JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

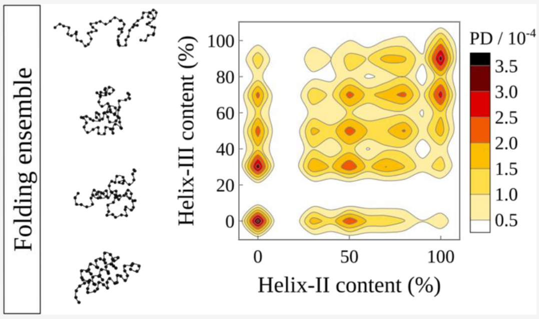
pubs.acs.org/jcim Article

Helical Content Correlations and Hydration Structures of the Folding Ensemble of the B Domain of Protein A

Ander Francisco Pereira and Leandro Martínez*

Cite This: *J. Chem. Inf. Model.* 2024, 64, 3350−3359 Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information



ABSTRACT: The B domain of protein A (BdpA), a small three-helix bundle, folds on a time scale of a few microseconds with heterogeneous native and unfolded states. It is widely used as a model for understanding protein folding mechanisms. In this work, we use structure-based models (SBMs) and atomistic simulations to comprehensively investigate how BdpA folding is associated with the formation of its secondary structure. The energy landscape visualization method (ELViM) was used to characterize the pathways that connect the folded and unfolded states of BdpA as well as the sets of structures displaying specific ellipticity patterns. We show that the native state conformational diversity is due mainly to the conformational variability of helix I. Helices I, II, and III occur in a weakly correlated manner, with Spearman's rank correlation coefficients of 0.1539 (I and II), 0.1259 (I and III), and 0.2561 (II and III). These results, therefore, suggest the highest cooperativity between helices II and III. Our results allow the clustering of partially folded structures of folding of the B domain of protein A on the basis of its secondary structure, paving the way to an understanding of environmental factors in the relative stability of the basins of the folding ensemble, which are illustrated by the structural dependency of the protein hydration structures, as computed with minimum-distance distribution functions.

1. INTRODUCTION

The B domain of protein A from staphylococcal bacteria (BdpA) is an attractive target for protein folding studies for having a simple three-helix bundle topology,[1] as shown in Figure 1, and a fast folding kinetics.[2,3] BdpA folding has been investigated using atomistic and coarse-grained simulations,[4−6] and experimental methods,[7,8] demonstrating agreement of the folding ensemble with a theoretical description based on a funneled energy landscape theory.[5,9−11] BdpA folding mechanism can be represented by a two-state kinetics, in which the native (N) and unfolded (U) ensembles coexist in equilibrium without significant population of intermediate states.[7,8,12,13] Short-lived intermediate states, of course, exist and have been characterized recently.[3,14]

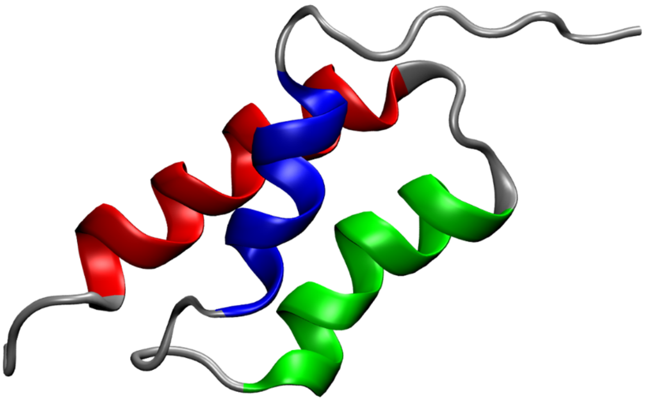


Figure 1. Cartoon representation of the B domain of staphylococcal protein A (BdPA), PDB: 1BDD.[15] Helix I (^{10}QQNAFYEILH^{19}), helix II (^{25}EEQRNGFIQSLKD^{37}), and helix III (^{42}SANLLAEAKKLNDAQ^{56}) are depicted in blue, green, and red, respectively.

In what concerns the characterization of the structural variability of native and denatured states, high-temperature simulations[6] and recent H/D-exchange experiments have shown that the residual structure observed in denatured states is associated with helix III.[14] The experimental study also suggested the possibility of a salt bridge between residues Lys50 and Asp54 as responsible for stability of helix III, and that this helix could act as a folding initiation site of BdpA.[14] At the same time, other experimental data from Φ-value analysis show that helix III is poorly formed in the transition state, while helix-II is well-formed.[7,8,16,17] Therefore, there is no clear

Received: November 10, 2023
Revised: March 19, 2024
Accepted: March 22, 2024
Published: April 3, 2024

https://doi.org/10.1021/acs.jcim.3c01822

consensus on the underlying mechanisms of the formation and deformation of BdpA helices during folding.

Here, we investigate the BdpA folding by combining coarse-grained simulations using C_α-structure-based models (SBMs)[18] and atomistic simulations. We add the helical content dimension to the projection of the folding ensemble, providing new insights into the distribution and correlation of the secondary structure in folding mechanisms. The simulations with the C_α-SBMs provided an exhaustive sampling of the folding landscape, including transition states. Atomistic simulations of each folding structure, in turn, allowed us to obtain equilibrated structures for a detailed analysis of their helical content and interaction with the solvent. Our results support the two-state mechanism. The analysis of the secondary structure ensembles revealed three significant basins, which are associated with the correlated formation of helices II and III, while helix I is consistently formed only in the native state. The characterization of these secondary structure elements allows a novel classification of the denatured state ensemble, paving the way for the investigation of environmental factors in the folding ensemble of the B domain of protein A, illustrated here by the conformational dependence of the protein hydration structures.

2. METHODS

2.1. Simulations with SBMs and Analysis. We initially obtained the folding ensemble of BdpA (PDB: 1BDD)[15] using $C\alpha$-SBMs.[18,19] In $C\alpha$-SBMs, each residue is represented as a single bead centered at the α-carbon ($C\alpha$).[18] The theoretical basis for the SBMs is the funneled energy landscape theory.[20,21] According to this model, the protein potential energy surface is minimally frustrated, and the folding occurs by successive conformational transformations that are monotonically biased toward the native state. Minimally frustrated energy surfaces can be reconstructed from the folded structure by the definition of biasing forces having equilibrium distances equal to those of the reference structure. SBMs are based on this assumption and construct a potential energy surface from the structural properties of the folded states, including bond, angle, dihedral, contact, and noncontact terms for the interactions of a $C\alpha$-only model.[18,20,22] The BdpA is a model protein for folding studies and has been studied thoroughly.[5,9−11,23] SBM simulations allow the exhaustive sampling of the conformational space of the model, providing, within the model approximations, a complete ensemble of the folding of small proteins such as BpdA. The convergence of the ensemble can be demonstrated by obtaining multiple folding-unfolding transitions at the critical temperature in the case of a two-state folding model, as the one studied here.

The contact map of BdpA (PDB: 1BDD) was determined with the contact of structural units (CSU) algorithm.[24] The $C\alpha$-SBM was generated using the SMOG web server (https://smog-server.org/).[25] All simulations with $C\alpha$-SBM were performed with Gromacs 4.6.7.[26] To determine the folding temperature, two sets of simulations were performed. The first set used Gromacs setup temperatures between 80 and 160 K (true units are not significant for SBM models) with a temperature step of 10 K. Once the temperature of maximum specific heat, C_v, was roughly identified, a new set of simulations with temperatures varying between 113 and 119 K with 1 K temperature steps was performed to localize within ~1 K the folding temperature. In reduced temperature units, the simulations were performed within 0.66 and 1.33 with 0.083 steps and within 0.94 and 0.99 with 0.0083 temperature unit steps for the first and second set of simulations. These simulations were analyzed with the weighted histogram analysis method (WHAM)[27] as implemented in SMOG2[28] for the determination of the temperature dependence of the specific heat, $C_v(T)$, and the potential of mean force, $F(Q)$, as a function of the fraction of native contacts. Each SBM simulation consisted of 5×10^8 steps with a time step of 0.0005 reduced units at constant temperature. A contact was considered native if the distance between the corresponding $C\alpha$ atoms was not greater than 20% of that observed in the experimentally reported NMR model.[15]

The temperature of the maximum specific heat is the model folding temperature (T_f). At T_f, folded and unfolded states have the same probability of occurrence. The SBM simulation performed at T_f was used to study the folding ensemble. The statistical convergence of the Q values at the folding temperature was confirmed by block-averaging (Figure S10 of the Supporting Information computed with the MolSimToolkit.jl package, version 1.3.4), which shows that the characteristic correlation time of Q is much lower than the total simulation time.

It is worth noting that the 10 additional NMR structures of the BdpA available in entry 1BDC[15] could also be adopted as initial configurations for the $C\alpha$-SBM simulations. Figure S11 illustrates that the contact maps of 1BDC models closely resemble those derived from the 1BDD structure. Also, all-atom SBMs are alternative models that could also be employed in protein folding studies, possibly capturing variations in relative probability among ensembles influenced by side chain dynamics.[11,22] Specifically for BdpA, a previous study showed that both models agree in the global description of its folding, despite the additional complexity and computational cost.[11]

2.2. Protein Folding Projection Maps. The protein folding landscape was visualized with the energy landscape visualization method (ELViM).[5,29−33] Given that protein folding occurs in a multidimensional space, its visualization depends on dimensionality reduction. The ELViM method uses the matrices of internal distances of the conformations to define a robust metric of the similarity between structures without an a priori definition of a reaction coordinate. The matrix of similarities between conformations is projected in 2D space. Here, we use the *Force Scheme* technique,[34] as originally proposed for ELViM.[5] This method consists, basically, of defining a potential in the projected space that is dependent on the similarity measure between the structures and optimizing this potential to obtain a distribution of points that optimally represents the distances in the multidimensional similarity space.

2.3. Secondary Structure Analysis from Atomistic Models. To compute the secondary structures, the all-atom representations of 5000 SBM models of the simulation performed at the T_f were reconstructed using the Pulchra software.[35] The simulation boxes were constructed with a minimum distance of 12.0 Å from the protein extrema using Packmol.[36,37] This resulted in box volumes ranging from 142848 to 423120 Å^3, with the number of water molecules varying between 4401 and 13,775. The protein structure was restrained by applying harmonic potentials with 25 kcal mol^{-1} force constants on $C\alpha$ atoms, such that the SBM topology was preserved, while allowing relaxation of the reconstructed atoms.

The CHARMM36 force field[38] for the protein and the TIP3P water model[39] were used. All atomistic simulations were performed in Gromacs 2021.2[40] at 298.15 K, 1 atm, and with a time step of 2 fs. Initially, the system was minimized by up to 20,000 steps using the steepest descent method and equilibrated by 1 ns in constant-volume and constant-temperature ensemble (NVT) followed by 1 ns of constant-volume and constant-pressure (NPT) simulation. Temperature and pressure were controlled using the modified Berendsen thermostat[41] and Parrinello−Rahman barostat.[42] Finally, 10 ns production simulations were performed in the NPT ensemble for each system, totaling 50 μs of simulation.

The secondary structure of the models was calculated with the DSSP method,[43,44] using the ProteinSecondaryStructures.jl interface.[45] Residues belonging to helices I, II, and III (Figure 1) were attributed according to the literature.[7,8,16,46] Assigning the α-helical content with DSSP required equilibrated all-atom structures, as the estimates obtained directly from Pulchra-reconstructed models failed, as shown in Table S1 of the Supporting Information. The all-atom reconstruction and equilibration procedures are described below. The distributions of the secondary structure structures obtained for each folding basin of the ensemble are shown in Figure S9 of the Supporting Information.

The helical content obtained from the folding ensemble was mapped on the ELViM projection to obtain contour plots with the *histogram2dcontour* function from the PlotlyJS.jl v0.18.10 package. Spearman's rank correlation coefficients[47] were calculated to identify the correlations between helices during folding.

2.4. Molecular Basis of Hydration.

The hydration of the models was studied using minimum-distance distribution functions (MDDFs)[48] and the Kirkwood−Buff (KB) theory of solvation.[49]

MDDFs are distribution functions computed from the minimum-distance counts between any solute and solvent atoms. They have the advantage over the radial distribution of taking the shapes of solute and solvent molecules automatically into account. That is, a peak associated with a minimum distance between the components is associated with an interaction at that precise distance, providing a picture of the solvation that matches the natural solvent-shell interpretation of the interactions.

To allow the computation of KB integrals (and thus thermodynamic properties), MDDFs have to be normalized by the minimum-distance count in an ideal distribution of the molecules with the same bulk density of the solvent. This requires generating random distributions of solvent molecules around the solute with correct density and molecular conformational distribution, which is tricky and computationally expensive.[48] Thus, we implemented these computations in the specialized ComplexMixtures.jl package,[50] which builds up on the efficient CellListMap.jl software for the computation of short-ranged interactions in particle systems.[51]

Formally,[48] MDDFs can be defined in terms of the average number density of the solvent $n_s(r)$ relative to the density of an ideal-gas distribution, $n_s{}^*(r)$:

$$g_{ps}(r) = \frac{n_s(r)}{n_s{}^*(r)} \tag{1}$$

where p refers to the protein, s is the solvent (in this case, water), and r is the minimum distance between any solvent and solute atoms. As with other distribution functions, MDDFs allow for the calculation of KB integrals, which can quantify the accumulation of the solvent around the solute. The KB integrals can be computed using $n_s(r)$ and $n_s^*(r)$:

$$G_{ps} = \frac{1}{\rho_s}\int_0^\infty [n_s(r) - n_s{}^*(r)]S(r)\mathrm{d}r \tag{2}$$

where $S(r)$ is the surface dependent on the solute's shape defined by the minimum-distance, and ρ_s is the molar concentration of the solvent. The integration of eq 2 in a finite subvolume of the system reduces to

$$G_{ps}(R) = \frac{1}{\rho_s}[N_{ps}(R) - N_{ps}{}^*(R)] \tag{3}$$

where $N_{ps}(R)$ is the number of minimum distances between the protein and the solvent at the R distance from the protein surface and $N_{ps}^*(R)$ is the minimum-distance count in an ideal-gas distribution.[48,52,53]

The KB integrals (eqs 2 and 3) quantify the excess volume occupied by the solvent in the domain of the solute, where solute−solvent interactions are significant, relative to the volume that the solvent would occupy in the absence of solute−solvent interactions.[54−56] For large solutes, such as proteins, KB integrals are generally negative as a consequence of the excluded volume of the solute.

MDDFs and KB integrals and solvation maps were computed with the ComplexMixtures.jl package[50] and plotted with Plots.jl. We compute the distribution functions and KB integrals for subsets of the ensembles with different folding characteristics independently, comprising averages over tenths or hundreds of structures in each case (Figure S8C and Table S3 of the Supporting Information).

The methodology described above is integrated into a comprehensive pipeline (Figure 2). The initial steps, highlighted in black, were performed to validate the BdpA folding based on previous works.[4,5,9,11] The incremental steps, highlighted in green, were implemented in this work to investigate the association among BdpA folding, the formation

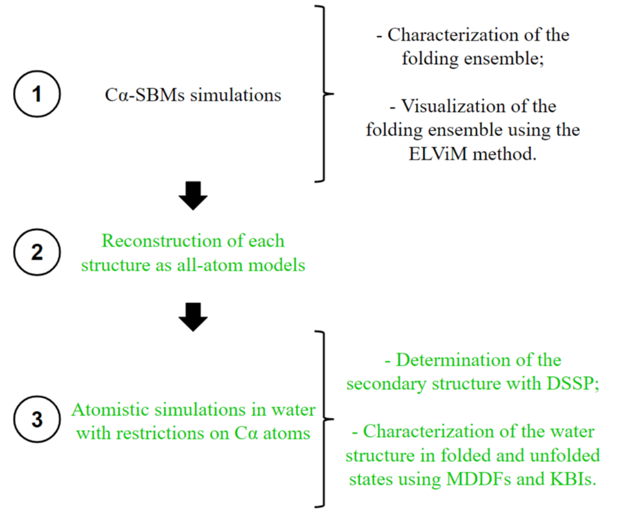

1
Cα-SBMs simulations
- Characterization of the folding ensemble;
- Visualization of the folding ensemble using the ELViM method.
2
Reconstruction of each structure as all-atom models
3
Atomistic simulations in water with restrictions on Cα atoms
- Determination of the secondary structure with DSSP;
- Characterization of the water structure in folded and unfolded states using MDDFs and KBIs.

Figure 2. Pipeline was used to study BdpA folding. Initial steps (black) validated BdpA folding based on previous works, while incremental steps (green) allowed investigation of the relation between BdpA folding, its secondary structure formation, and solvent structure around partially folded states.

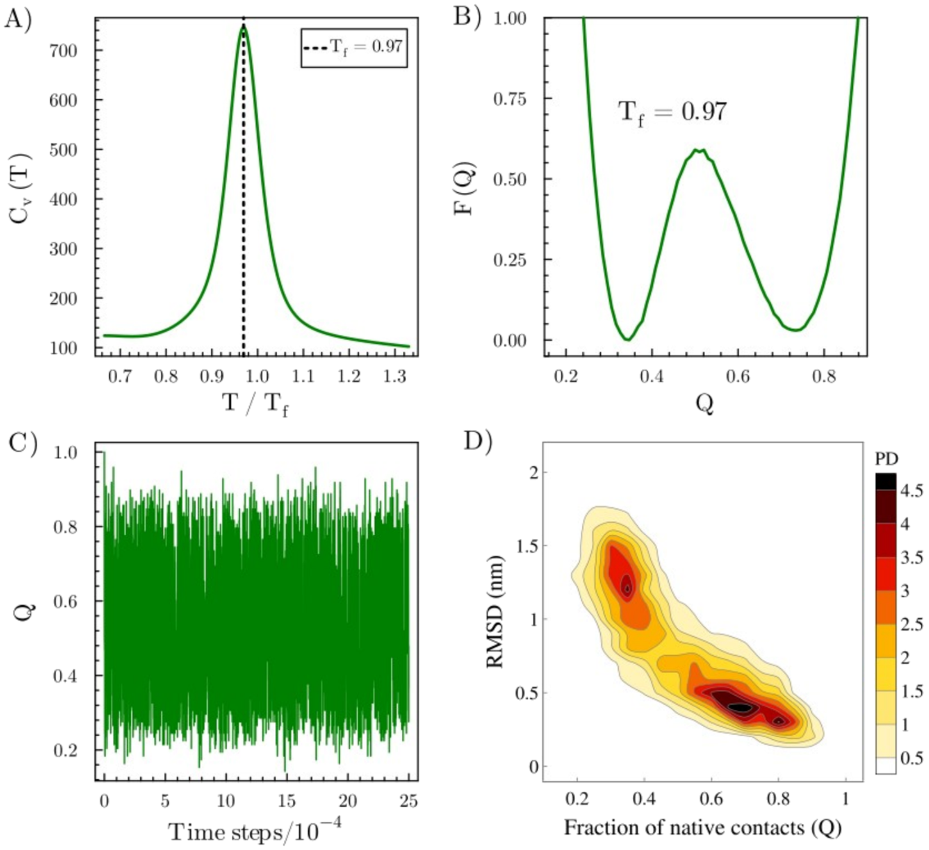


Figure 3. Characterization of BdPA folding. (A) Specific heat (C_v) as a function of temperature, allowing the identification of the folding temperature (T_f = 0.97 reduced units). From the simulation performed at the T_f: (B) Free energy as a function of the fraction of native contacts (Q). (C) Fraction of native contacts (Q) as a function of simulation time step. (D) Contour maps of the PD as a function of Q and RMSD.

of its secondary structure, and the solvent structure around partially folded states.

3. RESULTS AND DISCUSSION

The results here are divided into three parts: (1) the validation of the folding ensemble obtained, (2) the analysis of the ellipticity of the structure, and (3) insights into the solvation structures of partially folded states. Thus, in Section 3.1, we show the properties of the simulated ensemble and its consistency with previous simulations and experimental results. In Section 3.2, we characterize the formation of the helices of BdpA in unprecedented detail, revealing its correlation with the heterogeneous nature of the folded and unfolded ensembles and the correlation between helical propensities of each element of the structure. Finally, in Section 3.3, we show how the solvation structures of the protein vary in each of the unfolded basins relative to those of the folded state.

3.1. Protein A Folding in a 2D Phase Space.

Proteins with two-state ensembles exhibit a well-defined folding temperature, whereby the distribution of molecules over a measurable property is bimodal. The single sharp peak in the $C_v(T)$ profile in Figure 3A shows that the two-state model is a good representation of the folding of the BdpA SBM model.[57] There are no stable intermediate states in these cases, but rather a set of short-lived intermediate states that are distinct from one another.[58]

It is also possible to identify the folded (N) and unfolded (U) ensembles from the free energy profile as a function of the fraction of native contacts $F(Q)$ at T_f (Figure 3B). There is a single well-defined energy barrier connecting states N ($Q \sim 0.8$) and U ($Q \sim 0.3$). In Figure 3C, the Q values as a function of time steps at the folding temperature (T_f) clearly show several transitions between the folded and unfolded states, indicating a good sampling of the transition.[9] BdpA does not visit highly extended states ($Q \sim 0$), which was suggested to be a consequence of some high-affinity native contacts.[9] Here, the contact formation and the average distance between pairs of atoms (Table S2) revealed three contacts with the highest probability of being preserved (38−42, 83.52%, 38−45, 84.18%, and 20−31, 86.00%) mainly involving turn I (19−25) and turn II (37−42). Furthermore, the average distances of these pairs of atoms are close to those observed in the experimental structure.

In Figure 3D, we illustrate the folding ensemble by mapping the probability density (PD) of the ensemble as a function of the RMSDs (y-axis) relative to the native structure (PDB: 1BDD[15]) and to the fraction of native contacts (x-axis). The folded states span a range of Q values of roughly 0.5 to 0.9 and RMSD values of 0.2 to 0.6 nm. The unfolded ensemble displays Q values between 0.2 and 0.4, and larger RMSDs within 0.9 and 1.6 nm, approximately.

Even though it is possible to classify native and unfolded states in Figure 3, the conformational diversity of each state and of the intermediates is hidden from such representations. Therefore, we use here a coordinate-free method, the ELViM, to obtain a fine-grained visualization of the protein folding ensemble.[5,29,59]

Figure 4 shows a 2D projection of the phase space obtained with ELViM, colored with the fraction of native contacts (Q).

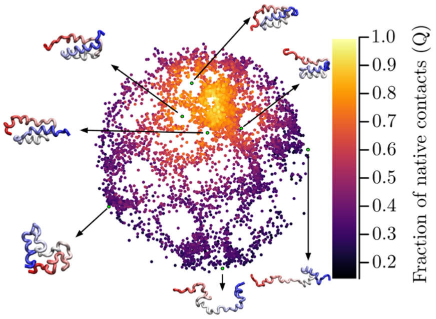


Figure 4. Projection of the folding ensemble of BdpA obtained with ELViM. Each structure is represented as a point, with a color associated with the fraction of native contacts (Q): yellowish and purplish regions depict the native (large Q) and unfolded (small Q) states, respectively.

Each point in the figure represents a structure from simulations with the SBMs. The projection attempts to map the structural dissimilarity between the structures to the Euclidean distance in the projection. Therefore, nearby points indicate similar structures, while distant points indicate different structures. Many dissimilar structures have similar Q values, illustrating the conformational variability of what is defined as the unfolded ensemble. The obtained map reproduces the ones calculated previously.[4,60] The structural variability of the unfolded ensemble is associated with the formation of the BdpA helices, as will be discussed.

3.2. Folding Ensemble of Helix Formation.

Figure 5 shows ELViM projections of the folding ensemble but colored according to the ellipticity of the peptide. The total ellipticity is shown in Figure 5A, and in comparison with Figure 4, it is clear that the regions of greater helical content are those associated with greater fraction of native contacts, that is, with the folded ensemble. Figure 4B–D shows the helical content of the peptide in the regions corresponding to helices I, II, and III, respectively, on top of the ELViM projection. It is possible to perceive in Figure 4A–C that the formation of helix I is a poorer indicative of the fold state than the formation of helices II and III. Thus, the maps suggest that helix I is less stable (less populated in the native states) than helices II and III, in agreement with experimental data.[2,7,8] In Figure S3 of the Supporting Information, we show the histogram of the probability of formation of each helix in the folding ensembles. The probability of helix I being completely unfolded (α-helix content of ≤0.25) is 43%. On the other side, the probabilities of helices II and III being unfolded are smaller, 32 and 22%, respectively. Similarly, previous simulations show the greatest instability of helix I and additionally suggest that it is particularly unstable in the absence of contacts with helix II.[4]

The histograms of the occurrence of each helical state can provide a perspective on the occupancy of the phase spaces as a function of the α-helix content of BdpA. Figure 5A–D shows the ELViM projections of the BpdA protein-folding ensembles for different extents of helix formation. The unfolded states will be named U_H^L, where the subindex H indicates the maximum amount of helical content and L is a label associated with the region in the landscape projection where the structures are found.

Figure 6A shows at least three distinct regions (U_{25}^1, $U_{25}^{2.1}$, and $U_{25}^{2.2}$) in phase space where unfolded BdpA is most likely to be found with a low ellipticity (<25%). These regions represent distinct unfolded states. In Figure 6B, several dissimilar sets of structures with significant density have α-helix contents within 25–50%. According to Garcia and Onuchic, the transition state of BdpA contains at least 40% of the α-helix,[4] thus being found in the structure ensemble of Figure 6B. On the other hand, many structures with α-helix content greater than 50–75% (Figure 6C) already show characteristics of native structures, as they almost exclusively occupy the native state region (N_{100}^8 region; Figure 6D).

The populations of each of the three helices in the 2D phase space (Figures 6 and Figures S4 and S5 of the Supporting Information) indicate that helix I is weakly correlated with the folding ensemble (Figure 7). Figure 7D shows that structures with higher α-helix I content occupy mainly the region of the

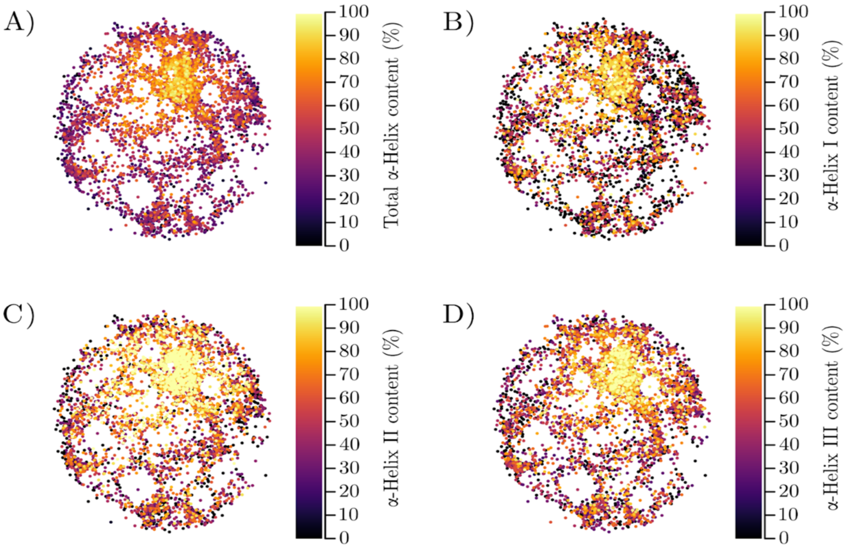


Figure 5. Helical content is projected into the folding ensemble. The color in each plot is a function of the content of the (A) total helical content of the protein, (B) α-helix I (^{10}QQNAFYEILH^{19}), (C) α-helix II (^{25}EEQRNGFIQSLKD^{37}), and (D) α-helix III (^{42}SANLLAEAKKLNDAQ^{56}).

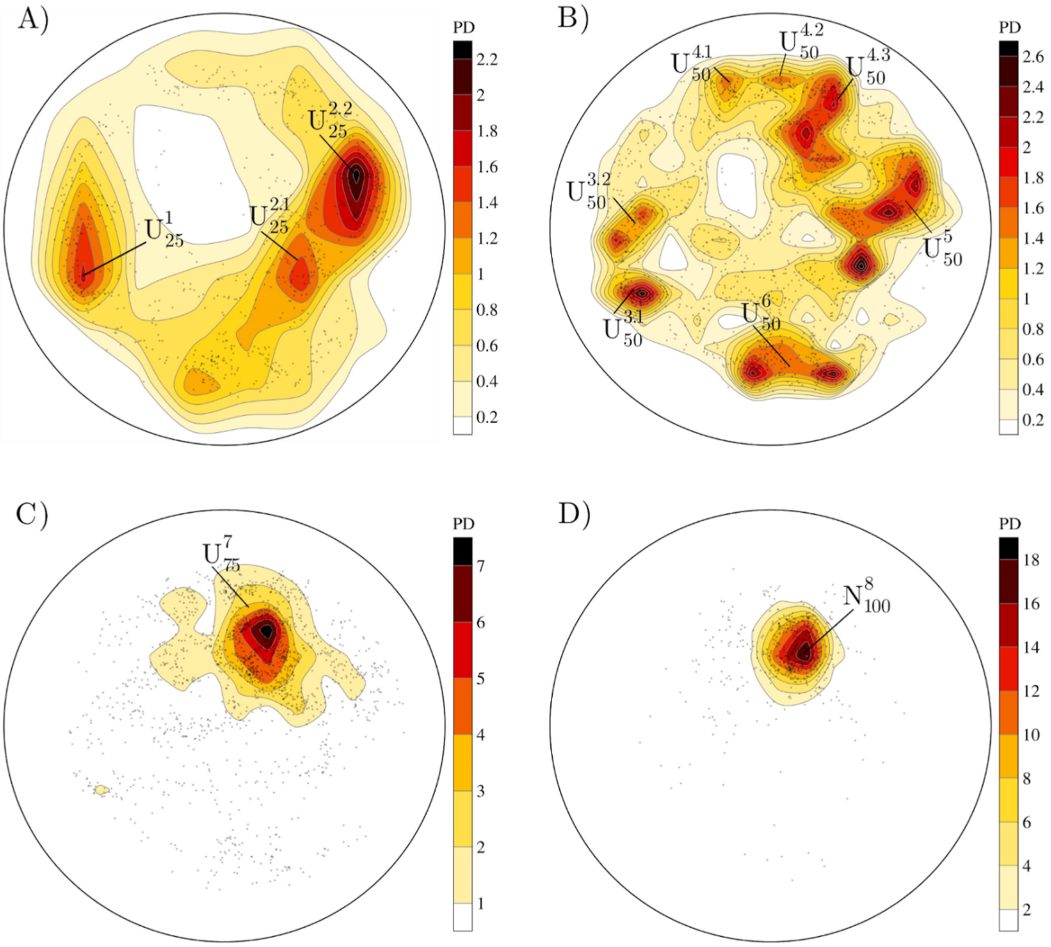


Figure 6. Contour maps of the PD of protein A structures with (A) 0−25%, (B) >25−50%, (C) >50−75%, and (D) >75−100% of the α-helix. The α-helix content is computed from the sum of the three helices of protein A.

folded protein (N^{8}_{100}), as expected. However, structures with low helix I content (Figure 7A) are also concentrated in the native state basin. Therefore, structures exhibiting mostly native contacts can still display an unfolded helix I. On the other hand, the probability of finding helices II or III unfolded in the native state basin (Figures S4A and S5A) is low. Thus, the heterogeneity of the native state arises from the variability of helix I, as suggested by Otosu et al. (2017).[2] In parallel, the higher probability regions associated to unfolded helices II and III (U^{1}_{25}, $U^{2.1}_{25}$, and $U^{2.2}_{25}$; Figures S4A and S5A) are associated to the unfolded state basins, which are dissimilar to one another.

In Figure 6B and Figures S4B and S5B, it is possible to visualize that there are structures at the edges of the 2D projection (thus unfolded structures) that have partially formed helices of all types. The gradual folding of the helices toward the native structure (N^{8}_{100} region) is consistent with the funnel-like energy landscape.[21]

In general, helices II and III appear to be correlated, as noted by the similarity of the histograms of occurrence of Figures S4 and S5. Because of this, we sought to evaluate the Spearman correlation coefficient of each helix during BdpA folding. The Spearman correlation coefficient is particularly useful here, given that the correlations between helical contents (Figure S6) or between the helices and the whole protein fold (Figure S7) are, apparently, nonlinear.

In Figure 8, we show the histograms of occurrence of each helix as a function of the helical content of each other helix to illustrate the possible correlations between their formation in the complete folding ensemble. Figure 7A,B shows that there is a low correlation between helices I and II, and between helices I and III: helices II and III can display a wide variety of structuration states, while helix I is unfolded. On the other hand, in Figure 8C, we see that the most probable states involving helices II and III are those where both helices have a low α-helix content (≤30%) or are structured (α-helix content ≥70%). The correlation coefficients for the formation of helices I and II (Figure 8A), I and III (Figure 8B), and II and III (Figure 8C) are 0.1539, 0.1259, and 0.2561, confirming that helices II and III are somewhat correlated, at least to a higher degree than the other pairs of helices. The higher correlation between helices II and III justifies the similarity in the histograms of Figures S4 and S5 and the probability distributions of the α-helix contents in Figure S3B,C.

Finally, we find partially formed helices of all types in the various unfolded states (Figure 6B and Figures S4B and S5B of the Supporting Information). We then calculated the most frequent contacts in each partially unfolded state, and the majority of these contacts (~64%) involve some helix-III residue (Table S4 of the Supporting Information). This supports helix III as the most important structure element in partially folded states, in some sense correlating with the study of BdpA chemical denaturation by Yanaka et al.,[14] which

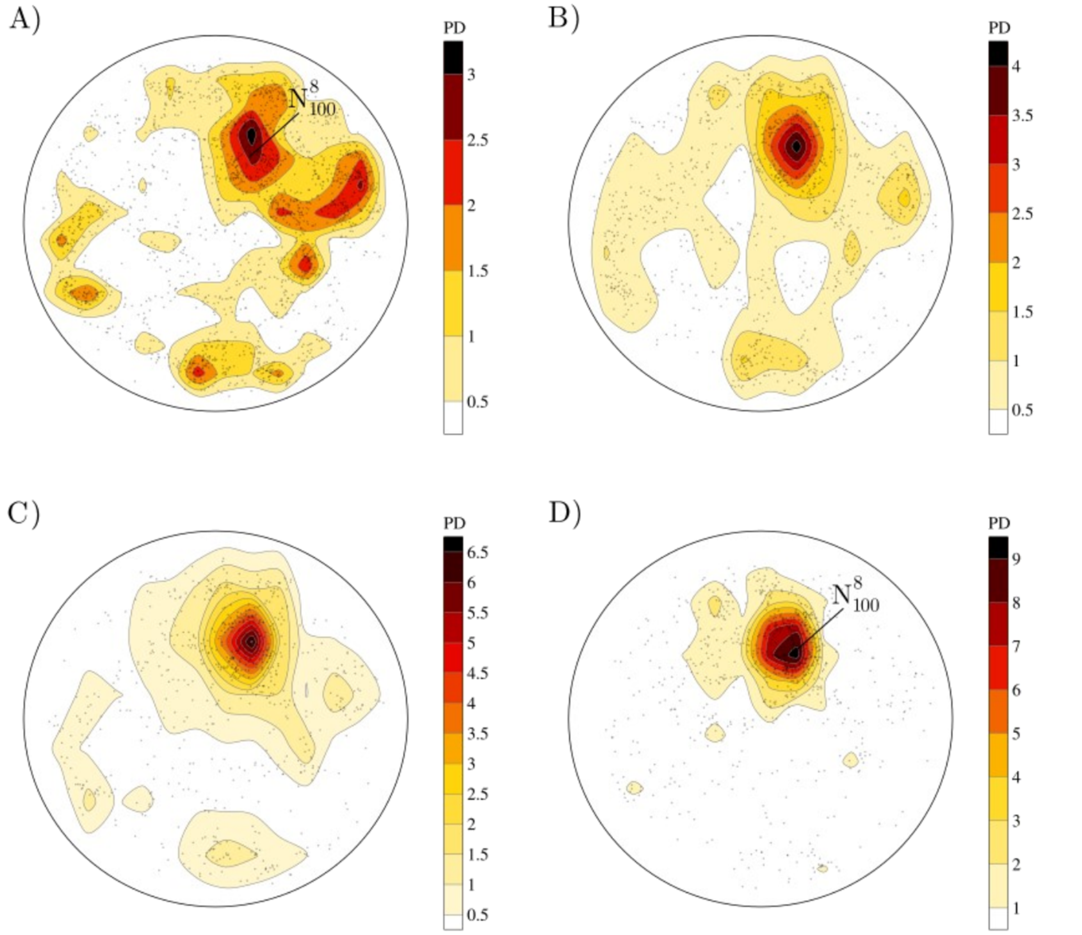


Figure 7. Contour maps of the PD of protein A structures with (A) 0−25%, (B) >25−50%, (C) >50−75%, and (D) >75−100% of α-helix I (^{10}QQNAFYEILH^{19}).

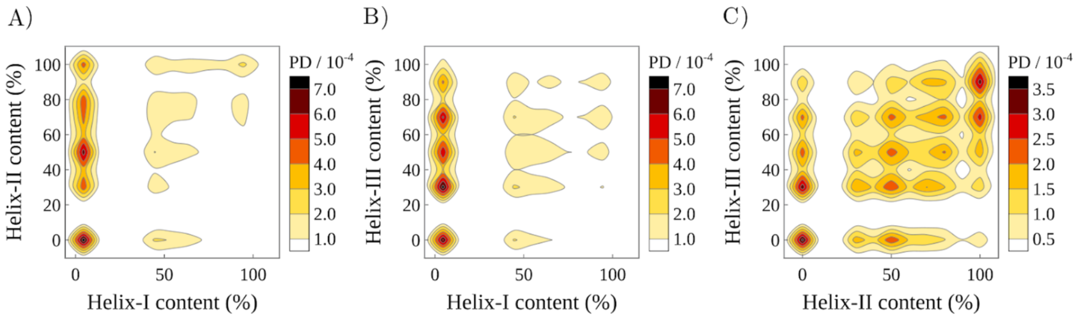


Figure 8. PD of states as a function of the contents of (A) α-helix I and α-helix II, (B) α-helix I and α-helix III, and (C) α-helix II and α-helix III of protein A.

showed that the most protected residues belong to helix III and might form a folding initiation site.

3.3. Hydration in Folded and Unfolded States.

As illustrated in Figure 6, we obtained a classification of various unfolded states of protein A, as a function of their distances in the 2D projection and of their secondary structure content. We performed simulations in water of each reconstructed all-atom model to provide a detailed characterization of how the folding state of the protein affects its solvation structure. This analysis was carried out using MDDFs and the KB theory of solvation.[48] MDDFs are adequate for representing solvation structures of irregularly shaped solutes, thus making possible obtaining a molecular and thermodynamic view of the correlation between the fold state and hydration structures. Distribution functions and KB integrals were averaged for the structures of each set represented in Figure 6.

Figure 9A shows the average MDDFs of water for selected sets of folded and unfolded states (the complete set of MDDFs and KBIs is available in Figures S8 and S9 and Table S3 of the Supporting Information). We choose here to illustrate the solvation of three sets: the native states of basin N_{100}^{8} and two unfolded states, $U_{50}^{4.2}$ and U_{50}^{6}, which displayed the maximum and minimum hydrogen-bonding peaks among the sets classified in Figure 6. In Figure 9B, the overlap of the structures of these ensembles are shown. The set $U_{50}^{04.2}$ has a

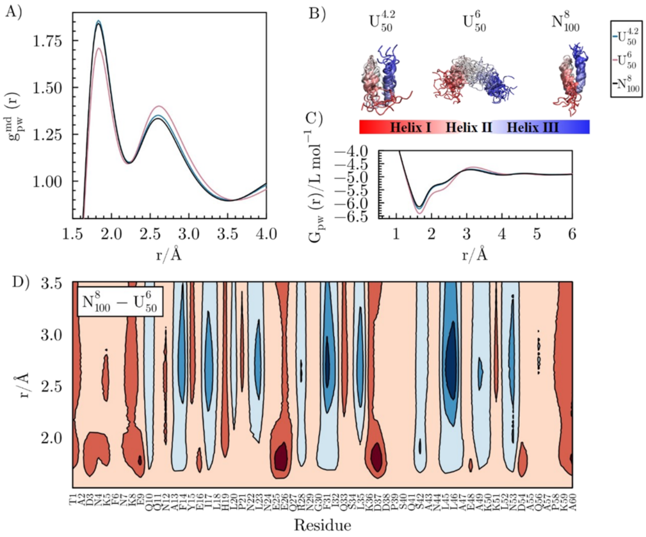


Figure 9. (A) MDDFs of water for selected folding ensemble subsets. (B) Structure sets taken from Figure 6. (C) Corresponding KB integrals for water. (D) Difference in the MDDF density of the water in the vicinity of native (N_{100}^{8}) and unfolded (U_{50}^{6}) states. The red color indicates greater densities of water in native N_{100}^{8}, while the blue colors are associated with greater densities of water in the unfolded U_{50}^{6} set, which highlights the interactions of the solvent with mostly hydrophobic residues that are protected from solvent in the native state.

tertiary arrangement that resembles that of the folded states but with ill-defined helices, while U_{50}^{6} structures are unwound and almost completely extended.

Overall, the density of water molecules increases at short distances and decreases in the second solvation layer as the protein folds (Figure S8 of the Supporting Information). The native N_{100}^{8} and the unfolded $U_{50}^{4.2}$ display similar distribution functions (Figure 9A, black and green) and KB integrals (Figure 9C). Thus, this illustrates that unfolded states may interact with the solvent similarly to the native state, if the conformations are such to expose similar surface areas and residues. On the other hand, states of basin U_{50}^{6} display a shorter hydrogen-bonding peak (at ∼1.9 Å) and a larger hydration peak associated to nonspecific interactions (at ∼2.7 Å).

The first dips in the KBIs (Figure 9C) are associated with the excluded protein volumes. The native set of structures is the most compact (shallower first dip). The U_{50}^{6} set has the deeper minimum, implying a larger exclusion volume. However, the density augmentation of water around the second peak compensates the initial dip, and the KB integrals converge to virtually the same value (Figure S8 and Table S3), implying that here the apparent molar volumes of the protein are similar independently of the folding state.

In Figure 9D, we illustrate the variations in the density of water molecules on the surface of individual residues in both the native and unfolded states by computing the by-residue contribution for the MDDFs. The intriguing aspect of this analysis is that it shows how water molecules interact with the protein residues, depending on the conformational states of the protein and, consequently, its exposure to the solvent. Regions marked in red denote higher water density in the native state, while those in blue indicate higher water density in the unfolded state. Notably, within the first solvation shell, the water density is significantly greater in the native state (N_{100}^{8}) compared to the unfolded state (U_{50}^{6}), particularly in proximity to charged residues (D3, K5, K8, E9, E25, E26, K36, D37, and D38). In the second shell, the density is in general greater in the unfolded state, particularly around hydrophobic residues (F14, I17, L23, F31, I32, L35, L45, and L46) that are exposed when the protein undergoes denaturation. The present approach, thus, can identify which residues are exposed or protected from solvent upon protein denaturation in a solvent-shell-dependent manner.

4. CONCLUSIONS

In this work, we combine C_α-SBM and atomistic simulations to provide insights into the equilibrium of BdpA helices in folded and unfolded ensembles. Simulations with C_α-SBMs and the ELViM method allowed obtaining a detailed picture of the BdpA folding ensembles with multiple folded, and particularly unfolded states of BdpA. Furthermore, we identified many sets of dissimilar structures with similar α-helix contents in the highest PD regions of the 2D phase spaces of BdpA. The formation of the individual helices were investigated from the structures of the atomistic simulations, which yielded equilibrated structures crucial for determining the secondary structure.

Helix I is the most unstable of the helices and is responsible for the heterogeneity of the native state of BdpA, while helix II is the most stable. We observed a gradual folding of the helices toward the native structure, consistent with the funnel-like landscape. However, we also identified that the helices are weakly but positively correlated with Spearman correlation coefficients equal to 0.15 (helices I and II), 0.13 (helices I and III), and 0.26 (helices II and III). This result indicates that helices are formed and deform numerous times in a weakly correlated manner.

Finally, with the precise characterization of the structural ensemble, we were able to describe the hydration structures of each structure basin using MDDFs. The hydration structure changes from folded to unfolded states, reducing the relative importance of hydrogen bonds and increasing the water density at distances associated with nonspecific interactions. The apparent molar volume of the structures, however, ends up being similar for all states. A detailed analysis of the contribution of each residue to hydration structures allowed the identification of the residues that are exposed to water upon denaturation. These results pave the way for the analysis of cosolvent effects on the protein folding structure and equilibrium.

■ ASSOCIATED CONTENT

Data Availability Statement

All simulation input scripts and data files necessary for the reproduction of the work presented here are available at the public repository: https://github.com/m3g/2023_AFP-LM_BpdA.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.3c01822.

Tables (Tables S1−S4) and figures (Figures S1−S11) with additional analyses of the simulation convergence, secondary structure content, and native contact maps of the BdpA; MDDFs, KBI, and error bars for all sets of structures of the folding ensemble (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Leandro Martínez − *Institute of Chemistry and Center for Computing in Engineering & Science, Universidade Estadual de Campinas (UNICAMP), 13083-861 Campinas, SP, Brazil;* orcid.org/0000-0002-6857-1884; Email: lmartine@unicamp.br

Author

Ander Francisco Pereira − *Institute of Chemistry and Center for Computing in Engineering & Science, Universidade Estadual de Campinas (UNICAMP), 13083-861 Campinas, SP, Brazil*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jcim.3c01822

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge the financial support of Fapesp (2010/16947-9, 2018/24293-0, 2013/08293-7, 2018/14274-9, and 2020/04549-0) and CAPES 206-04/092018. Research developed with the help of CENAPAD-SP (National Center for High Performance Processing in São Paulo), project UNICAMP/FINEP - MCTI.

■ REFERENCES

(1) Wolynes, P. G. Latest Folding Game Results: Protein A Barely Frustrates Computationalists. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (18), 6837−6838.

(2) Otosu, T.; Ishii, K.; Oikawa, H.; Arai, M.; Takahashi, S.; Tahara, T. Highly Heterogeneous Nature of the Native and Unfolded States of the B Domain of Protein A Revealed by Two-Dimensional Fluorescence Lifetime Correlation Spectroscopy. *J. Phys. Chem. B* **2017**, *121* (22), 5463−5473.

(3) Davis, C. M.; Cooper, A. K.; Dyer, R. B. Fast Helix Formation in the B Domain of Protein A Revealed by Site-Specific Infrared Probes. *Biochemistry* **2015**, *54* (9), 1758−1766.

(4) García, A. E.; Onuchic, J. N. Folding a Protein in a Computer: An Atomic Description of the Folding/unfolding of Protein A. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100* (24), 13898−13903.

(5) Oliveira, A. B., Jr.; Yang, H.; Whitford, P. C.; Leite, V. B. P. Distinguishing Biomolecular Pathways and Metastable States. *J. Chem. Theory Comput.* **2019**, *15* (11), 6482−6490.

(6) 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.* **2000**, *97* (1), 133−138.

(7) 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.* **2004**, *101* (18), 6952−6956.

(8) 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.* **2006**, *360* (4), 850−864.

(9) Lammert, H.; Schug, A.; Onuchic, J. N. Robustness and Generalization of Structure-Based Models for Protein Folding and Function. *Proteins* **2009**, *77* (4), 881−891.

(10) Noel, J. K.; Schug, A.; Verma, A.; Wenzel, W.; Garcia, A. E.; Onuchic, J. N. Mirror Images as Naturally Competing Conformations in Protein Folding. *J. Phys. Chem. B* **2012**, *116* (23), 6880−6888.

(11) Whitford, P. C.; Noel, J. K.; Gosavi, S.; Schug, A.; Sanbonmatsu, K. Y.; Onuchic, J. N. An All-Atom Structure-Based Potential for Proteins: Bridging Minimal Models with All-Atom Empirical Forcefields. *Proteins* **2009**, *75* (2), 430−441.

(12) Myers, J. K.; Oas, T. G. Preorganized Secondary Structure as an Important Determinant of Fast Protein Folding. *Nat. Struct. Biol.* **2001**, *8* (6), 552−558.

(13) Bai, Y.; Karimi, A.; Dyson, H. J.; Wright, P. E. Absence of a Stable Intermediate on the Folding Pathway of Protein A. *Protein Sci.* **1997**, *6* (7), 1449−1457.

(14) Yanaka, S.; Yagi-Utsumi, M.; Kato, K.; Kuwajima, K. The B Domain of Protein A Retains Residual Structures in 6 M Guanidinium Chloride as Revealed by Hydrogen/deuterium-Exchange NMR Spectroscopy. *Protein Sci.* **2023**, *32* (3), No. e4569.

(15) Gouda, H.; Torigoe, H.; Saito, A.; Sato, M.; Arata, Y.; Shimada, I. Three-Dimensional Solution Structure of the B Domain of Staphylococcal Protein A: Comparisons of the Solution and Crystal Structures. *Biochemistry* **1992**, *31* (40), 9665−9672.

(16) 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.* **2007**, *372* (1), 254−267.

(17) Baxa, M. C.; Freed, K. F.; Sosnick, T. R. Quantifying the Structural Requirements of the Folding Transition State of Protein A and Other Systems. *J. Mol. Biol.* **2008**, *381* (5), 1362−1381.

(18) Clementi, C.; Nymeyer, H.; Onuchic, J. N. Topological and Energetic Factors: What Determines the Structural Details of the Transition State Ensemble and "En-Route" Intermediates for Protein Folding? An Investigation for Small Globular Proteins. *J. Mol. Biol.* **2000**, *298* (5), 937−953.

(19) Taketomi, H.; Ueda, Y.; Gō, N. Studies on Protein Folding, Unfolding and Fluctuations by Computer Simulation. I. The Effect of Specific Amino Acid Sequence Represented by Specific Inter-Unit Interactions. *Int. J. Pept. Protein Res.* **1975**, *7* (6), 445−459.

(20) Onuchic, J. N.; Nymeyer, H.; García, A. E.; Chahine, J.; Socci, N. D. The Energy Landscape Theory of Protein Folding: Insights into Folding Mechanisms and Scenarios. *Adv. Protein Chem.* **2000**, *53*, 87−152.

(21) Onuchic, J. N.; Wolynes, P. G. Theory of Protein Folding. *Curr. Opin. Struct. Biol.* **2004**, *14* (1), 70−75.

(22) Noel, J. K.; Onuchic, J. N. The Many Faces of Structure-Based Potentials: From Protein Folding Landscapes to Structural Characterization of Complex Biomolecules. In *Computational Modeling of Biological Systems; Biological and Medical Physics, Biomedical Engineering*; Springer US: Boston, MA, 2012; pp 31−54.

(23) Noel, J. K.; Whitford, P. C.; Onuchic, J. N. The Shadow Map: A General Contact Definition for Capturing the Dynamics of Biomolecular Folding and Function. *J. Phys. Chem. B* **2012**, *116* (29), 8692−8702.

(24) Sobolev, V.; Sorokine, A.; Prilusky, J.; Abola, E. E.; Edelman, M. Automated Analysis of Interatomic Contacts in Proteins. *Bioinformatics* **1999**, *15* (4), 327−332.

(25) Noel, J. K.; Whitford, P. C.; Sanbonmatsu, K. Y.; Onuchic, J. N. SMOG@ctbp: Simplified Deployment of Structure-Based Models in GROMACS. *Nucleic Acids Res.* **2010**, *38*, W657−W661.

(26) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4* (3), 435−447.

(27) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. THE Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules I. The Method. *J. Comput. Chem.* **1992**, *13* (8), 1011−1021.

(28) Noel, J. K.; Levi, M.; Raghunathan, M.; Lammert, H.; Hayes, R. L.; Onuchic, J. N.; Whitford, P. C. SMOG 2: A Versatile Software Package for Generating Structure-Based Models. *PLoS Comput. Biol.* **2016**, *12* (3), No. e1004794.

(29) Oliveira Junior, A. B.; Lin, X.; Kulkarni, P.; Onuchic, J. N.; Roy, S.; Leite, V. B. P. Exploring Energy Landscapes of Intrinsically Disordered Proteins: Insights into Functional Mechanisms. *J. Chem. Theory Comput.* **2021**, *17* (5), 3178−3187.

(30) Dias, R. V. R.; Pedro, R. P.; Sanches, M. N.; Moreira, G. C.; Leite, V. B. P.; Caruso, I. P.; de Melo, F. A.; de Oliveira, L. C. Unveiling Metastable Ensembles of GRB2 and the Relevance of Interdomain Communication during Folding. *J. Chem. Inf. Model.* **2023**, *63* (20), 6344−6353.

(31) Sanches, M. N.; Parra, R. G.; Viegas, R. G.; Oliveira, A. B., Jr; Wolynes, P. G.; Ferreiro, D. U.; Leite, V. B. P. Resolving the Fine Structure in the Energy Landscapes of Repeat Proteins. *QRB Discov* **2022**, *3*, No. e7.

(32) Viegas, R. G.; Sanches, M. N.; Chen, A. A.; Paulovich, F. V.; Garcia, A. E.; Leite, V. B. P. Characterizing the Folding Transition-State Ensembles in the Energy Landscape of an RNA Tetraloop. *J. Chem. Inf. Model.* **2023**, *63* (17), 5641−5649.

(33) da Silva, F. B.; Simien, J. M.; Viegas, R. G.; Haglund, E.; Leite, V. B. P. Exploring the Folding Landscape of Leptin: Insights into Threading Pathways. *J. Struct. Biol.* **2024**, *216* (1), No. 108054.

(34) Hardin, C.; Eastwood, M. P.; Prentiss, M. C.; Luthey-Schulten, Z.; Wolynes, P. G. Associative Memory Hamiltonians for Structure Prediction without Homology: Alpha/beta Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100* (4), 1679−1684.

(35) Rotkiewicz, P.; Skolnick, J. Fast Procedure for Reconstruction of Full-Atom Protein Models from Reduced Representations. *J. Comput. Chem.* **2008**, *29* (9), 1460−1465.

(36) Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. PACKMOL: A Package for Building Initial Configurations for Molecular Dynamics Simulations. *J. Comput. Chem.* **2009**, *30* (13), 2157−2164.

(37) Martínez, J. M.; Martínez, L. Packing Optimization for Automated Generation of Complex System's Initial Configurations for Molecular Dynamics and Docking. *J. Comput. Chem.* **2003**, *24* (7), 819−825.

(38) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102* (18), 3586−3616.

(39) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *The Journal of Chemical Physics.* **1983**, *79*, 926−935.

(40) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, Flexible, and Free. *J. Comput. Chem.* **2005**, *26* (16), 1701−1718.

(41) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81* (8), 3684−3690.

(42) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52* (12), 7182−7190.

(43) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22* (12), 2577−2637.

(44) Joosten, R. P.; te Beek, T. A. H.; Krieger, E.; Hekkelman, M. L.; Hooft, R. W. W.; Schneider, R.; Sander, C.; Vriend, G. A Series of PDB Related Databases for Everyday Needs. *Nucleic Acids Res.* **2011**, *39* (Databaseissue), D411−D419.

(45) Martínez, L.; Matthies, M. ProteinSecondaryStructures.jl: This Package Parses STRIDE and DSSP Secondary Structure Prediction Outputs, to Make Them Convenient to Use from Julia, Particularly for the Analysis of MD Simulations. 2023. DOI: 10.5281/zenodo.8192529.

(46) Itoh, K.; Sasai, M. Flexibly Varying Folding Mechanism of a Nearly Symmetrical Protein: B Domain of Protein A. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (19), 7298−7303.

(47) Daniel, W. W. *Applied Nonparametric Statistics*; Wadsworth Publishing Company, 1990.

(48) Martínez, L.; Shimizu, S. Molecular Interpretation of Preferential Interactions in Protein Solvation: A Solvent-Shell Perspective by Means of Minimum-Distance Distribution Functions. *J. Chem. Theory Comput.* **2017**, *13* (12), 6358−6372.

(49) Shimizu, S.; Boon, C. L. The Kirkwood-Buff Theory and the Effect of Cosolvents on Biochemical Reactions. *J. Chem. Phys.* **2004**, *121* (18), 9147−9155.

(50) Martínez, L. ComplexMixtures.jl: Investigating the Structure of Solutions of Complex-Shaped Molecules from a Solvent-Shell Perspective. *J. Mol. Liq.* **2022**, *347*, No. 117945.

(51) Martínez, L. CellListMap.jl: Efficient and Customizable Cell List Implementation for Calculation of Pairwise Particle Properties within a Cutoff. *Comput. Phys. Commun.* **2022**, *279*, No. 108452.

(52) Piccoli, V.; Martínez, L. Correlated Counterion Effects on the Solvation of Proteins by Ionic Liquids. *J. Mol. Liq.* **2020**, *320* (114347), 114347.

(53) de Oliveira, I. P.; Martínez, L. The Shift in Urea Orientation at Protein Surfaces at Low pH Is Compatible with a Direct Mechanism of Protein Denaturation. *Phys. Chem. Chem. Phys.* **2020**, *22* (1), 354−367.

(54) Harries, D.; Rösgen, J. A Practical Guide on How Osmolytes Modulate Macromolecular Properties. *Methods Cell Biol.* **2008**, *84*, 679−735.

(55) Oprzeska-Zingrebe, E. A.; Smiatek, J. Aqueous Ionic Liquids in Comparison with Standard Co-Solutes: Differences and Common Principles in Their Interaction with Protein and DNA Structures. *Biophys. Rev.* **2018**, *10* (3), 809−824.

(56) Schroer, M. A.; Michalowsky, J.; Fischer, B.; Smiatek, J.; Grübel, G. Stabilizing Effect of TMAO on Globular PNIPAM States: Preferential Attraction Induces Preferential Hydration. *Phys. Chem. Chem. Phys.* **2016**, *18* (46), 31459−31470.

(57) Jackson, S. E.; Fersht, A. R. Folding of Chymotrypsin Inhibitor 2. 1. Evidence for a Two-State Transition. *Biochemistry* **1991**, *30* (43), 10428−10435.

(58) Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. Theory of Protein Folding: The Energy Landscape Perspective. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545−600.

(59) da Silva, F. B.; Martins de Oliveira, V.; de Oliveira Junior, A. B.; de G. Contessoto, V.; Leite, V. B. P. Probing the Energy Landscape of Spectrin R15 and R16 and the Effects of Non-Native Interactions. *J. Phys. Chem. B* **2023**, *127* (6), 1291−1300.

(60) Guo, Z.; Brooks, C. L., 3rd; Boczko, E. M. Exploring the Folding Free Energy Surface of a Three-Helix Bundle Protein. *Proc. Natl. Acad. Sci. U. S. A.* **1997**, *94* (19), 10161−10166.