Atomistic Simulations of a Protein Unfolding Process at Room Temperature Using the Weighted Ensemble Path Sampling Strategy

by

Andrew G. Bellesis

Master of Science, Clark University, 2017

Bachelor of Arts, Clark University, 2016

Submitted to the Graduate Faculty of the

Dietrich School of Arts and Sciences

in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This thesis was presented

by

Andrew G. Bellesis

It was defended on

February 21, 2020

and approved by

Angela Gronenborn, PhD, Professor and Chair, Department of Structural Biology

Kenneth Jordan, PhD, Distinguished Professor of Chemistry, Department of Chemistry

Jacob Durrant, PhD, Assistant Professor, Department of Biological Sciences

Lillian Chong, PhD, Associate Professor, Department of Chemistry

Atomistic Simulations of a Protein Unfolding Process at Room Temperature Using the Weighted Ensemble Path Sampling Strategy

Andrew G. Bellesis, MS

University of Pittsburgh, 2020

Developing a detailed understanding of protein folding and unfolding processes, including the conformational details of the unfolded state ensemble, has been a longstanding challenge in biophysics. In particular, the transient states of these processes have been difficult to capture using experimental techniques such as NMR spectroscopy and the timescales associated with both processes are beyond the microsecond timescales that can be accessed by standard molecular dynamics (MD) simulations. Here, the weighted ensemble (WE) path sampling strategy was employed in conjunction with MD simulations to enable the simulation of protein unfolding events at room temperature. The WE strategy has enabled the simulation of rare events such as ligand binding, large protein conformational transitions, and protein unfolding. I present atomically detailed simulations of the unfolding process of the G29A mutant of the B-domain of protein A from the virulence factor, *Streptococcal aureus* (BdpA) at room temperature and characterize the conformational details of representative unfolding pathways.

Table of Contents

List of Tables

List of Figures

Preface

Acknowledgements: I would like to humbly thank my advisor, Prof. Lillian Chong, as well as the following current and former members of the Chong Lab for their advisement, support, and camaraderie: Anthony Bogetti, Audrey Pratt, Ali Saglam, Hannah Piston, Paul Torrillo, Page Harrison, and Jeremy Leung.

1.0 Introduction

The BdpA protein is an ideal system for atomistic simulations of protein folding due to its small size (58 residues) and relatively rapid folding on the microsecond timescale[1]. BdpA consists of a three-helix bundle separated by short flexible linkers. Helix 2 and helix 3 are antiparallel to each other; helix 1 is tilted at approximately a 30° angle to the other helices[2] (**Figure** 1). Although the folding and unfolding processes of BdpA have been studied by both experiment and simulation[1,3-14], atomistic pathways of the unfolding process at room temperature have not yet been reported.

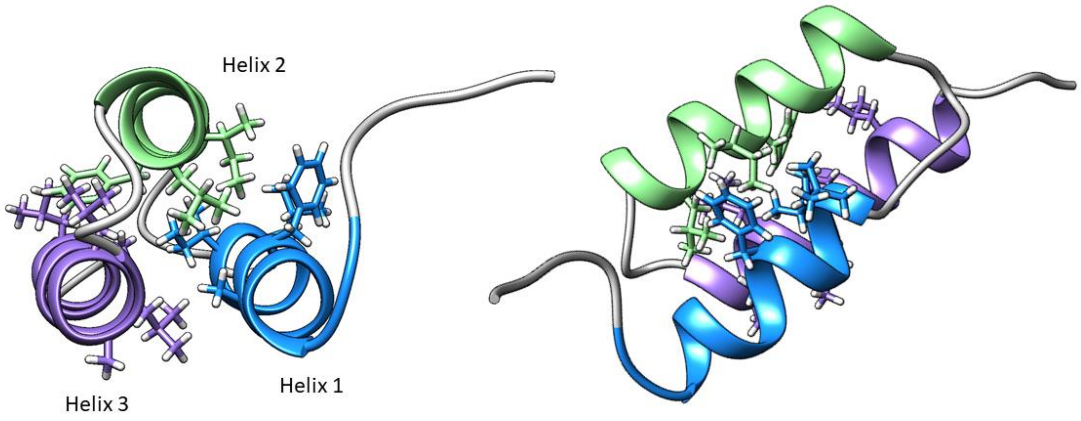

Figure 1: An unpublished G29A BdpA structure determined by Prof. Seth Horne at the University of Pittsburgh shows the same three-helix bundle tertiary fold as that published for wildtype BdpA by Gouda *et al.* in 1992[3]. In both panels, helix 1 (residues 6-17) is shown in blue; helix 2 (residues 24-36) is shown in green, and helix 3 (residues 41-54) is shown in purple. The flexible loop regions on the termini and between the helices are shown in grey. The sidechains of the hydrophobic core residues A12, F13, I16, L17, F30, I31, L34, A42, L44, L45, A48, and L51 are highlighted in stick representation.

The focus of my thesis is on the G29A mutant of BdpA, which is one of the fastest folding proteins yet reported[4]. In particular, the G29A mutation was predicted to increase the folding rate constant[4,15] by stabilizing helix 2[16] due to the higher helical propensity of alanine with respect to glycine[17]. The following rate constants for folding and unfolding for the G29A mutant of BdpA were obtained by lineshape fitting of dynamic NMR spectra at 310 K[4]: k_f = 370,000 s^{-1} and k_u = 37 s^{-1}. Hence, the mean first passage time for the folding process is 3 μs, which is three-fold faster than

the wildtype[1], and the mean first passage time for the unfolding process is 27 ms. For the F13W/G29A mutant, which enables the monitoring of the folding/unfolding process via intrinsic tryptophan fluorescence, the rate constants are k_f = 450,000 s^{-1} and k_u = 100 s^{-1}, which corresponds to mean first passage times of 2.2 μs and 10 ms, respectively [4]. Dimitriadis *et al.* measured folding rate constants of F13W/G29A as a function of temperature through laser-induced temperature-jump kinetics[9]. At 298 K, k_f = 177,000 s^{-1} and k_u = 30 s^{-1}. At 310 K, k_f = 235,000 s^{-1} and k_u = 211 s^{-1}. Although the values of the rate constants differ from those of Arora *et al*[4], both studies reveal that the F13W/G29A mutant folds faster than the wildtype protein. Importantly, the G29A and the F13W/G29A double-mutant each have folding rate constants over threefold higher than the wildtype protein[1]. The G29A mutant has an unfolding rate constant that is twofold lower than the wildtype protein, but threefold lower than the F13W/G29A double mutant. For both the G29A and F13W/G29A mutants, all kinetic traces fit well to a single exponential, supporting the two-state folding hypothesis.

Several additional studies have been carried out on the G29A mutant to better contextualize the kinetics elucidated by Arora *et al*[4]*.* and Dimitriadis *et al*[9]. Sato and Fersht applied extensive Φ-value analysis on the Y15W/G29A double mutant of the BdpA protein using intrinsic tryptophan fluorescence[11]. Φ-value analysis is an experimental technique to quantify how “native-like” interactions in the transition state are by comparing the energy of activation required and equilibrium of mutant proteins with the wildtype protein[18,19]. They determined that the G29A mutation stabilizes the protein by approximately 0.7 kcal/mol and observed an acceleration in both the folding and unfolding rate constants[11]. For the Y15W/G29A mutant, k_f =276,000 s^{-1} and k_u=31.4 s^{-1} at 298 K. For the Y15W single mutant, k_f =98,000 s^{-1} and k_u=25.0 s^{-1}, which is similar to wildtype values. While there is variance in the folding rate constants of the G29A, F13W/G29A, and Y15W/G29A mutants, in part due to differing experimental conditions, the results reveal that the G29A mutation accelerates folding compared to the wildtype and that the G29A mutation

affects folding rate constants more strongly than the Y15W/G29A mutation, which is illustrated in Table 1.

Table 1: A summary of folding and unfolding rate constants elucidated by varied experimental techniques is shown above.

Construct	k_f	k_u	Temperature	Technique
Wildtype	120,000 ± 36,000 s^{-1}	68 ± 18 s^{-1}	310 K	Lineshape fitting NMR/chemical denaturation[1]
G29A	370,000 s^{-1}	37 s^{-1}	310 K	Lineshape fitting NMR/chemical denaturation[4]
F13W/G29A	450,000 s^{-1}	100 s^{-1}	310 K	Lineshape fitting NMR/chemical denaturation[4]
F13W/G29A	177,000 s^{-1}	30 s^{-1}	298 K	Laser-induced temp-jump kinetics/chemical denaturation[9]
F13W/G29A	235,000 s^{-1}	211 s^{-1}	310 K	Laser-induced temp-jump kinetics/chemical denaturation[9]
Y15W/G29A	276,000 s^{-1}	31.4 s^{-1}	298 K	Circular dichroism/chemical denaturation[11]
Y15W	98,000 s^{-1}	25.0 s^{-1}	298 K	Circular dichroism/chemical denaturation[11]

Multiple insights have been made regarding folding and unfolding mechanisms of the G29A mutant from both experiments and simulations. Sato and Fersht hypothesized that the G29A mutation induces “structural strain” in the native state from its methyl sidechain that is released in the transition state[11]. They also concluded that Φ-value analysis measured via temperature-jump experiments and circular dichroism show F15W/G29A BdpA unfolds through a single transition state. On the simulation side, Lei *et al.* conducted extensive atomistic MD

simulations on the folding of the G29A mutant[10]. Starting from the fully-extended unfolded state, a series of conventional MD simulations were carried out at 300 K and 20 replica-exchange MD simulations were carried out with target temperatures between 250 K and 550 K. Lei *et al*. achieved folding to the native state within experimental error of the backbone RMSD of 0.7 $\pm$ 0.3 Å of the high-resolution NMR structure (PDB code 1Q2N)[20]. They observed that folding started with the formation of helix 3 followed by the folding of the helix 2/helix 3 segment and completed by the docking of helix 1[10]. Cheng *et al.* also explored continuous folding MD trajectories from the extended state to the native state of wildtype BdpA[7,10]. They observed helix 2 form first, followed by helix 1, which dock to each other before the formation and docking of helix 3. This is contradictory to the results of Lei *et al*. Although Cheng *et al.* worked with the wildtype construct, their results give some insights into the fast-folding behavior of the G29A mutant. They hypothesize that the long, positively-charged sidechain of R28 may play an important role in docking helix 1 and helix 2 and rigidifying the helix turn from R28-I32, which is already rigidified by the mutation of G29 to an alanine residue, thereby speeding up the folding process. Lastly, Chowdhury *et al.* reported MD simulations of the G29A mutant of BdpA starting from a fully extended protein chain, revealing an initial hydrophobic collapse that involves the formation of both helix 1 and helix 3, followed by the slower formation of helix 2[8]. These results indicated that both native and non-native hydrophobic interactions may play a role in the folding kinetics and that non-native hydrophobic interactions may stabilize the denatured, unfolded state.

Unfolding studies using both experimental and atomistic simulation techniques have also indicated multiple pathways in the wildtype protein. In 1994, Bottomley *et al.* demonstrated that helix 1 unfolds first followed by helix 2 and helix 3 together[6]. Here, the authors inserted tryptophan mutations in each helix and conducted tryptophan fluorescence measurements at increasing concentrations of guanidine HCl denaturant. The helix 1 mutants were less stable under denaturing conditions than the helix 2 and helix 3 mutants, which is consistent with the observation

that helix 2 and helix 3 interact more intimately with each other than they do with helix 1. However, inserting mutations can alter the stability and unfolding pathway of a protein. Alonso and Daggett performed two high-temperature (498 K) unfolding MD simulations of wildtype BdpA and observed two unfolding pathways[3]. In one pathway, the Glu16-Lys50 salt bridge was disrupted, leading to the denaturation of helix 1 followed by helix 2. In a second simulation using the same methodology and starting structure, helix 2 dissociated first, followed by helix 1. In both simulations, helix 3 retained the most helical character. Additionally Bai *et al.* created a series of BdpA fragments and used circular dichroism to determine that helix 3 is the most stable fragment, whereas the helix 1 and helix 2 fragments lose their helical character[5]. Furthermore, the helix 2/helix 3 fragment retains 50% helix whereas the helix 1/helix 2 fragment loses all helical character. Together, these results suggest that helix 3 remains helical even when helix 1 and helix 2 unfold.

Theoretical and experimental studies have shown that BdpA is a two-state folder. Proteins that fold via a two-state mechanism fold cooperatively by crossing a single free-energy barrier and do not form intermediate species along the folding pathway[21]. Meyers and Oas showed that wildtype BdpA denaturation using guanidine HCl measured by circular dichroism produces a curve that fits the two-state model[1]. Dimitriadis *et al.* used ns laser-induced temperature-jump kinetics to report the folding kinetics of the F13W/G29A mutant[9]. They demonstrated that relaxation kinetics were well-described by a single exponential decay and hence a two-state model is supported. Furthermore, two-state unfolding was observed both by varying denaturant and temperature. Importantly, BdpA was shown to fold according to the diffusion-collision theory[1,4,14], which is compatible with two-state folding. Diffusion-collision theory divides proteins into elementary microdomains that are composed of only a few amino acids that explore limited conformational space and move diffusively until colliding with each other in the correct orientation to associate, quickly coalescing into longer-range secondary and finally native tertiary structures that include long-range interactions[15,22,23].

Like most proteins, BdpA adopts a heterogeneous ensemble of conformations in the unfolded state and the nature of this ensemble depends on the experimental conditions. Dimitriadis *et al.* showed that the F13W/G29A mutant lacks residual secondary structure under strongly denaturing conditions (6M guanidine HCl), but that the unfolding pathway transition state is relatively compact, independent of guanidine HCl l concentration[9]. Additionally, Sato *et al.* showed through Φ-value analysis that there is residual secondary structure in the unfolded state at 298 K and 0 M guanidine HCl but no secondary structure at 2M guanidine HCl [9,11-13]. At low guanidine HCl concentrations, there are non-local interactions, which are disrupted at higher concentrations. While experimental techniques using chemical denaturation have elucidated some conformational features of the unfolded state ensemble, no atomistic models of unfolded BdpA under non-denaturing conditions exist – the low population of the unfolded state ensemble under non-denaturing conditions highlights the need for atomistic simulation for elucidating the adopted structures that cannot be detected experimentally.

Here, I present atomically detailed simulations of the BdpA unfolding process at room temperature and characterize the conformational details of representative unfolding pathways using the weighted ensemble (WE) strategy[24] in conjunction with molecular dynamics. In the WE strategy, configurational space is divided into bins along a progress coordinate towards a target state, which need not be defined in advance[24,25]. Below, I will discuss efforts to identify an appropriate progress coordinate for generating unfolding events, define the folded and unfolded states within the context of the progress coordinate, and analyze representative unfolding pathways. Together, these efforts lay the groundwork for characterizing the unfolding mechanism of the BdpA G29A mutant.

2.0 Methods

2.1 The Weighted Ensemble (WE) Path Sampling Strategy

The cost of standard MD simulations remains prohibitive when simulating long-timescale biological processes such as protein folding and unfolding, conformational changes, and ligand binding. Path sampling strategies, including the WE strategy, enhance the sampling of long-timescale processes by focusing computing power on the transitions between stable states rather on the stable states themselves[24,25]. In WE, configurational space is divided into bins along a progress coordinate that lead sequentially to a target state which can be defined post-hoc. N stochastic trajectories are initiated in parallel from the starting state and are given equal statistical weights of 1/N. Dynamics are then propagated for fixed time intervals (τ) and after each interval, a resampling procedure is carried out as shown in **Figure 2**. The resampling procedure involves the replication and pruning of trajectories with the goal of populating each bin along the progress coordinate with a target number of N trajectories. Due to rigorous tracking of trajectory weights, no bias is introduced into the dynamics thereby enabling the calculation of rate constants[25-28]. The spacing of the bins and even the entire progress coordinate can be updated on-the-fly throughout the simulation without introducing bias to facilitate bin transitions toward the target state[27]. WE can be carried out either under steady state or equilibrium conditions[29]. Importantly, WE can be orders of magnitude more efficient than running standard MD simulations in sampling various molecular processes[27-32].

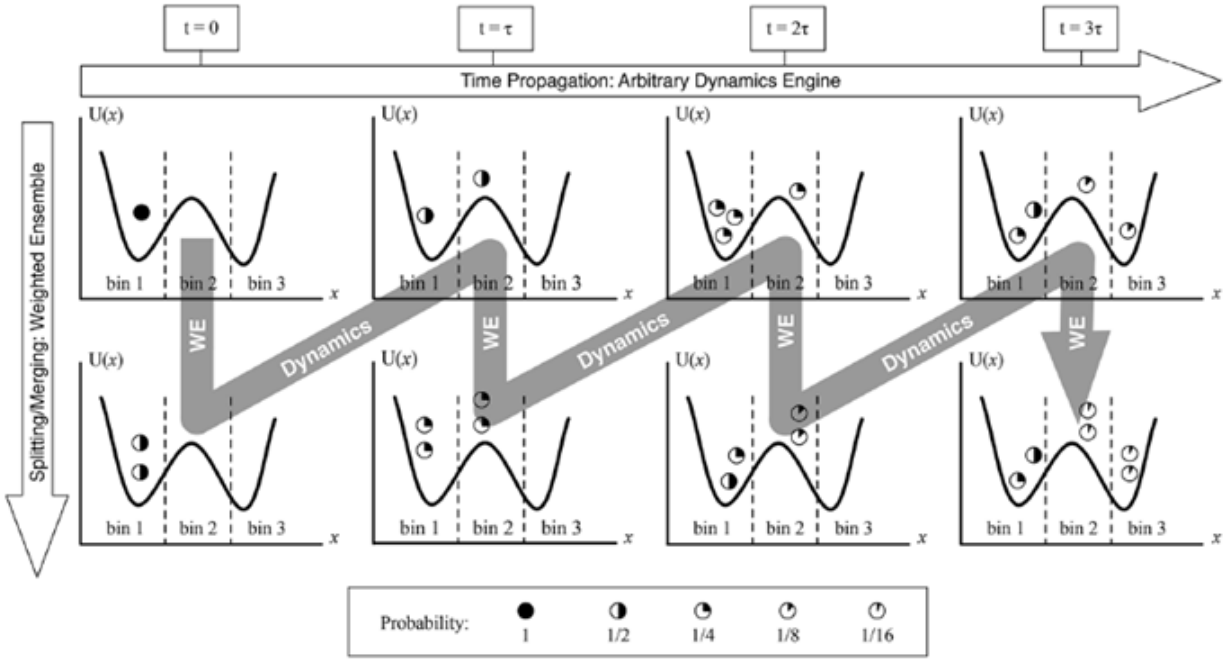

Figure 2 (Above): A detailed schematic of the WE strategy is illustrated for a simple double-well potential consisting of two alternate stable states. The progress coordinate is seen on the x-axis and is divided into user-defined bins. Bin 3 constitutes the target state. In the upper left-hand corner, a starting structure is prepared and placed in the starting bin. The first iteration of WE creates two trajectories with an equal probabilistic weight. A short MD simulation with a fixed time interval of τ is then propagated, at which point one trajectory enters a new bin. The next WE iteration then splits the statistical weight of each trajectory appropriately. Ideally, this process is repeated until a sufficient number of trajectories reach the target state to estimate a rate constant. **Figure reprinted from Donovan *et. al.* (2013)[33]**.

2.2 Preparation of the Simulation System

An unpublished NMR structure of the BdpA G29A mutant solved by collaborator Prof. Seth Horne was used as the starting model **(Figure 1)**. The quality of the model was assessed using the MolProbity webserver[34]. Model 1 scored in the 83rd percentile of structures and showed no Ramachandran outliers and was hence picked as the starting structure for the simulation. Residue 1 was mutated from valine to alanine using the SCAP software program[35]. The C-terminus was capped with an amide group to remove the negative charge using Avogadro[36]. Protonation states were chosen to be consistent with the pH 5.0 experimental conditions, including a positively charged His18 that had been observed experimentally[4]. The PACKMOL software package[37] was used to create an 85 Å truncated octahedral box with BdpA surrounded by 5 acetate ions that

were parameterized for the Amber ff15ipq forcefield[38]. The size of the box was designed such that a high-temperature unfolded conformation of the protein would have 12 Å clearance from the edge of the box; the unfolded conformation was generated by a previous 20-ns MD simulation at 600 K. The protein was modeled using the Amber ff15ipq force field, solvated with SPC/E_B water molecules, and the system was neutralized with 34 Na+ and 28 Cl- ions[39]. The Na+ and Cl- ions used parameters derived by Joung and Cheatham intended for the SPC/E water model[40]. The acetates, Na+ ions, and Cl- ions were added to be consistent with the corresponding concentrations used in experiments. After solvation and neutralization, the total system contained 47,830 atoms.

2.3 Energy Minimization, Equilibration, and Propagation of Dynamics

The solvated, neutralized system was energy-minimized and equilibrated in three stages. In the first stage, the system was equilibrated for 20 ps under NVT conditions at 298 K with a harmonic restraint of 1.0 *kcal/(mol-Å^2)* applied to the heavy atoms of the protein. In the second stage, the system was equilibrated for 1 ns under NPT conditions at 298 K and 1.0 atm using the Langevin thermostat and Monte Carlo barostat. The Langevin thermostat was used because WE simulations require a stochastic thermostat to be used in the underlying dynamics[27]. Again, the protein heavy atoms were restrained. In the third stage, the system was equilibrated for 1 ns without restraints under the same NPT conditions. In all MD simulations, all bonds to hydrogens were constrained to their equilibrium values using the SHAKE algorithm[41], enabling a 2-fs timestep. Non-bonded interactions were truncated at 10 Å and long-range electrostatic interactions were calculated using the particle mesh Ewald method. In all NPT simulations, the Monte Carlo barostat relaxation time was set to 1.0 ps. In all equilibration simulations, the collision frequency for the Langevin thermostat was set to 1.0 ps^{-1} and during production simulations, the collision frequency was reduced to 0.001 ps^{-1}; by reducing the collision frequency, and hence the coupling strength, the perturbation of the dynamics caused by the thermostat is reduced[42]. In all

MD simulations, the Amber ff15ipq forcefield[38] was used in conjunction with the three-point SPC/E_b water model[43], which reproduces experimentally measured tumbling times of proteins in solution[43].

2.4 Weighted Ensemble Simulations of Protein Unfolding

All WE simulations were carried out using the open-source WESTPA software package[44] and dynamics were propagated using the Amber18 software package[39] with the forcefield and parameters described above. A two-dimensional progress coordinate was applied, consisting of (i) the backbone RMSD of helix 1 (residues 6-17) from the folded protein (*i.e.* energy minimized NMR structure) and (ii) the backbone RMSD of the of helix 2 (residues 24-36) from the folded protein. RMSD calculations for each individual helix involved aligning on residues 6-17, 24-36, and 41-54 of the three-helix bundle, excluding loop regions and the flexible N- and C-termini. Each of the two dimensions of the progress coordinate was divided into 10 bins. The bin spacing was automatically adapted after each iteration to encourage bin transitions. The simulation was carried out for N=101 WE iterations each with a fixed interval τ-value of 100 ps, yielding a maximum molecular time (Nτ) of 10.1 ns and an aggregate simulation time of ~3 μs, requiring ~196 hours of wallclock time using 1 to 4 NVIDIA GTX1080 GPUs at a time.

2.5 Assessment of Simulation Convergence

The convergence of the simulation was assessed by monitoring the instantaneous probability distribution at different timepoints during the simulation as a function of the progress coordinate. As shown in **Figure 3**, the probability distribution evolved throughout the simulation. Between iterations 1 and 25, all trajectories remain below 6 Å backbone RMSD for helix 1 and 4 Å backbone RMSD for helix 2. Higher RMSDs are rapidly reached between iterations 50 and 75. The difference in the probability distribution between iterations 75 and 101 is less pronounced, but some unfolded trajectories reach higher RMSDs. Importantly, there are no regions of relatively

high probability where both helix 1 and helix 2 attained high RMSDs, suggesting that unfolded state has not been fully populated and that the simulation did not converge.

2.6 State Definitions

The folded and unfolded states were defined based on the probability distribution as a function of the two-dimensional progress coordinate **(Figure 3)**. The folded state was defined as any conformation having an RMSD of the helix 1 backbone (residues 6-17) of < 5 Å from the folded protein and an RMSD of the helix 2 backbone (residues 24-36) of < 2 Å from the folded protein. Likewise, the unfolded state was defined as any conformation having a helix 1 backbone RMSD of > 6 Å from the folded protein and a helix 2 backbone RMSD of > 4 Å from the folded protein. For both the folded and unfolded state definitions, RMSD calculations for each individual helix involved aligning on residues 6-17, 24-36, and 41-54 of the three-helix bundle, excluding loop regions and the flexible N- and C-termini. This definition of the unfolded state is preliminary and will need to be refined in future simulations that can provide more extensive sampling of the unfolded state.

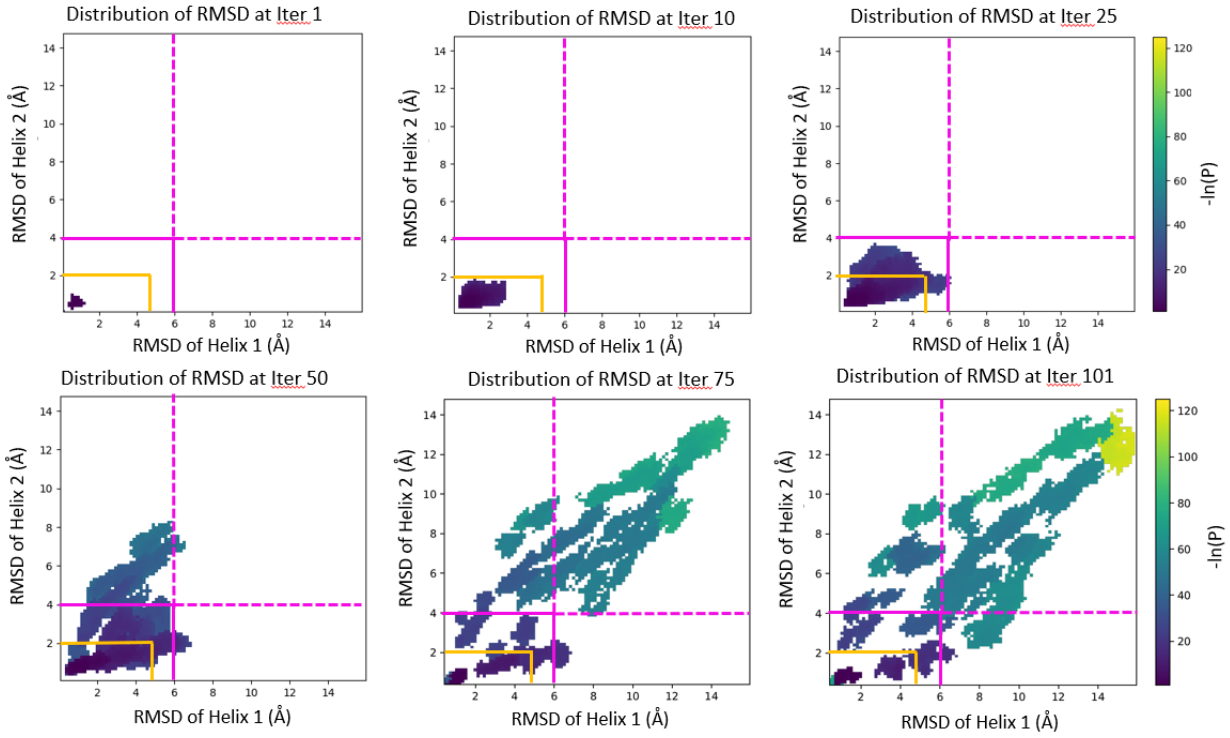

Figure 3: State definitions and assessment of simulation convergence. **A. (Top)**: The probability distribution was visualized after 1, 10, 25, 50, 75, and 101 WE iterations, corresponding to 0.1, 1, 2.5, 5, 7.5 and 10.1 ns of molecular time, respectively. The color bar of each histogram plot corresponds to the probability distribution on an inverted natural log scale corresponding to the free energy. The folded state is defined as being backbone RMSDs below 5 Å for helix 1 and below 2 Å for helix 2 and is delineated by the orange lines. The unfolded state is defined as backbone RMSDs above 6 Å for helix 1 and above 4 Å for helix 2 and is delineated by the pink dashed lines. The RMSDs were calculated for each individual helix by aligning on residues 6-17, 24-36, and 41-54 of the three-helix bundle. The probability continues to evolve at every chosen timepoint. The most probability remains below 6 Å RMSD for helix 1 and 4 Å RMSD for helix 2.

3.0 Results and Discussion

In this study, the WE technique was employed in conjunction with molecular dynamics to simulate unfolding pathways of the BdpA G29A mutant at 298 K. The RMSD of the helix 1 backbone atoms and the RMSD of the helix 2 backbone atoms aligned to the three-helix bundle (residues 6-17, 24-36, and 41-54) of the starting structure were used as the progress coordinate to generate unfolding events. Preliminary definitions of the folded and unfolded states involved the backbone RMSDs of helices 1 and 2. The highest backbone RMSD achieved for helix 1 is 15.5 Å and the helix 2 reaches a maximum backbone RMSD of 14.2 Å. It was observed that helix 1 reached approximately 6 Å backbone RMSD within 25 WE iterations (2.5 ns molecular time) while helix 2 reached approximately 3.5 Å within the same time. After 50 WE iterations (5 ns molecular time) helix 2 reached approximately 8 Å backbone RMSD, but with lower probability than below 4 Å; during this time, helix 1 remained near 6 Å. Hence, helix 2 reached the defined unfolded state before helix 1. After 75 iterations (7.5 ns), helix 1 and helix 2 had both reached approximately 14 Å backbone RMSD, indicating that both helices had undocked from the three-helix bundle. After 10 ns, helix 1 had explored backbone RMSDs over 15 Å, but with very low probability.

Figure 4A illustrates in more detail the conformational space explored by each helix during the course of the simulation. Here, instead of plotting the evolution, the average probability distribution of all 101 iterations is shown. It is observed that the highest probability trajectories remain below 6 Å for helix 1 and 4 Å for helix 2, helping to validate the choice of definitions for the folded and unfolded states. Interestingly, some trajectories enter regions where helix 1 is in the unfolded state while helix 2 remains in the folded state and vice versa. However, trajectories reaching the highest backbone RMSDs show that the two progress coordinates are highly correlated in the unfolded state.

A representative unfolding pathway (Pathway 1), selected because it explores the highest RMSD areas of conformational space explored during the simulation (above 14 Å from the folded three-helix bundle for both helices), is traced in **Figure 4A** and indicates that helix 2 reaches the unfolded state before helix 1. The numbers that are overlaid along the traced pathway indicate the backbone RMSDs of snapshot configurations along the pathway depicted in **Figure 4B.** These configurations illustrate important events along the unfolding pathway. Configuration 1 is the starting structure of the folded state; configurations 2, 3, 4, 5, and 6 correspond to 1 ns, 2.5 ns, 5 ns, 7.5 ns, and 9.8 ns of molecular time, respectively. Configurations 1 and 2 are in the folded state; configuration 3 shows helix 2 in the unfolded state but helix 1 in the folded state; configuration 4 shows helix 2 in the unfolded state and helix 1 between the folded and unfolded states; configurations 5 and 6 are in the fully unfolded state. After 1 ns, the C-terminal end of helix 2 bends outwards from the hydrophobic core. The majority of native contacts – or residue-residue contacts within 4.5 Å formed in the native, folded state – in the hydrophobic core remain formed. Dynamic fluctuations within the folded state ensemble routinely break native contacts observed in the starting structure of a protein. After 2.5 ns, helix 2 begins to lose helical character from its C-terminal end, but the hydrophobic core remains largely intact, retaining 58% of its native contacts. After 5 ns, helix 2 has lost most of its helical character and has become undocked from helix 1 but still interacts with helix 3 through the F30 sidechain. The undocking of helix 2 drastically reduces native interactions in the hydrophobic core to 43%. Helix 1 and helix 3 remain docked and retain some native interactions. After 7.5 ns, helix 1 has undocked from helix 3 and helix 1 has lost some helical structure. Helix 2 has become largely unstructured but undergoes some hydrophobic sidechain interactions with helix 3. Therefore, despite helix 1 and helix 2 both undocking, attaining backbone RMSDs of 14.1 and 13.6 Å from the folded three-helix bundle respectively, 36% of native hydrophobic core interactions are retained. After 9.8 ns, helix 1 remains undocked and has lost some of its structure. Helix 2 remains largely unstructured but

continues to interact with helix 3 and hence 35% of native hydrophobic core interactions remain. Helix 3 retains its helical structure throughout the unfolding pathway.

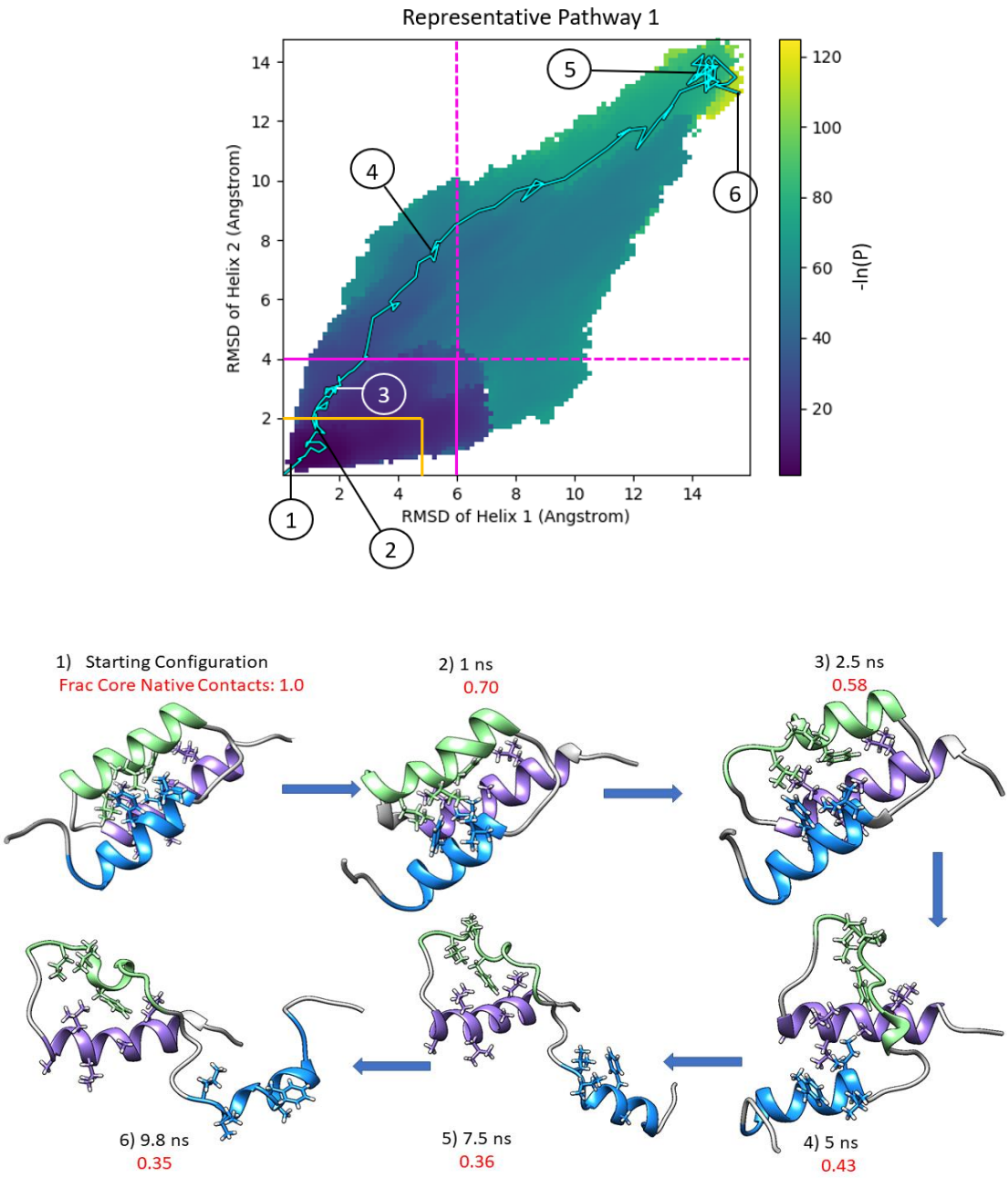

Figure 4A (Top): The cumulative probability distribution is shown. Helix 1 backbone RMSD reaches a maximum of 15.5 Å and helix 2 backbone RMSD reaches a maximum of 14.2 Å from the folded three-helix bundle. A representative unfolding pathway (Pathway 1) is traced in cyan. The folded state is delineated as being within the orange lines; unfolded state is delineated as being outside of the orange lines. Numbers 1-6 indicate the place along the progress coordinate the representative configurations from the unfolding pathway shown in part B were taken from. **B (Bottom)**: Representative configurations from the unfolding

pathway traced in cyan in figure 5A are shown. The fraction of native contacts between the hydrophobic core residues (defined as interactions within 4.5 Å between residues A12, F13, I16, L17, F30, I31, L34, A42, L44, L45, A48, and L51) is noted in red text. Times are given as molecular time.

Additional unfolding pathways were analyzed to understand pathways that entered regions where helix 1 displayed unfolded backbone RMSDs while helix 2 backbone RMSDs remains below the unfolded state cutoff. A representative pathway (Pathway 2) traced over the average probability distribution of all 101 iterations is shown in **Figure 5A**. Helix 2 crosses the unfolded state cutoff earlier in Pathway 2 than helix 1, but later returns to lower backbone RMSDs while helix 1 reaches unfolded state backbone RMSDs. Snapshot configurations of this pathway are shown in **5B** and illustrate important events along the unfolding pathway. Configuration 1 is the starting structure of the folded state; configuration 2, 3, 4, 5, and 6 correspond to 1 ns, 2.5 ns, 4.4 ns, 6.9 ns, and 10.1 ns, respectively. Configurations 2 and 3 are the same as configurations 2 and 3 from Pathway 1 in **Figure 4B**, showing that the two pathways are correlated and do not diverge within the first 25 iterations of the simulation. After 4.4 ns, Pathways 1 and 2 have diverged. Helix 2 becomes dissociated from helix 1 and helix 3, entering the unfolded state but only partially undocking. Helix 1 remains at the cusp of the folded state and retains its helical character. Nonetheless, the hydrophobic core has only retained 42% of native interactions. After 6.9 ns, helix 2 remains in the unfolded state, shows increased flexibility, but still undergoes some sidechain interactions with helix 3. Meanwhile, helix 1 undocks from the three-helix bundle, entering the unfolded state. Despite undocking, helix 1 retains its helical structure and 40% of native hydrophobic core interactions are preserved. After 10.1 ns, helix 2 has left the unfolded state and interacts closely with helix 3 but the backbone RMSD remains above the folded state cutoff. Helix 1 remains undocked in the unfolded state. This movement partially reassembles the hydrophobic core, which is reflected with an uptick to 49% native interactions. Interestingly, the helices largely retain their helical character in each snapshot.

No pathway could be identified where helix 1 unfolded before helix 2. These results suggest that the unfolding of helix 2 before helix 1 may be requisite for the unfolding of BdpA G29A at room temperature. It is possible that helix 1 loses interactions with helix 2 before it can become undocked from helix 3 and that the “breathing” motion of helix 2 dissociating and then reassociating with helix 3 as illustrated here allows for this to happen.

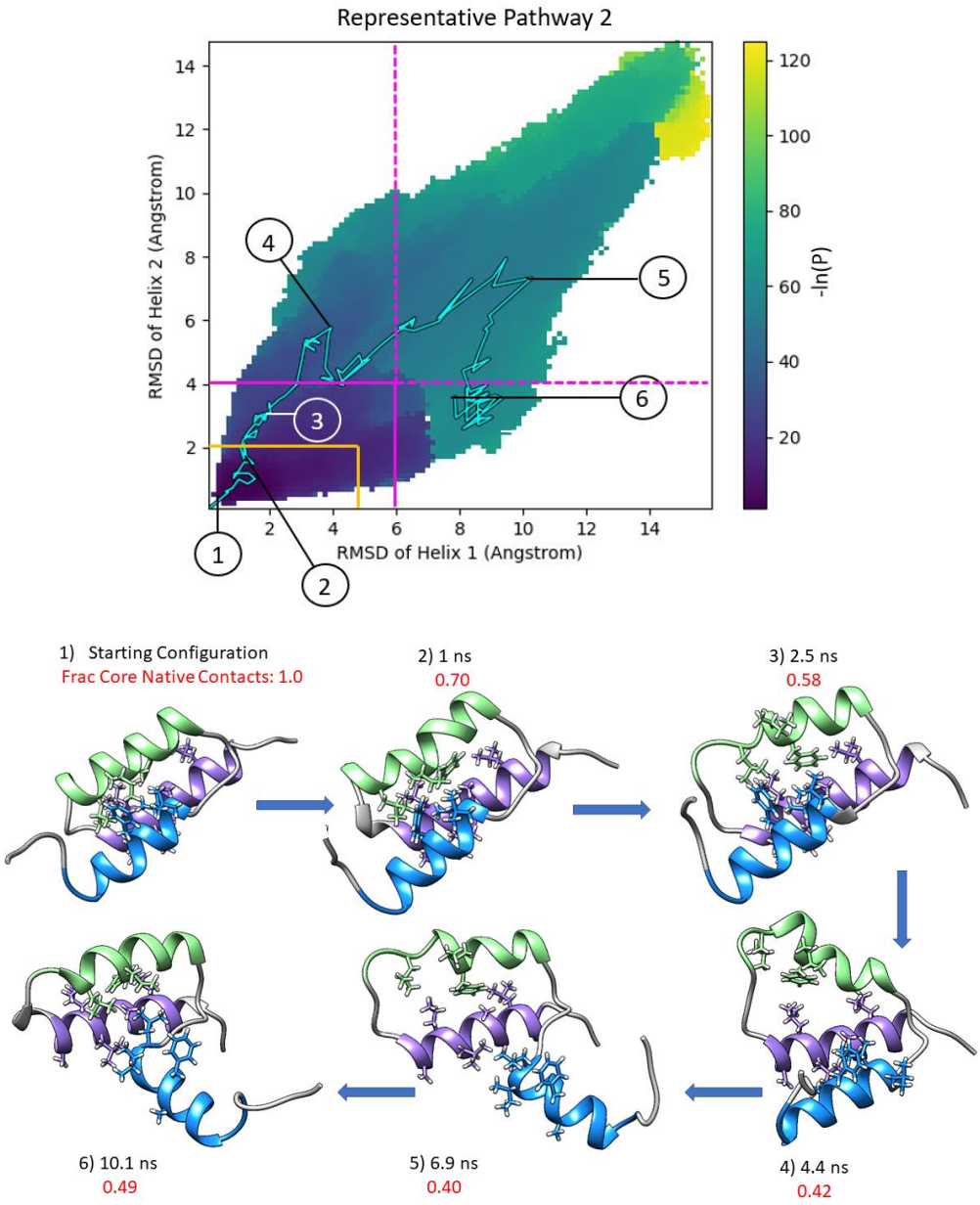

Figure 5A (Above, Top): A second representative pathway (Pathway 2) was traced in cyan. The folded state is delineated as being within the orange lines; the unfolded state is delineated as being outside of the

orange lines. Numbers 1-6 indicate the places along the progress coordinate that the representative configurations from the unfolding pathway shown in part B were taken from. **B (Above, Bottom)**: Representative configurations from the unfolding pathway traced in cyan in figure 6A are shown. The fraction of native contacts between the hydrophobic core residues (defined as interactions within 4.5 Å between residues A12, F13, I16, L17, F30, I31, L34, A42, L44, L45, A48, and L51) is noted in red text. Times are given as molecular time.

To our knowledge, no atomistic simulations have been reported at room temperature for the unfolding process of the BdpA G29A mutant- until now. However, published unfolding studies on the wildtype construct can give indirect insight into the similarities and differences in the unfolding pathways of the wildtype protein vs. the G29A mutant. It is also important to note that the published experiments were conducted under chemically denaturing conditions, unlike our simulation study. The unfolding pathways depicted in **Figures 4** and **5** contradict Bottomley ***et al's*** 1994 tryptophan fluorescence unfolding study on wildtype BdpA that indicated under increasing denaturant conditions, helix 1 unfolds first followed by helix 2 and helix 3 together[6]. This differs from the representative unfolding pathways described here. Alonso and Daggett performed two high-temperature (498 K) unfolding MD simulations on BdpA and observed different unfolding pathways[3]. In one pathway, the Glu16-Lys50 salt bridge was disrupted early in the simulation, leading to the denaturation of helix 1 followed by helix 2. In the second simulation using the same methodology and starting structure, helix 2 dissociated first, followed by helix 1. In both simulations, helix 3 retained the most helical character, like in the representative unfolding trajectories presented here. Importantly, while the simulation presented here shows that helix 1 can be in the unfolded state while helix 2 remains folded, helix 2 needs to at least partially unfold first for helix 1 to undock at room temperature without denaturant. Lastly, the residual secondary structures within the unfolded state observed here reflect experimental results that show residual secondary structure under mildly denaturing conditions[9]. However, because the simulation presented here is not converged, it is impossible to extrapolate fully about the range of pathways that are possible at room temperature without denaturant.

4.0 Conclusions and Future Directions

Here, to our knowledge, the first atomistic unfolding pathways at room temperature of the BdpA G29A mutant have been simulated. A key observation from representative trajectories is that helix 2 partially dissociates from the three-helix bundle, before helix 1 can enter the unfolded state and subsequently undock from helix 3. Afterwards, helix 2 can either fully undock to form an extended unfolded state, as seen in Pathway 1 **(Figure 4)** or re-dock to helix 3 to form a more compact, partially-unfolded state as seen in Pathway 2 **(Figure 5)**. Helix 1 and helix 2 reach higher backbone RMSDs from the folded three-helix bundle in the unfolded state than helix 3, which retains more helical structure than helices 1 and 2 in the representative pathways.

This initial WE simulation lays the groundwork for further room temperature unfolding simulations of BdpA using the weighted ensemble strategy. A promising two-dimensional progress coordinate consisting of the backbone RMSD of helix 1 and the backbone RMSD of helix 2 from the three-helix bundle (residues 6-17, 24-36, and 41-54) of the folded protein structure was identified for generating unfolding events in which the majority of native contacts in the hydrophobic core were broken and the three-helix bundle dissociated upon helices 1 and 2 undocking. This progress coordinate also allowed for a preliminary definition of the unfolded state, characterized by helix 1 having a backbone RMSD > 6 Å and helix 2 having an RMSD > 4 Å from the three-helix bundle of the folded protein structure. The next step is to initiate WE simulations from the sampled unfolded conformations to extensively sample the unfolded state ensemble.

5.0 Works Cited

1. Myers, J.K. & Oas, T.G. Preorganized secondary structure as an important determinant of fast protein folding. *Nat Struct Biol* **8**, 552-8 (2001).

2. Gouda, H. et al. Three-dimensional solution structure of the B domain of staphylococcal protein A: comparisons of the solution and crystal structures. *Biochemistry* **31**, 9665-72 (1992).

3. Alonso, D.O. & Daggett, V. Staphylococcal protein A: unfolding pathways, unfolded states, and differences between the B and E domains. *Proc Natl Acad Sci U S A* **97**, 133-8 (2000).

4. Arora, P., Oas, T.G. & Myers, J.K. Fast and faster: a designed variant of the B-domain of protein A folds in 3 microsec. *Protein Sci* **13**, 847-53 (2004).

5. Bai, Y., Karimi, A., Dyson, H.J. & Wright, P.E. Absence of a stable intermediate on the folding pathway of protein A. *Protein Sci* **6**, 1449-57 (1997).

6. Bottomley, S.P. et al. The stability and unfolding of an IgG binding protein based upon the B domain of protein A from Staphylococcus aureus probed by tryptophan substitution and fluorescence spectroscopy. *Protein Eng* **7**, 1463-70 (1994).

7. Cheng, S., Yang, Y., Wang, W. & Liu, H. Transition state ensemble for the folding of B domain of protein A: a comparison of distributed molecular dynamics simulations with experiments. *J Phys Chem B* **109**, 23645-54 (2005).

8. Chowdhury, S., Lei, H. & Duan, Y. Denatured-state ensemble and the early-stage folding of the G29A mutant of the B-domain of protein A. *J Phys Chem B* **109**, 9073-81 (2005).

9. Dimitriadis, G. et al. Microsecond folding dynamics of the F13W G29A mutant of the B domain of staphylococcal protein A by laser-induced temperature jump. *Proc Natl Acad Sci U S A* **101**, 3809-14 (2004).

10. Lei, H., Wu, C., Wang, Z.X., Zhou, Y. & Duan, Y. Folding processes of the B domain of protein A to the native state observed in all-atom ab initio folding simulations. *J Chem Phys* **128**, 235105 (2008).

11. Sato, S. & Fersht, A.R. Searching for multiple folding pathways of a nearly symmetrical protein: temperature dependent phi-value analysis of the B domain of protein A. *J Mol Biol* **372**, 254-67 (2007).

12. Sato, S., Religa, T.L., Daggett, V. & Fersht, A.R. Testing protein-folding simulations by experiment: B domain of protein A. *Proc Natl Acad Sci U S A* **101**, 6952-6 (2004).

13. Sato, S., Religa, T.L. & Fersht, A.R. Phi-analysis of the folding of the B domain of protein A using multiple optical probes. *J Mol Biol* **360**, 850-64 (2006).

14. Vu, D.M., Myers, J.K., Oas, T.G. & Dyer, R.B. Probing the folding and unfolding dynamics of secondary and tertiary structures in a three-helix bundle protein. *Biochemistry* **43**, 3582-9 (2004).

15. Karplus, M. & Weaver, D.L. Protein folding dynamics: the diffusion-collision model and experimental data. *Protein Sci* **3**, 650-68 (1994).

16. Myers, J.K., Pace, C.N. & Scholtz, J.M. A direct comparison of helix propensity in proteins and peptides. *Proc Natl Acad Sci U S A* **94**, 2833-7 (1997).

17. Pace, C.N. & Scholtz, J.M. A helix propensity scale based on experimental studies of peptides and proteins. *Biophys J* **75**, 422-7 (1998).

18. Fersht, A.R., Leatherbarrow, R.J. & Wells, T.N. Structure-activity relationships in engineered proteins: analysis of use of binding energy by linear free energy relationships. *Biochemistry* **26**, 6030-8 (1987).

19. Fersht, A.R., Matouschek, A. & Serrano, L. The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J Mol Biol* **224**, 771-82 (1992).

20. Zheng, D., Aramini, J.M. & Montelione, G.T. Validation of helical tilt angles in the solution NMR structure of the Z domain of Staphylococcal protein A by combined analysis of residual dipolar coupling and NOE data. *Protein Sci* **13**, 549-54 (2004).

21. Jackson, S.E. How do small single-domain proteins fold? *Fold Des* **3**, R81-91 (1998).

22. Karplus, M. & Weaver, D.L. Protein-folding dynamics. *Nature* **260**, 404-6 (1976).

23. Karplus, M. & Weaver, D.L. Diffusion-Collision Model for Protein Folding. *Biopolymers* **18**, 1421-1437 (1979).

24. Huber, G.A. & Kim, S. Weighted-ensemble Brownian dynamics simulations for protein association reactions. *Biophys J* **70**, 97-110 (1996).

25. Zuckerman, D.M. & Chong, L.T. Weighted Ensemble Simulation: Review of Methodology, Applications, and Software. *Annu Rev Biophys* **46**, 43-57 (2017).

26. Suarez, E. et al. Simultaneous Computation of Dynamical and Equilibrium Information Using a Weighted Ensemble of Trajectories. *J Chem Theory Comput* **10**, 2658-2667 (2014).

27. Zhang, B.W., Jasnow, D. & Zuckerman, D.M. The "weighted ensemble" path sampling method is statistically exact for a broad class of stochastic processes and binning procedures. *J Chem Phys* **132**, 054107 (2010).

28. Zwier, M.C. et al. Efficient Atomistic Simulation of Pathways and Calculation of Rate Constants for a Protein-Peptide Binding Process: Application to the MDM2 Protein and an Intrinsically Disordered p53 Peptide. *J Phys Chem Lett* **7**, 3440-5 (2016).

29. Bhatt, D., Zhang, B.W. & Zuckerman, D.M. Steady-state simulations using weighted ensemble path sampling. *J Chem Phys* **133**, 014110 (2010).

30. Adelman, J.L. & Grabe, M. Simulating rare events using a weighted ensemble-based string method. *J Chem Phys* **138**, 044105 (2013).

31. Saglam, A.S. & Chong, L.T. Highly Efficient Computation of the Basal kon using Direct Simulation of Protein-Protein Association with Flexible Molecular Models. *J Phys Chem B* **120**, 117-22 (2016).

32. Zwier, M.C., Kaus, J.W. & Chong, L.T. Efficient Explicit-Solvent Molecular Dynamics Simulations of Molecular Association Kinetics: Methane/Methane, Na(+)/Cl(-), Methane/Benzene, and K(+)/18-Crown-6 Ether. *J Chem Theory Comput* **7**, 1189-97 (2011).

33. Donovan, R.M., Sedgewick, A.J., Faeder, J.R. & Zuckerman, D.M. Efficient stochastic simulation of chemical kinetics networks using a weighted ensemble of trajectories. *J Chem Phys* **139**, 115105 (2013).

34. Williams, C.J. et al. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci* **27**, 293-315 (2018).

35. Xiang, J., Honig, B. Jackal: A protein structure modeling package. (2002).

36. Hanwell, M.D. et al. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J Cheminform* **4**, 17 (2012).

37. Martinez, L., Andrade, R., Birgin, E.G. & Martinez, J.M. PACKMOL: a package for building initial configurations for molecular dynamics simulations. *J Comput Chem* **30**, 2157-64 (2009).

38. Debiec, K.T. et al. Further along the Road Less Traveled: AMBER ff15ipq, an Original Protein Force Field Built on a Self-Consistent Physical Model. *J Chem Theory Comput* **12**, 3926-47 (2016).

39. D.A. Case, I.Y.B.-S., S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R Harris, N. Homeyer, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D.J. Mermelstein, K.M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, L. Xiao, D.M. York and P.A. Kollman. AMBER 2018. (2018).

40. Joung, I.S. & Cheatham, T.E., 3rd. Molecular dynamics simulations of the dynamic and energetic properties of alkali and halide ions using water-model-specific ion parameters. *J Phys Chem B* **113**, 13279-90 (2009).

41. Ryckaert, J.-P.C., Giovanni; Berendsen, Herman J C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **23**, 327-341 (1977).
42. Basconi, J.E. & Shirts, M.R. Effects of Temperature Control Algorithms on Transport Properties and Kinetics in Molecular Dynamics Simulations. *J Chem Theory Comput* **9**, 2887-99 (2013).

43. Takemura, K. & Kitao, A. Water model tuning for improved reproduction of rotational diffusion and NMR spectral density. *J Phys Chem B* **116**, 6279-87 (2012).

44. Zwier, M.C. et al. WESTPA: an interoperable, highly scalable software package for weighted ensemble simulation and analysis. *J Chem Theory Comput* **11**, 800-9 (2015).