Electronic Supplementary Material (ESI) for Chemical Science.

Implications of the Unfolded State in the Folding Energetics of Heterogeneous-Backbone Protein Mimetics

Jacqueline R. Santhouse†, Jeremy M. G. Leung†, Lillian T. Chong*, W. Seth Horne*

Department of Chemistry, University of Pittsburgh, Pittsburgh, PA 15260.

† These authors contributed equally to this work.

* ltchong@pitt.edu, horne@pitt.edu

SUPPORTING INFORMATION

Materials and Methods

No unexpected or unusually high safety hazards were encountered in the course of the experimental work described below.

Protein synthesis and purification. BdpA variants were prepared by automated Fmoc solid phase methods on NovaPEG Rink Amide Resin (0.1 mmol scale) using a Biotage Alstra synthesizer. All reactions were carried out at room temperature. Couplings were performed with Fmoc-protected amino acids in DMF (4 equiv. relative to resin, 0.2 M), HCTU in NMP (3.9 equiv., 0.2 M), and DIEA (6 equiv.) for 45 min. Fmoc deprotections were performed using 20% 4-methylpiperidine in DMF twice for 5 min each. Resin was washed 3 times with DMF between each step. After synthesis, the resin was washed with DCM and methanol and dried via vacuum desiccation for a minimum of 30 min. After drying, the peptide was cleaved from resin by treatment with a solution of trifluoracetic acid (TFA) / ethanedithiol / water / triisopropylsilane (94 / 2.5 / 2.5 / 1 by volume) followed by agitation for 3.5 h. The resulting mixture was filtered and excess TFA evaporated under a stream of nitrogen. Protein was precipitated by addition of cold ether, centrifuged, and the liquid decanted. The resulting pellets were dissolved in solutions of 0.1% TFA in water (solvent A) and 0.1% TFA in acetonitrile (solvent B) and purified via preparative HPLC on a Hitachi LaChrom Elite system equipped with a Phenomenex Jupiter C18 column (250 x 21.2 mm, 300 Å pore size, 10 μm particle size) using gradients between solvents A and B. Fractions containing pure protein as determined by MALDI MS (Bruker Daltronics UltrafleXtreme MALDI TOF-TOF instrument) were combined and lyophilized. Identity of purified material was confirmed by ESI MS (Thermo Fisher Q Exactive Orbitrap instrument) and purity assessed by analytical HPLC on a Hitachi LaChrom Elite system equipped with a Phenomenex Jupiter C18 column (250 x 4.6 mm, 300 Å pore size, 5 μm particle size) (Figures S1-S5). Stock solutions of each purified protein were prepared in water and concentrations determined by UV absorbance.[1]

NMR data acquisition and structure determination. All NMR experiments were performed on a Bruker Avance 700 MHz spectrometer. NMR samples were prepared from concentrated stocks in water to a final composition of 0.5-2.5 mM protein and 0.2 mM sodium 3-(trimethylsilyl)propane-1-sulfonate (DSS) in H_2O / D_2O (9 / 1) at pH 5.0 ± 0.1 (uncorrected). Spectra were acquired at 303 K for **WT**, **Aib-H2**, and **Aib-H3** and at 283 K for **β^3-H2** and **β^3-H3**. The following two-dimensional homonuclear spectra were acquired for each protein: NOESY (150 ms mixing time), TOCSY (70 ms mixing time), magnitude COSY, and DQF-COSY. FID sizes for each acquisition were 512 t_1-increments of 2048 or 4096 data points. Water signal was suppressed using excitation sculpted gradient pulse sequences with parameters optimized for each variant and SPNAM1 set to Sinc1.1000. All spectra were processed in Topspin and chemical shifts referenced to DSS. Resonances were assigned manually using NMRFAM-SPARKY.[2]

NMR structure determination was carried out by simulated annealing using the program ARIA (Ambiguous Restraints for Iterative Assignment, version 2.3)[3] in conjunction with CNS (Crystallography & NMR System, version 1.2)[4] adapting methods described previously.[5-6] Briefly, ARIA was patched to handle chains containing the artificial monomers, and parameter and topology definitions for each were generated based on analogous atom types already present. Program settings for the structure calculations were modified from program defaults to improve model quality and convergence, as described.[7] H-bond restraints for helical regions were generated based on contiguous medium range NOEs. Backbone φ dihedral restraints were prepared based on $^3J_{H\alpha\text{-}HN}$ and $^3J_{H\beta\text{-}HN}$ coupling constants for α- and β-residues, respectively (φ = -65° ± 25° for $J \leq$

6.0 Hz and φ = -120° ± 40° for $J \geq 8.0$ Hz) when J values could be determined from well-resolved signals in the 1D ^{1}H spectrum or phase-sensitive COSY.[8] NOE distance restraints were generated automatically by the ARIA program, starting from a list of ^{1}H resonances and an unassigned set of integrated NOESY peaks. The set of ten lowest energy structures resulting from the calculation was taken as the final NMR ensemble of the BdpA variant and used in subsequent analysis. Ensemble coordinates and additional experimental data are deposited in the PDB (accession codes **WT**: 7TIO, **β^{3}-H2**: 7TIP, **Aib-H2**: 7TIQ, **β^{3}-H3**: 7TIR, **Aib-H3**: 7TIS) and BMRB (accession codes **WT**: 30980, **β^{3}-H2**: 30981, **Aib-H2**: 30982, **β^{3}-H3**: 30983, **Aib-H3**: 30984).

Circular dichroism spectroscopy. All CD experiments were performed on an Olis DSM17 Spectrophotometer. Samples for CD analysis consisted of 50 µM protein, 10 mM phosphate buffer at pH 7.0. CD scans were collected at 20 °C in the range 200-260 nm with 1 nm increment and 3 s averaging time and corresponding cell-matched buffer blanks subtracted. Thermal melts were monitored at the minimum closest to 220 nm in the range 4-98 °C with a 2 °C increment, 2 min equilibration at each new temperature, and 3 s averaging time. Melts were fit to a two-state folding model to generate population normalized unfolding curves and reported T_{m} values.[9] Coupled thermal / chemical denaturation experiments were carried out using methods detailed previously.[10-11] Briefly, samples of each protein were prepared under conditions described above with differing concentrations of guanidinum chloride. Optimal denaturant concentration ranges for each variant were determined based on observed stability, and a series of evenly spaced concentrations within that range utilized for the resulting thermal unfolding experiments. Final data points were collected at 25 °C to check for irreversible aggregation during unfolding. Data for each protein were globally fit to a two-state folding model using the program Mathematica to produce reported thermodynamic parameters and uncertainties for folding. The parameter describing the linear dependence of folded baseline ellipticity as a function of guanidinium was set to zero for β^{3}-residue containing variants due to their high sensitivity to chemical denaturation.

Weighted ensemble simulations. All weighted ensemble (WE) simulations were run using the WE path sampling strategy,[12-13] as implemented in the WESTPA 2.0 software package.[14] The WE strategy enhances the sampling of stable states or transitions between stable states by running a large number of properly weighted trajectories in parallel and iteratively replicating trajectories at short time intervals τ that have made transitions to less-visited regions of configurational space. The relevant configurational space is typically defined by a progress coordinate that has been divided into bins. Trajectory weights are rigorously tracked such that no statistical bias is introduced into the dynamics. WE simulation can be run under either non-equilibrium steady state or equilibrium conditions.[15]

Simulation workflow. To extensively sample the folded and unfolded states of each BdpA protein, WE simulations were run in two stages at 25 °C under equilibrium conditions. In Stage 1, unfolding simulations were run to sample the folded state ensemble and generate unfolding transitions to provide initial, representative conformations of the unfolded state ensemble. In Stage 2, simulations were initiated from the unfolded conformations generated in Stage 1 to extensively sample the unfolded state ensemble. Further details of these simulations are provided below.

Stage 1: Simulation of the folded state and generation of unfolded conformations. For each BdpA variant, a single unfolding simulation was initiated from the corresponding, equilibrated NMR structure. These simulations used a one-dimensional "nested" WE progress coordinate that initially consisted of an all-atom RMSD of helix 2 after alignment on the three-helix bundle and then upon reaching an RMSD value > 8 Å, was switched to an all-atom RMSD of helix 3 after alignment on the helix bundle. This progress coordinate was chosen based on our findings from exploratory simulations that helix 1 unfolds the most easily and helix 3 appears to unfold only after helix 2 unfolds. While this progress coordinate would not be ideal for monitoring the progress of a folding process, the coordinate is effective for sampling the folded state ensemble and generating *unfolding* events at room temperature, providing unfolded conformations for initiating Stage 2 simulations of the unfolded state ensemble (Figure S10). To adaptively position bins along the progress coordinate, the minimal adaptive binning (MAB) scheme [16] was used with 10 equally spaced bins between the trailing and leading trajectories. A τ value of 100 ps was used along with a target number of 5 trajectories per bin to provide reasonably even coverage along the progress coordinate. Each unfolding simulation was run for 450 WE iterations, which is equivalent to 45 ns of "molecular time", defined as Nτ where N is the number of WE iterations and τ is the fixed-time interval for WE resampling (100 ps), yielding an aggregate simulation time of 3.2 μs for the **WT** and 5.3 μs for each BdpA variant. Each simulation was completed within 10 days using 16 NVIDIA Tesla V100 GPUs on Pittsburgh Supercomputing Center (PSC)'s Bridges-2 supercomputer. Based on these unfolding simulations, 5-8 representative unfolded conformations with the largest RMSD-progress-coordinate values and lowest number of native contacts in the hydrophobic core were selected for use as initial structures for Stage 2 simulations of the unfolded state ensemble, prioritizing structures from trajectories that do not share a common parent.

Stage 2: Simulations of the unfolded state. To extensively sample the unfolded state ensemble of each BdpA variant, five independent WE simulations were initiated from the set of representative unfolded conformations generated by Stage 1 simulations of the unfolding process. These Stage 2 simulations of the unfolded state employed a one-dimensional progress coordinate consisting of the all-atom RMSD from the equilibrated NMR structure of the corresponding folded state. Each simulation was run for 300 WE iterations with the same τ (100 ps), target number of trajectories per bin (5 trajectories/bin), and MAB scheme settings (10 bins between the trailing and leading trajectories) used for the Stage 1 unfolding simulations. This number of iterations is equivalent to 30 ns of molecular time (defined above), yielding a total simulation time of ~90 μs over all 25 Stage 2 WE simulations. These simulations were completed within 10 days using 40 NVIDIA Tesla V100 GPUs at a time on PSC's Bridges-2 supercomputer. Conformations were saved every 100 ps for analysis.

Dynamics propagation. Dynamics were propagated using the AMBER 18 software package[17] with the Amber ff15ipq-m force field for protein mimetics[18] (https://github.com/chonglab-pitt/force-fields/tree/main/ff15ipq-m) and SPC/E$_b$ water model.[19] Heavy-atom coordinates for initial models of the folded proteins were extracted from the corresponding NMR structures determined in this study. Hydrogen atoms were added to each model using ionization states present in solution at pH 7. Each system was immersed in a sufficiently large truncated octahedral box of explicit water molecules to provide a minimum clearance of 15 Å between the protein and box walls. Counterions were added to neutralize and achieve concentrations of 100 mM NaCl and 20 mM NaOAc used in experimental conditions.

Prior to running WE simulations, each solvated system was first subjected to energy minimization followed by two stages of solvent equilibration while applying harmonic positional restraints to the proteins with a force constant of 1 kcal mol^{-1} Å^{-2}. In the first stage, the restrained system was heated from 0 to 25 °C for 25 ps in an NVT ensemble. In the second stage, the solvent was subjected to 1 ns equilibration in an NPT ensemble, constraining the positions of all heavy atoms of the protein. The entire system was then equilibrated without any restraints for 1 ns. Given that the WE strategy is rigorous with stochastic dynamics,[20] a weak stochastic thermostat was used (i.e. Langevin thermostat with a collision frequency of 1 ps^{-1}) to maintain a constant temperature of 25 °C. To maintain a constant pressure of 1 atm, a Monte Carlo barostat was applied with a coupling constant of 100 steps. To enable a 2-fs timestep, all bonds to hydrogens were restrained to their equilibrium values using the SHAKE algorithm.[21] Short-range nonbonded interactions were calculated using a cutoff of 10 Å and the particle mesh Ewald method [22] was applied to treat long-range electrostatics.

State definitions. For all analysis, folded and unfolded states of each BdpA variant were defined as regions with -lnP < 4 where P is the probability as a function of the fraction of native contacts and radius of gyration (R_g) from the first and second stages of WE simulations, respectively (Figure S10). For the folded state, this region corresponds to >60% native contacts and an R_g of 11-14 Å. For the unfolded state, this region corresponds to 30-60% native contacts and an R_g between 11 and 18 Å. Helices were defined as residues 6-17 for helix 1, residues 24-36 for helix 2, and residues 41-54 for helix 3.

Reweighting trajectories for equilibrium conditions. To obtain probability distributions of alternate conformations in the folded and unfolded state ensembles of each BdpA protein, state populations from Stage 2 WE simulations were reweighted for equilibrium conditions by applying a Markov state model (MSM) analysis procedure [23] using a customized version of the msm_we Python package (https://github.com/westpa/msm_we) and the haMSM plugin for the WESTPA 2.0 software package.[14] In this analysis procedure, trajectories were first "featurized" using a stratified clustering procedure to discretize the configurational space into "microbins" followed by the construction of a transition matrix among the microbins.

The stratified clustering procedure involved two steps: (i) group structures into individual "strata" corresponding to each of 20 bins used for resampling during the Stage 2 WE simulation, and (ii) cluster the structures within each bin to generate 10 microbins using the mini-batch k-means algorithm (as implemented in the scikit-learn package [24]) along with a pairwise all-atom RMSD as the similarity metric and a tolerance of 10^{-5}. The clustering procedure yielded a total of 200 microbins among all 20 WE bins. The WE bins were positioned along a progress coordinate consisting of the all-atom RMSD from the equilibrated folded structure, as specified by the array [-inf, -0.5, 0, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, inf]. If there were no transitions between microbins, the structures of the disconnected clusters were reassigned to the nearest connected cluster. The final set of microbins with associated statistical weights from the WE simulation was then used to construct a transition matrix to estimate steady-state populations. The updated populations (statistical weights) were then redistributed to individual trajectories based on their statistical weights from the WE simulation using Algorithm 5.3 in reference[25]:

$$\omega_{new}^{i} = pSS_p \cdot \ \omega^i / \sum_{j \in p} \omega^j$$

where ω^i is the statistical weight of trajectory i, p is the microbin occupied by trajectory i, and pSS_p is the estimated equilibrium state population of microbin p.

To assess the convergence of the state populations, reweighted probability distributions and heavy-atom RMSD values from the folded state of each MSM model were compared with those of MSM models constructed using two subsets of the data (groups 1 and 2), each with six equally-sized blocks. The resulting R_g probability distributions from the MSM models of the two groups are comparable to the MSM model for the entire dataset (Figure S16) regardless of whether the same or different cluster centers are used, demonstrating reasonable convergence of the equilibrium state populations.

Clustering of simulated unfolded state ensembles. To better characterize the simulated ensembles of the unfolded state for each protein, structures were clustered using a "bottom-up" agglomerative hierarchical clustering algorithm, as implemented in the cpptraj module of the AMBER 18 software package.[17] The unfolded state ensemble of each protein was clustered separately using two different metrics: (i) the pairwise "best-fit" C_α RMSD of the three helices, and (ii) the mass-weighted R_g of the entire protein. Using the RMSD metric, clusters were merged "bottom-up" until the average distance between any pair of structures between two clusters (average-linkage) was ≥ 4Å. Using the R_g metric, clusters were merged until the average-linkage was ≥ 0.5 Å. The probability of each resulting cluster was determined by summing over the statistical weights of each structure in the cluster, as provided by the WE simulations.

Probability maps of residue-level tertiary contacts. To generate probability maps of pairwise residue tertiary contacts for each simulated unfolded state ensemble, a heavy-atom distance matrix for each protein conformation in the ensemble was first calculated using cpptraj. Next, each matrix was assigned the corresponding statistical weight from the WE simulation to generate a weighted-average probability map of tertiary contacts for the entire ensemble of unfolded conformations, defining a tertiary contact as a pair of residues with $|i-j| \geq 6$ and ≤ 5 Å distance between heavy atoms. Any contacts absent in the equilibrated reference NMR structure are designated as non-native.

Other analysis of simulations. All observables except for the solvent-accessible surface area (SASA) were calculated using the cpptraj module of the AMBER 18 software package. SASA values were calculated using the Shrake-Rupley algorithm,[26] as implemented in the MDTraj analysis suite.[27] To facilitate the extraction of structures and/or trajectories from a probability distribution of a given order parameter (or set of order parameters), the k-d tree "nearest neighbors" search algorithm was applied. We have provided a Python script for the k-d tree search algorithm at https://github.com/chonglab-pitt/bdpa_scripts.

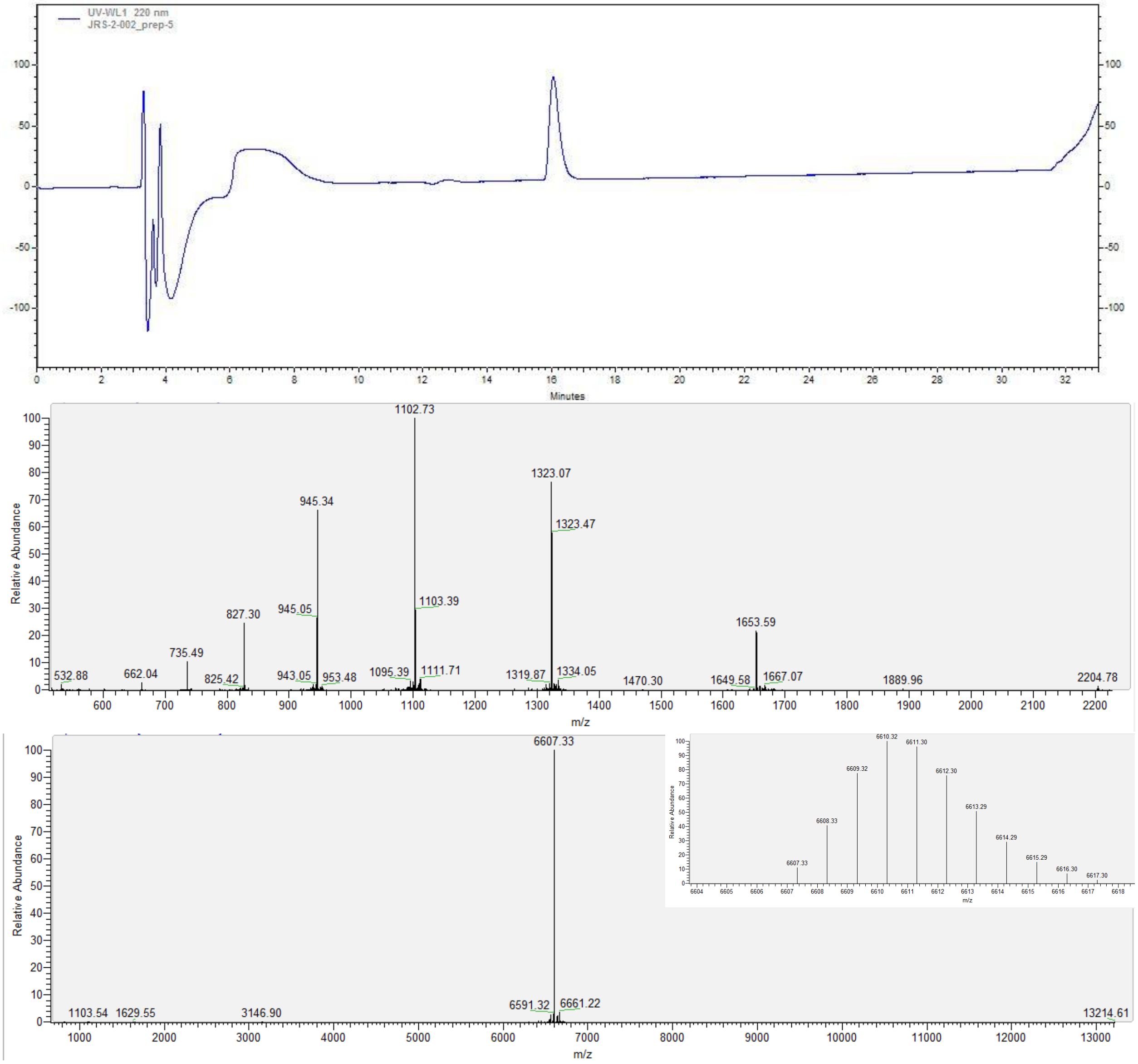

Figure S1. Analytical HPLC (top; gradient: 20%-20%-40% solvent B, 0-3-33 min), raw ESI-MS (middle), and deconvoluted ESI-MS for species $[M]^+$ (bottom) for purified **WT** (monoisotopic $[M]^+$ m/z calc. = 6607.4).

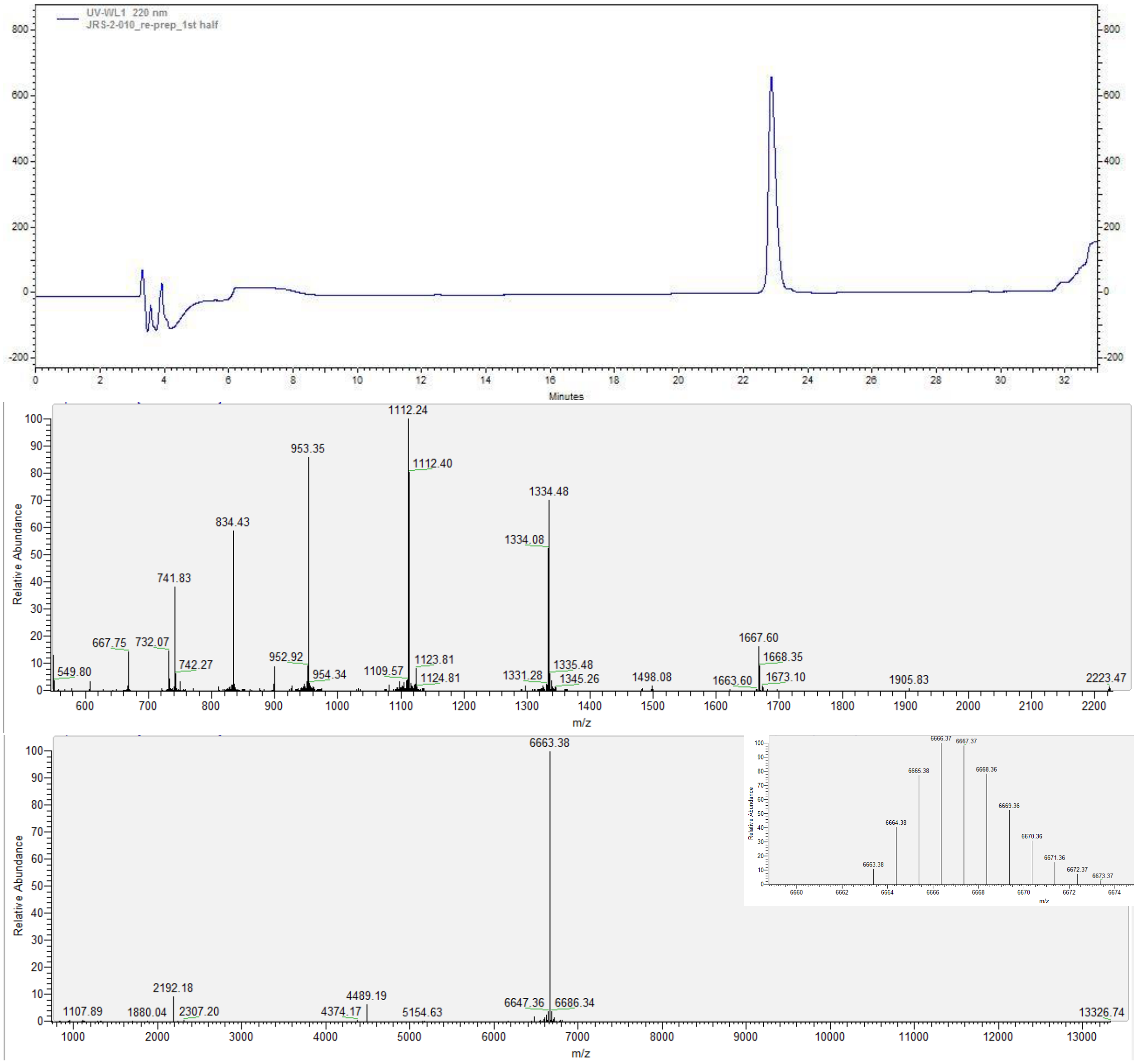

Figure S2. Analytical HPLC (top; gradient: 20%-20%-40% solvent B, 0-3-33 min), raw ESI-MS (middle), and deconvoluted ESI-MS for species $[M]^+$ (bottom) for purified **β^3-H2** (monoisotopic *m*/*z* calc. = 6663.4).

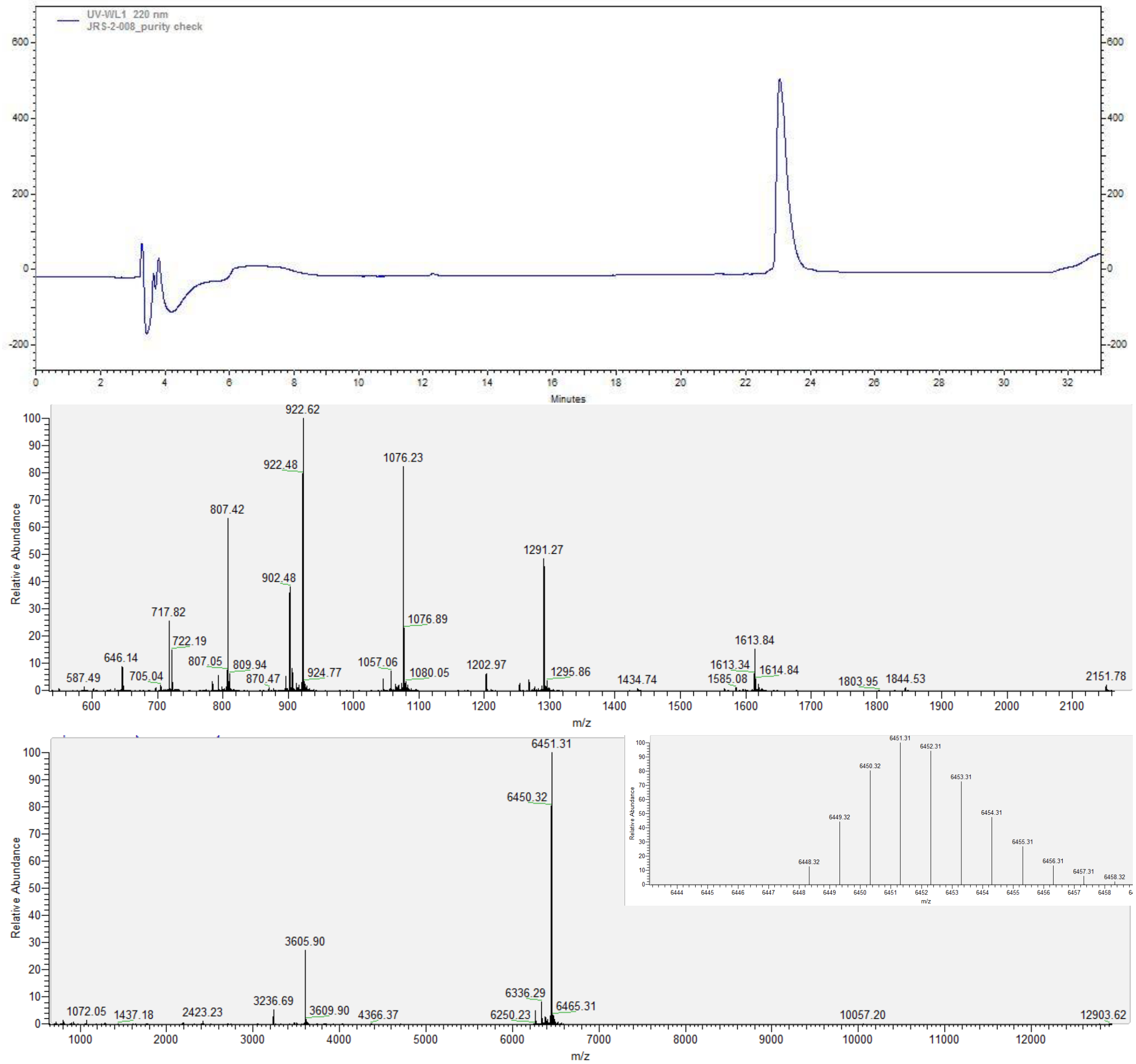

Figure S3. Analytical HPLC (top; gradient: 20%-20%-40% solvent B, 0-3-33 min), raw ESI-MS (middle), and deconvoluted ESI-MS for species $[M]^+$ (bottom) for purified **Aib-H2** (monoisotopic *m*/*z* calc. = 6448.3). Minor species with deconvoluted masses 3605.9 and 3236.7 are both attributed to x type ions resulting from in-source fragmentation between Arg27-Aib28 and Glu24-Aib25, respectively.

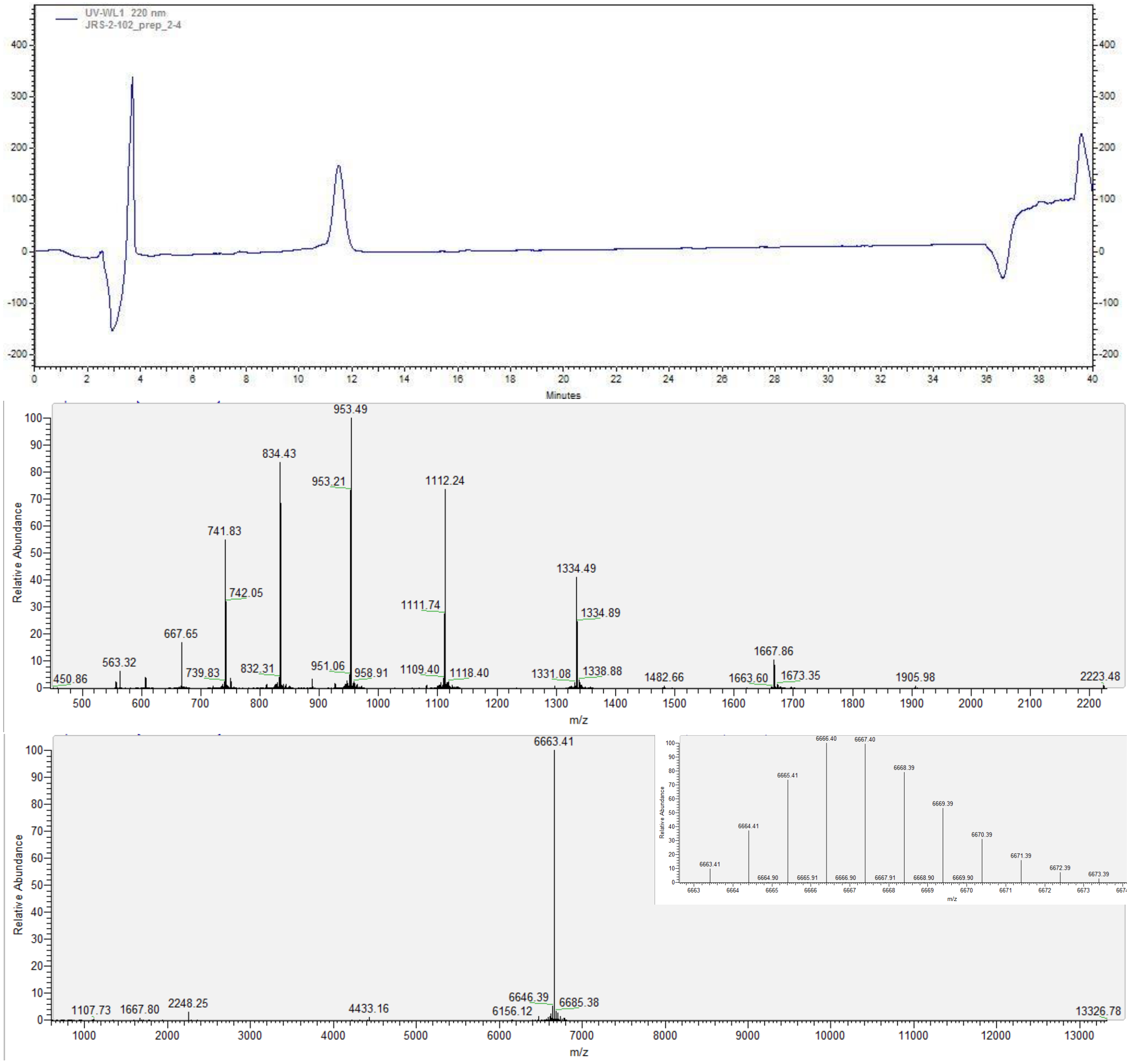

Figure S4. Analytical HPLC (top; gradient: 30%-30%-50% solvent B, 0-3-33 min), raw ESI-MS (middle), and deconvoluted ESI-MS for species $[M]^+$ (bottom) for purified **β^3-H3** (monoisotopic *m*/*z* calc. = 6663.4).

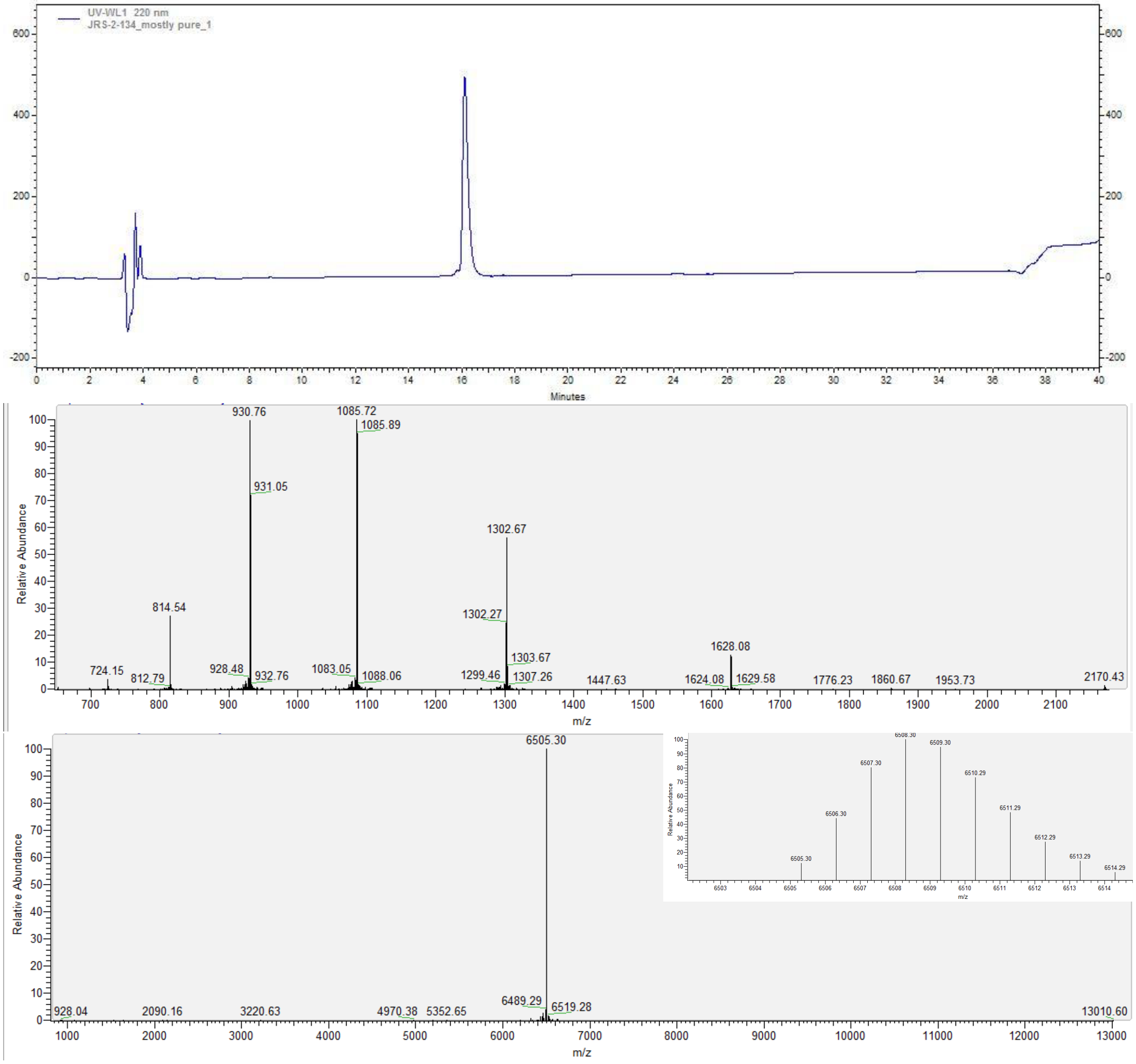

Figure S5. Analytical HPLC (top; gradient: 25%-25%-50% solvent B, 0-3-33 min), raw ESI-MS (middle), and deconvoluted ESI-MS for species $[M]^+$ (bottom) for purified **Aib-H3** (monoisotopic *m*/*z* calc. = 6505.3).

Table S1. Statistics from NMR structure calculations for **WT**.

PDB Accession Code	7TIO
Experimental restraints	
Unambiguous NOEs	905
Intra-residue	417
Sequential ($\|i - j\| = 1$)	185
Medium-range ($1 < \|i - j\| < 5$)	162
Long-range ($\|i - j\| \geq 5$)	141
Ambiguous NOEs	270
Total NOEs	1175
H-bonds	56
Dihedrals	0
Violations	
NOE >0.5 Å	21.8 ± 4.2
NOE rmsd (Å)	0.15 ± 0.02
H-bond >0.5 Å	0
Dihedral >5°	n/a
Ensemble rmsd	
Backbone heavy atoms	1.35 ± 0.52
All heavy atoms	1.45 ± 0.42
Geometry analysis	
rmsd bonds (Å)	0.00280 ± 0.00008
rmsd angles (°)	0.42 ± 0.01
rmsd impropers (°)	1.1 ± 0.1
Ramachandran analysis[a]	
Favored (%)	94.7
Allowed (%)	5.3
Disallowed (%)	0

[a] Performed using the MolProbity server;[28] artificial residues excluded.

Table S2. Statistics from NMR structure calculations for **β^3-H2**.

PDB Accession Code	7TIP
Experimental restraints	
Unambiguous NOEs	794
Intra-residue	416
Sequential ($\lvert i-j \rvert = 1$)	140
Medium-range ($1 < \lvert i-j \rvert < 5$)	111
Long-range ($\lvert i-j \rvert \geq 5$)	127
Ambiguous NOEs	245
Total NOEs	1039
H-bonds	56
Dihedrals	5
Violations	
NOE >0.5 Å	34.1 ± 2.8
NOE rmsd (Å)	0.19 ± 0.02
H-bond >0.5 Å	0
Dihedral >5°	0
Ensemble rmsd	
Backbone heavy atoms	0.65 ± 0.13
All heavy atoms	1.10 ± 0.11
Geometry analysis	
rmsd bonds (Å)	0.00361 ± 0.00008
rmsd angles (°)	0.49 ± 0.01
rmsd impropers (°)	1.26 ± 0.08
Ramachandran analysis[a]	
Favored (%)	85.5
Allowed (%)	13.7
Disallowed (%)	0.8

[a] Performed using the MolProbity server;[28] artificial residues excluded.

Table S3. Statistics from NMR structure calculations for **Aib-H2**.

PDB Accession Code	7TIQ
Experimental restraints	
Unambiguous NOEs	684
Intra-residue	350
Sequential ($\|i - j\| = 1$)	130
Medium-range ($1 < \|i - j\| < 5$)	98
Long-range ($\|i - j\| \geq 5$)	106
Ambiguous NOEs	223
Total NOEs	907
H-bonds	56
Dihedrals	0
Violations	
NOE >0.5 Å	21.3 ± 3.1
NOE rmsd (Å)	0.15 ± 0.01
H-bond >0.5 Å	0
Dihedral >5°	n/a
Ensemble rmsd	
Backbone heavy atoms	1.25 ± 0.46
All heavy atoms	1.49 ± 0.41
Geometry analysis	
rmsd bonds (Å)	0.00278 ± 0.00006
rmsd angles (°)	0.437 ± 0.006
rmsd impropers (°)	1.02 ± 0.06
Ramachandran analysis[a]	
Favored (%)	94.2
Allowed (%)	5.2
Disallowed (%)	0.6

[a] Performed using the MolProbity server;[28] artificial residues excluded.

Table S4. Statistics from NMR structure calculations for **β^3-H3**.

PDB Accession Code	7TIR
Experimental restraints	
Unambiguous NOEs	774
Intra-residue	414
Sequential ($\|i - j\| = 1$)	143
Medium-range ($1 < \|i - j\| < 5$)	98
Long-range ($\|i - j\| \geq 5$)	119
Ambiguous NOEs	206
Total NOEs	980
H-bonds	56
Dihedrals	3
Violations	
NOE >0.5 Å	17.7 ± 2.4
NOE rmsd (Å)	0.12 ± 0.01
H-bond >0.5 Å	0
Dihedral >5°	0
Ensemble rmsd	
Backbone heavy atoms	1.12 ± 0.15
All heavy atoms	1.44 ± 0.17
Geometry analysis	
rmsd bonds (Å)	0.00331 ± 0.00006
rmsd angles (°)	0.445 ± 0.007
rmsd impropers (°)	1.17 ± 0.08
Ramachandran analysis[a]	
Favored (%)	91.7
Allowed (%)	7.9
Disallowed (%)	0.4

[a] Performed using the MolProbity server;[28] artificial residues excluded.

Table S5. Statistics from NMR structure calculations for **Aib-H3**.

PDB Accession Code	7TIS
Experimental restraints	
Unambiguous NOEs	941
Intra-residue	423
Sequential ($|i - j| = 1$)	195
Medium-range ($1 < |i - j| < 5$)	164
Long-range ($|i - j| \geq 5$)	159
Ambiguous NOEs	321
Total NOEs	1262
H-bonds	56
Dihedrals	0
Violations	
NOE >0.5 Å	31.4 ± 5.2
NOE rmsd (Å)	0.18 ± 0.02
H-bond >0.5 Å	0
Dihedral >5°	n/a
Ensemble rmsd	
Backbone heavy atoms	0.86 ± 0.24
All heavy atoms	1.07 ± 0.18
Geometry analysis	
rmsd bonds (Å)	0.0029 ± 0.0001
rmsd angles (°)	0.47 ± 0.02
rmsd impropers (°)	1.14 ± 0.07
Ramachandran analysis[a]	
Favored (%)	92.3
Allowed (%)	6.8
Disallowed (%)	0.9

[a] Performed using the MolProbity server;[28] artificial residues excluded.

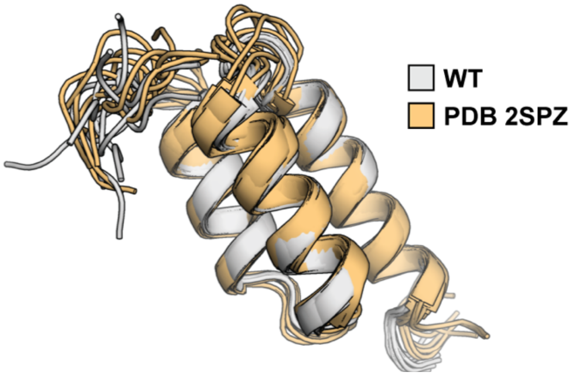

Figure S6. Overlay of the NMR structure ensemble for **WT** determined in the present study with a previously reported NMR structure of the same sequence (PDB 2SPZ).[29] Backbone atom RMSD for a representative model from each ensemble is 0.8 Å, excluding the disordered N-terminal tail.

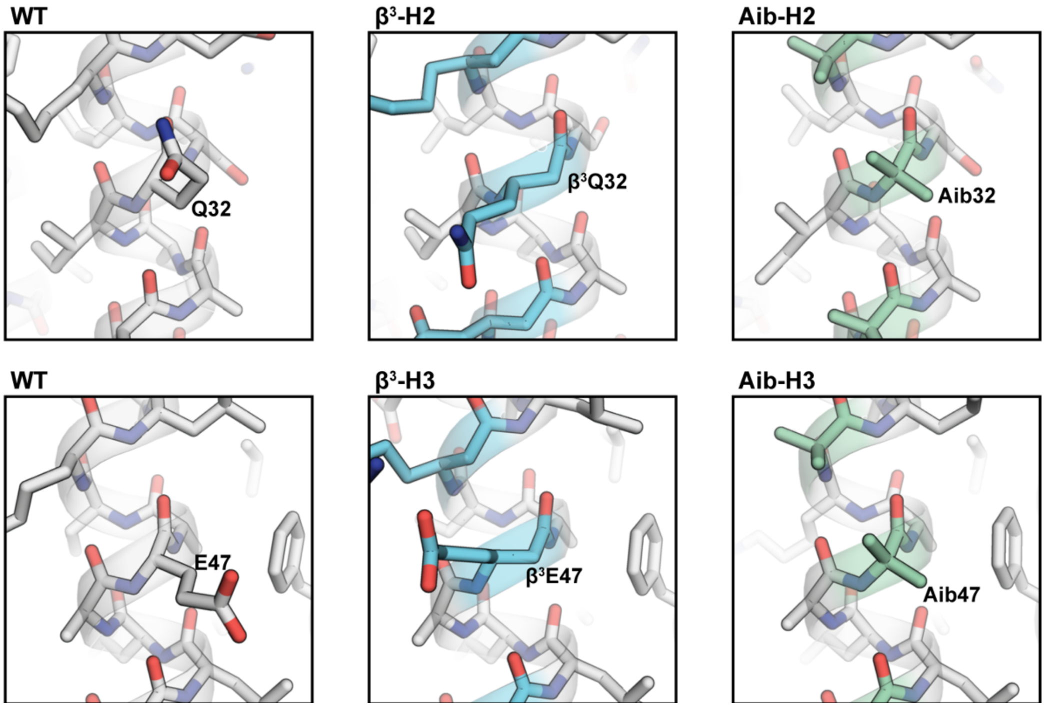

Figure S7. Zoomed view of a representative artificial monomer from the NMR structure of each heterogeneous-backbone BpdA variant alongside the corresponding sequence position from the native-backbone **WT**.

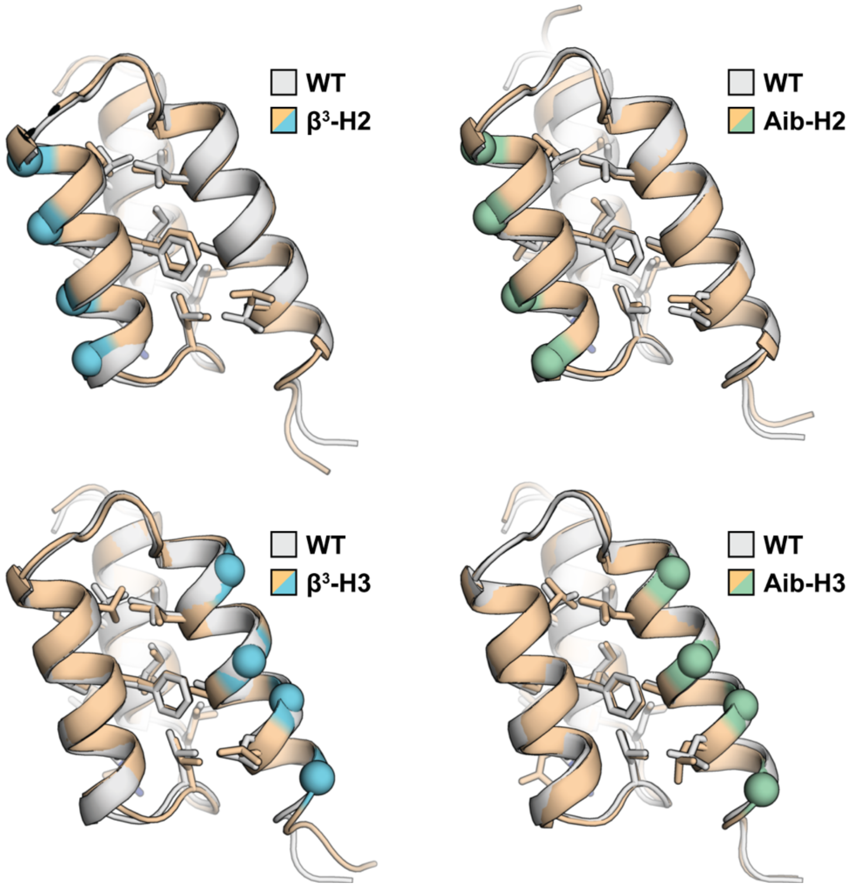

Figure S8. Comparison of the hydrophobic core packing in the NMR structures for native backbone **WT** and heterogeneous-backbone variants. Positions bearing backbone modification are shown as spheres. One representative model from each ensemble is shown. Positions bearing artificial residues are shown as spheres and colored according to the scheme in Figure 1.

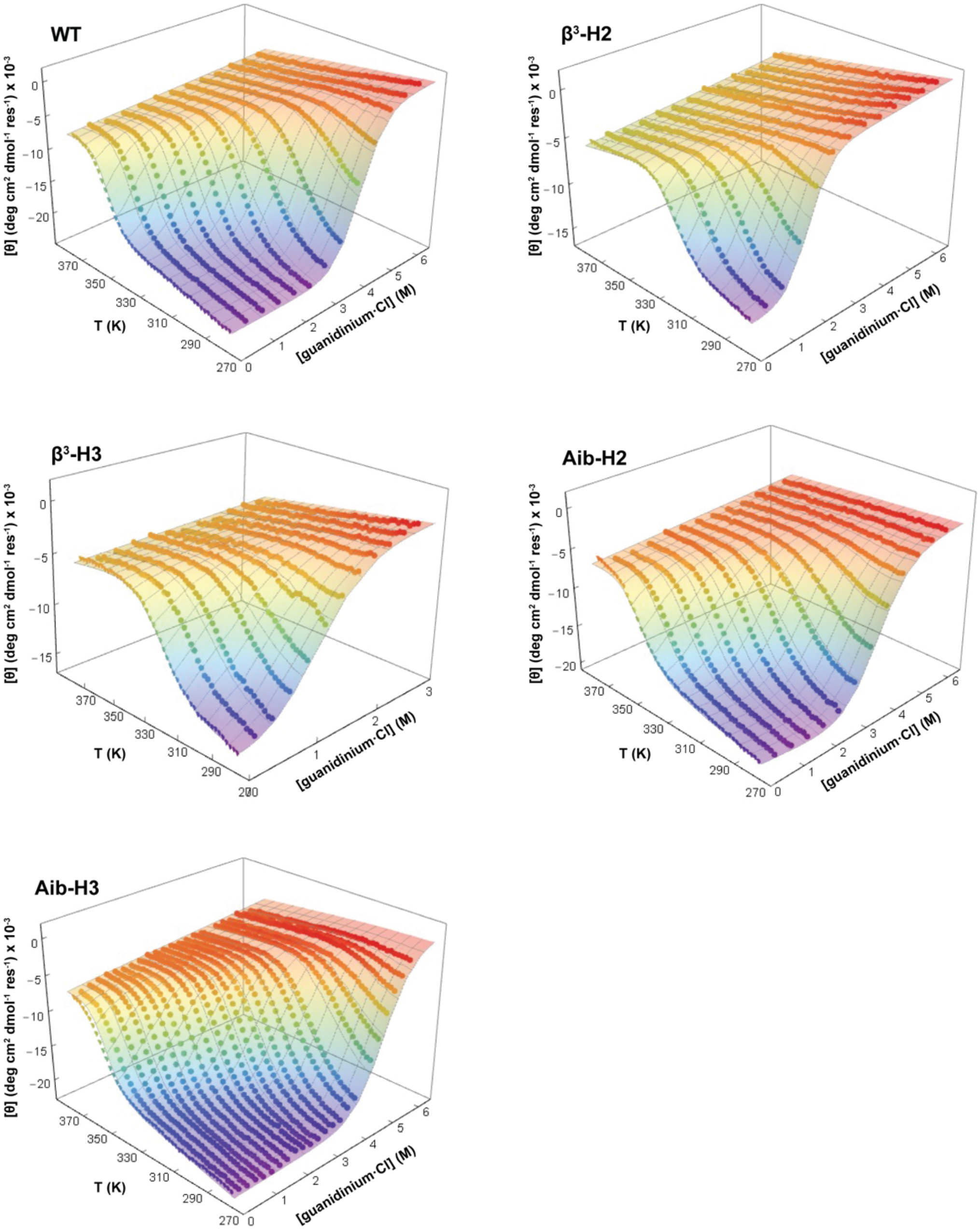

Figure S9. Coupled thermal / chemical denaturation of the unfolding transition monitored by CD for **WT**, **β^3-H2**, **Aib-H2**, **β^3-H3**, and **Aib-H3**. Data points are experimental observations and the surface the result of fitting the data to a two-state folding model.

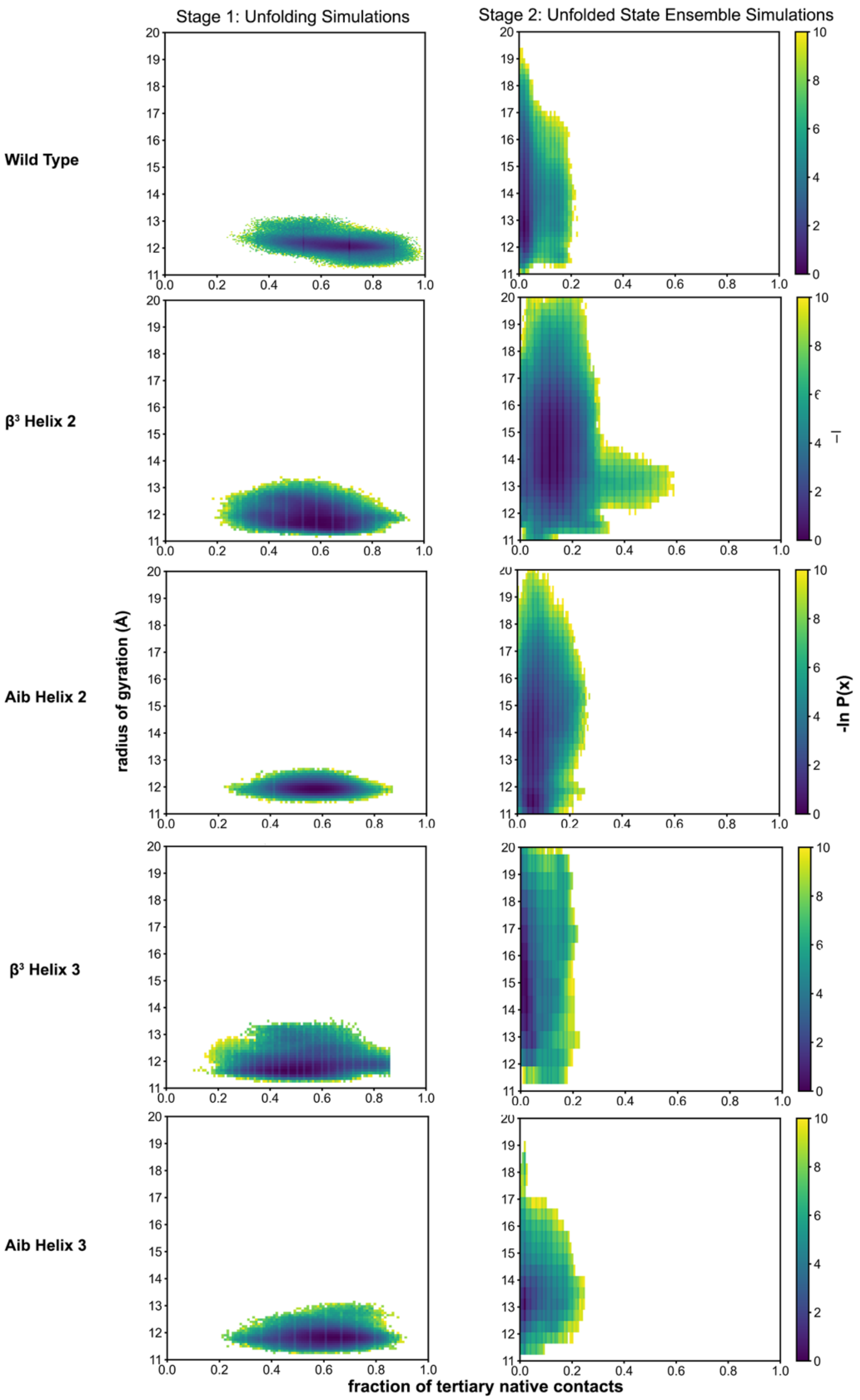

Figure S10. Definitions of the folded and unfolded states for each BdpA variant based on probability distributions from Stage 1 and 2 simulations, respectively, as a function of percent native contacts and radius of gyration. States are defined as regions where -lnP < 4 where P is the statistical weight (probability).

Table S6. Most probable values for selected features of the folded state and unfolded state ensembles resulting from simulations of BdpA and variants.[a]

	(%) Native contacts	R_g (Å)	SASA ($Å^2$)		
			All atom	Hydrocarbon	BB Amide
WT					
Folded	91	12.13	4540	3070	400
Unfolded	75	12.7	5340	3630	620
β^3-H2					
Folded	87	11.65	4460	3050	420
Unfolded	75	14.5	5430	3670	640
Aib-H2					
Folded	89	11.93	4320	3030	380
Unfolded	77	11.4, 13.8	5220	3730	560
β^3-H3					
Folded	85	11.68	4460	2930	420
Unfolded	75	14.8	5610	3790	620
Aib-H3					
Folded	85	11.85	4380	3010	410
Unfolded	73	13.3	5190	3550	610

[a] Heavy atom pairs within 5 Å of each other in the folded reference structures are classified as native contacts. See Figure S12 for full probability distributions.

Table S7. Extent of helicity, inter-helical contacts, and hydrophobic core (H-core) contacts in the folded state and unfolded state ensembles resulting from simulations of BdpA and variants.

	(%) Helicity	(%) Per-helix helicity			% Inter-helical contacts	% H-core contacts
		Helix 1	Helix 2	Helix 3		
WT						
Folded	91 ± 4	79 ± 24	95 ± 18	98 ± 8	67 ± 10	81 ± 3
Unfolded	51 ± 10	26 ± 19	46 ± 17	78 ± 10	3 ± 3	
β^3-H2						
Folded	91 ± 5	87 ± 12	91 ± 22	94 ± 16	57 ± 9	79 ± 3
Unfolded	54 ± 9	16 ± 17	66 ± 20	75 ± 10	13 ± 5	
Aib-H2						
Folded	95 ± 4	95 ± 14	91 ± 18	98 ± 8	57 ± 7	80 ± 2
Unfolded	57 ± 8	68 ± 14	61 ± 10	45 ± 16	7 ± 4	
β^3-H3						
Folded	89 ± 7	87 ± 7	82 ± 17	97 ± 6	52 ± 11	80 ± 3
Unfolded	62 ± 9	81 ± 13	54 ± 11	53 ± 24	2 ± 3	
Aib-H3						
Folded	93 ± 4	88 ± 12	94 ± 20	97 ± 10	60 ± 9	81 ± 3
Unfolded	49 ± 7	52 ± 15	6 ± 11	87 ± 13	4 ± 4	

Values shown are average ± one standard deviation. A residue is defined as helical if categorized as 3_{10}, α-, or π-helices using the DSSP secondary structure assignment program. The large standard deviations of per-helix helicity are due to the highly discrete nature of the dataset. Tertiary native contacts are defined as heavy atoms pairs ≤ 5Å apart in the helical residues of the folded reference structure (See Methods).

Table S8. Percent of inter-helical contacts from conventional simulations of WT folded state using other force fields and water models. Values shown are average ± one standard deviation.

	(%) Native Inter-helical Contacts		
	ff15ipq-m [18] + SPC/E_b	ff03 [30] + TIP3P	ff19SB [31] + OPC
WT			
Folded	53 ± 10	50 ± 7	50 ± 8

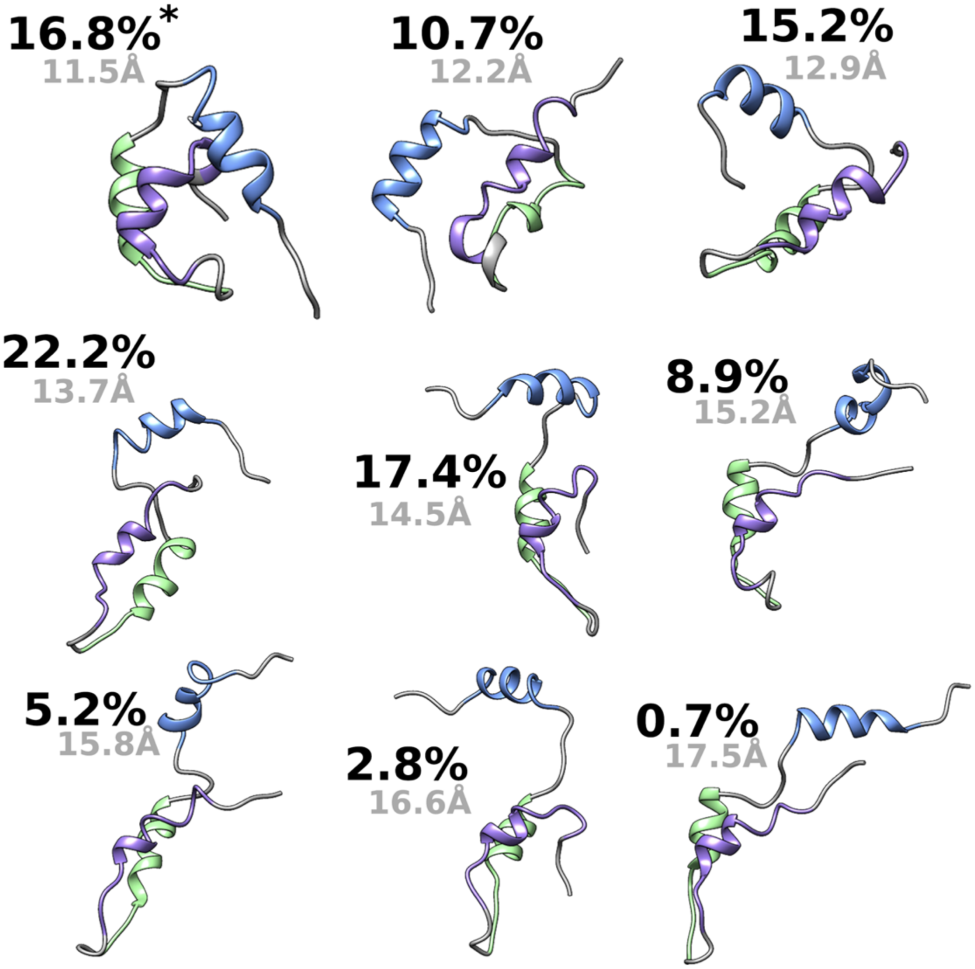

Figure S11. Representative structures of each cluster of the **Aib-H2** BdpA variant using the radius of gyration as a similarity metric. Structures shown are closest to the average structure of the cluster. The (*) indicates the cluster corresponding to the peak in **Figure 1C** with a radius of gyration of ~11.5 Å.

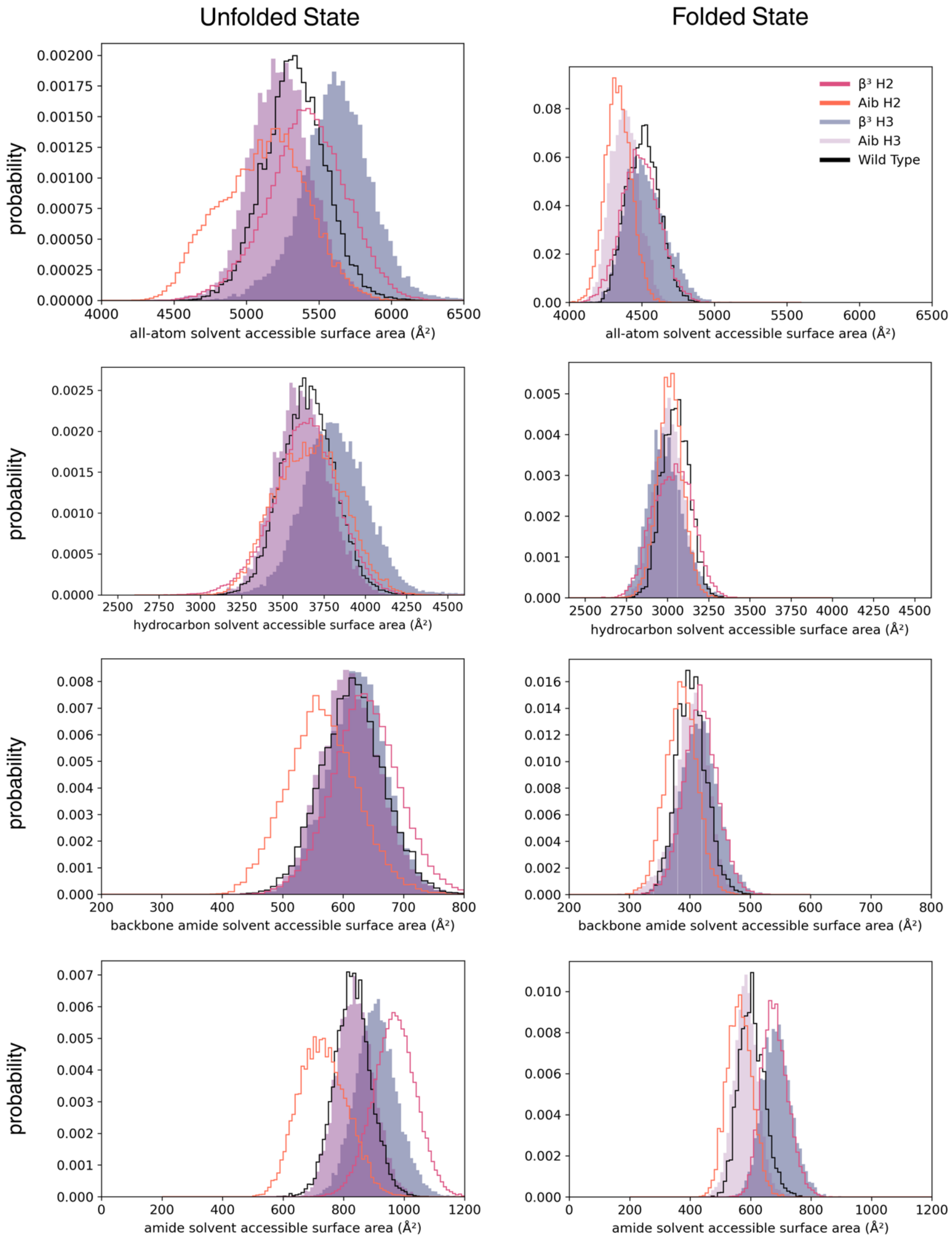

Figure S12. SASA probability distributions for all atoms, hydrocarbon, backbone amide, and all amide functional groups. The backbone amide SASA exclude contributions from the Gln and Asn sidechains. Clear trend of Aib variants < **WT** < β^3 variants is observed in the all-atom SASA, but not in backbone amide.

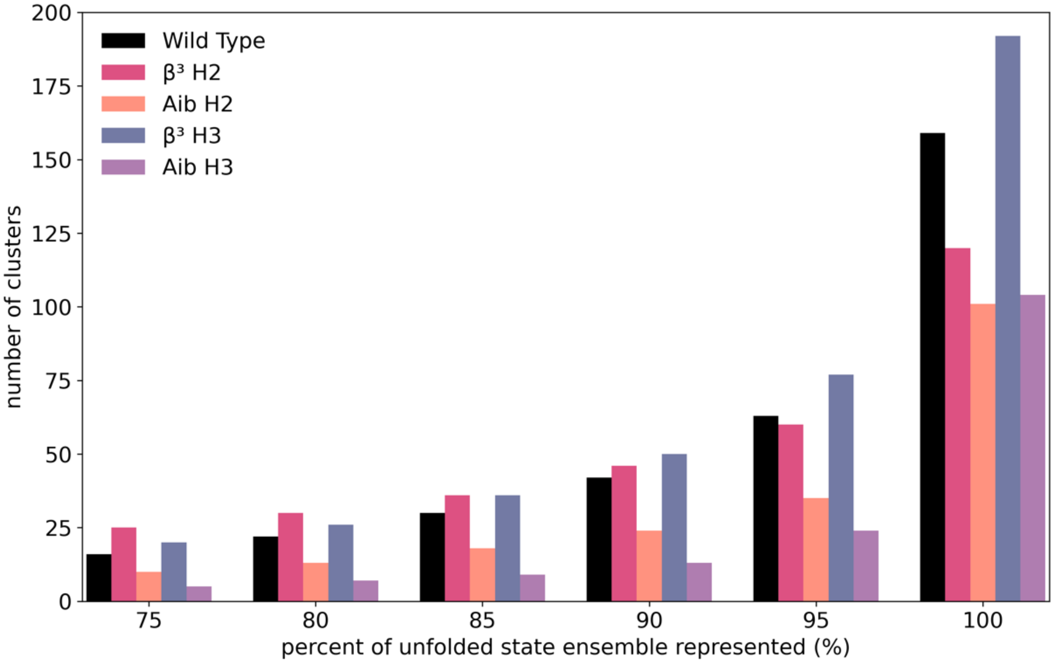

Figure S13. Number of clusters vs. percentage of the unfolded state ensemble for each protein, as determined by hierarchical-agglomerate clustering.

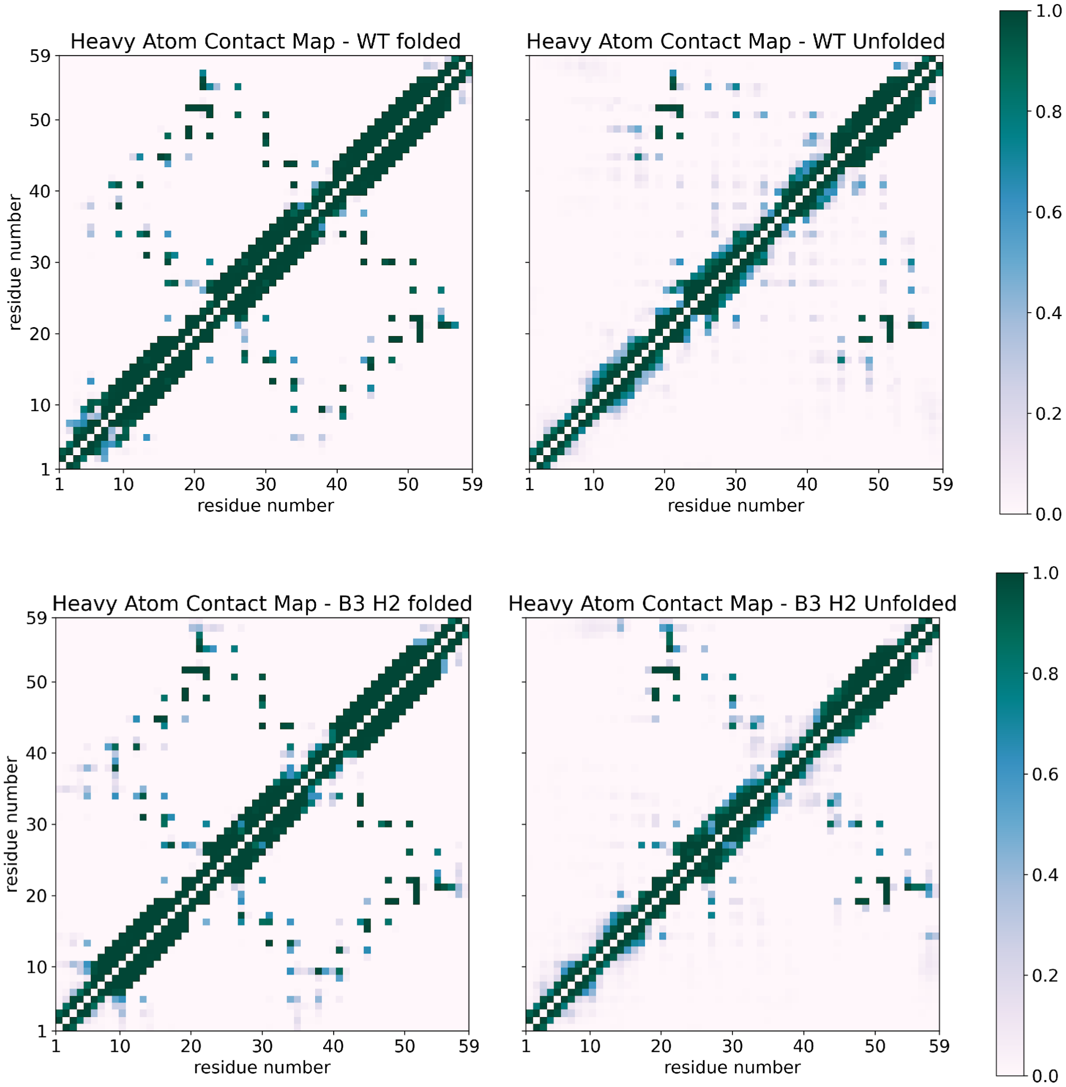

Figure S14. Raw contact maps for the folded and unfolded state of **WT** and **β^3-H2**.

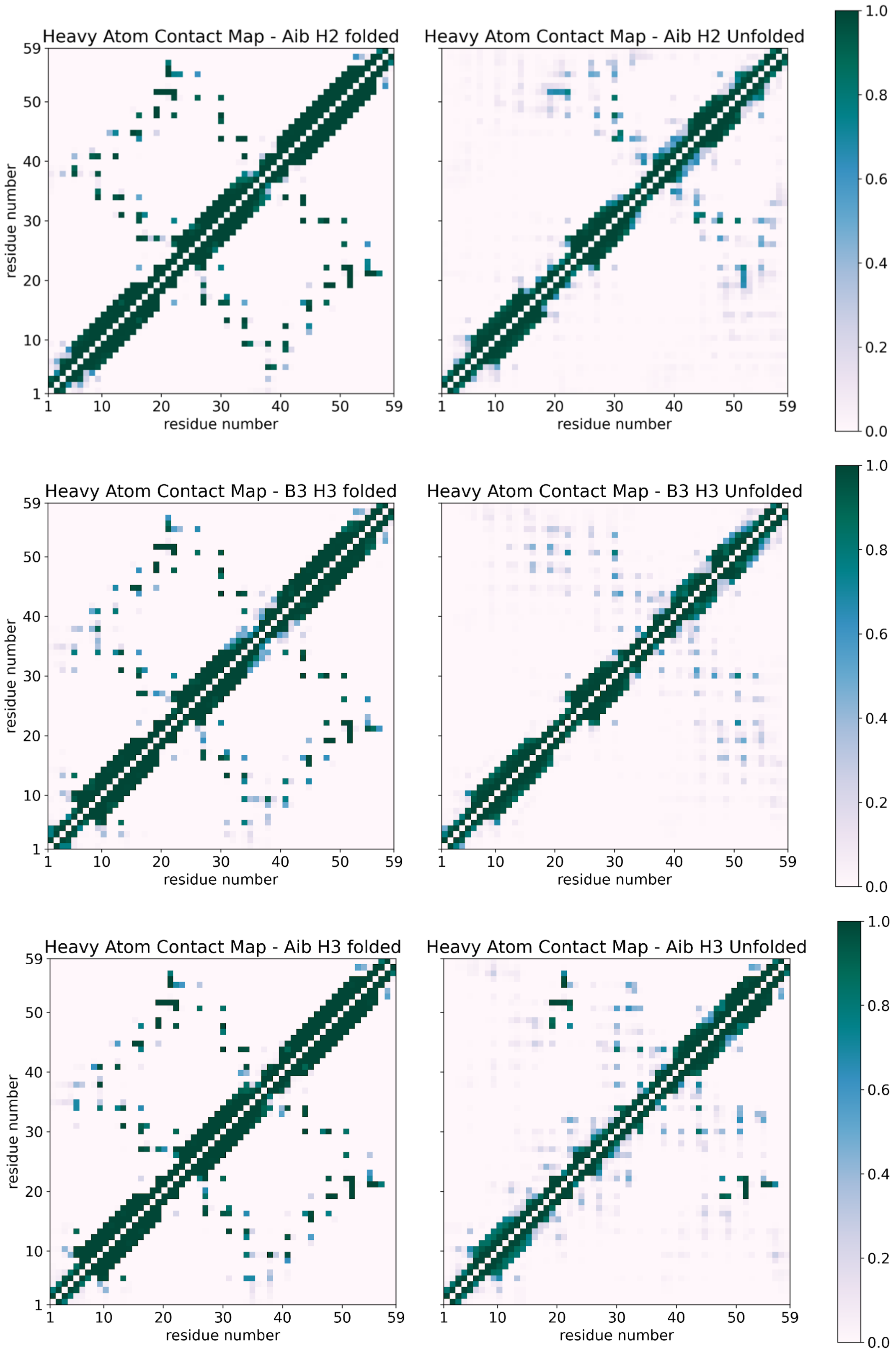

Figure S15. Raw contact maps for the folded and unfolded state of **Aib-H2, β^3-H3**, and **Aib-H3**.

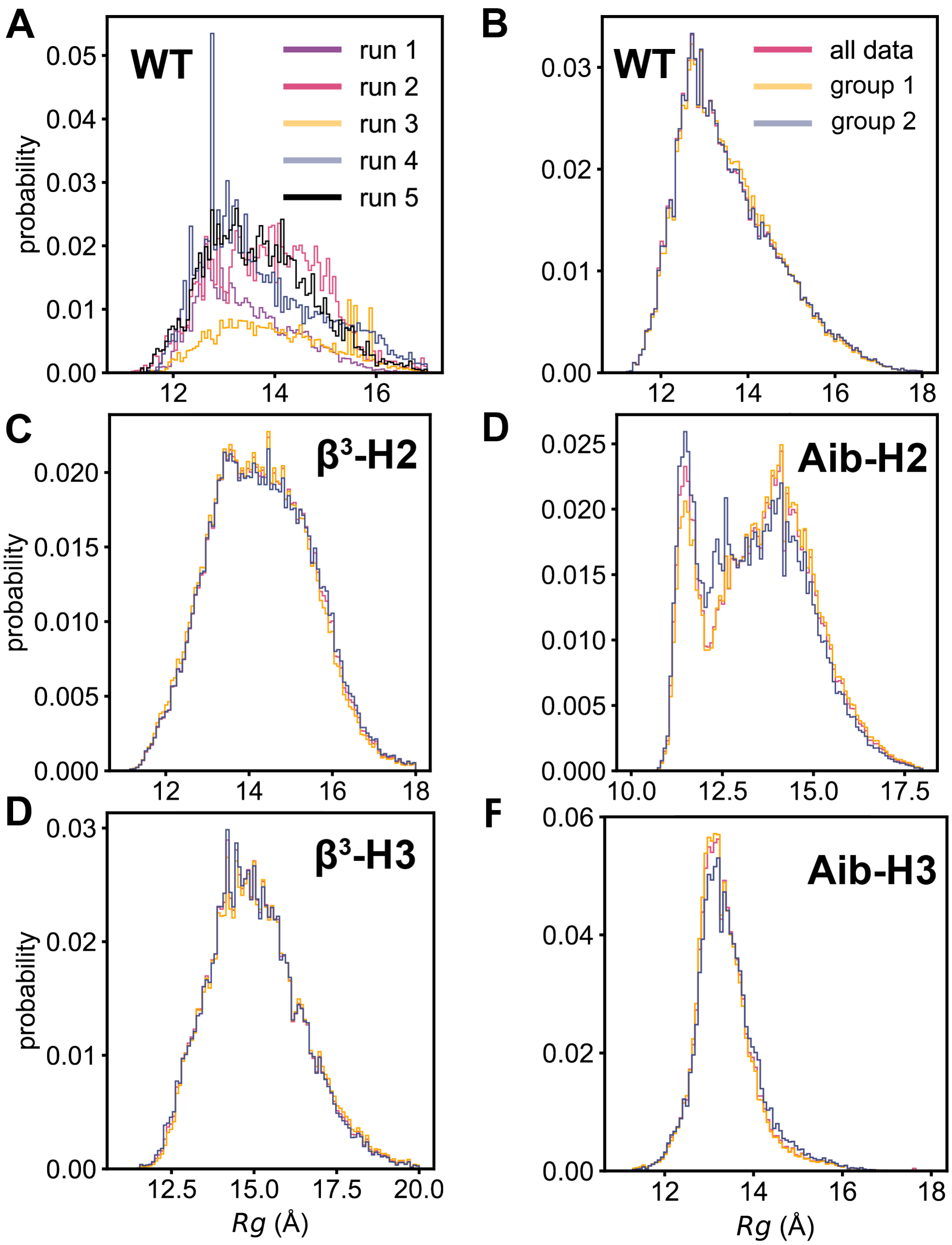

Figure S16. Radius of gyration (R_g) probability distributions generated from different Markov state models (MSMs). (A) Distributions generated by each of five WE simulations (runs) of the **WT** unfolded state. Due to insufficient data for runs 1 and 3, the corresponding distributions were generated from an MSM based an aggregate of data from runs 1 and 3. (B-F) Distributions for **WT** and each of the four BdpA variants generated using either the entire dataset (from all five WE simulations; we refer the corresponding MSM as the "final model") or one of two subsets of the dataset that were created by (i) dividing the dataset into 12 equal-sized blocks of WE iterations, and (ii) grouping the odd-numbered blocks into one subset (group 1) and its even-numbered blocks into another subset (group 2). The MSM for each subset was generated from the same set of cluster centers as that of the corresponding final model.

References

(1) Gill, S. C.; von Hippel, P. H. Calculation of protein extinction coefficients from amino acid sequence data. *Anal. Biochem.* **1989,** *182*, 319-326.

(2) Lee, W.; Tonelli, M.; Markley, J. L. NMRFAM-SPARKY: Enhanced software for biomolecular NMR spectroscopy. *Bioinformatics* **2015,** *31*, 1325-1327.

(3) Rieping, W.; Habeck, M.; Bardiaux, B.; Bernard, A.; Malliavin, T. E.; Nilges, M. ARIA2: Automated NOE assignment and data integration in NMR structure calculation. *Bioinformatics* **2007,** *23*, 381-382.

(4) Brunger, A. T. Version 1.2 of the Crystallography and NMR system. *Nat. Protoc.* **2007,** *2*, 2728.

(5) Cabalteja, C. C.; Mihalko, D. S.; Horne, W. S. Heterogeneous-Backbone Foldamer Mimics of a Computationally Designed, Disulfide-Rich Miniprotein. *ChemBioChem* **2019,** *20*, 103-110.

(6) Rao, S. R.; Horne, W. S. Proteomimetic zinc finger domains with modified metal-binding β-turns. *Pept. Sci.* **2020,** *112*, e24177.

(7) Mareuil, F.; Malliavin, T. E.; Nilges, M.; Bardiaux, B. Improved reliability, accuracy and quality in automated NMR structure calculation with ARIA. *J. Biomol. NMR* **2015,** *62*, 425-438.

(8) Kim, Y.; Prestegard, J. H. Measurement of vicinal couplings from cross peaks in COSY spectra. *J. Magn. Reson.* **1989,** *84*, 9-13.

(9) Shortle, D.; Meeker, A. K.; Freire, E. Stability mutants of staphylococcal nuclease: large compensating enthalpy-entropy changes for the reversible denaturation reaction. *Biochemistry* **1988,** *27*, 4761-4768.

(10) Santhouse, J. R.; Rao, S. R.; Horne, W. S. Analysis of folded structure and folding thermodynamics in heterogeneous-backbone proteomimetics. *Methods Enzymol.* **2021,** *656*, 93-122.

(11) Kuhlman, B.; Raleigh, D. P. Global analysis of the thermal and chemical denaturation of the N-terminal domain of the ribosomal protein L9 in H2O and D2O. Determination of the thermodynamic parameters, ΔH°, ΔS°, and ΔC°p, and evaluation of solvent isotope effects. *Protein Sci.* **1998,** *7*, 2405-2412.

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

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

(14) Russo, J. D.; Zhang, S.; Leung, J. M. G.; Bogetti, A. T.; Thompson, J. P.; DeGrave, A. J.; Torrillo, P. A.; Pratt, A. J.; Wong, K. F.; Xia, J.; Copperman, J.; Adelman, J. L.; Zwier, M. C.; LeBard, D. N.; Zuckerman, D. M.; Chong, L. T. WESTPA 2.0: High-Performance Upgrades for Weighted Ensemble Simulations and Analysis of Longer-Timescale Applications. *J. Chem. Theory. Comput.* **2022,** *18*, 638-649.

(15) Suarez, E.; Lettieri, S.; Zwier, M. C.; Stringer, C. A.; Subramanian, S. R.; Chong, L. T.; Zuckerman, D. M. Simultaneous Computation of Dynamical and Equilibrium Information Using a Weighted Ensemble of Trajectories. *J. Chem. Theory. Comput.* **2014,** *10*, 2658-2667.

(16) Torrillo, P. A.; Bogetti, A. T.; Chong, L. T. A Minimal, Adaptive Binning Scheme for Weighted Ensemble Simulations. *J. Phys. Chem. A* **2021,** *125*, 1642-1649.

(17) Case, D. A.; Aktulga, H. M.; Belfon, K.; Ben-Shalom, I. Y.; Brozell, S. R.; Cerutti, D. S.; III, T.E. Cheatham ; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E.; Giambasu, G.; Gilson, M. K.; Gohlke, H.; Goetz, A. W.; Harris, R.; Izadi, S.; Izmailov, S. A.; Jin, C.; Kasavajhala, K.; Kaymak,

M. C.; King, E.; Kovalenko, A.; Kurtzman, T.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Machado, M.; Man, V.; Manathunga, M.; Merz, K. M.; Miao, Y.; Mikhailovskii, O.; Monard, G.; Nguyen, H.; O'Hearn, K. A.; Onufriev, A.; Pan, F.; Pantano, S.; Qi, R.; Rahnamoun, A.; Roe, D. R.; Roitberg, A.; Sagui, C.; Schott-Verdugo, S.; Shen, J.; Simmerling, C. L.; Skrynnikov, N. R.; Smith, J.; Swails, J.; Walker, R. C.; Wang, J.; Wei, H.; Wolf, R. M.; Wu, X.; Xue, Y.; York, D. M.; Zhao, S.; Kollman, P. A. *Amber 2018*, University of California, San Francisco: 2018.

(18) Bogetti, A. T.; Piston, H. E.; Leung, J. M. G.; Cabalteja, C. C.; Yang, D. T.; DeGrave, A. J.; Debiec, K. T.; Cerutti, D. S.; Case, D. A.; Horne, W. S.; Chong, L. T. A twist in the road less traveled: The AMBER ff15ipq-m force field for protein mimetics. *J. Chem. Phys.* **2020,** *153*, 064101.

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

(20) 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.* **2010,** *132*, 054107.

(21) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977,** *23*, 327-341.

(22) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995,** *103*, 8577-5893.

(23) Copperman, J.; Zuckerman, D. M. Accelerated Estimation of Long-Timescale Kinetics from Weighted Ensemble Simulation via Non-Markovian "Microbin" Analysis. *J. Chem. Theory. Comput.* **2020,** *16*, 6763-6775.

(24) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011,** *12*, 2825-2830.

(25) Aristoff, D.; Zuckerman, D. M. Optimizing Weighted Ensemble Sampling of Steady States. *Multiscale Model. Simul.* **2020,** *18*, 646-673.

(26) Shrake, A.; Rupley, J. A. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* **1973,** *79*, 351-371.

(27) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernandez, C. X.; Schwantes, C. R.; Wang, L. P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015,** *109*, 1528-32.

(28) Williams, C. J.; Headd, J. J.; Moriarty, N. W.; Prisant, M. G.; Videau, L. L.; Deis, L. N.; Verma, V.; Keedy, D. A.; Hintze, B. J.; Chen, V. B.; Jain, S.; Lewis, S. M.; Arendall III, W. B.; Snoeyink, J.; Adams, P. D.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **2018,** *27*, 293-315.

(29) Tashiro, M.; Tejero, R.; Zimmerman, D. E.; Celda, B.; Nilsson, B.; Montelione, G. T. High-resolution solution NMR structure of the Z domain of staphylococcal protein A. *J. Mol. Biol.* **1997,** *272*, 573-590.

(30) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003,** *24*, 1999-2012.

(31) Tian, C.; Kasavajhala, K.; Belfon, K. A. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q.; Simmerling, C. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory. Comput.* **2020,** *16*, 528-552.