace, and the marginal distributions are shown on the right axis, in each of panels a, c, d, e, and f.

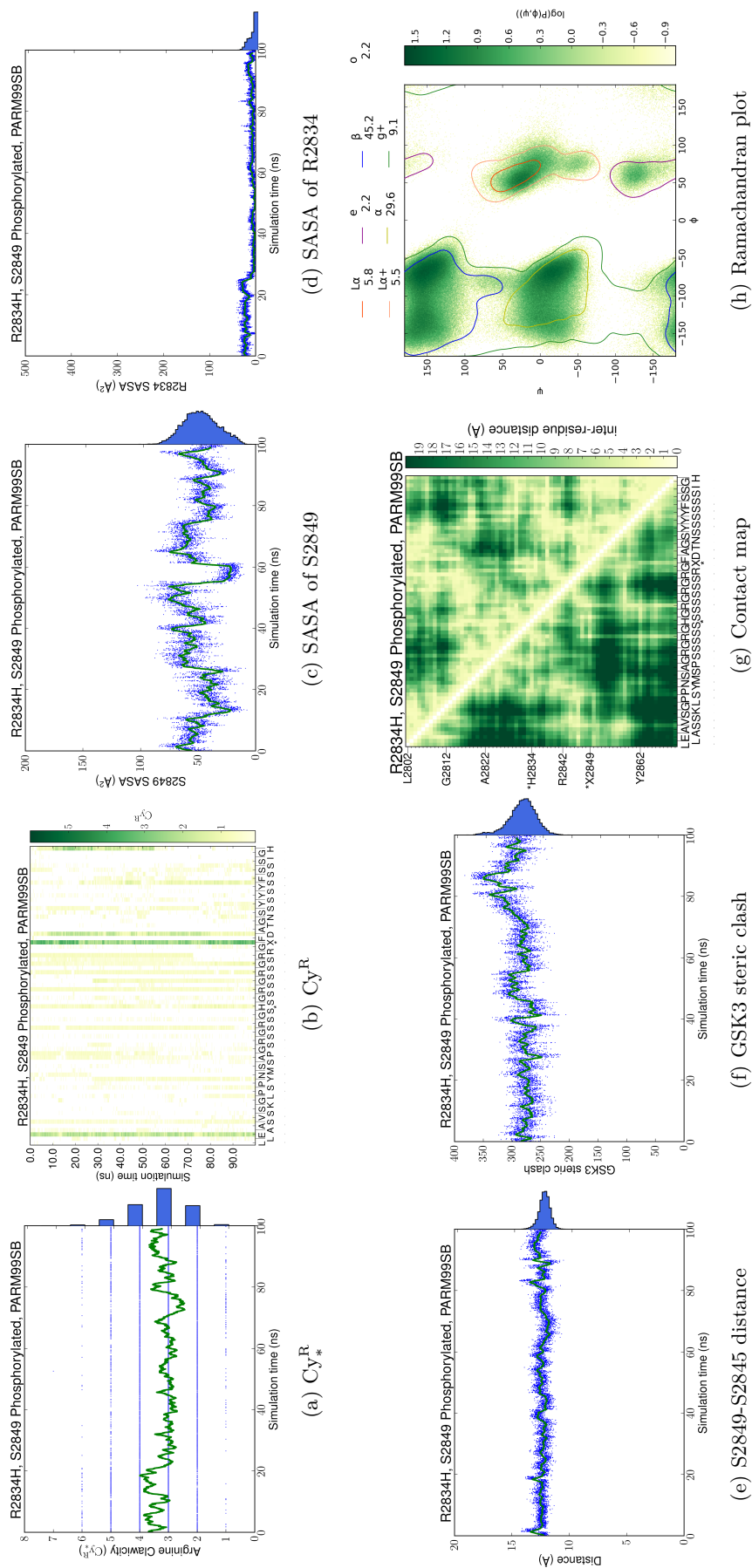


Figure S2.11: Behavior of **R2834H_S2849S2P_PARM99SB_cycle2**: Time-series plots mark each observation as a blue point and contain a 1-ns running average as a green trace, and the marginal distributions are shown on the right axis, in each of panels a, c, d, e, and f.

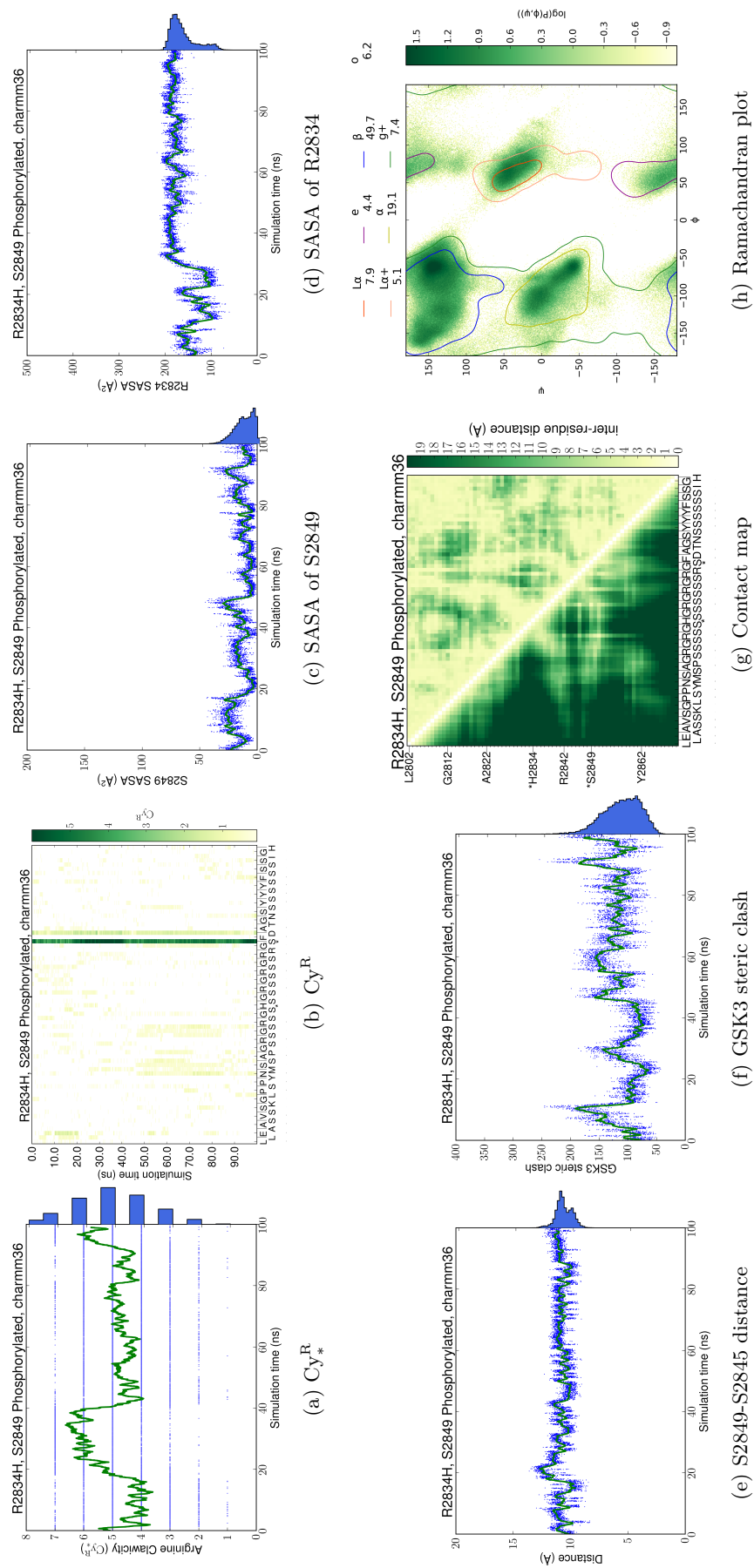


Figure S2.12: Behavior of **R2834H_S2849S2P_CHARMM36**: Time-series plots mark each observation as a blue point and contain a 1-ns running average as a green trace, and the marginal distributions are shown on the right axis, in each of panels a, c, d, e, and f.

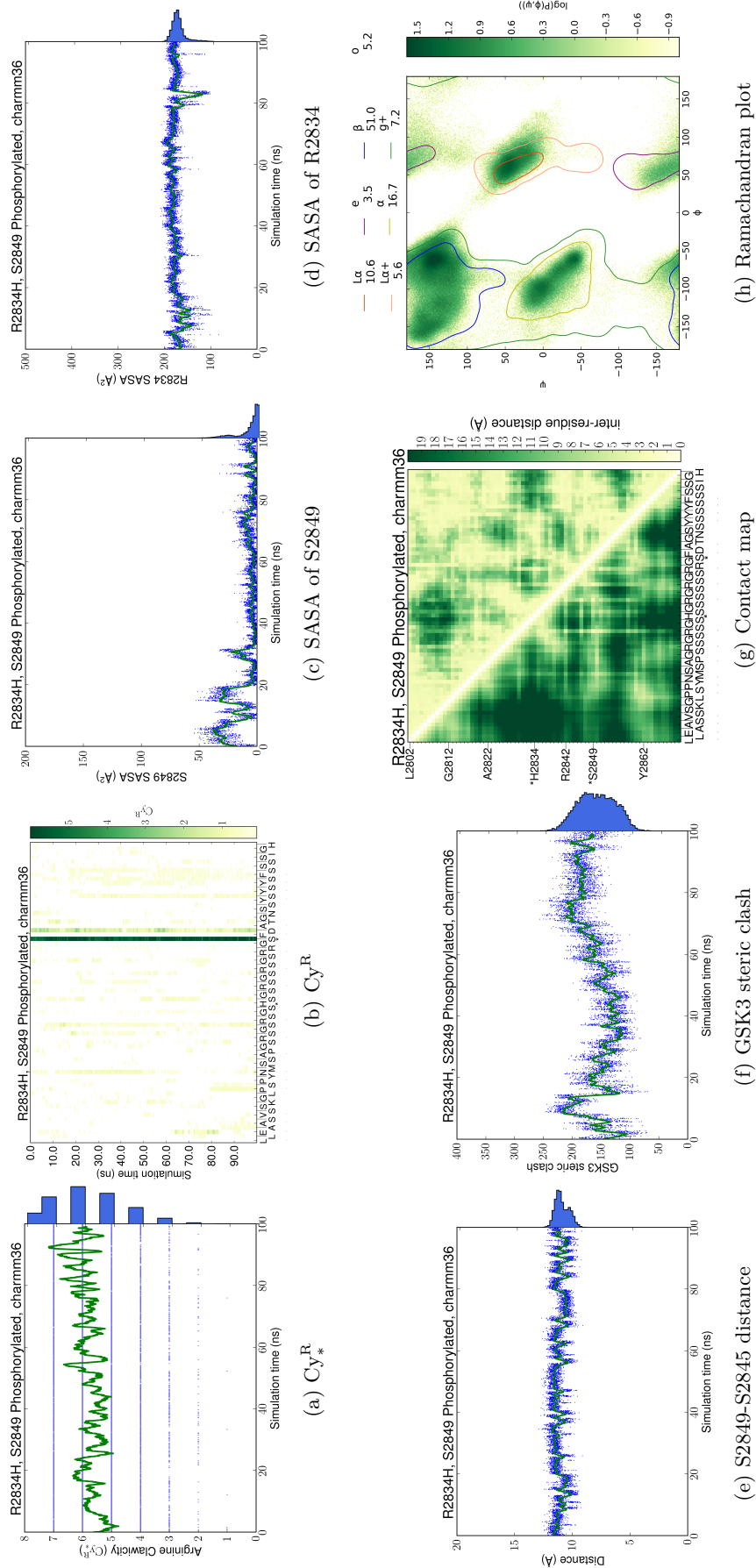


Figure S2.13: Behavior of **R2834H_S2849S2P_CHARM36_cycle2**: Time-series plots mark each observation as a blue point and contain a 1-ns running average as a green trace, and the marginal distributions are shown on the right axis, in each of panels a, c, d, e, and f.

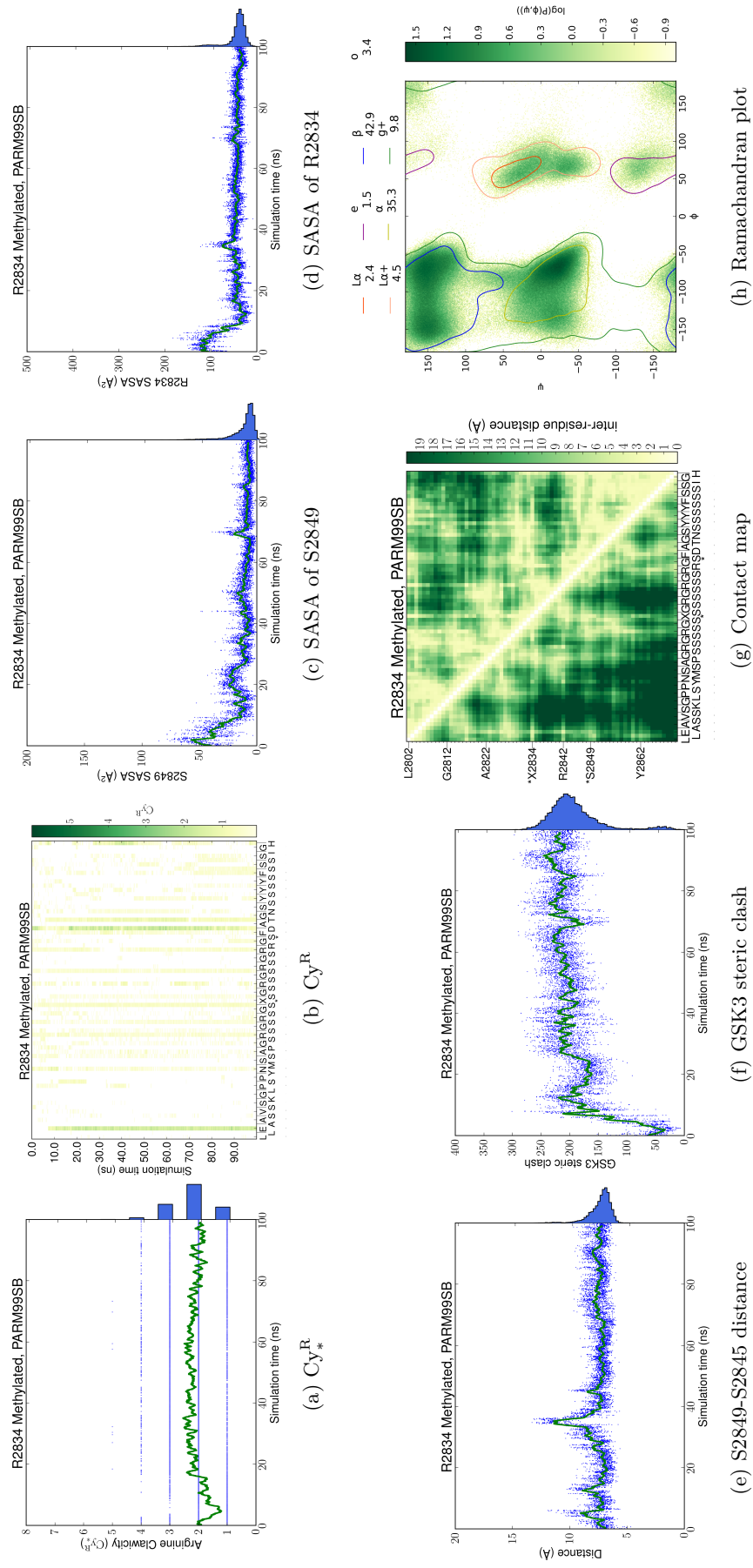


Figure S2.14: Behavior of **R2834MeMe_PARM99SB**: Time-series plots mark each observation as a blue point and contain a 1-ns running average as a green trace, and the marginal distributions are shown on the right axis, in each of panels a, c, d, e, and f.

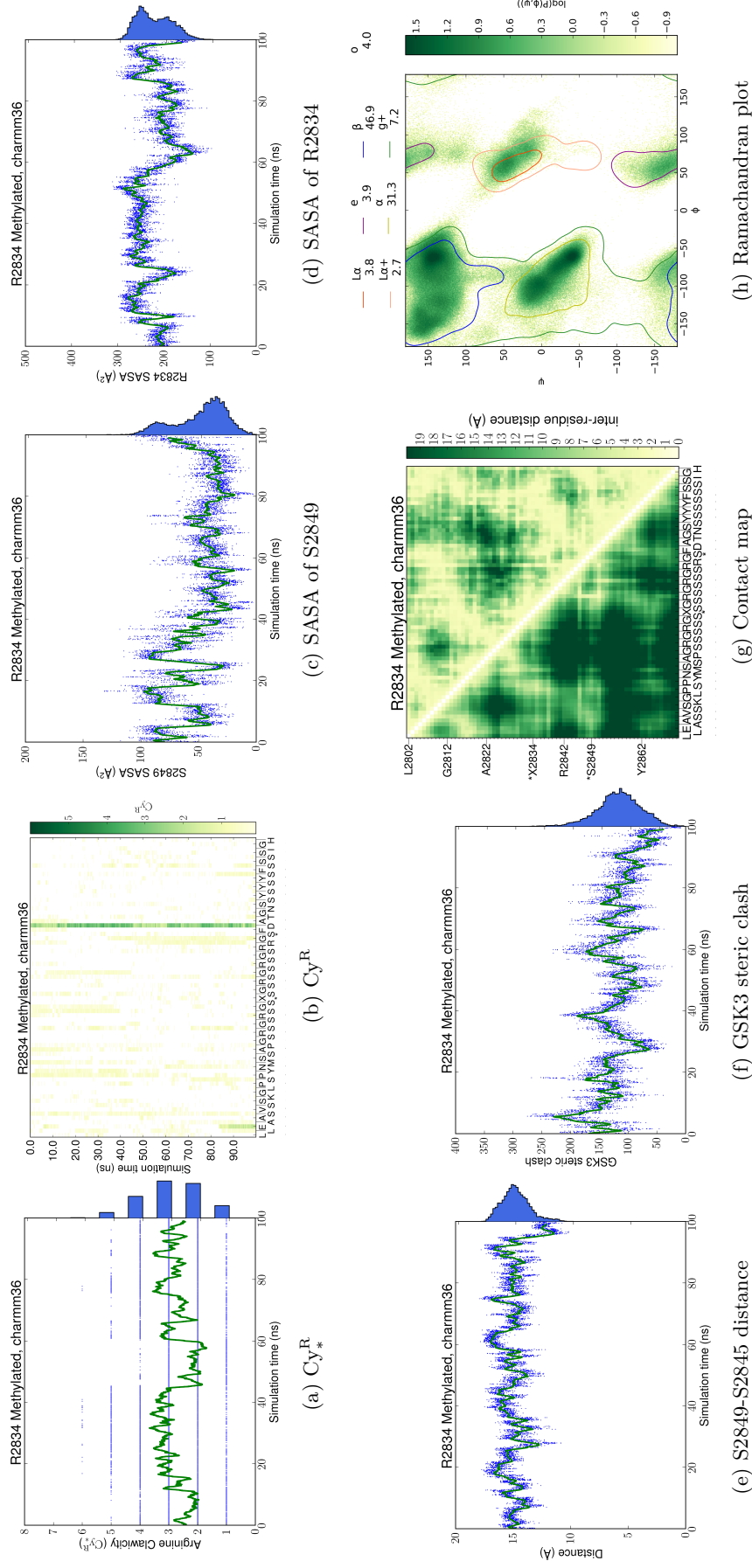


Figure S2.15: Behavior of **R2834MeMe-CHARMM36**: Time-series plots mark each observation as a blue point and contain a 1-ns running average as a green trace, and the marginal distributions are shown on the right axis, in each of panels a, c, d, e, and f.

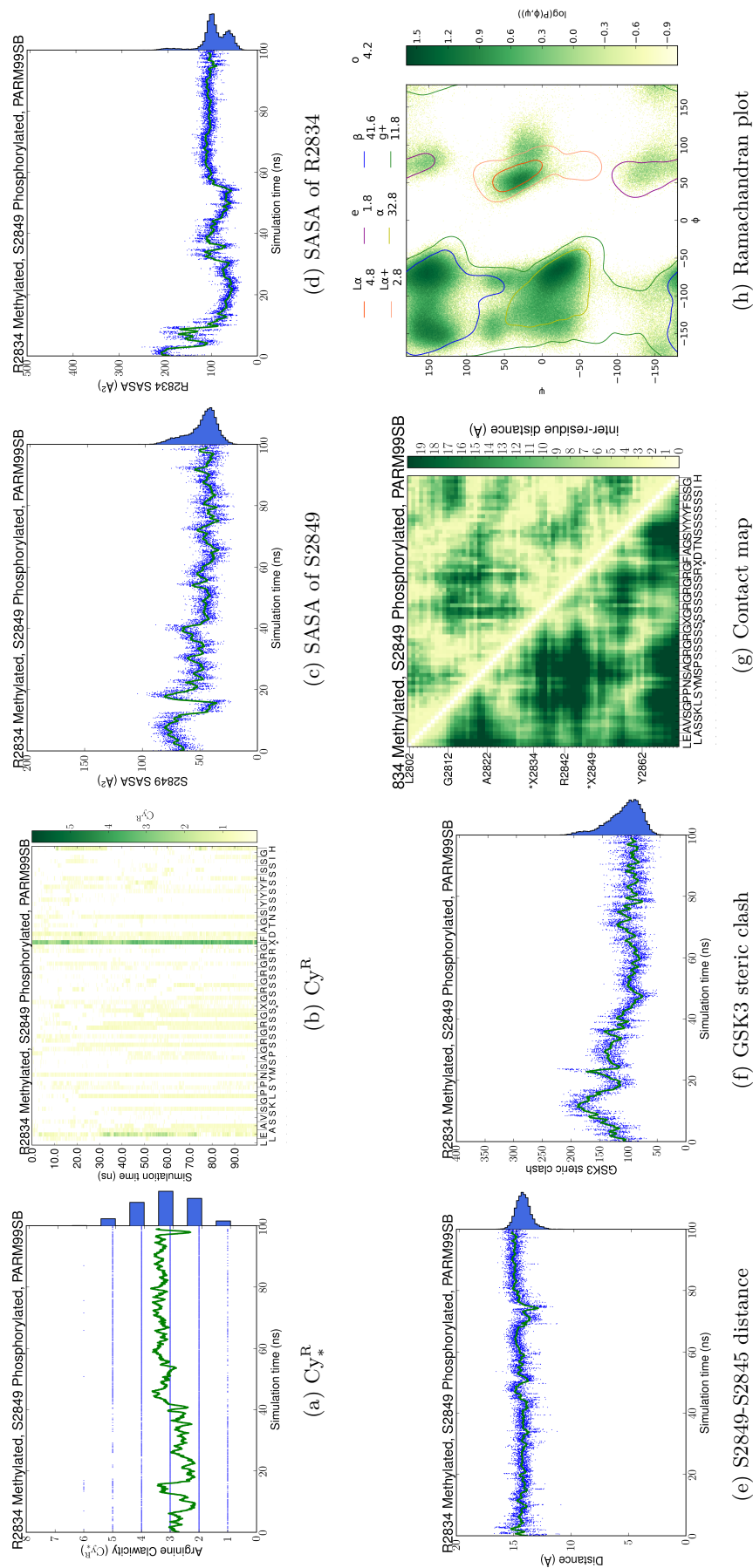


Figure S2.16: Behavior of **R2834MeMe-S2849S2P_PARM99SB**: Time-series plots mark each observation as a blue point and contain a 1-ns running average as a green trace, and the marginal distributions are shown on the right axis, in each of panels a, c, d, e, and f.

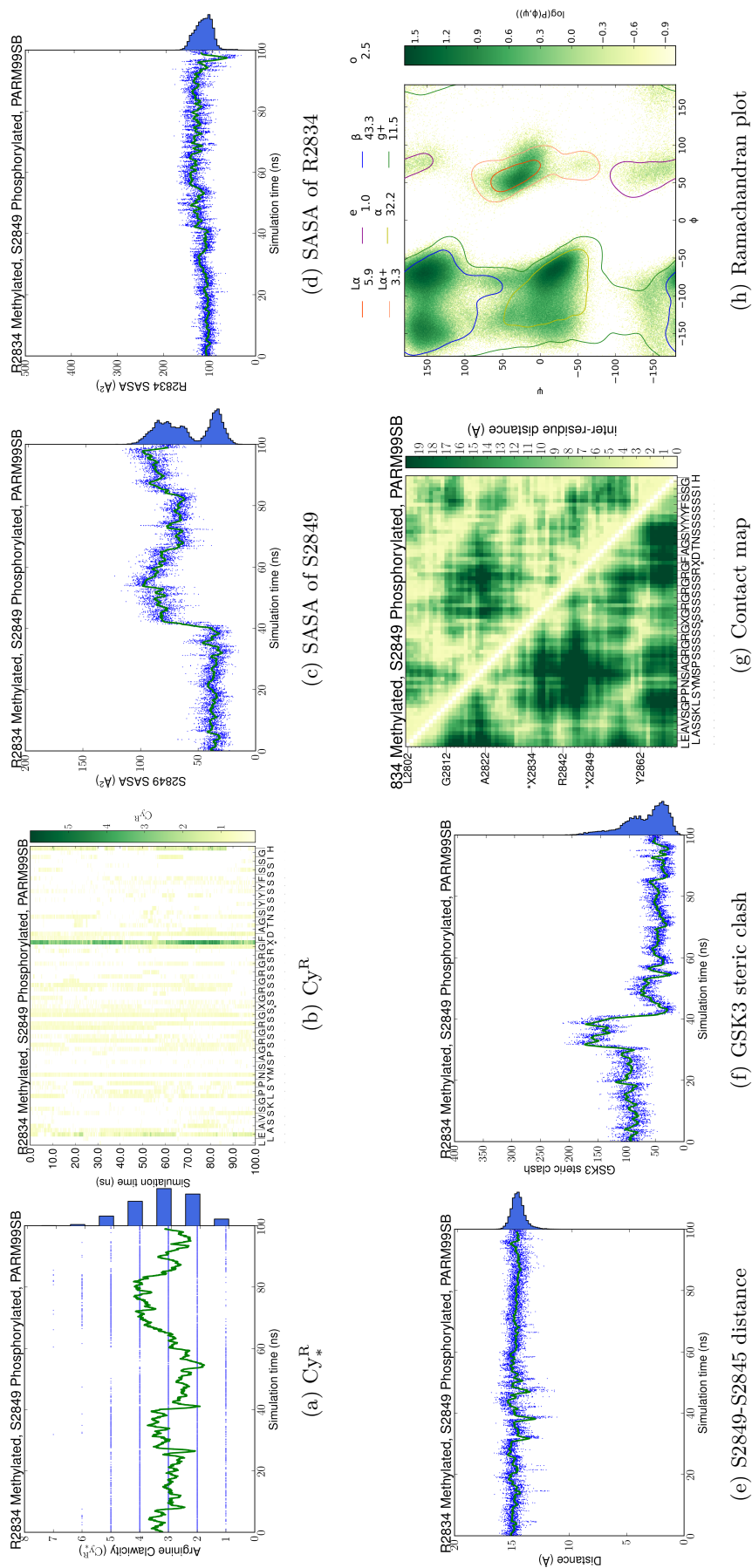


Figure S2.17: Behavior of **R2834MeMe_S2849S2P_PARM99SB_cycle2**: Time-series plots mark each observation as a blue point and contain a 1-ns running average as a green trace, and the marginal distributions are shown on the right axis, in each of panels a, c, d, e, and f.

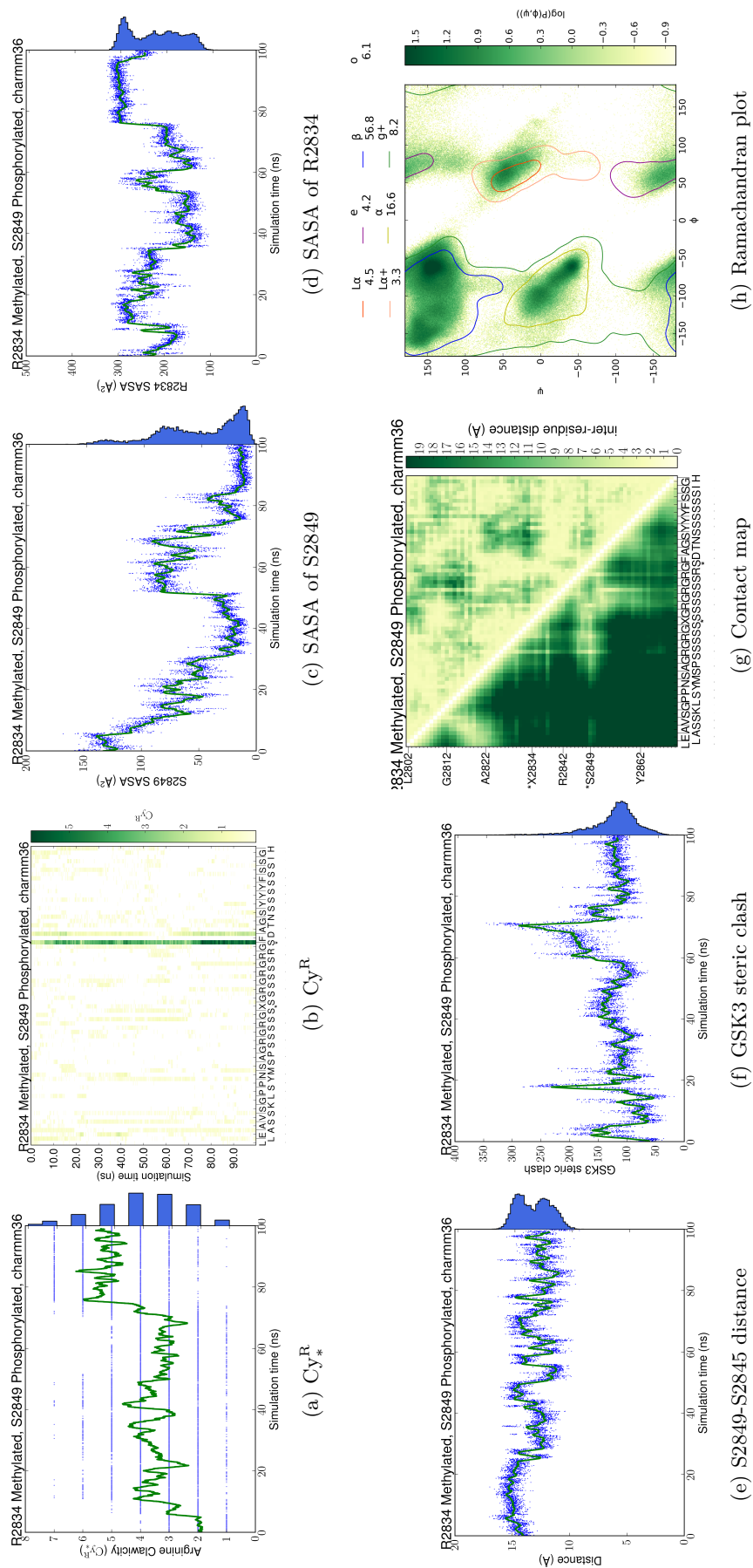


Figure S2.18: Behavior of **R2834MeMe_S2849S2P_CHARMM36**: Time-series plots mark each observation as a blue point and contain a 1-ns running average as a green trace, and the marginal distributions are shown on the right axis, in each of panels a, c, d, e, and f.

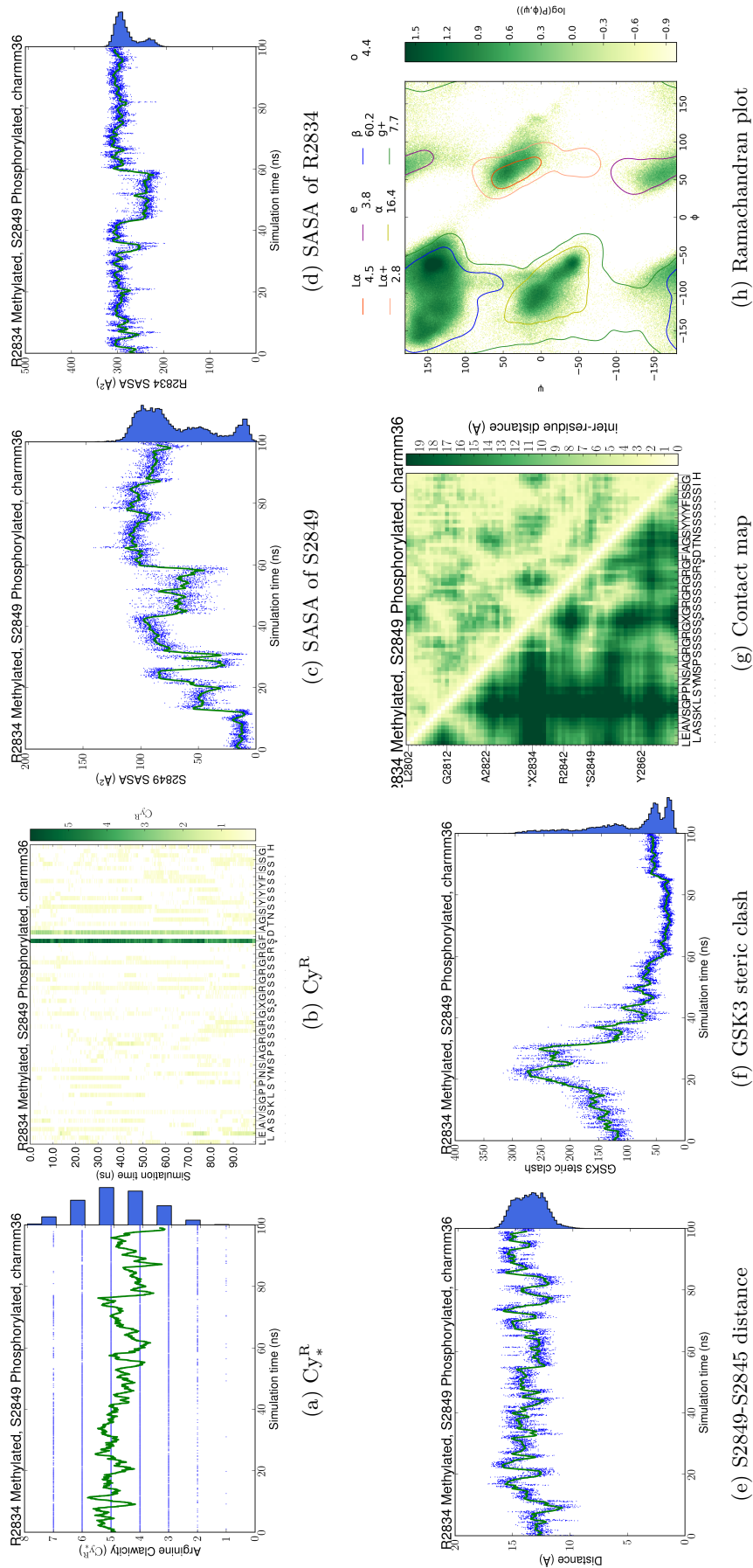


Figure S2.19: Behavior of **R2834MeMe_S2849S2P_CHARMM36_cycle2**: Time-series plots mark each observation as a blue point and contain a 1-ns running average as a green trace, and the marginal distributions are shown on the right axis, in each of panels a, c, d, e, and f.

Chapter S3

Supplementary Information for LG-ELP

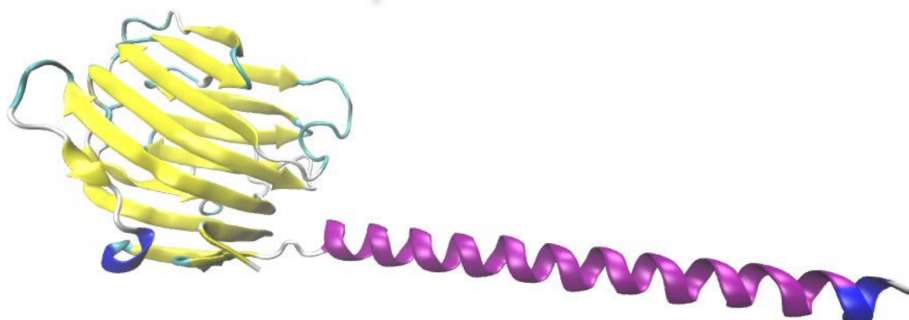


Figure S3.1: **Initial starting structure of the LG-ELP protein.** The N-terminal LG5 domain was drawn from the crystal structure of the mouse homolog of the laminin $\alpha 2$ chain (1DYK), while the C-terminal ELP domain was built using Avogadro's peptide builder, and was modelled as a canonical α -helix starting structure, with backbone torsion angles $\phi = -60^\circ$, $\psi = -40^\circ$.

Table S3.1: **Summary of MD simulation systems of the engineered LG-ELP fusion protein.** Atomistic MD simulations were performed using the NAMD 2.9 code and the CHARMM36 force-field for the protein system. The protein was solvated in explicit water with periodic boundary conditions and simulated as described in the Methods section.

System	Trajectory type, length
LG-ELP	Solvation, Minimization
LG-ELP	Equilibration, 10 ns
LG-ELP 290 K	Production, 100 ns
LG-ELP 295 K	Production, 100 ns
LG-ELP 300 K	Production, 100 ns
LG-ELP 305 K	Production, 100 ns
LG-ELP 310 K	Production, 200 ns
LG-ELP 315 K	Production, 100 ns
LG-ELP 320 K	Production, 100 ns
LG-ELP 310K at 290 K	Production, 40 ns
LG-ELP 310K at 300 K	Production, 40 ns
LG-ELP 310K at 320 K	Production, 40 ns

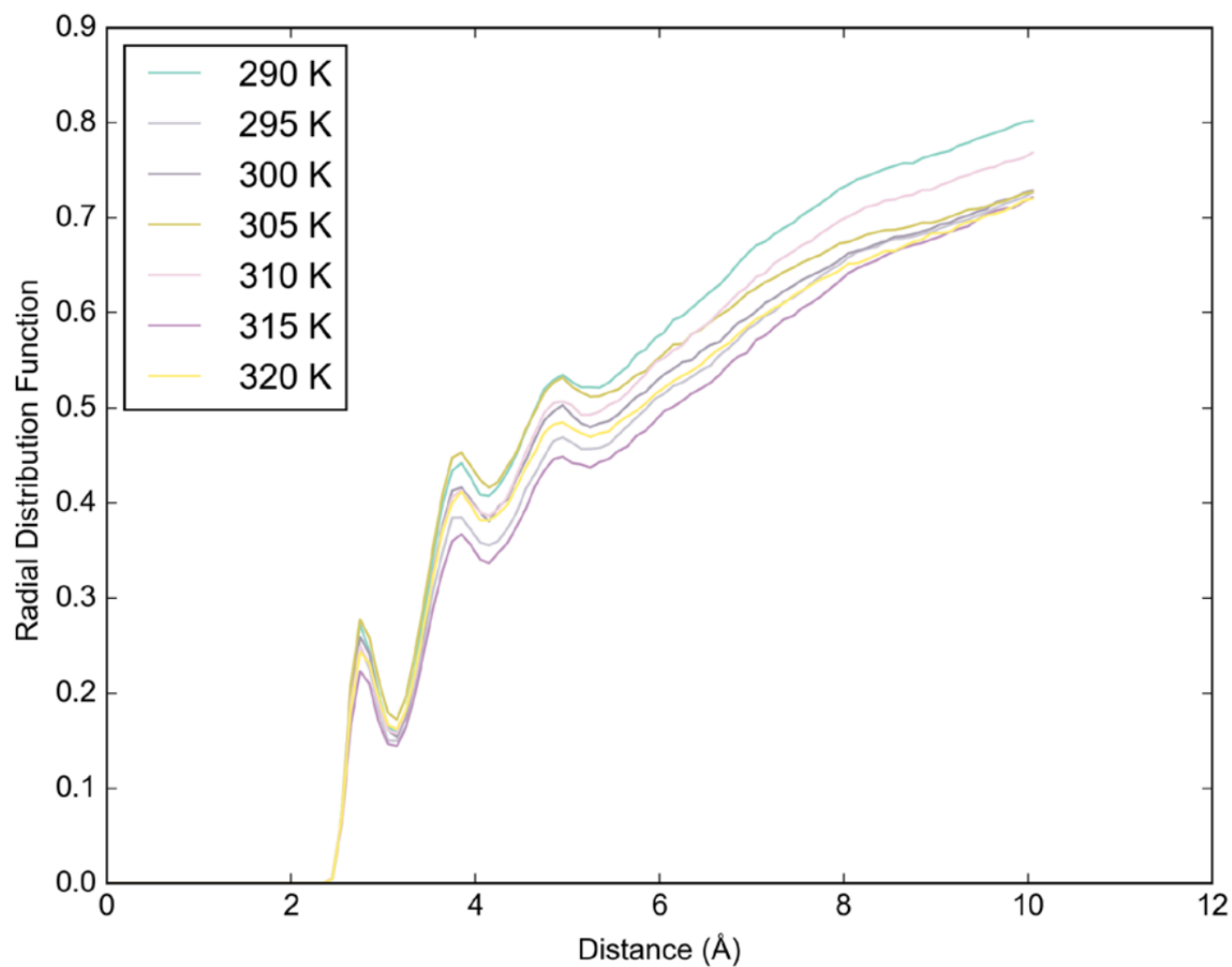


Figure S3.2: **RDF of oxygen atoms around the ELP backbone.** The RDF is plotted as a function of temperature, and the first hydration shell was chosen to be the minimum for subsequent analysis in determining the number of surrounding water molecules (Figure 3.7a).

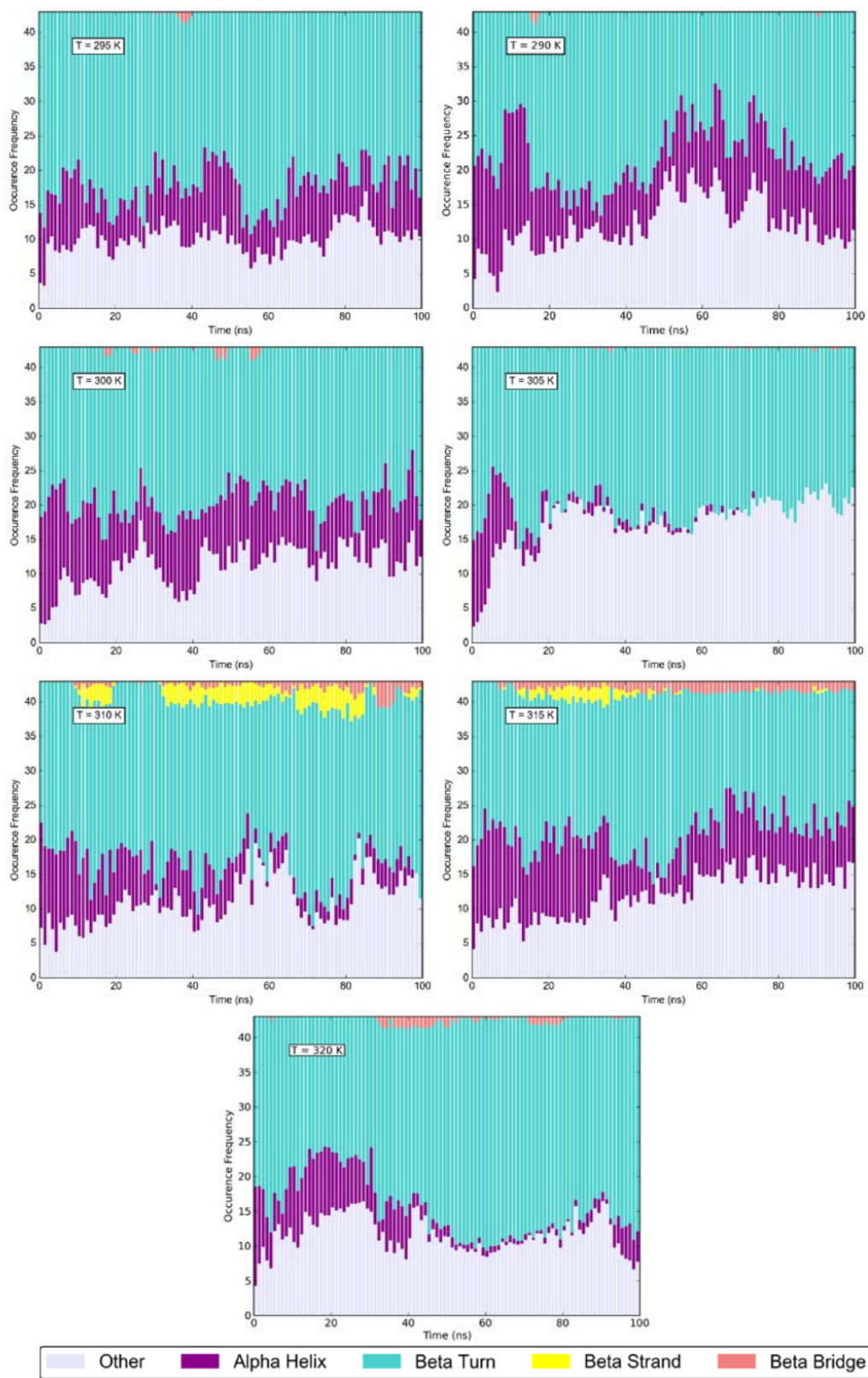


Figure S3.3: **Secondary structural content across a range of temperatures as a function of time.** Simulated systems were sampled at different temperatures (five degree increments from 290 K to 315 K).

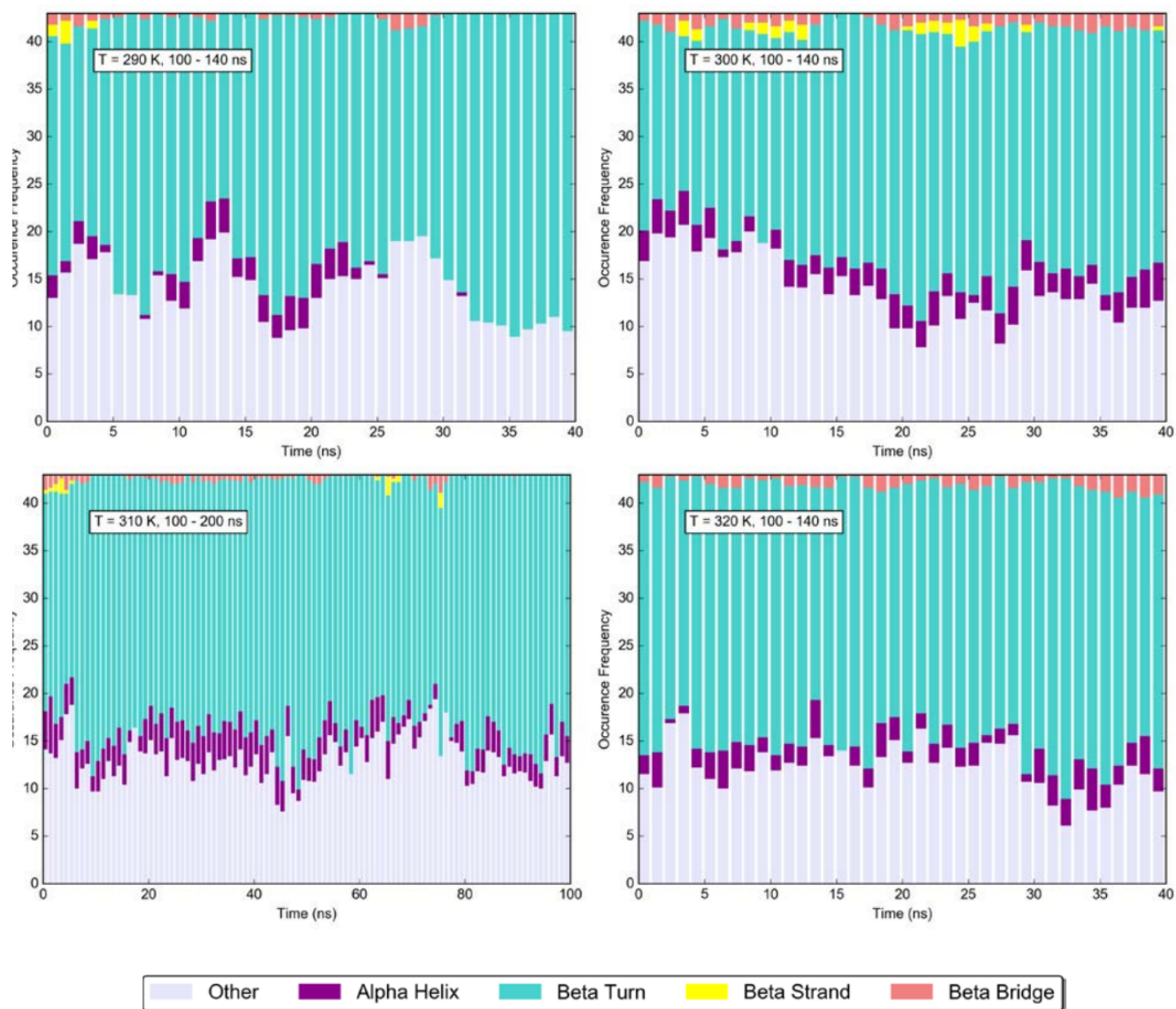


Figure S3.4: **Secondary structural content in extended simulations across a range of temperatures as a function of time** (100 - 140 ns for 290 K, 300 K, and 320 K and 100 - 200 ns for 310 K)

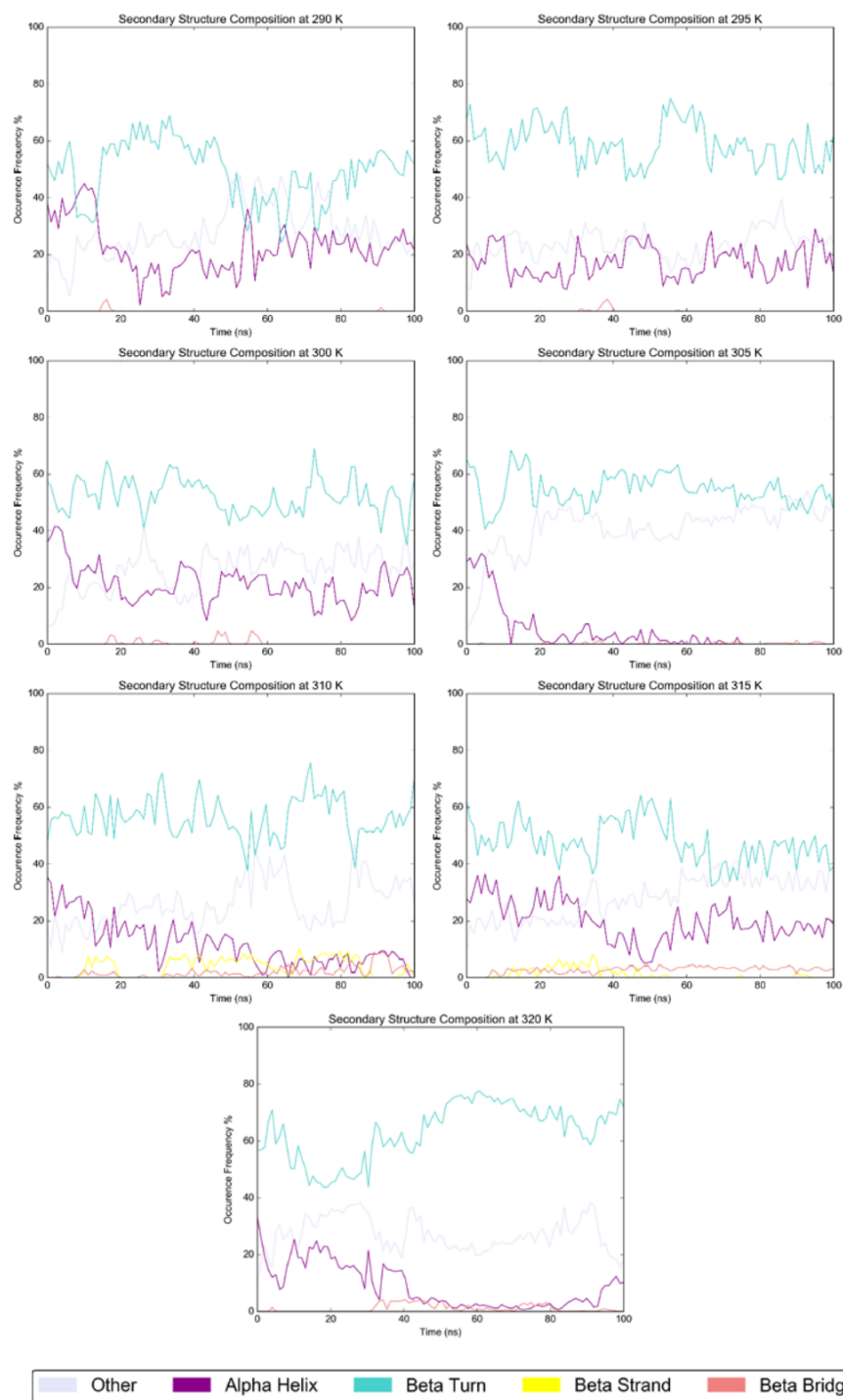


Figure S3.5: **Frequency of occurrence for secondary structural content** across a range of temperatures as a function of time.

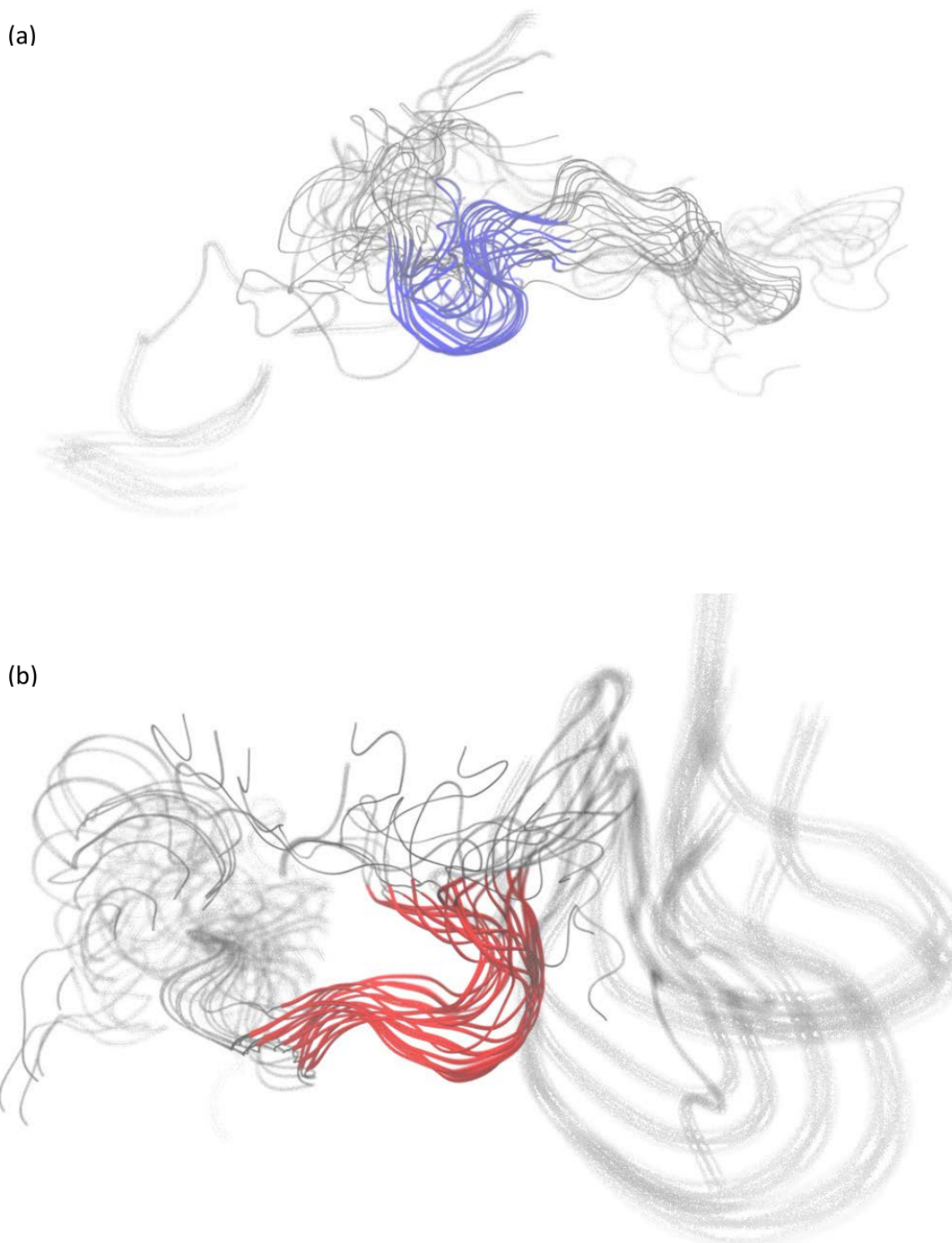


Figure S3.6: **The time evolution of the secondary structure ensemble** in the extended simulation at 310 K showed four distinct regions of persistent β -sheet like conformations, (a) Leu4 – Gly_{5200–201} – Leu4 – Gly_{5205–206} (highlighted by blue ribbons) and (b) Ile4 – Gly_{5210–211} – Ile4 – Gly_{5214–215} (red ribbons) with reduced conformational flexibility. The trajectory was rendered overlaying multiple frames at 10-ns intervals.

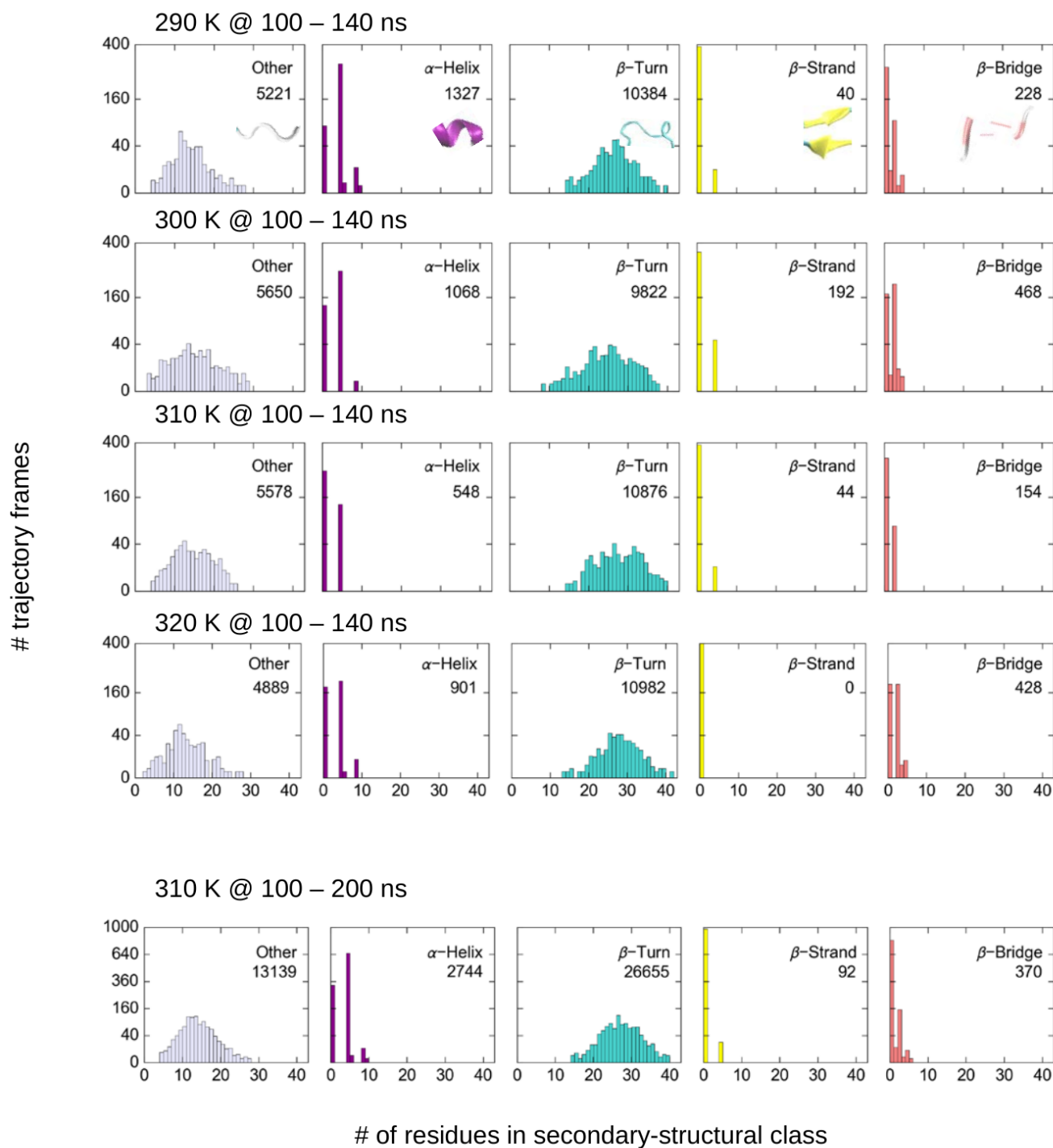


Figure S3.7: **Secondary structural content of the ELP region** across the 290-320 K temperature series for extended simulations (100 - 140 ns for 290 K, 300 K, and 320 K and 100 - 200 ns for 310 K). Numbers within the plot represent the total number of times that the secondary structure was observed in the simulation. Secondary structure cartoon representations in the thumbnails displayed in the first row match the colors in the histogram.

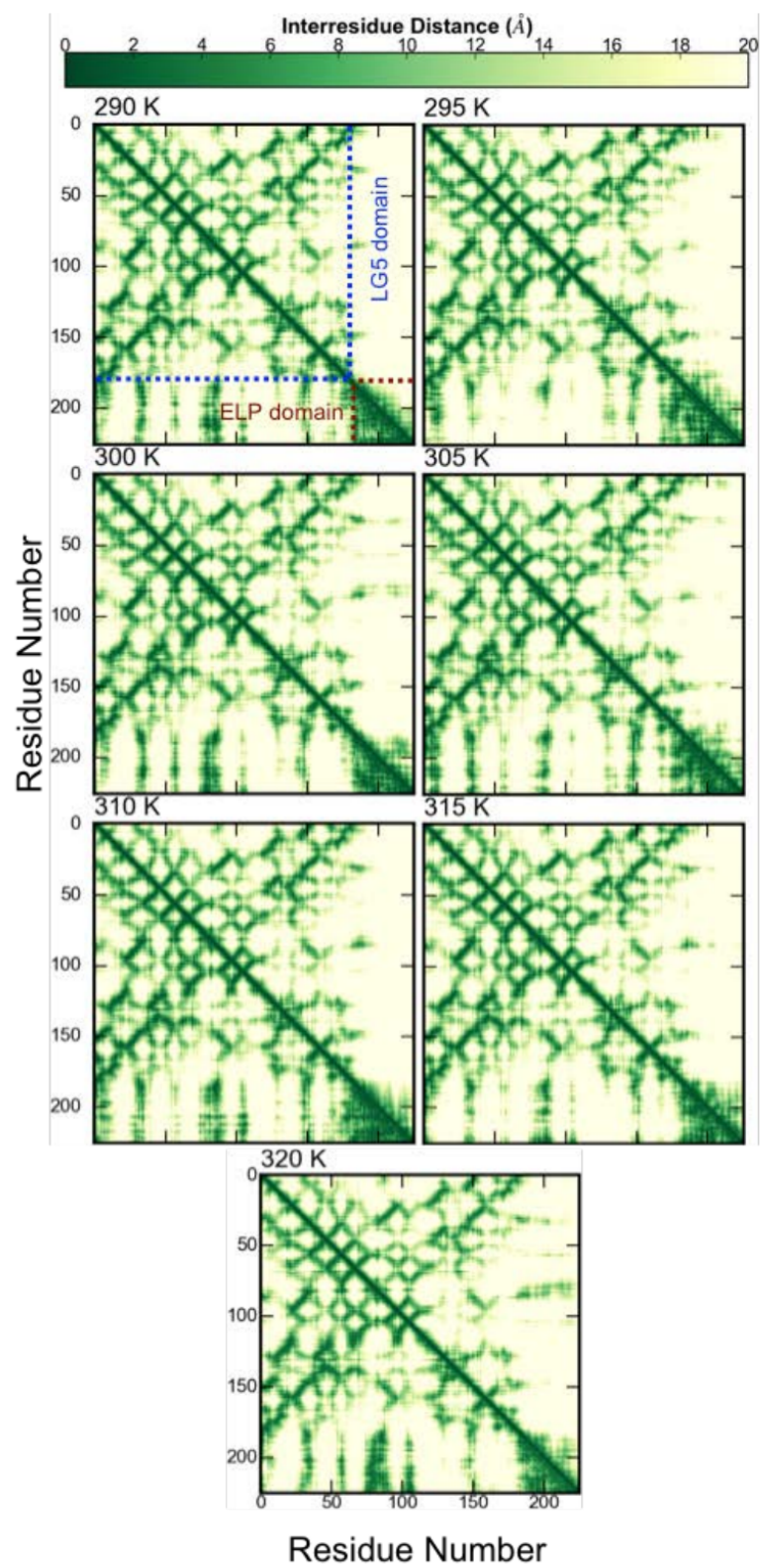


Figure S3.8: **Protein contact maps of the dynamical interactions** in the designed fusion suggest a lack of persistent LG-ELP interactions (for 290 K - 320 K).

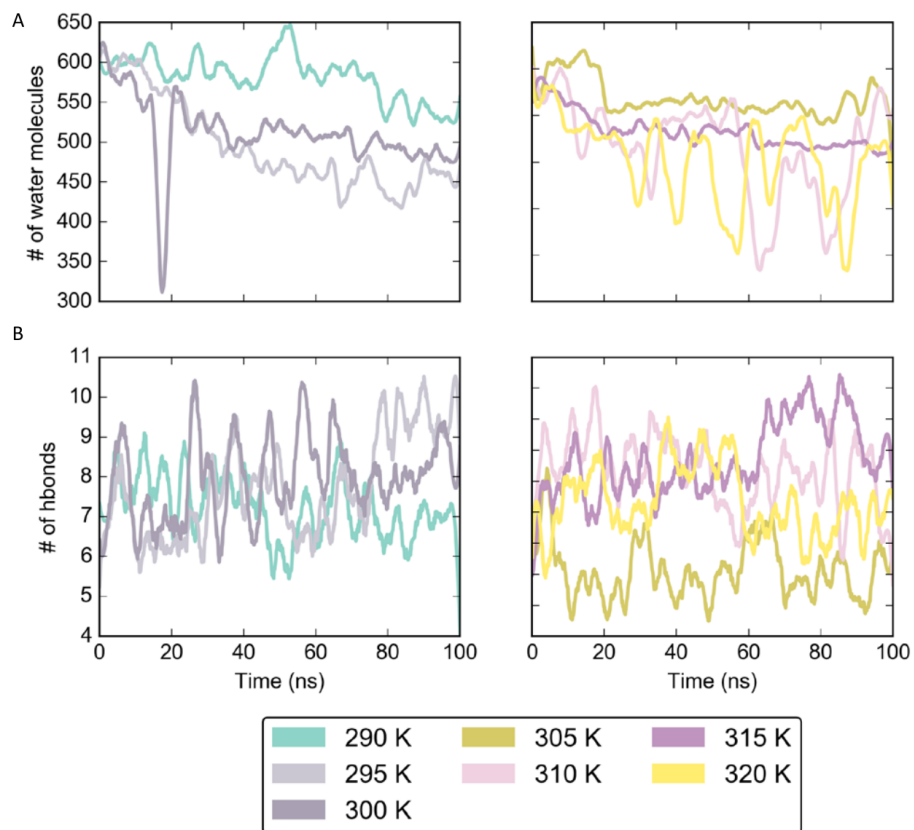


Figure S3.9: **Hydration of the ELP region.** (A) Number of water molecules surrounding the ELP region as a function of time. The abrupt drop in water molecules at 64 ns and 82 ns for 310 K corresponds to the formation of β -sheets. (B) Number of intramolecular hydrogen bonds as a function of time. All data was smoothed using a Savzky-Golay filter with a window size of 51 and 3rd order polynomial.

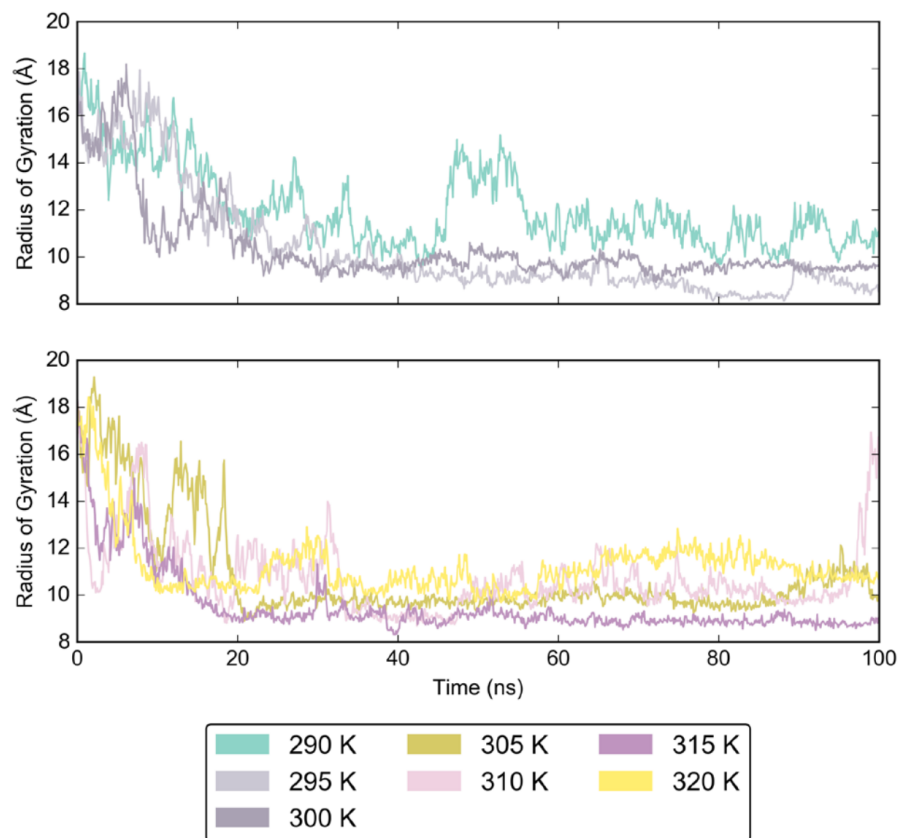


Figure S3.10: **Time evolution of the radius of gyration** simulated at different temperatures.

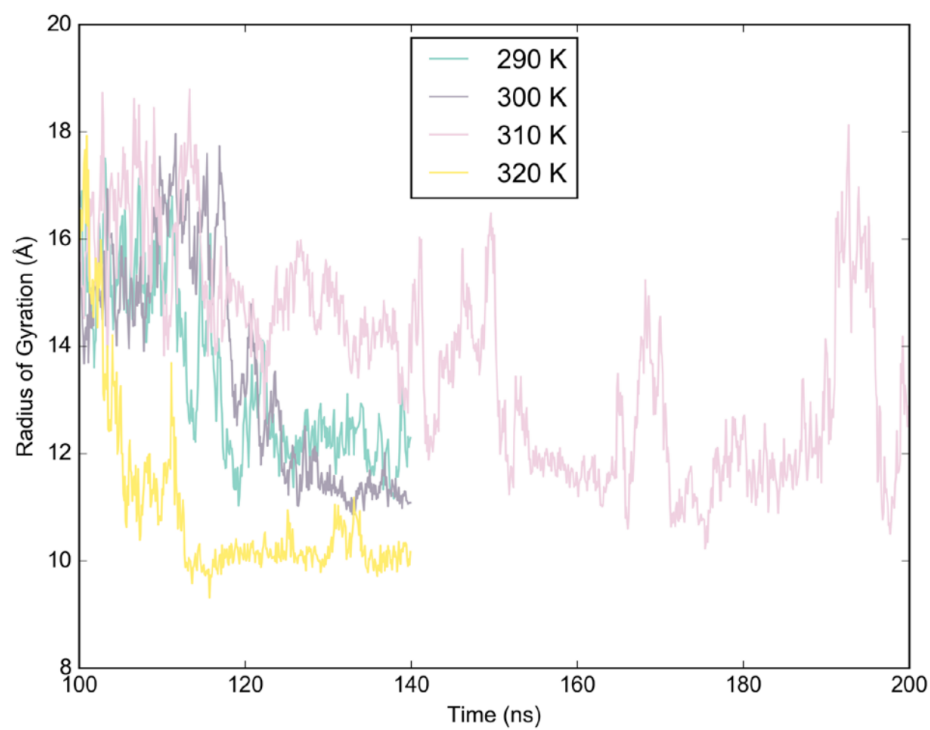


Figure S3.11: **Time evolution of the radius of gyration for extended simulations** at different temperatures (100 - 140 ns for 290 K, 300 K, and 320 K, and 100 - 200 ns for the 310 K trajectory)

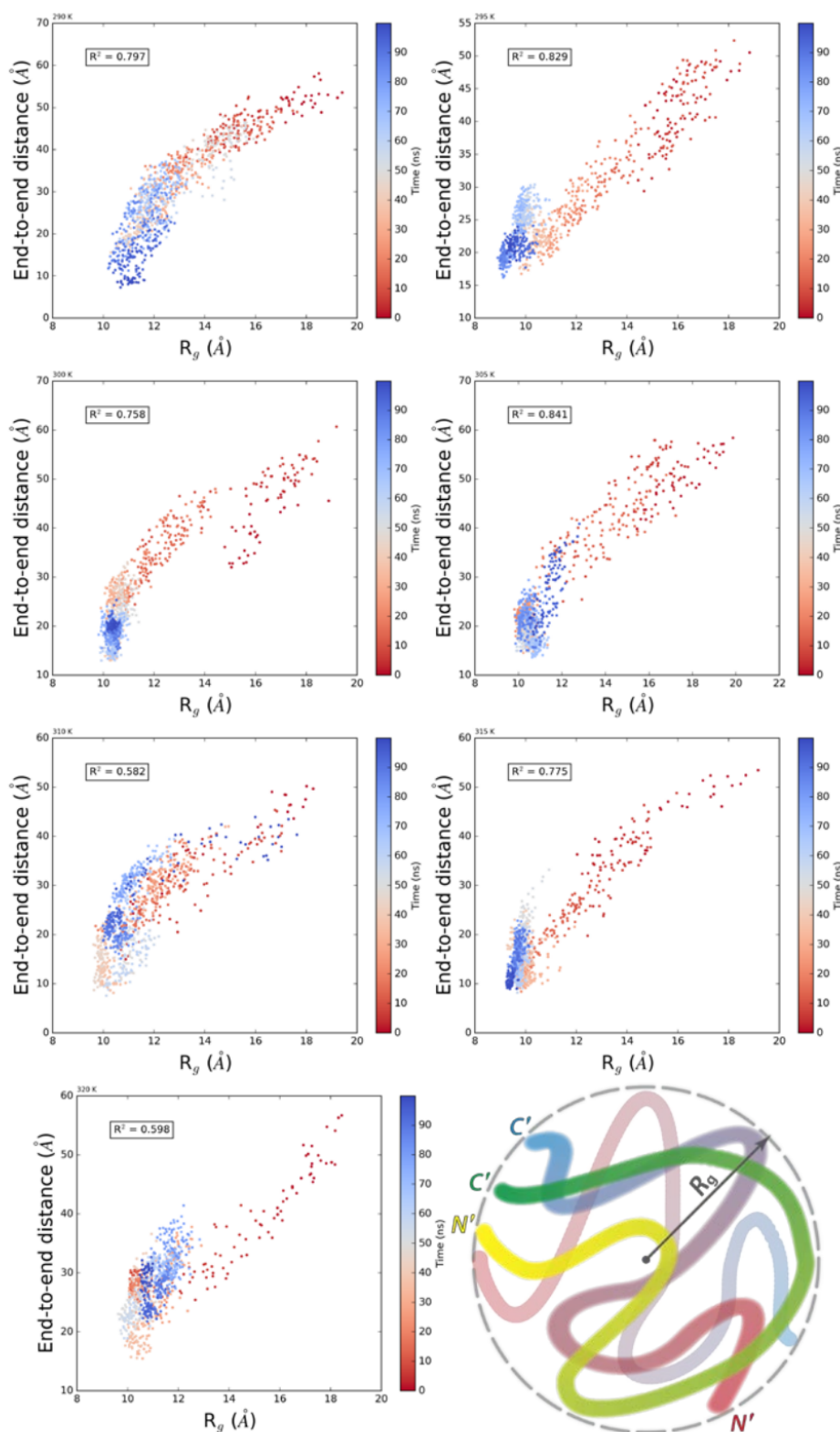


Figure S3.12: **Correlation between the radius of gyration and end-to-end distance of the ELP region.** We computed a least-squares regression using SciPy's stats function. R^2 is the coefficient of determination. Colors correspond to the time steps in the simulation (red indicates first time step, blue is the last time step of the MD simulation). The schematic shown at lower-right represents the differences between the end-to-end distance (Euclidean distance between N' - and C' - termini) and radius of gyration (R_g) of random coils. These two arbitrary chains have similar R_g values but quite dissimilar end-to-end distances.

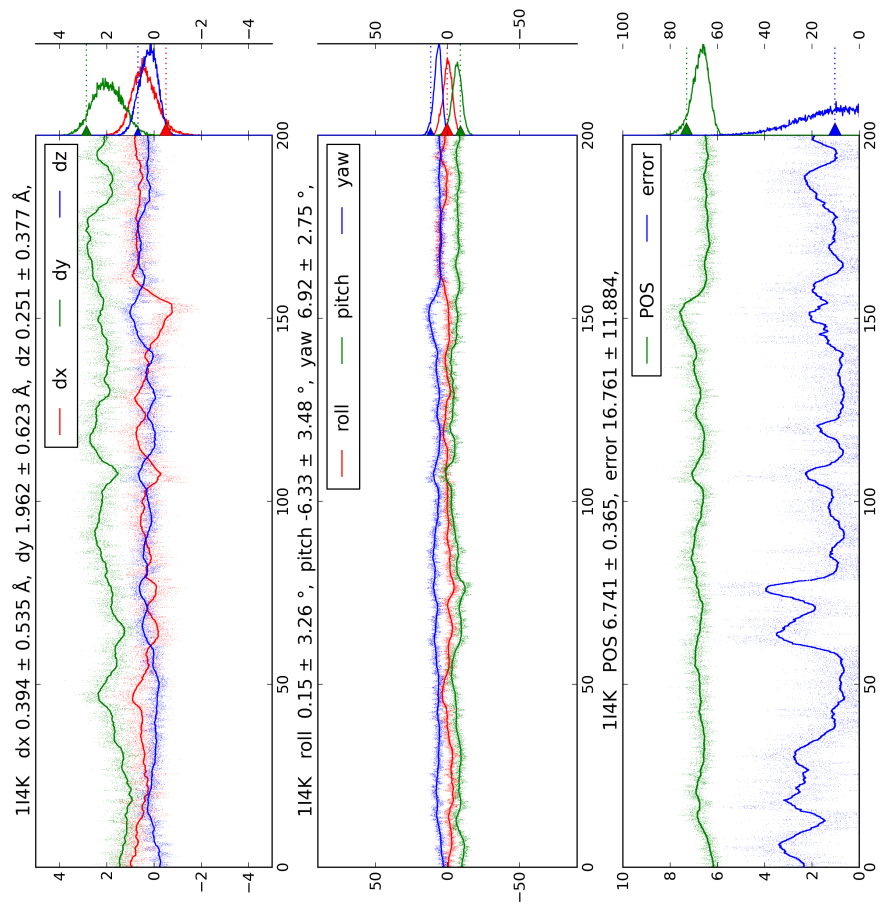
Chapter S4

Supplementary Information for Sm Oligomeric Plasticity

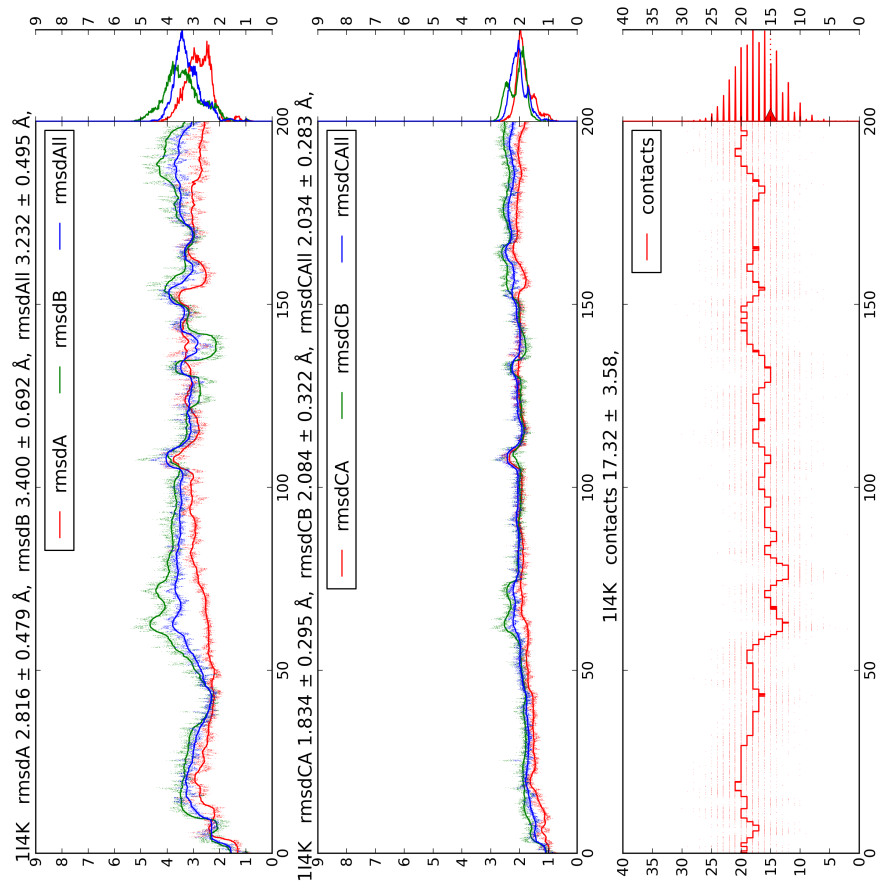
S4.1 PT, POS, and Validation statistics for all simulations

For each simulation, the left panel shows pizza tensor (PT) and predicted oligomeric state (POS) values at each time step, and thick lines represent a 5-ns sliding median filter. Supertitles on each figure present the mean and standard deviation for each of the various measures, over the course of the entire trajectory. Marginal distributions are provided on the right side of each plot, and the colored markers (small triangles) indicate the value of the corresponding quantity in the original crystal structure.

In the right panel, we present RMSD values and intermonomer contact numbers. In the top panel of each figure, the RMSD is shown for all atoms in the system for chain A (red trace), chain B (green trace) and the complete dimer (blue trace). These panels include atoms which are outside of the Sm core, such as N- and C- terminal extensions. These regions were not used in the calculation of the PT or POS, and often contain unstructured regions that were modelled essentially as straight chains extending from the protein surface if they were not observed in the crystal structure. It is unsurprising, therefore, that systems containing such extensions exhibit high RMSD values. In the center panel, we show the RMSD for the atoms that comprise the Sm core, again with chain A (red trace), chain B (green trace) and both core regions together (blue trace). In the bottom panel, we show the number of intermonomer contacts, which is defined as the number of atoms (excluding hydrogen) that are within 3 Å of the other chain in the dimer (that is, “(chain A and within 3 of chain B) or (chain B and within 3 of chain A)”). In all panels, individual pixels represent the values at each frame, the thick lines represent a 5-ns running median filter, and marginal distributions are shown on the right. The marker in the marginal distribution of the contact plot shows the number of contacts measured for the crystal structure.

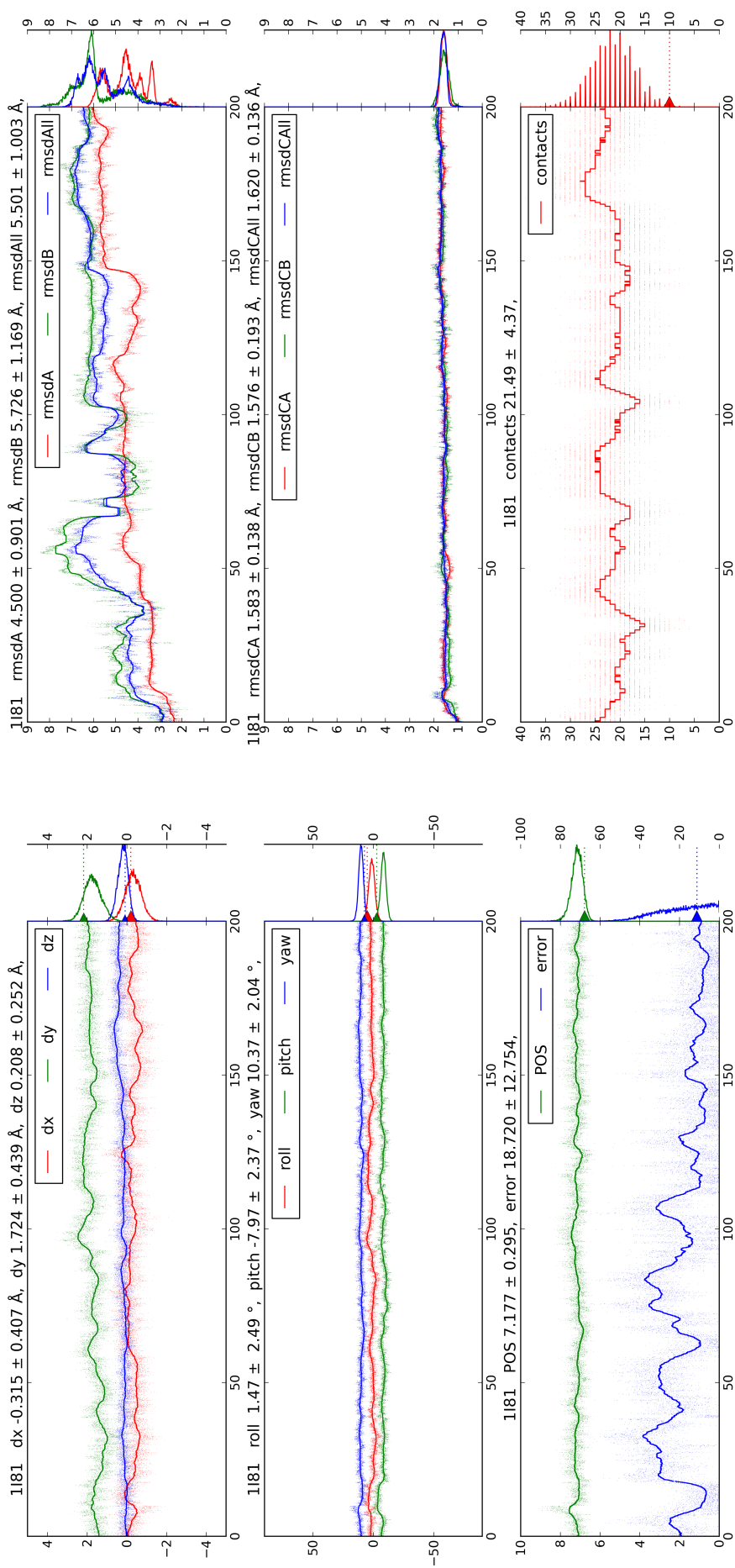


(a) PT, POS



(b) Validation

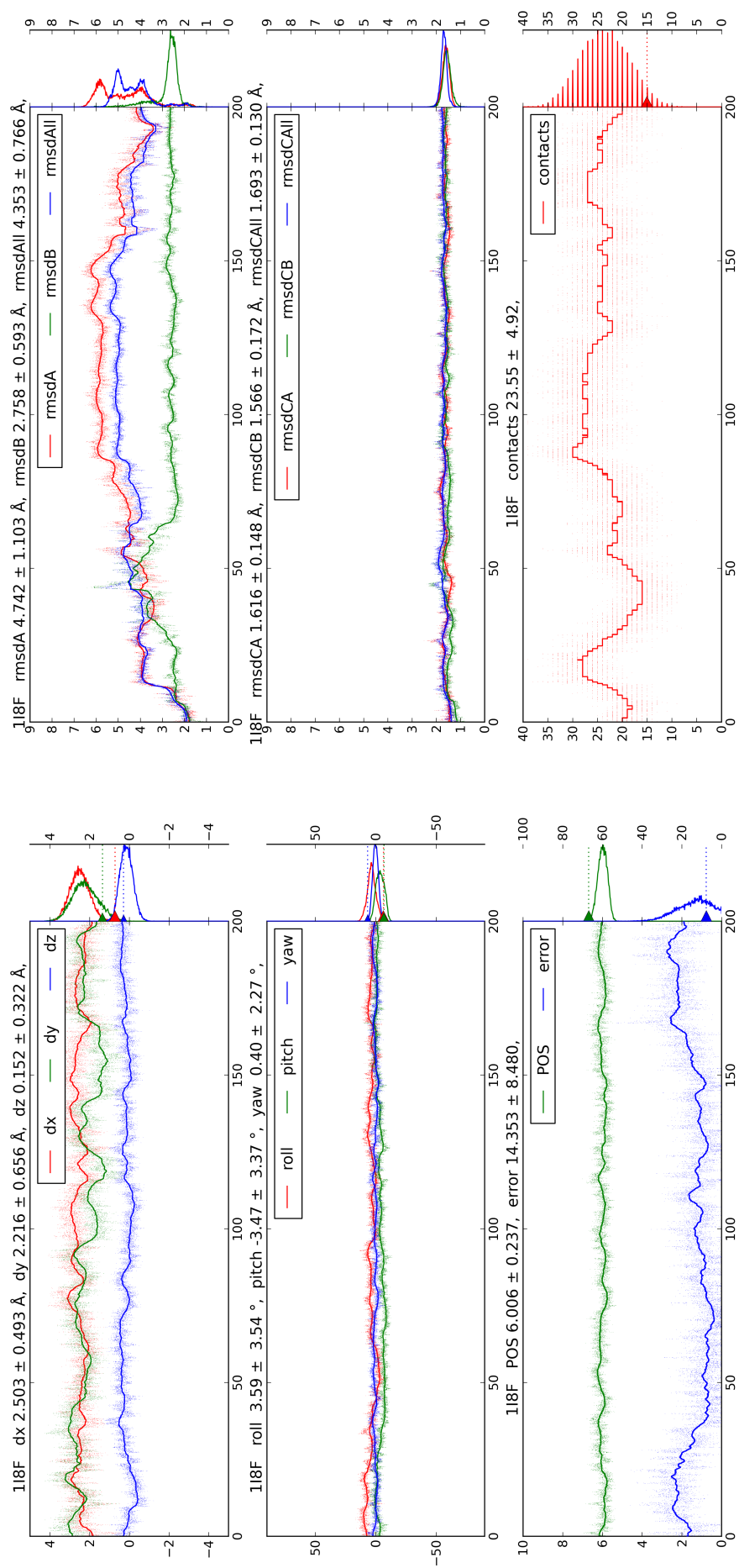
Figure S4.1: **114K** (*Archaeoglobus fulgidus*)



(a) PT, POS

(b) Validation

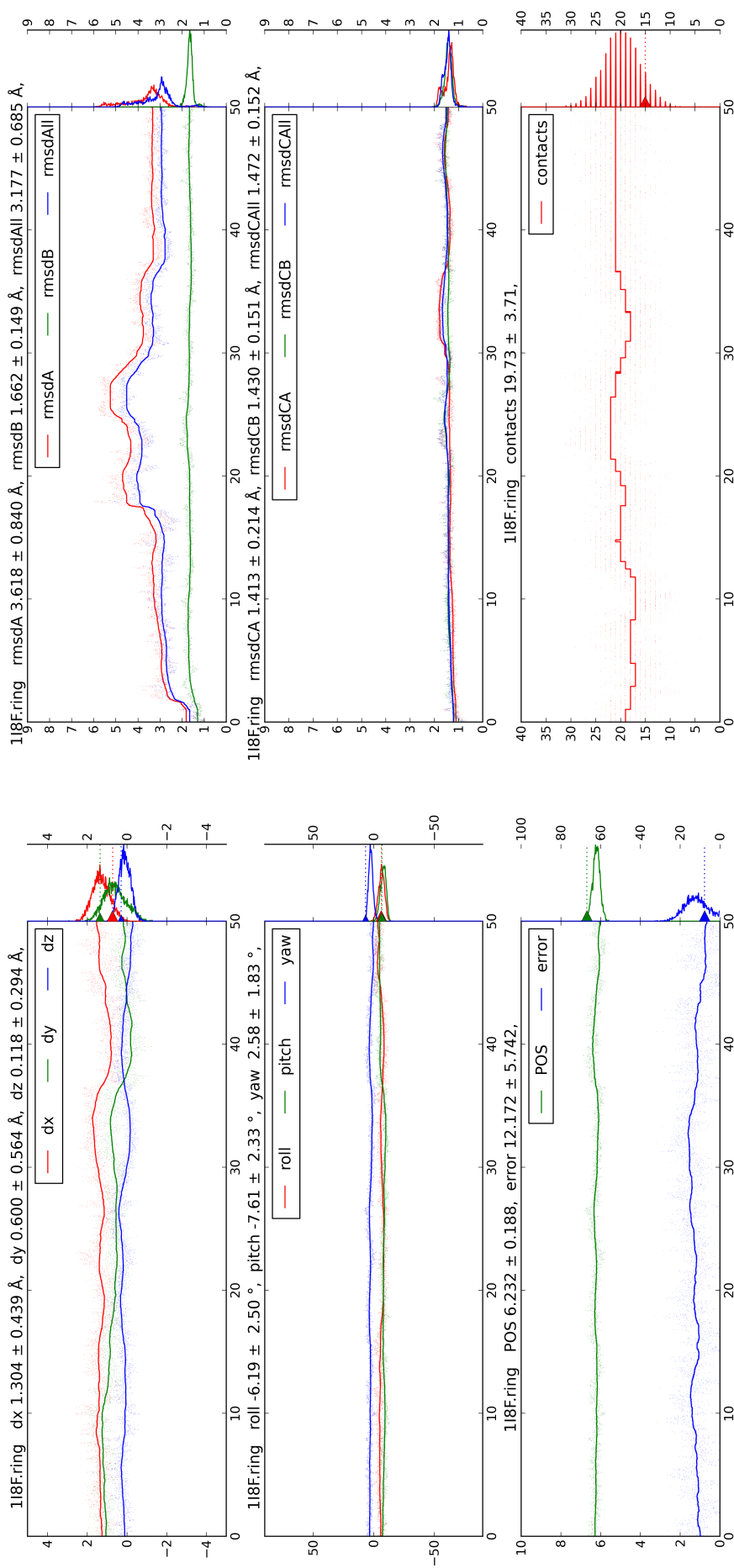
Figure S4.2: **1I81** (*Methanobacterium thermoautotrophicus*)



(a) PT, POS

(b) Validation

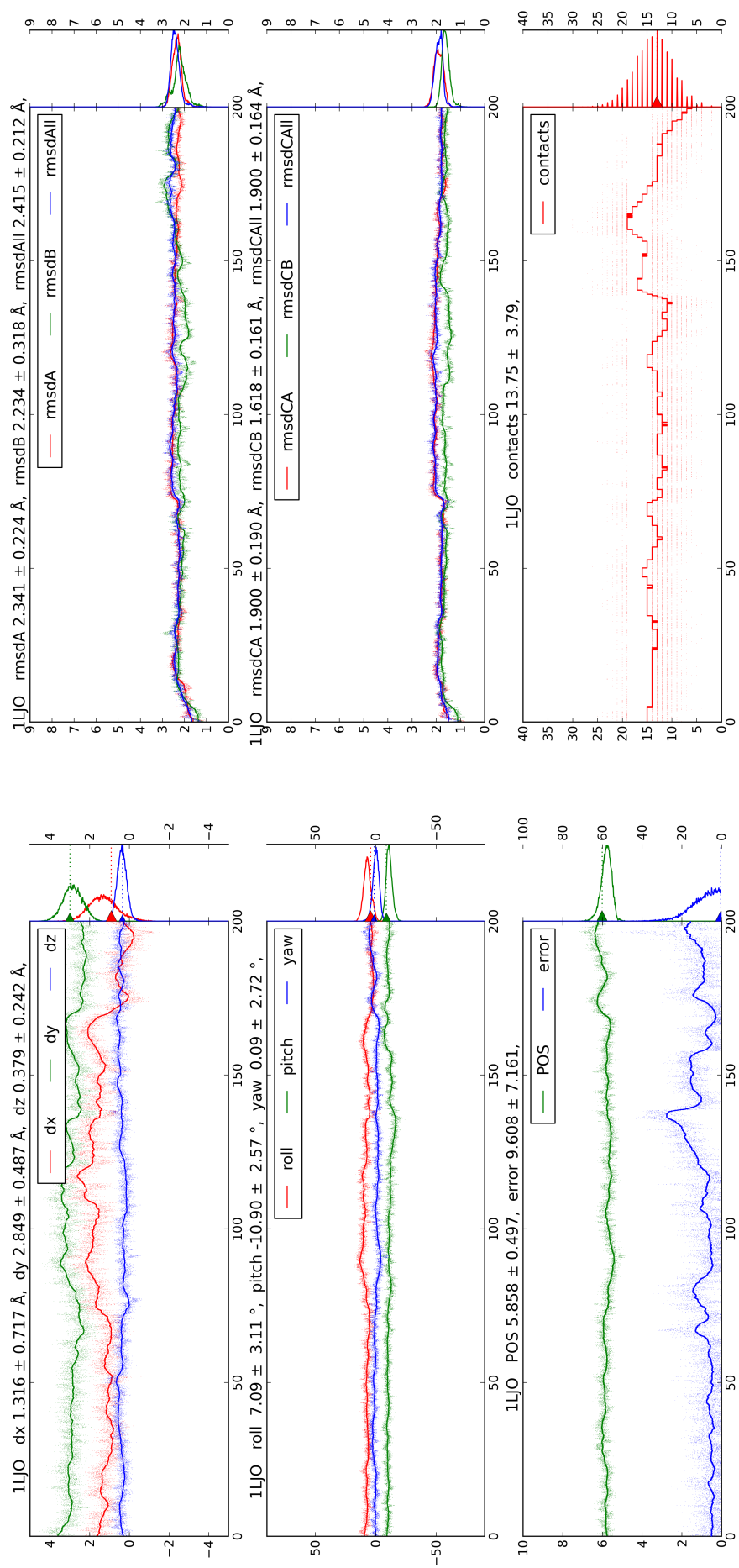
Figure S4.3: **118F** (*Pyrobaculum aerophilum*)



(a) PT, POS

(b) Validation

Figure S4.4: **118F.ring** (*Pyrobaculum aerophilum*)



(a) PT, POS

(b) Validation

Figure S4.5: **1LJO** (*Archaeoglobus fulgidus*)

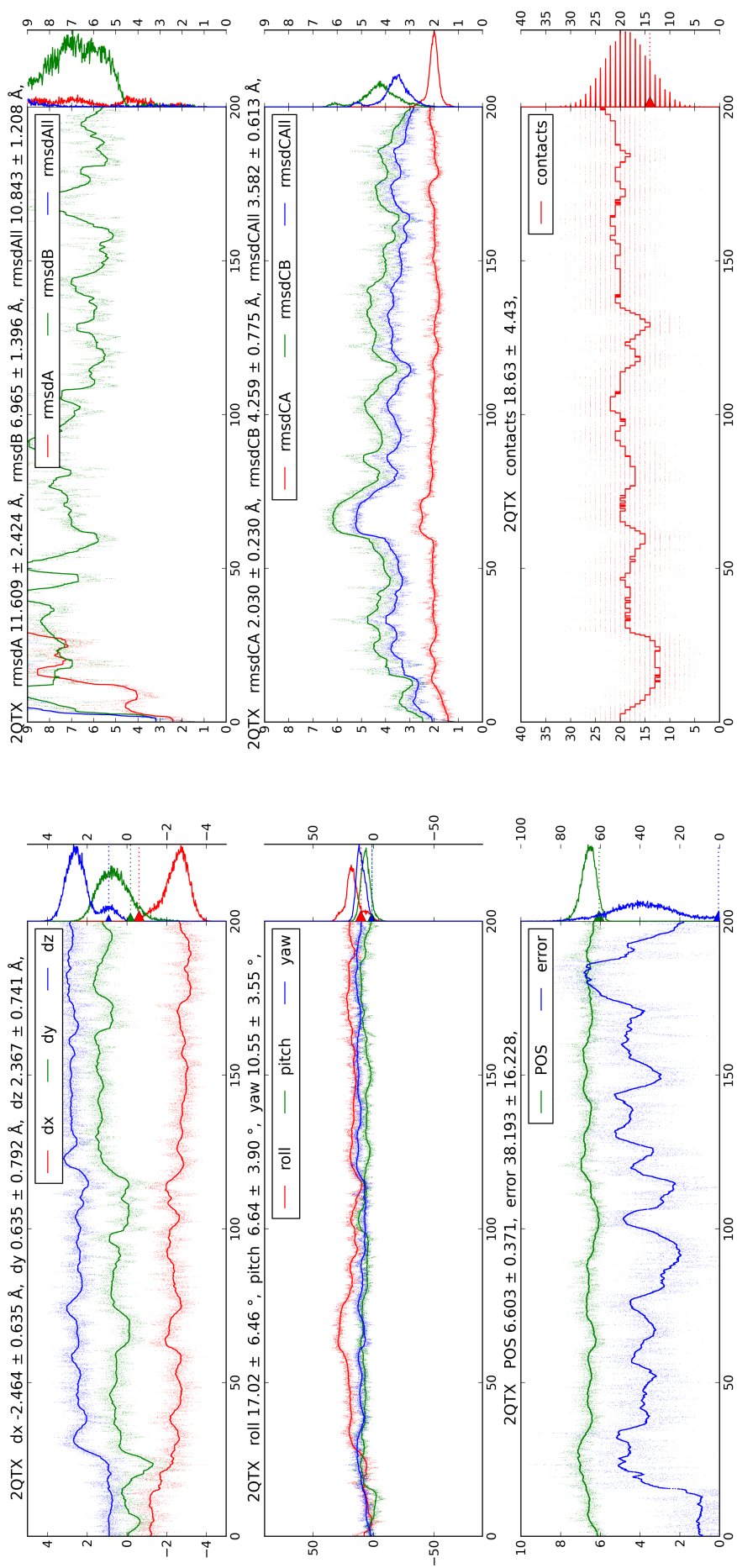
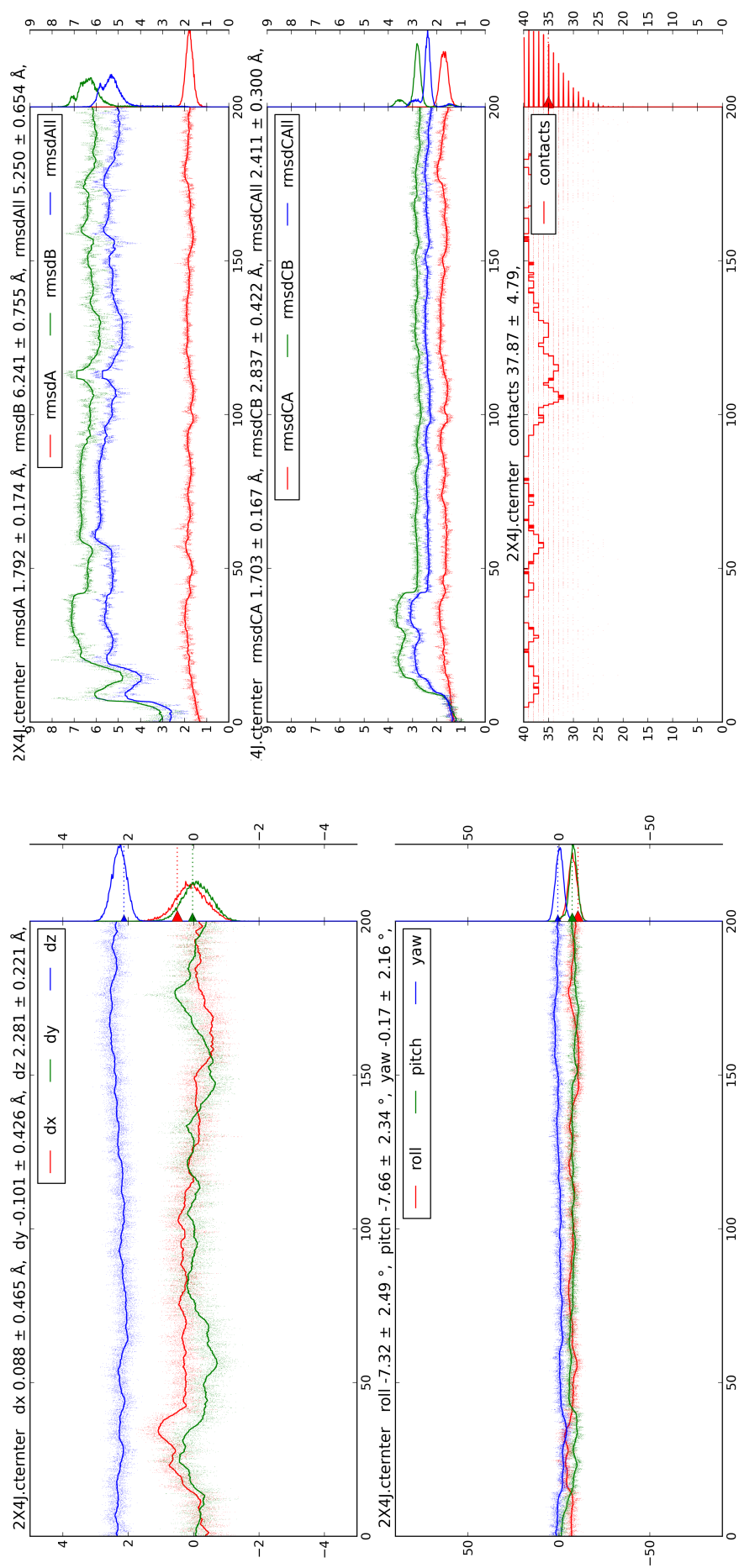
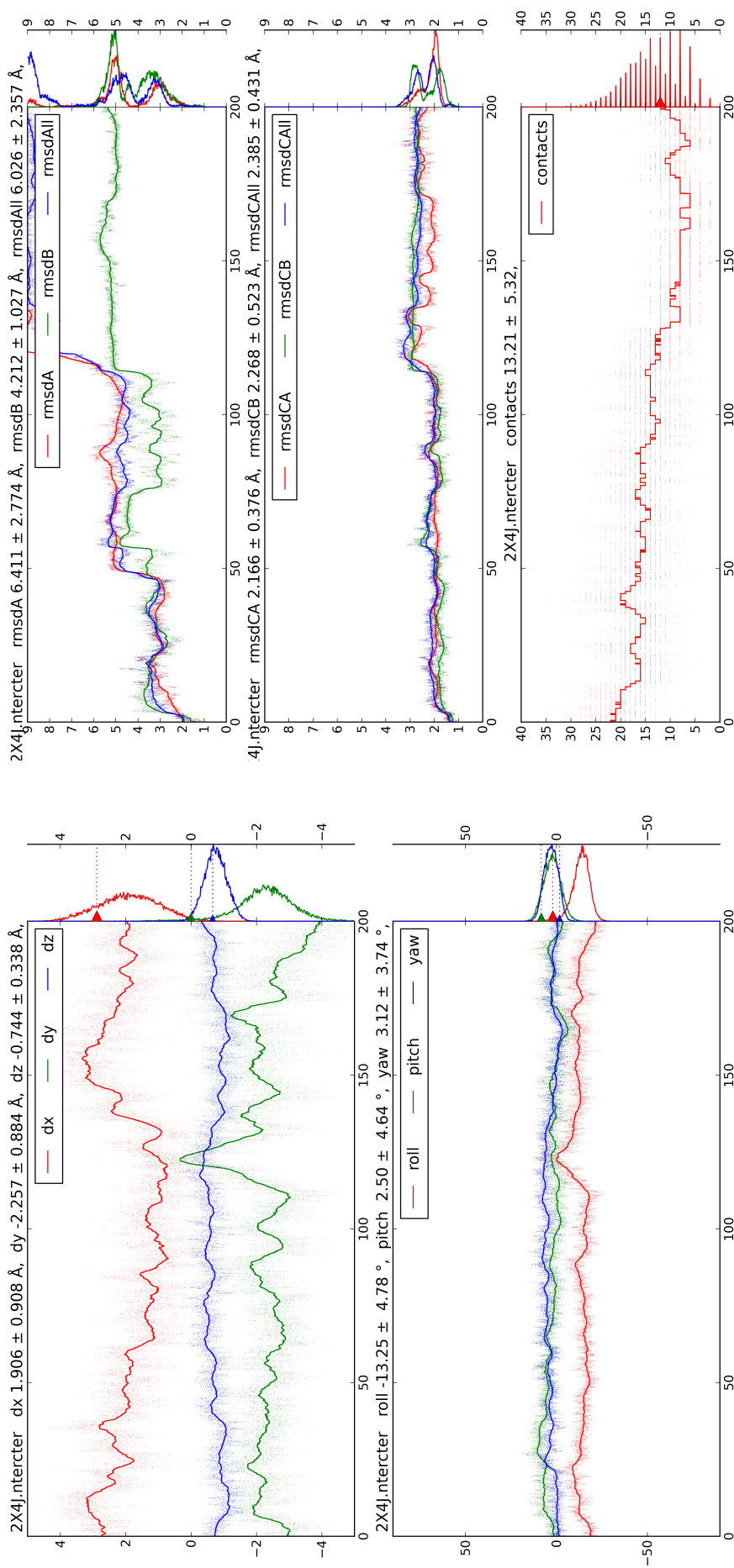


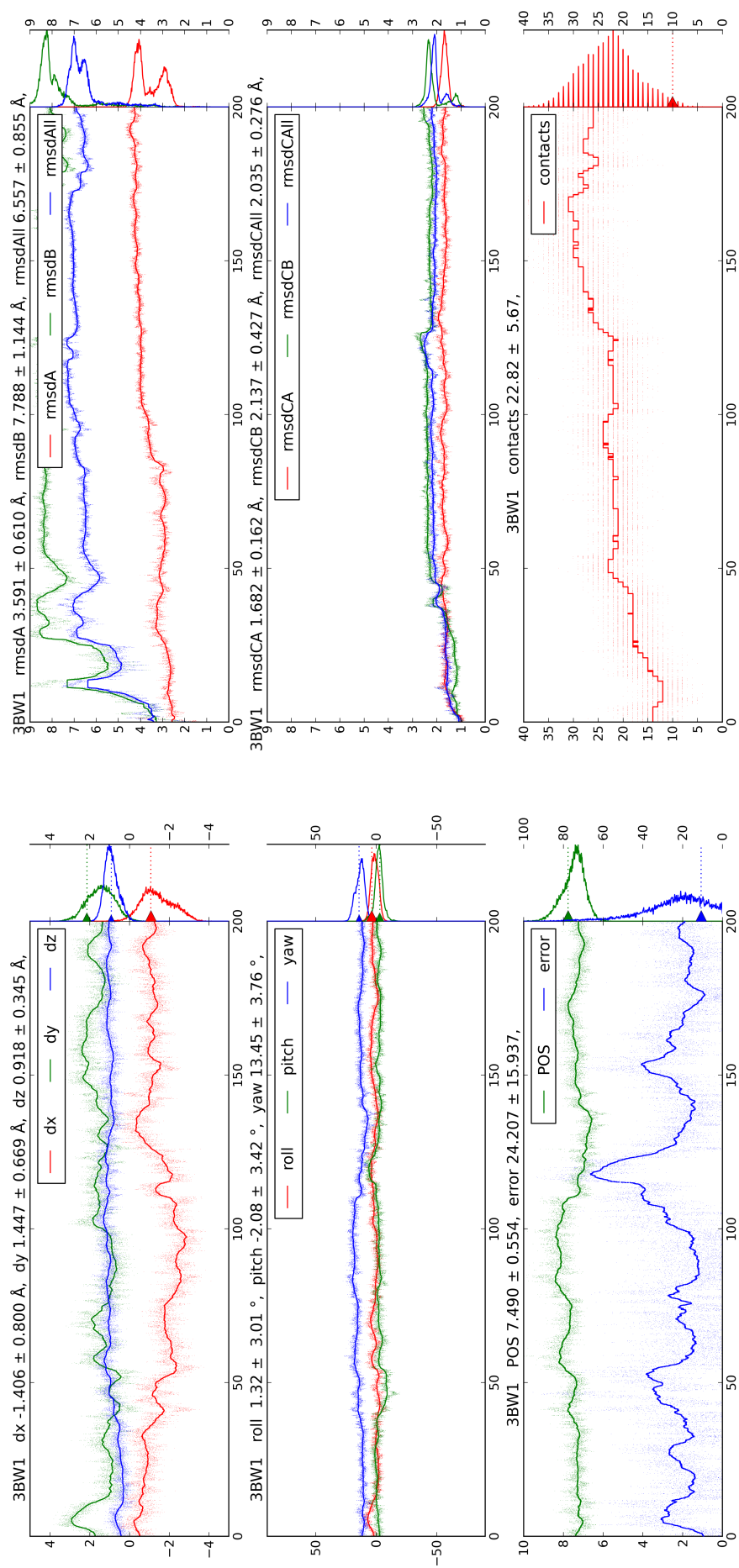
Figure S4.6: **2QTX** (*Methanococcus jannaschii*)

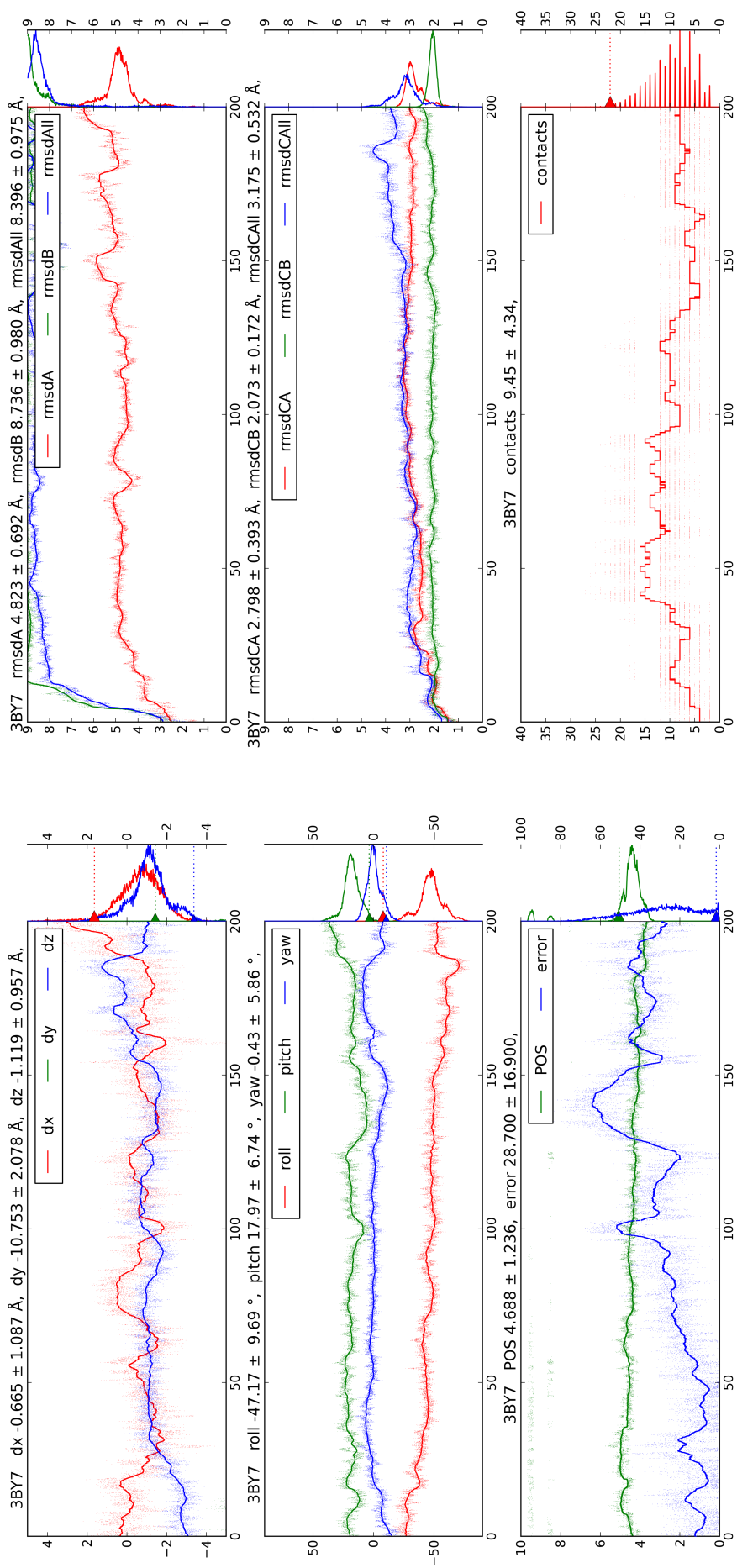
Figure S4.7: **2X4J Cter, Nter** (Pyrobaculum spherical virus)



(a) PT

(b) Validation

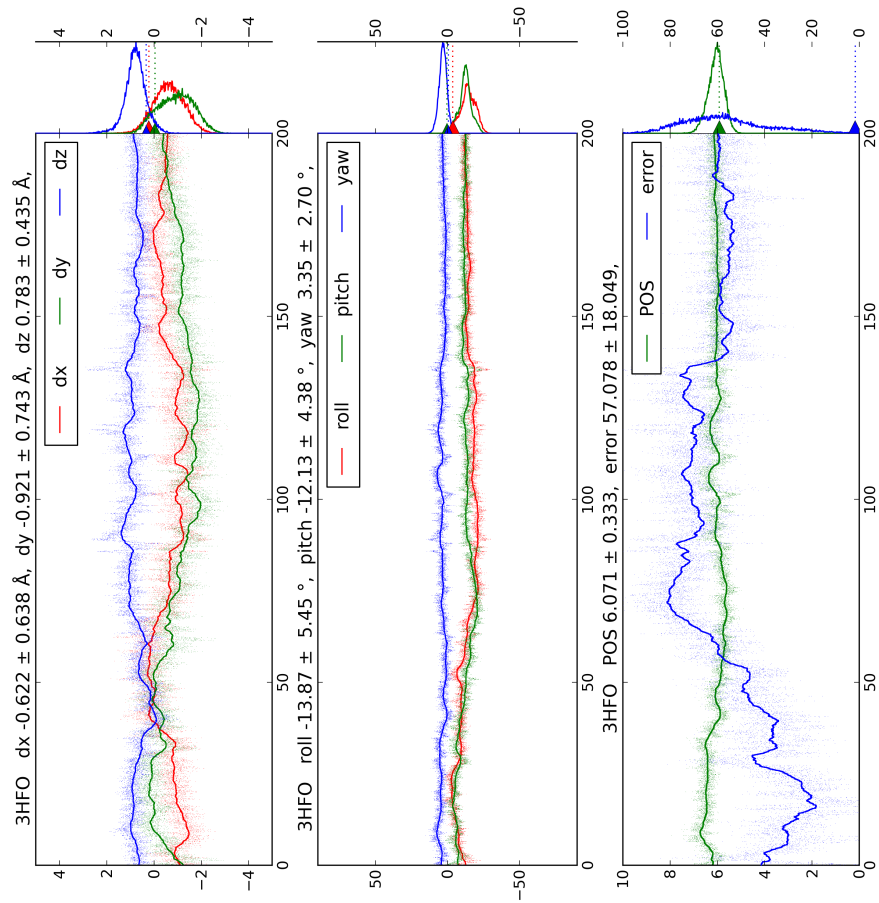
Figure S4.9: **3BW1** (*Saccharomyces cerevisiae*)



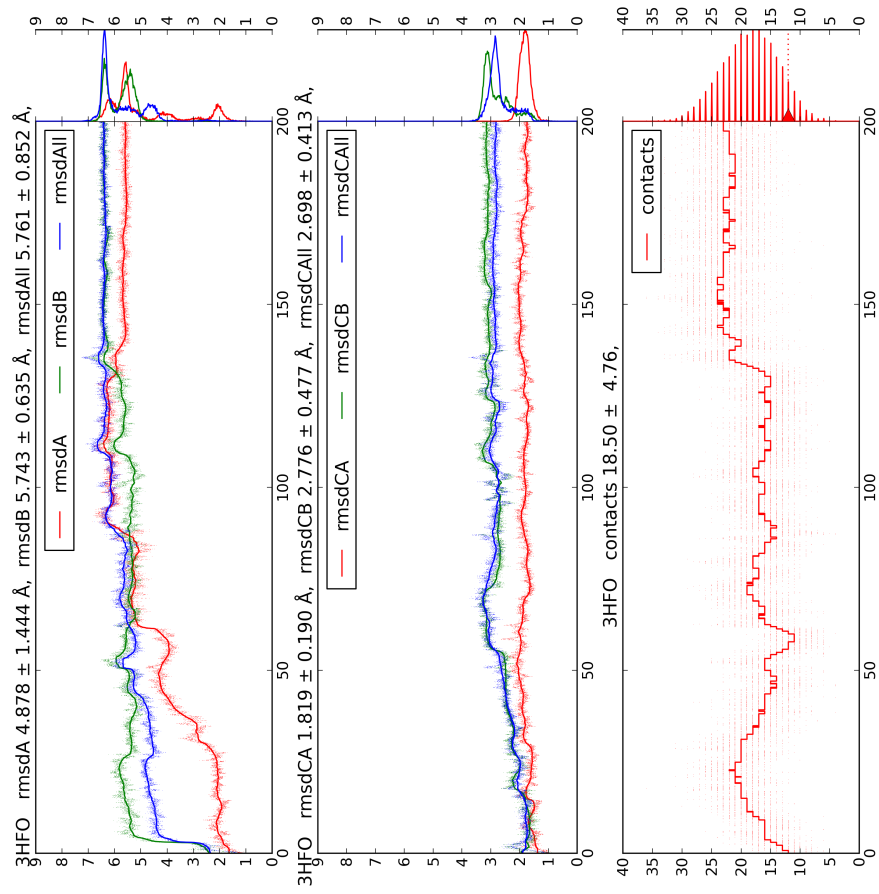
(a) PT, POS

(b) Validation

Figure S4.10: **3BY7** (Unknown organism)



(a) PT, POS



(b) Validation

Figure S4.11: **3HFO** (*Synechocystis sp.*)

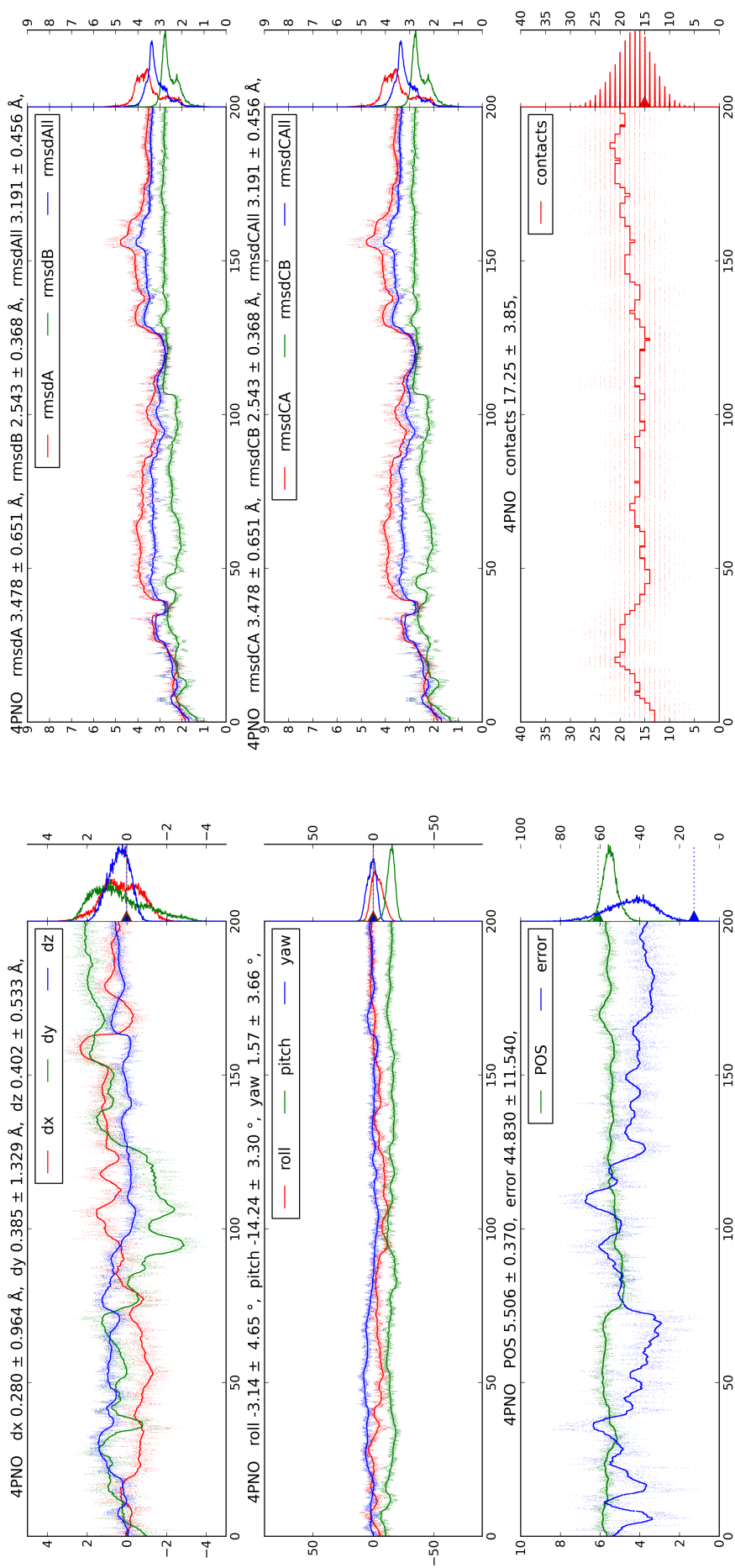
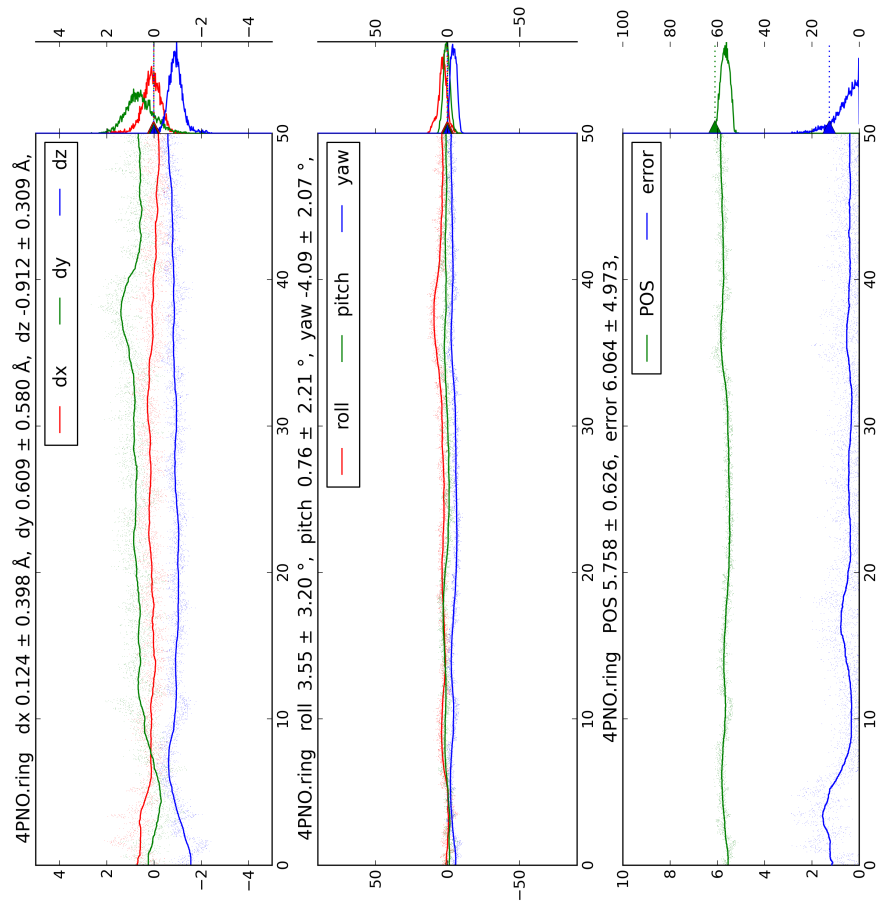
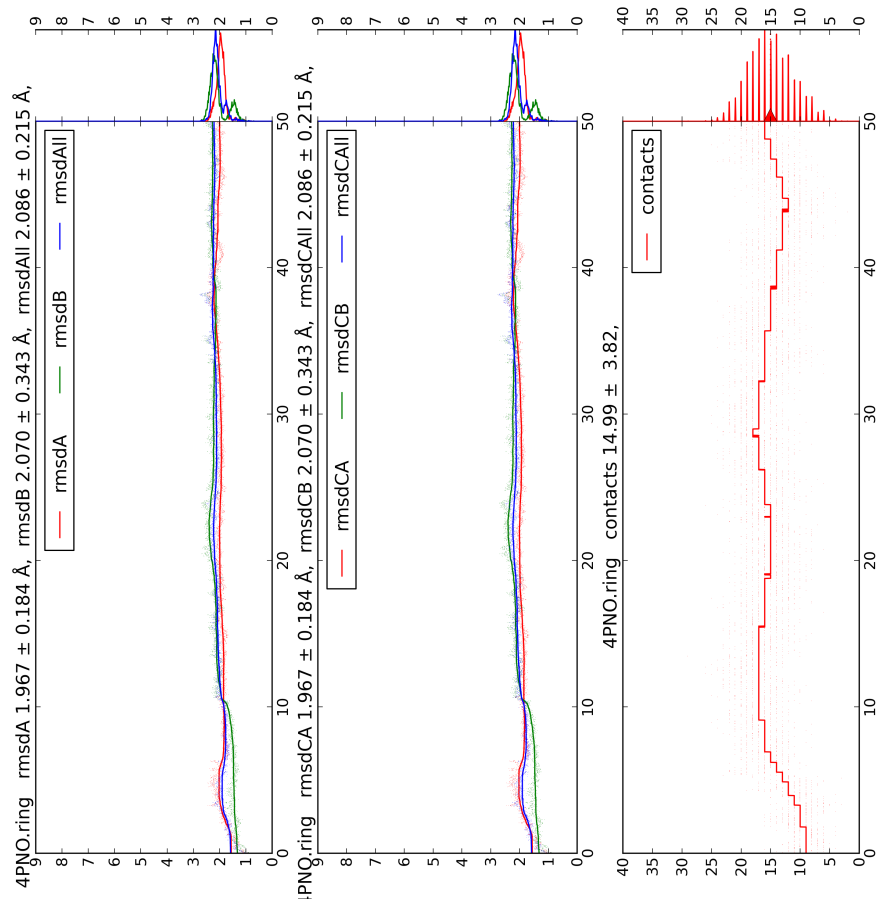


Figure S4.12: 4PNO (*Escherichia coli*)



(a) PT, POS



(b) Validation

Figure S4.13: 4PNO ring (*Escherichia coli*)

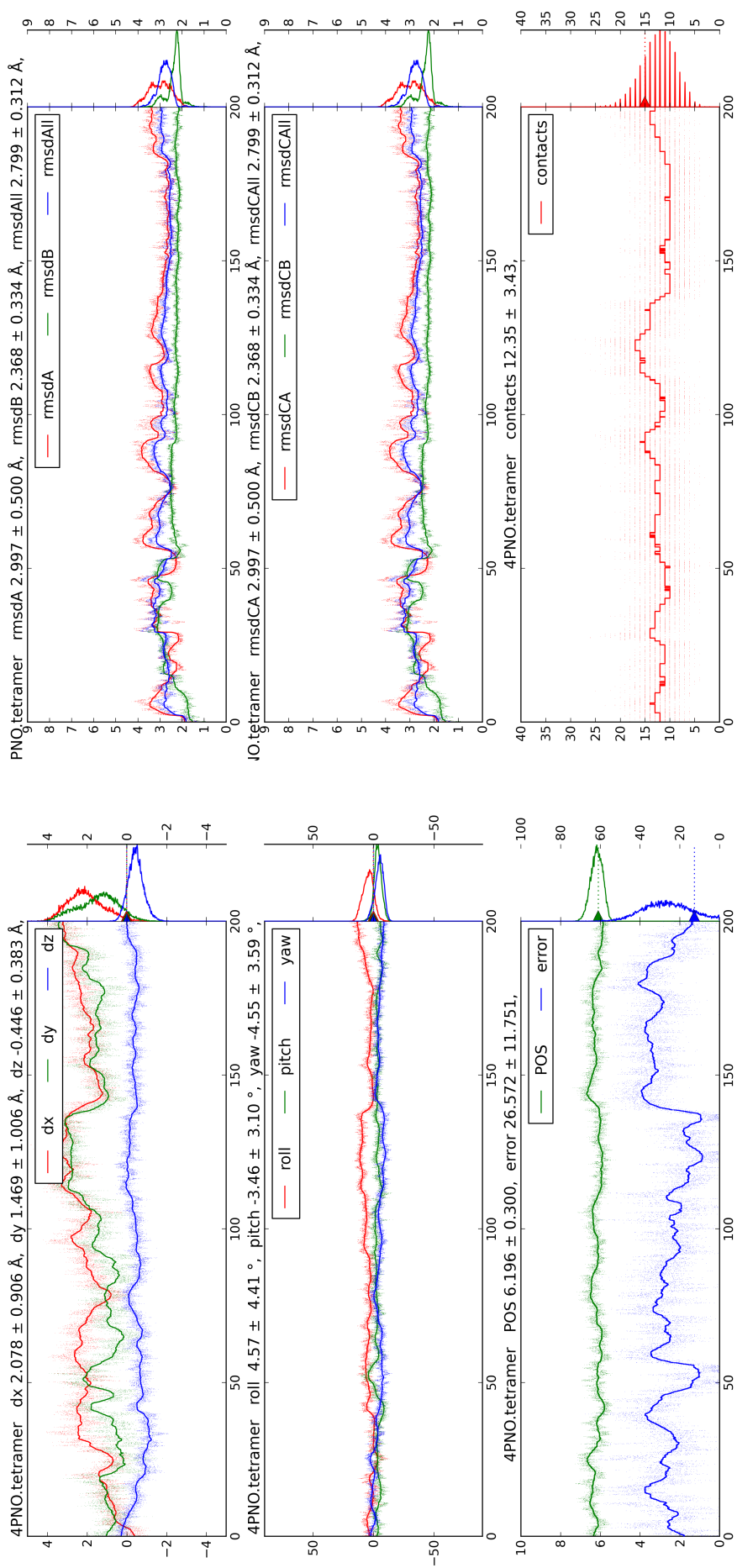
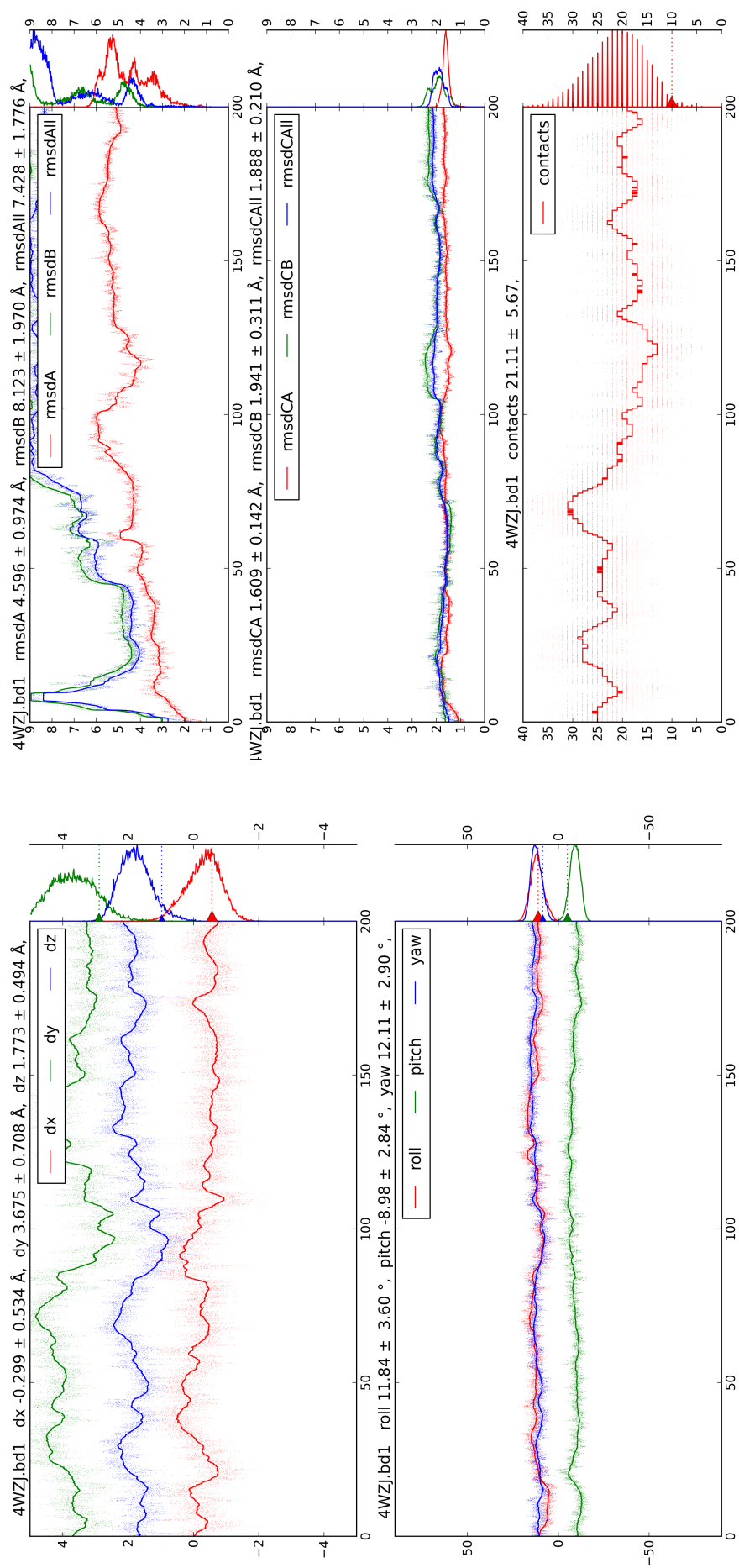


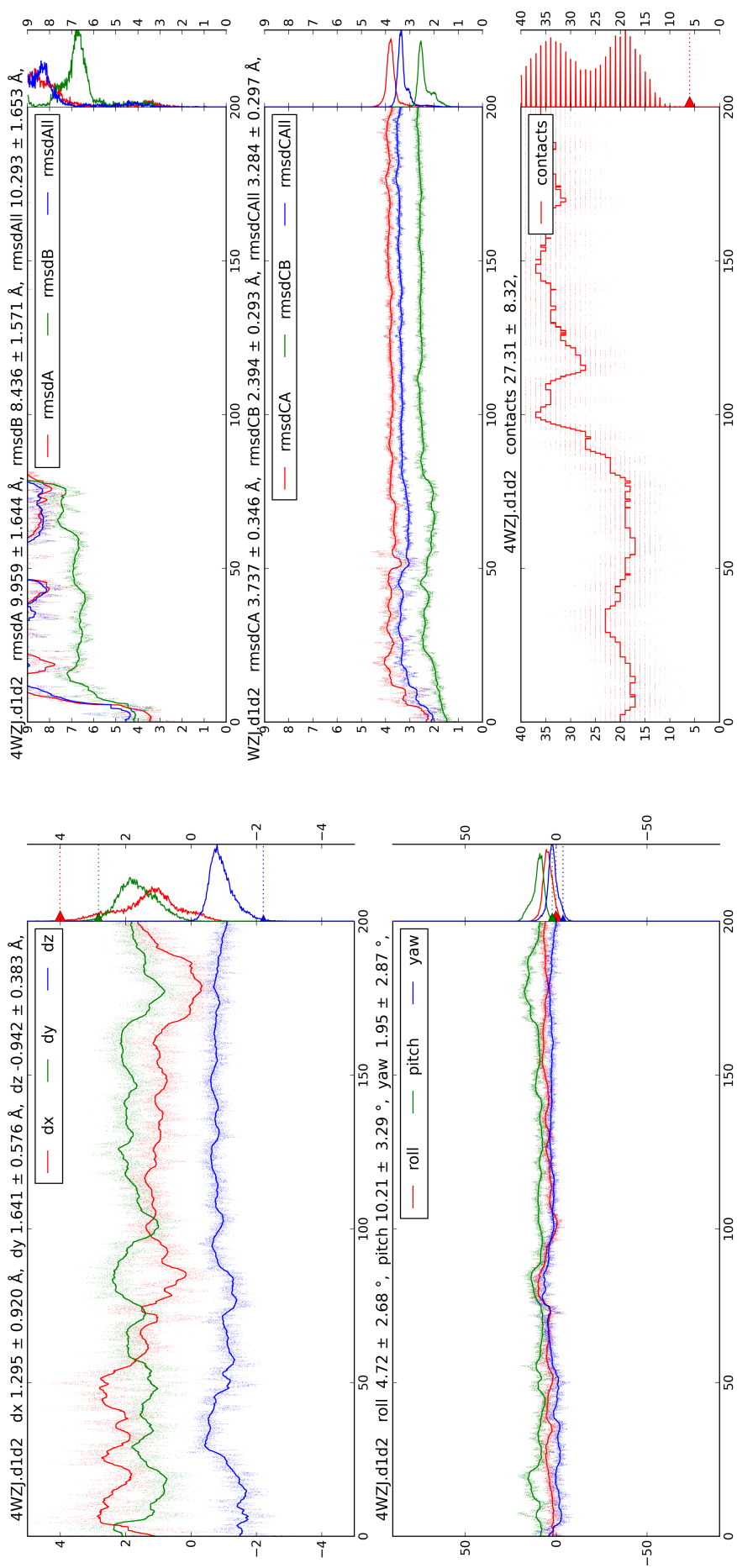
Figure S4.14: 4PNO tetramer (*Escherichia coli*)



(a) PT

Figure S4.15: **4WZJ bd1** (*Homo sapiens*)

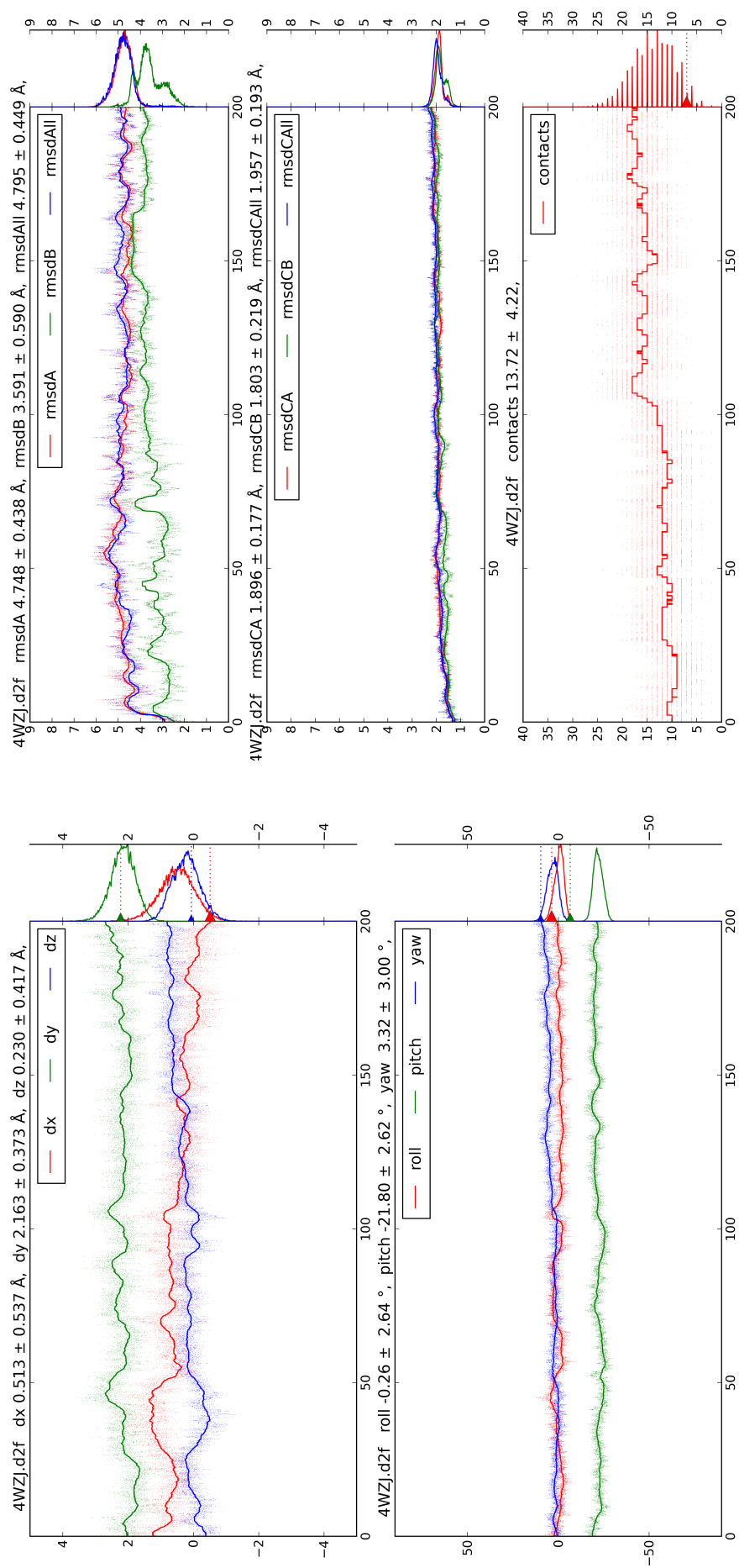
(b) Validation



(a) PT

(b) Validation

Figure S4.16: **4WZJ d1d2** (*Homo sapiens*)



(a) PT

(b) Validation

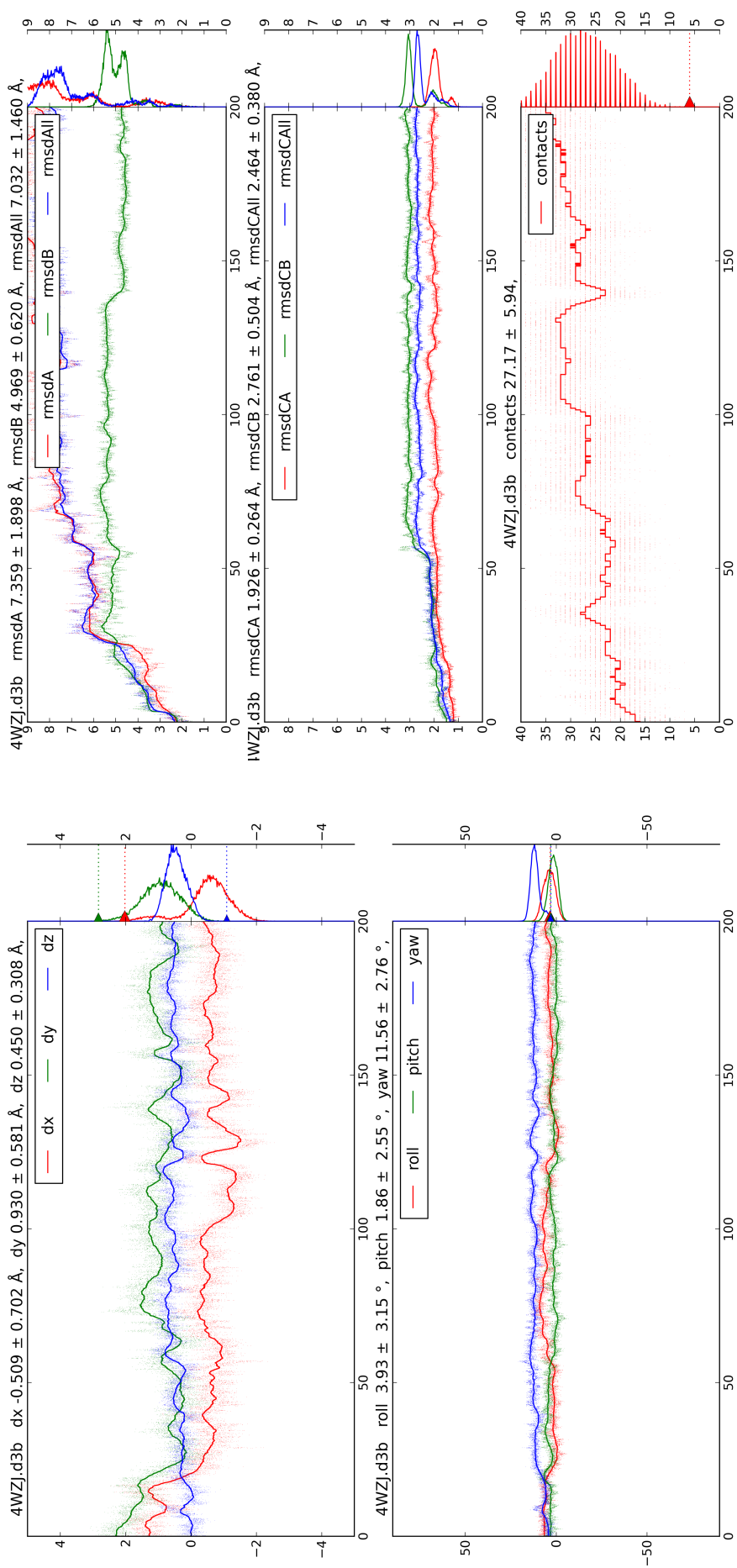
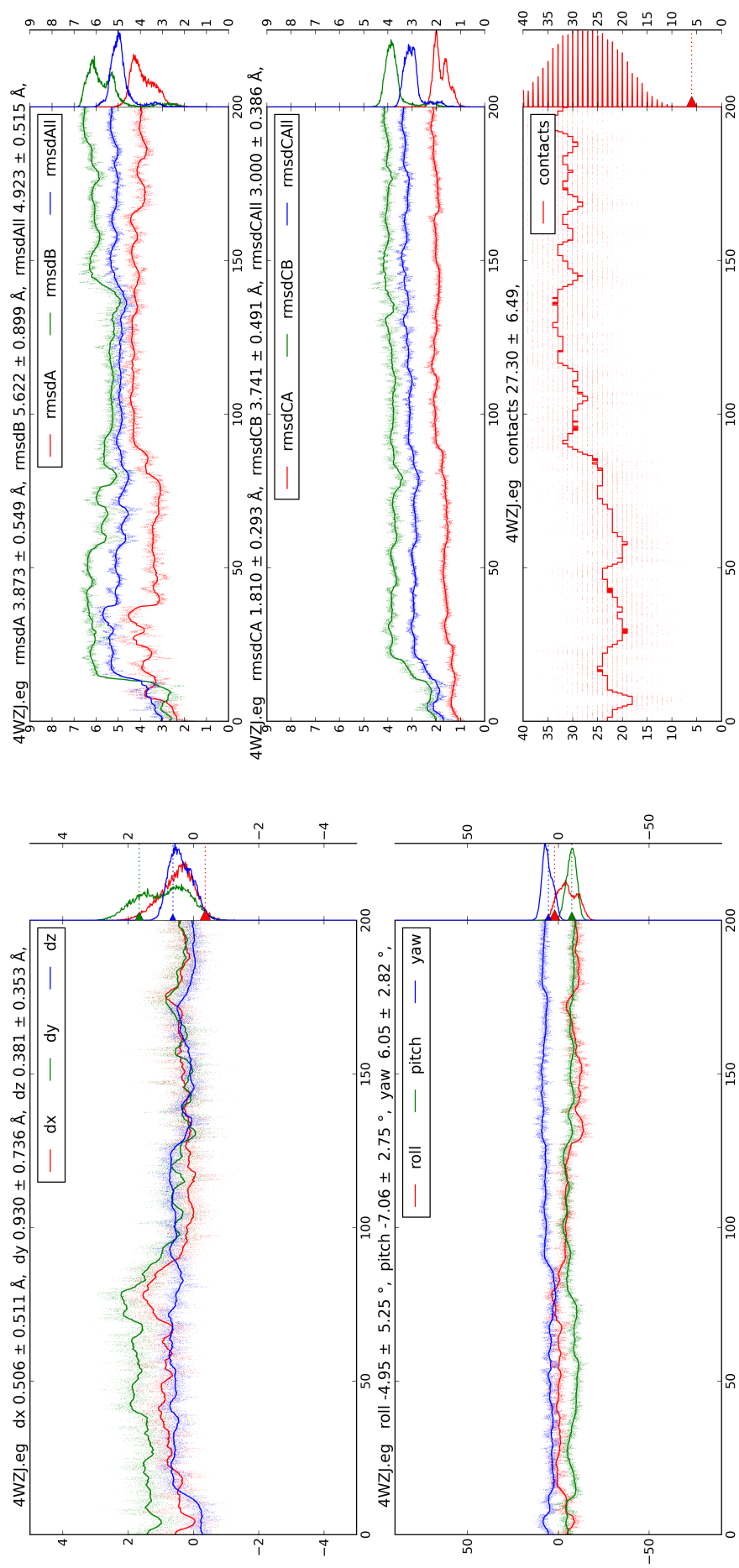


Figure S4.18: **4WZJ d3b** (*Homo sapiens*)



(a) PT

Figure S4.19: **4WZJ eg** (*Homo sapiens*)

(b) Validation

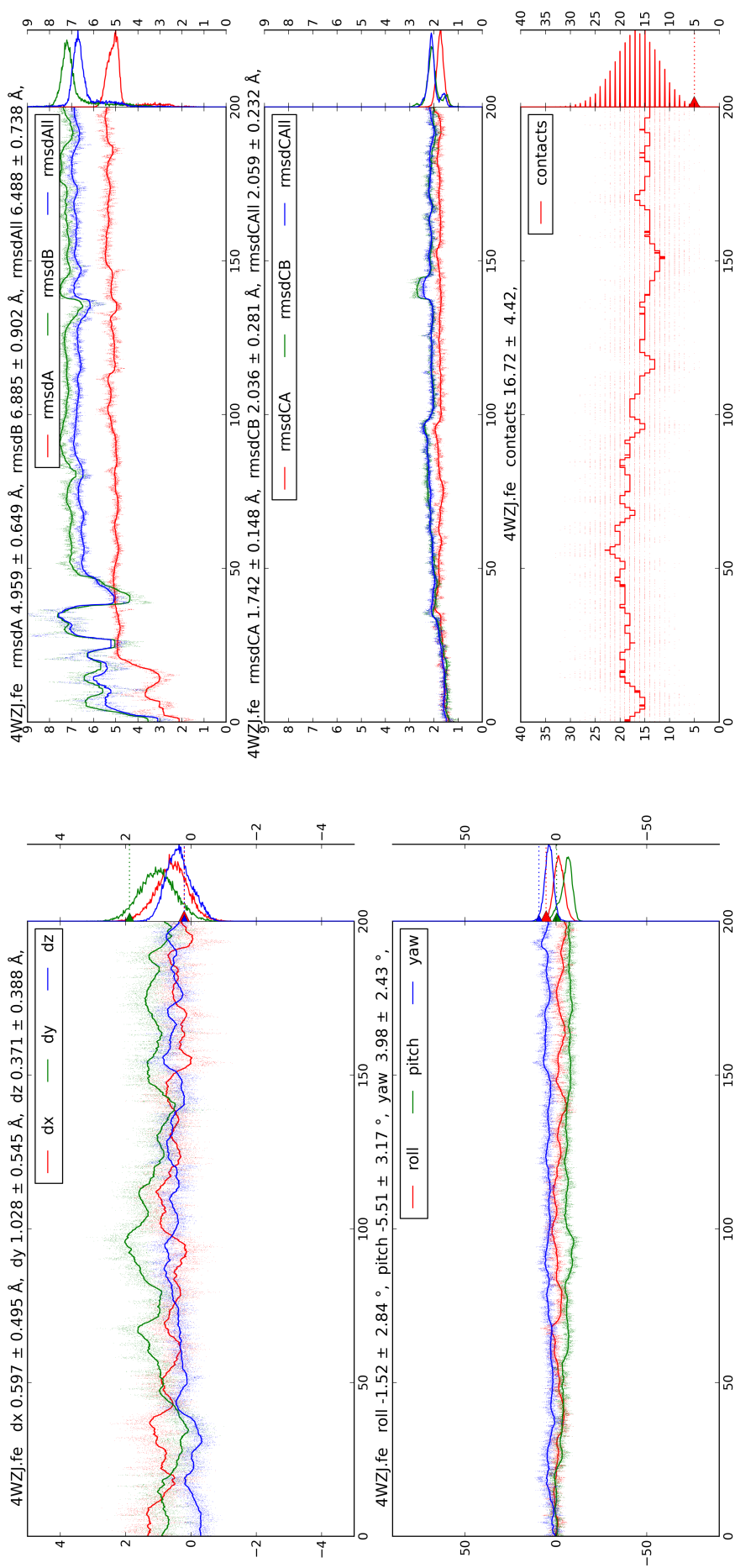
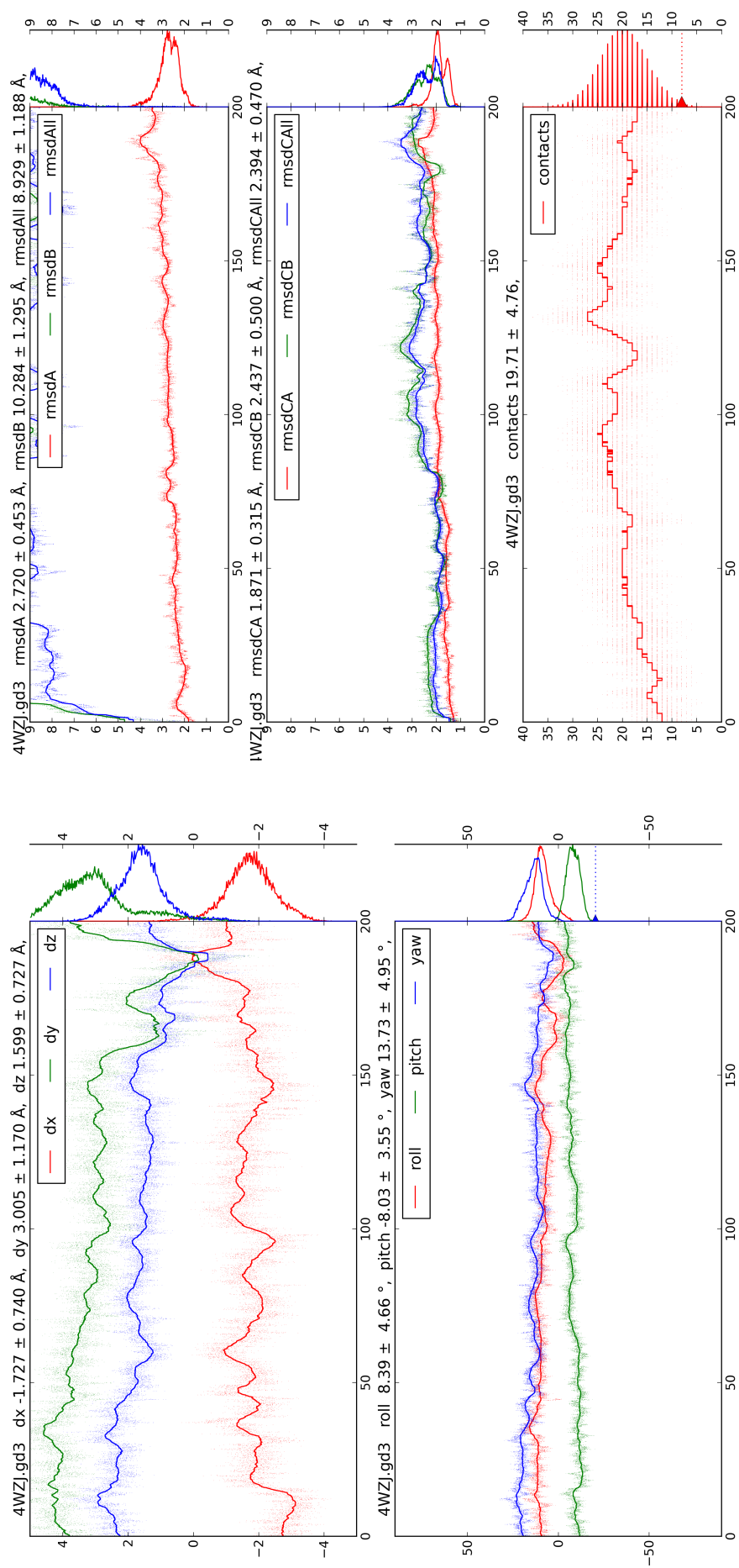


Figure S4.20: **4WZJ fe** (*Homo sapiens*)



(a) PT

(b) Validation

Figure S4.21: **4WZJ gd3** (*Homo sapiens*)

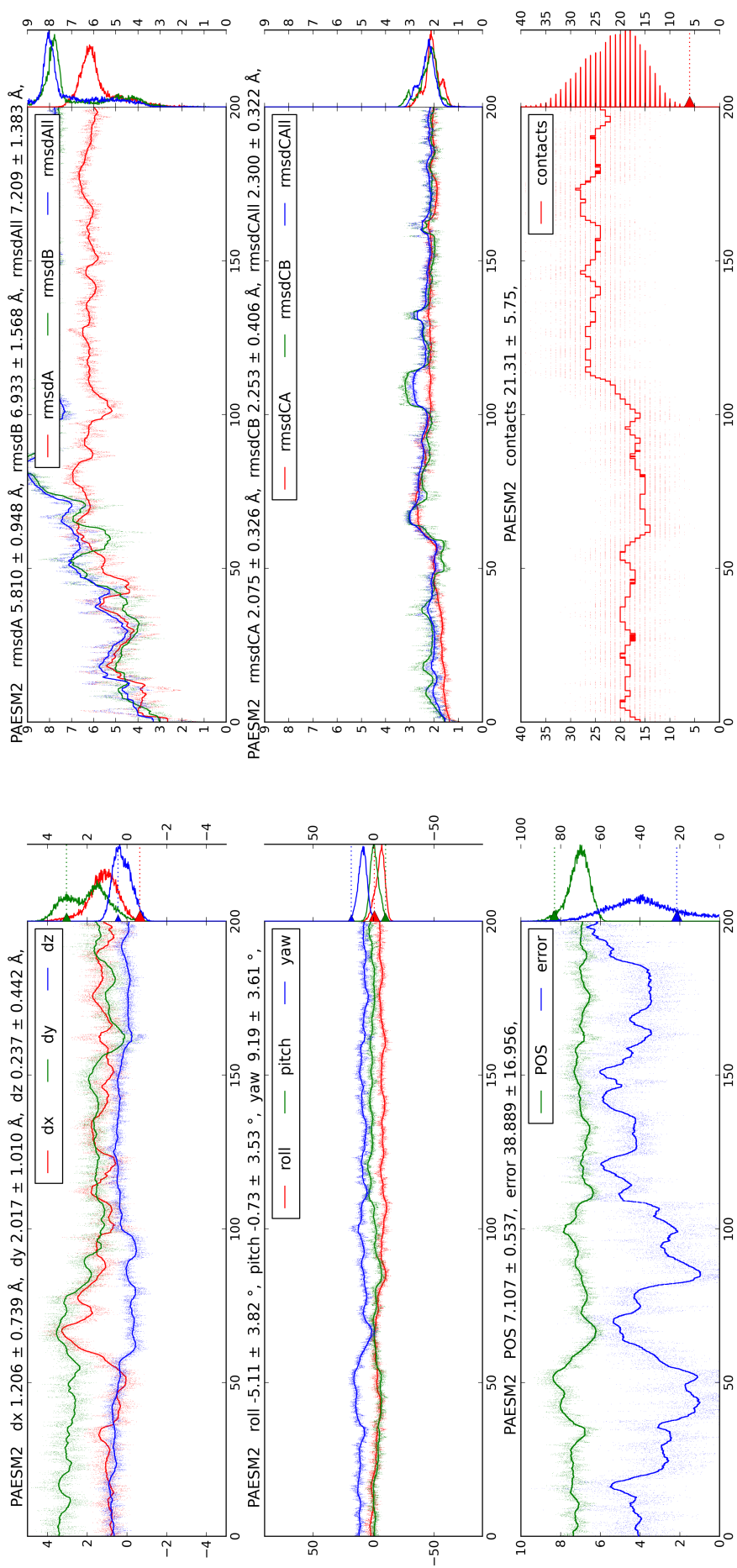
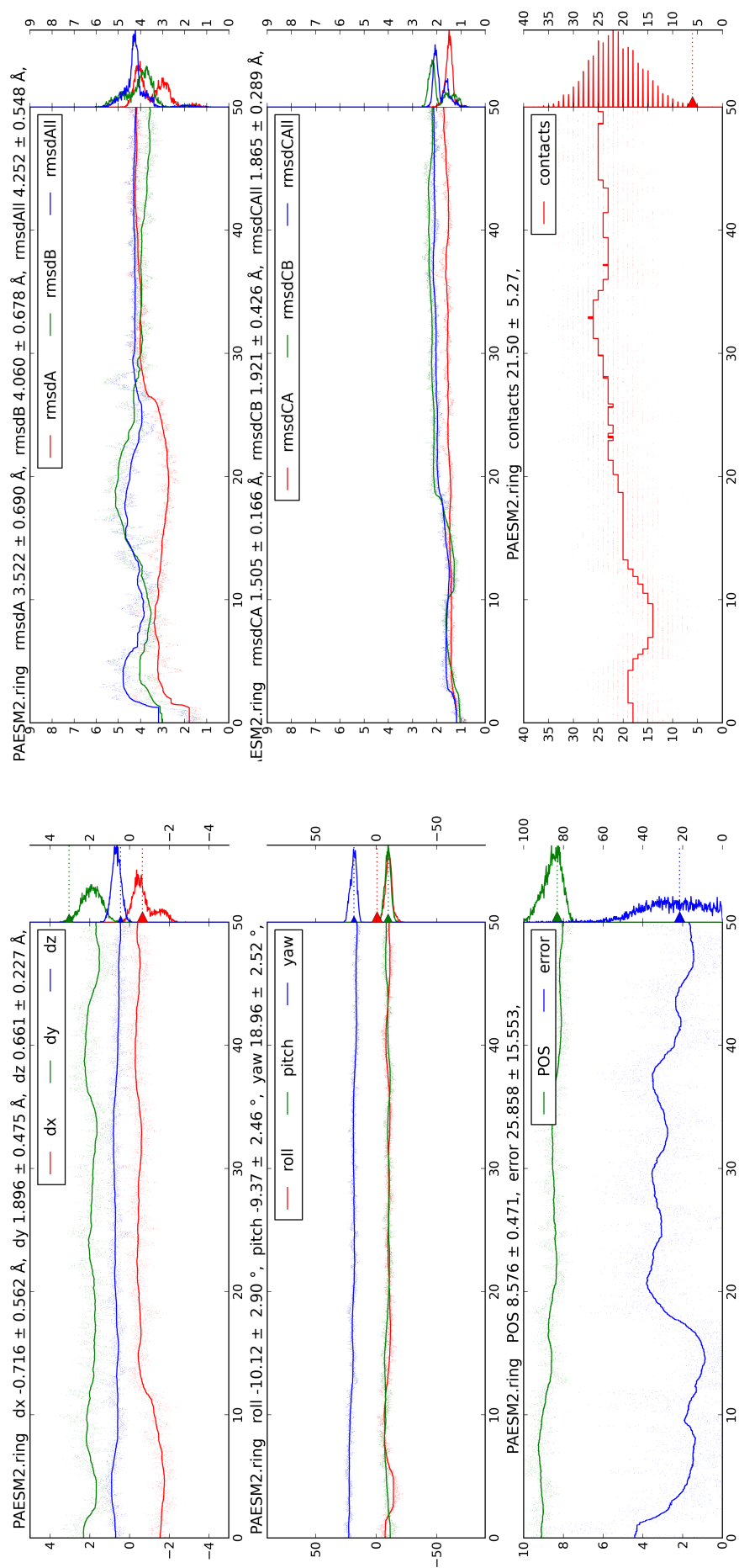


Figure S4.22: **PAESM2** (*Pyrobaculum aerophilum*)

Figure S4.23: **PAESM2 ring** (*Pyrobaculum aerophilum*)

S4.2 POS dependence on atom selection

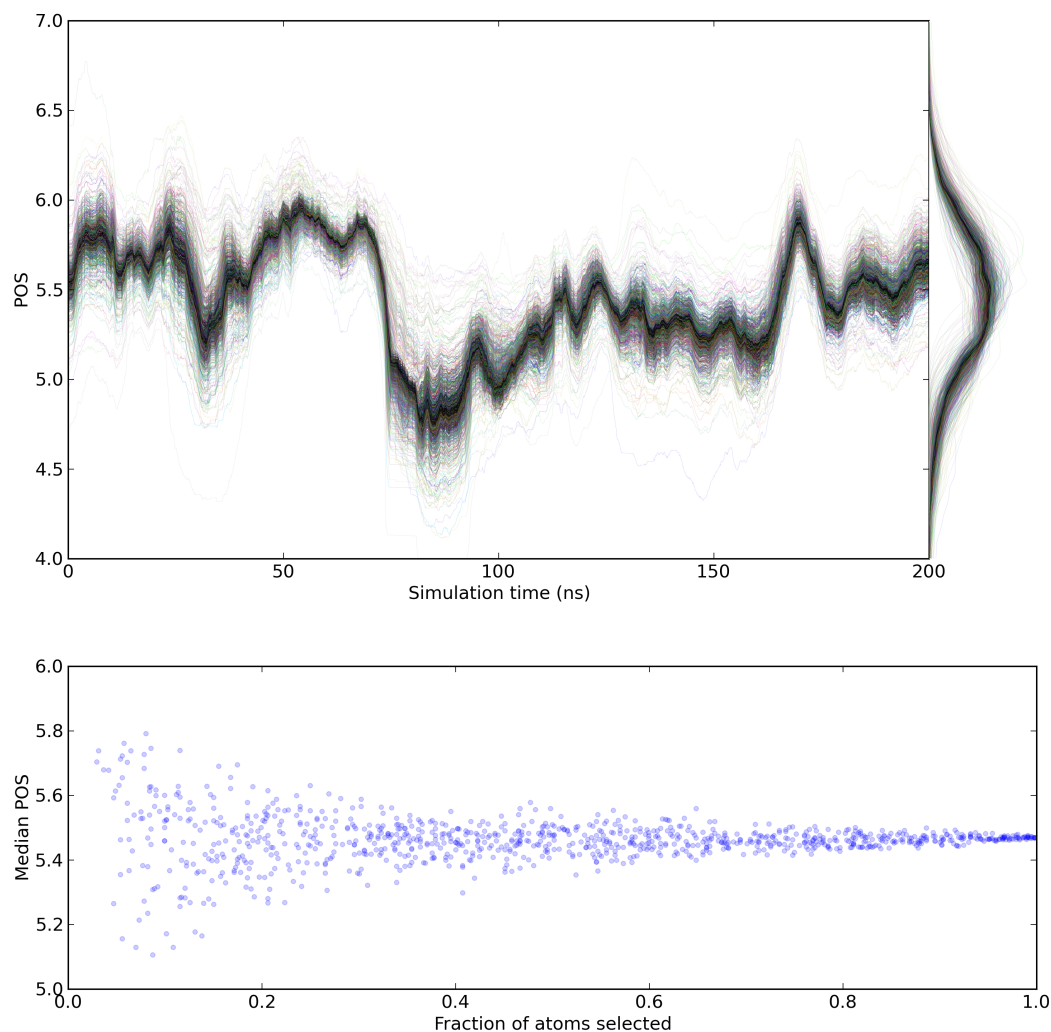


Figure S4.24: **POS calculated for random subsets of atoms**, to show that the POS is not strongly dependent on the particular atoms selected for the calculation. One thousand random subsets of the atoms in 4PNO were used to perform the POS analysis. In the top panel, all of the results are shown. Lighter-colored lines correspond to runs with fewer atoms selected, while darker lines correspond to runs with more atoms selected. In the bottom panel, the median value of the POS is shown as a function of the fraction of atoms selected. After the majority of atoms are used in the calculation, the POS never differs by more than 0.2 from the median with all atoms selected. The outliers tend to include side-chain atoms (only backbone atoms are used in the POS calculations in this paper) and residues at the N and C termini of the protein (data not shown).

S4.3 Principal components analysis of PT values

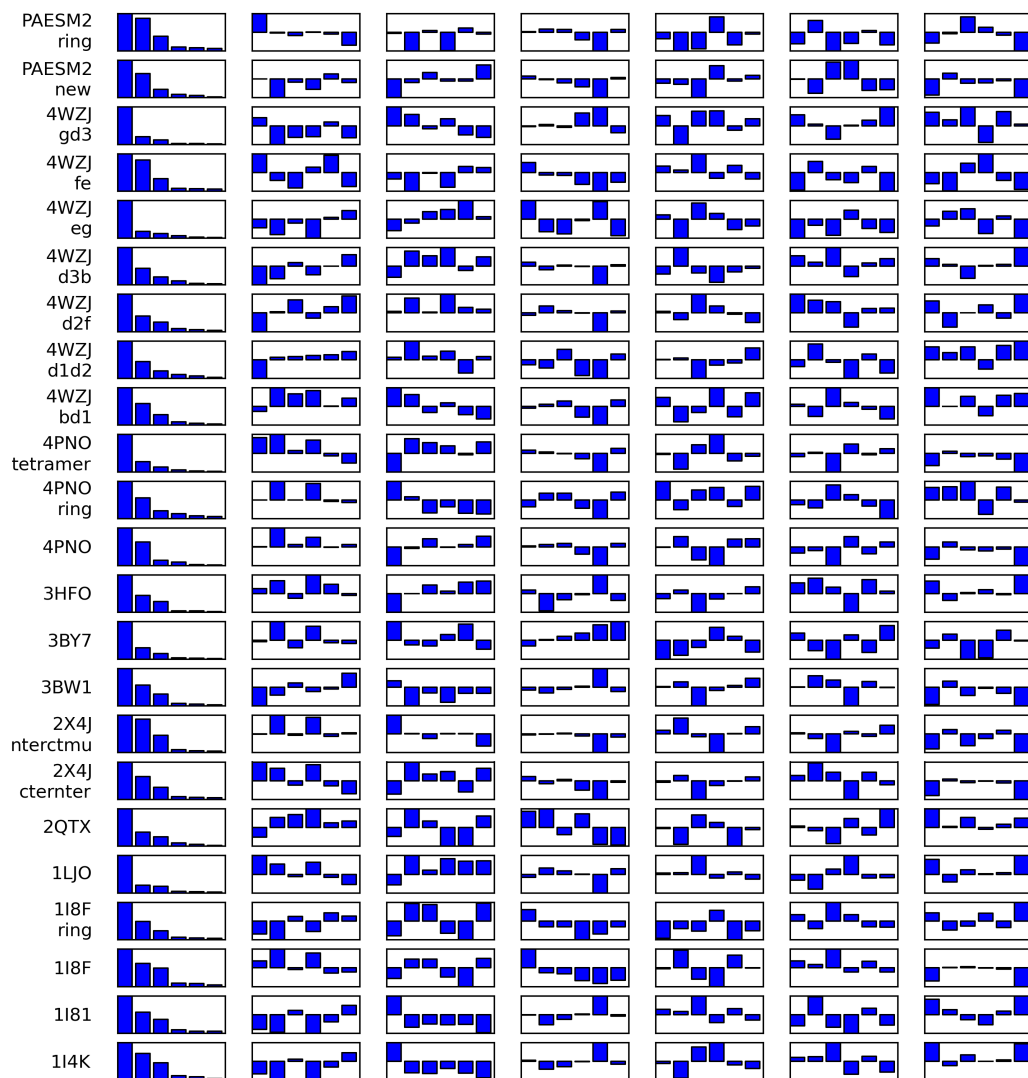


Figure S4.25: **Principal components analysis of the components of the PT.** For each simulation, we have looked for common trends in PT values by performing principal components analysis. The first column shows the relative contribution of each of the six principal components for each system (normalized so that the first component is at 100 %.) The remaining columns show which components of the PT contribute to each mode, in $[dx, dy, dz, roll, pitch, yaw]$ order. For example, the first principal component of “PAESM2 ring” consists of a large positive dx and negative yaw . No trend is readily apparent in the components extracted from this analysis, suggesting that no single combination of PT terms can describe the motions seen in Sm proteins.