JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

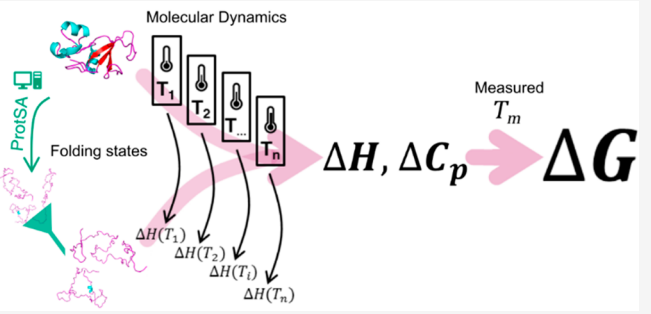
pubs.acs.org/jcim Article

Calculation of Protein Folding Thermodynamics Using Molecular Dynamics Simulations

Juan J. Galano-Frutos, Francho Nerín-Fonz, and Javier Sancho*

Cite This: J. Chem. Inf. Model. 2023, 63, 7791−7806 Read Online

ABSTRACT: Despite advances in artificial intelligence methods, protein folding remains in many ways an enigma to be solved. Accurate computation of protein folding energetics could help drive fields such as protein and drug design and genetic interpretation. However, the challenge of calculating the state functions governing protein folding from first-principles remains unaddressed. We present here a simple approach that allows us to accurately calculate the energetics of protein folding. It is based on computing the energy of the folded and unfolded states at different temperatures using molecular dynamics simulations. From this, two essential quantities (ΔH and ΔCp) are obtained and used to calculate the conformational stability of the protein (ΔG). With this approach, we have successfully calculated the energetics of two- and three-state proteins, representatives of the major structural classes, as well as small stability differences ($\Delta\Delta G$) due to changes in solution conditions or variations in an amino acid residue.



1. INTRODUCTION

Proteins are very versatile biological molecules,[1] and thermodynamics can greatly help to understand how they fold and perform useful tasks.[2,3] Molecular Dynamics (MD) simulation has become a powerful tool to study protein folding and other related processes.[4−12] However, despite great efforts in developing algorithms and methods to enable longer and better sampled simulations and in improving the accuracy of force fields and water models, significant challenges remain.[13] On one hand, simulating the protein folding time (from microseconds up to tens of seconds) in explicit solvent remains inaccessible, except for small fast-folding proteins.[5,8,10,11] On the other hand, work on improving the accuracy of MD force fields seems to have focused on reproducing structural, dynamic, and mechanistic aspects of protein behavior[14−17] and paid less attention to try to reproduce protein potential energy. One reason for this is the difficulty of obtaining accurate structural models of unfolded ensembles, which has prevented comprehensive studies of this side of the problem, making fine-tuning of the force field parameters challenging. The experimental limitations inherent in quantifying individual atomic interactions and the massive cancellation of interactions that takes place in a protein folding reaction[18] add to the complexity of the goal.[2] All of the above has perhaps frustrated the interest of scientists in the use of MD simulations to quantitatively study protein thermodynamics, hindering progress in many applied fields, such as protein design,[19] drug design,[20] genetic interpretation,[21] protein engineering,[22] or cell engineering.[23]

Recently, we addressed this issue by carrying out accurate, quantitative calculations of conformational stability on two two-state model proteins (barnase and nuclease) through an all-atom MD simulation approach.[24] The approach circumvents the simulation of the whole folding/unfolding time and is based on separately simulating the two relevant conformations. The folded state is modeled starting from an experimentally determined structure that is conveniently solvated and sampled conformationally. The unfolded state is modeled and sampled from an ensemble of completely unfolded conformations generated by the ProtSA server[25] that are similarly solvated. From the simulations, the enthalpy change of unfolding (ΔH_{unf}) is calculated by the difference (unfolded state minus folded state enthalpy averages), while the heat capacity change at constant pressure (ΔC_{punf}) is obtained from the temperature dependence of the calculated enthalpy change. As a final step, the calculated thermodynamic quantities (ΔH_{unf} and ΔCp_{unf}) are combined with the experimentally determined melting temperature (T_m) to calculate the conformational stability of the protein (ΔG_{unf}) as a function of temperature by means of the Gibbs−Helmholtz equation.[26]

Received: July 20, 2023
Revised: October 18, 2023
Accepted: October 18, 2023
Published: November 13, 2023

https://doi.org/10.1021/acs.jcim.3c01107
J. Chem. Inf. Model. 2023, 63, 7791−7806

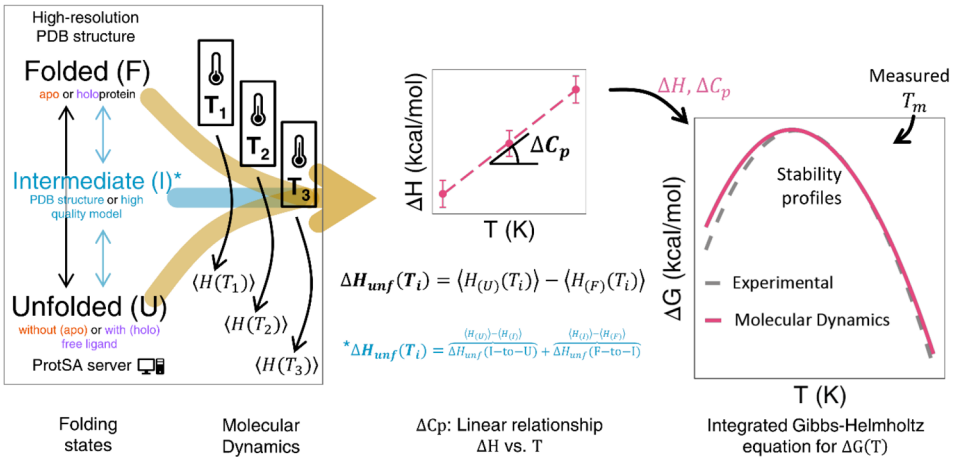


Figure 1. General workflow of the devised MD-based approach. The enthalpy of simulation boxes containing either folded (e.g., $H_{apo(F)}$ or $H_{holo(F)}$) or unfolded (e.g., $H_{apo(U)}$ or $H_{apo(U)+cofactor}$ in holoproteins) protein or, when applicable, a structure representative of an intermediate state (e.g., $H_{apo(I)}$) is directly computed and averaged from MD simulations. The unfolding enthalpy change (ΔH_{unf}) of interest is obtained as the difference between the enthalpies of the appropriate simulation boxes. The simulations are performed at three temperatures, and the change in heat capacity (ΔCp_{unf}) is obtained as the slope of a linear plot of enthalpy change versus temperature. The two calculated thermodynamic changes (ΔH_{unf} and ΔCp_{unf}) are combined with the experimental T_m of the protein to calculate the conformational stability by using the Gibbs−Helmholtz equation (eq 1). For holoproteins, a similar equation, SI eq 5 in the Supporting Information, is used that applies a correction to Gibbs free-energy to account for the ligand concentration and uses the van't Hoff approximation to describe the temperature dependence of the binding constant, $K_b(T)$. The number of water molecules and ions present in the folded and unfolded (or intermediate, if applicable) boxes must be identical. Forty replicas of the folded box (normally built from a high-resolution PDB structure) and 100 replicas of the unfolded one (built from a filtered sample of completely unfolded conformations generated by the ProtSA server[25]) are simulated. For intermediate states, 100 simulation replicas were built from a representative structural ensemble. For holoproteins, the unfolded box is built by placing an unfolded protein molecule generated with ProtSA and one molecule of the cofactor at a given minimum distance of the protein. The rest of the general details can be found in Methods and in panel a of Figures 2−4 and Figures S1−S4.

One initial goal of the approach was testing the ability of classical force fields, e.g. Charmm22-CMAP[15] and AmberSB99-ILDN[16] (or the more recently released AmberSB99-disp[14]), to yield accurate folding energetics by difference, using systems solvated with explicit water. Thus, the indicated force fields were combined with seven explicit water models, Tip3p,[27] Tip4p,[27] Tip4p-d,[28] Tip4-d-mod,[14] Tip5p,[29] Spc,[30] and Spc/E.[31] Results obtained from short MD simulations (2 ns productive trajectories per replica) and the combinations of either Charmm22-CMAP or AmberSB99-ILDN with Tip3p allowed, for the two proteins indicated, to finely capture the energy balance between the numerous interactions established between protein and solvent atoms in both the native state and the unfolded ensemble.[32]

In this work, we generalize the described methodology using the most accurate combination of force field and water model found[24] and a larger conformational sampling (see Methods) and demonstrate the precise correspondence of the thermodynamic quantities calculated on a set of two-state, three-state, apo, holo, wild-type (WT), or mutated proteins with their experimentally determined values. In addition to barnase[33,34] and nuclease[35,36] (that are here calculated anew with higher precision[24]), we present the calculation for additional two-state proteins: barley chymotrypsin inhibitor 2 (CI2, truncated variant)[37,38] and phage T4 lysozyme[39] (WT and pseudo-WT variant), for a three-state protein: apoflavodoxin from *Anabaena PCC 7119*[40−43] (for which the energetics involved in the two unfolding transitions, F-to-I and I-to-U, is obtained), and for a holoprotein: flavodoxin from *Anabaena PCC 7119* (which contains a flavin mononucleotide (FMN) cofactor, noncovalently bound). Furthermore, we evaluate the capability and limits of the approach to capture small stability changes or small differences between similar systems, e.g. those associated with mutation ($\Delta\Delta H_{mut\text{-}nat}$ and $\Delta\Delta G_{mut\text{-}nat}$), changes in pH ($\Delta\Delta H_{pH1\text{-}pH2}$ and $\Delta\Delta G_{pH1\text{-}pH2}$), or individual steps within a multistate unfolding ($\Delta H_{unf(F\text{-}to\text{-}I)}$, $\Delta Cp_{unf(F\text{-}to\text{-}I)}$, and $\Delta G_{unf(F\text{-}to\text{-}I)}$ or $\Delta H_{unf(I\text{-}to\text{-}U)}$, $\Delta Cp_{unf(I\text{-}to\text{-}U)}$, and $\Delta G_{unf(I\text{-}to\text{-}U)}$). Although the method requires a reliable structural model for the folded conformation, which sometimes may not be available, advances in high resolution AI-based protein modeling[44,45] will likely allow the application of the method to the entire proteome.

2. METHODS

2.1. General MD Simulation Workflow for Calculation of Unfolding Energetics (ΔH_{unf}, ΔCp_{unf}, and ΔG_{unf}) in Apoproteins. A previous version of the workflow here described has been reported.[24] The current version (Figure 1) relies on a higher sampling of the folded and unfolded states. Briefly, X-ray crystal structures with the highest resolution and sequence coverage have been retrieved from the RCSB Protein Data Bank (https://www.rcsb.org/,[46,47] see PDB codes below) and taken as the starting structures for modeling the native (folded) state. When needed, the initial crystal structure has been used to model the amino acid replacement leading to the mutant simulated (e.g., the CI2 Ile76Ala and lysozyme Ile3Glu variants).

Forty replicas of the folded structure have been simulated, each consisting of a single protein molecule solvated with water molecules in a specified simulation box additionally containing, when required, ions (Na^+ and/or Cl^-). On the other hand, a random sample of 100 unfolded structures has been extracted from a large unfolded ensemble (~2000 structures) generated by the ProtSA server[25] from the protein

https://doi.org/10.1021/acs.jcim.3c01107
J. Chem. Inf. Model. 2023, 63, 7791−7806

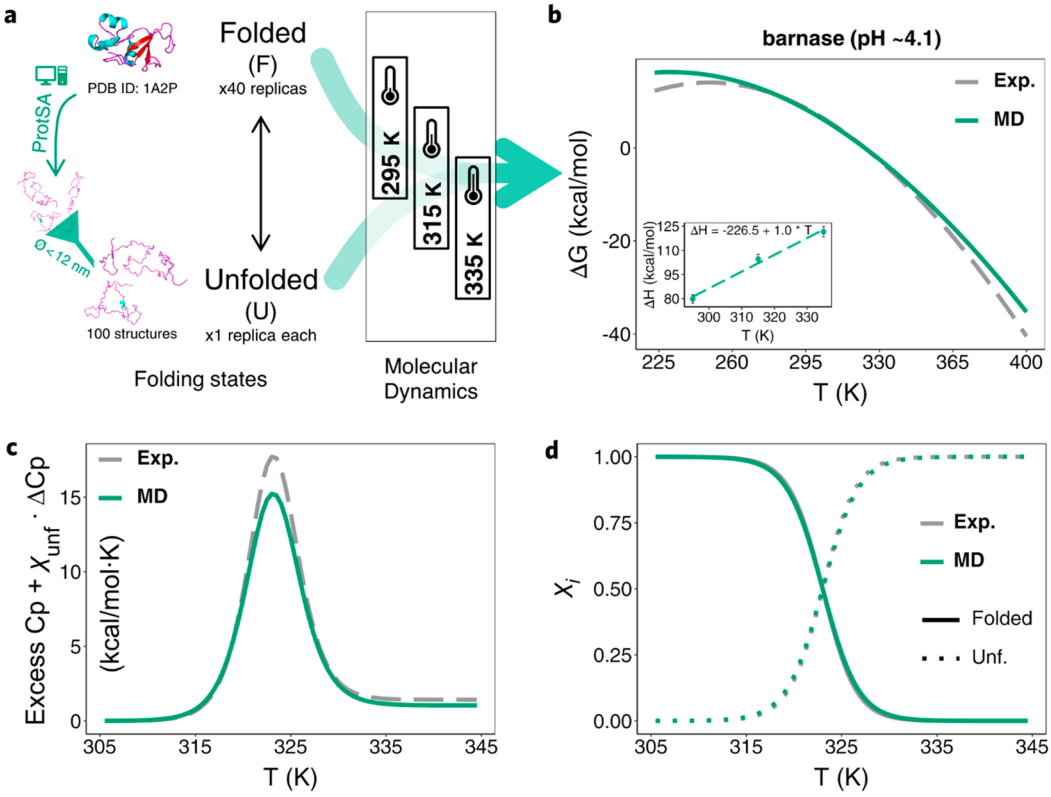


Figure 2. Simplified MD-based scheme and comparison with experimental results for a two-state protein example: barnase. a) The protein models, the number of structures (unfolded) and replicas (folded) simulated, the diameter cutoff used to filter too-elongated unfolded structures obtained from ProtSA[25] (left, see also Figure S5), and temperatures selected for the MD-based calculation (Charmm22-CMAP) of thermodynamics of barnase. b-d) Stability curves ($\Delta G_{unf}(T)$), thermograms (Excess Cp + $\chi_{unf} \times \Delta$Cp vs T), and protein molar fractions (χ_i) vs T plot (*in silico* vs experimental), respectively, obtained for barnase simulated at pH ~ 4.1. Inset in b depicts the calculated ΔH_{unf} vs T linear plot with the fitted equation (the slope being ΔCp_{unf}) obtained from the MD simulations. The color coding is indicated in the legends of the panels.

sequence (see Figure 1 and panel a in Figures 2−4 and Figures S1−S4). ProtSA uses the Flexible-Meccano algorithm[48] to generate the backbone-conformation and Sccomp[49] to add the side chains. Flexible-Meccano uses a coil-library and a simple volume exclusion term to perform conformational sampling, so that the protein unfolded ensembles generated successfully describe backbone fluctuations typically observed in intrinsically disordered proteins (probed by NMR and SAXS experiments).[25]

To avoid using too large simulation boxes, which would increase the simulation time as well as add noise to the results, the most extended unfolded conformations (~10%) generated by ProtSA have been previously identified and removed as described[24] (using a diameter-based filtering, Figure S5). The selected 100 unfolded conformations have been simulated in boxes containing one unfolded molecule and exactly the same number of water molecules, ions, and cofactors—when it is the case— as in the corresponding boxes used to simulate the folded conformations of the same protein. For three-state proteins, in addition to the overall enthalpy change, those of the individual steps (F-to-I and I-to-U) can be obtained if the absolute enthalpy of an additional box containing one molecule of protein in the intermediate conformation and the same number of water and ion entities is calculated (see Figure 1 and Figure 3a). To model the intermediate conformation, a suitable structural model is needed. In three-state apoFld, a 20-model NMR ensemble previously described[50] has been used. In this case, five replicas have been simulated for each of the 20 structures, totaling 100 replicas, the same number of unfolded conformations modeled (Figure 3a).

For each replica, a short 2 ns productive trajectory (see Table S1) has been run, and the individual time-averaged enthalpy (H^i_F, H^i_U, or H^i_I) has been retrieved. The individual enthalpies of replicas of the same conformational state (i.e., folded, unfolded or intermediate) have been ensemble-averaged to obtain the enthalpy corresponding to each folding state ($\langle H_F \rangle$, $\langle H_U \rangle$, or $\langle H_I \rangle$). Subsequently, the unfolding enthalpy change, ΔH_{unf}, has been calculated by difference, i.e. by subtracting the calculated ensemble-averaged enthalpy obtained from simulations of the folded state from the ensemble-averaged enthalpy obtained from simulations of the unfolded state: $\Delta H_{unf} = \langle H_U \rangle - \langle H_F \rangle$. For three-state proteins, enthalpy changes corresponding to the first unfolding transition (F-to-I) and the second one (I-to-U) have been calculated likewise: $\Delta H_{unf(F\text{-to-}I)} = \langle H_I \rangle - \langle H_F \rangle$ and $\Delta H_{unf(I\text{-to-}U)} = \langle H_U \rangle - \langle H_I \rangle$ (Figure 1).

The use of multiple short 2 ns simulations in this study is motivated by the well-known overcompaction problem associated with Charmm22-CMAP when long simulations are performed.[24] We believe that although the sampling of conformational space achieved in an individual 2 ns simulation is limited, the overall sampling obtained by simulating a large and diverse set of starting unfolded structures, as done here (see Section 2.6 below), is adequate.

The calculation of the heat capacity change upon unfolding (ΔCp_{unf}) relies on the linear dependency of ΔH_{unf} with

https://doi.org/10.1021/acs.jcim.3c01107
J. Chem. Inf. Model. 2023, 63, 7791−7806

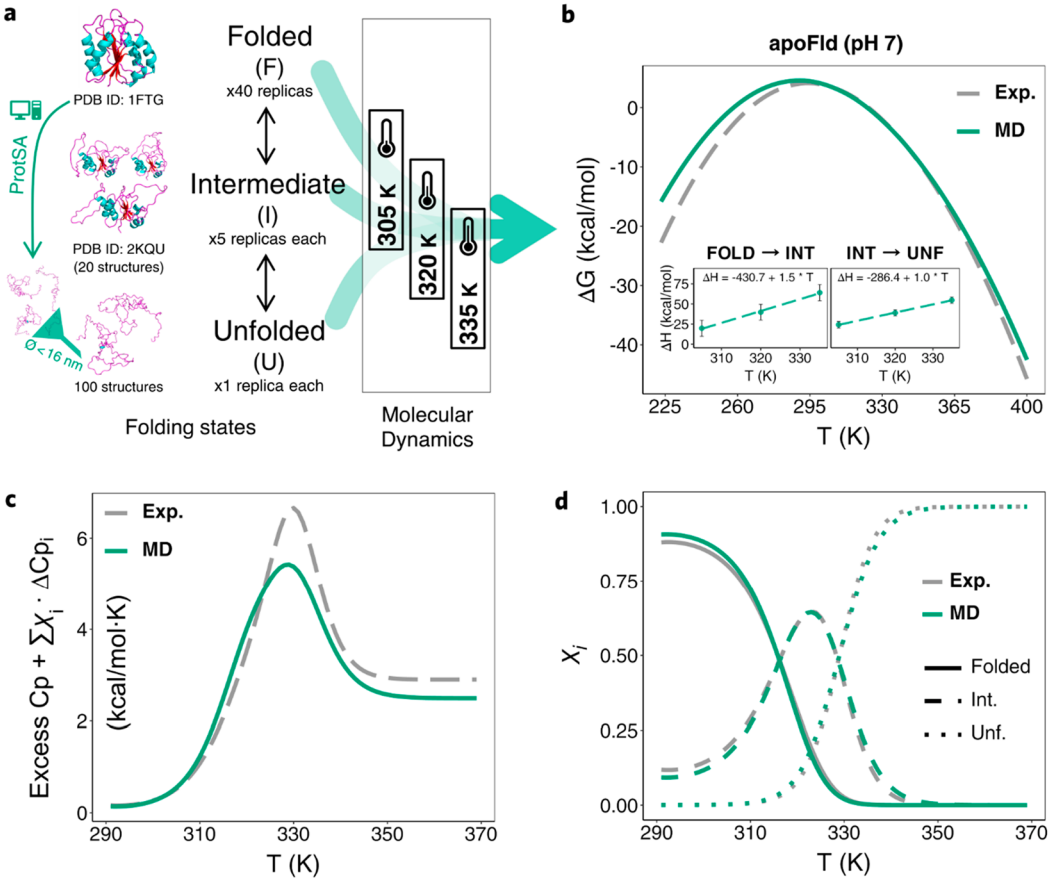


Figure 3. Simplified MD-based scheme and comparison with experimental results for a three-state protein example: apoFld. a) Protein models, number of structures (unfolded) and replicas (folded) simulated, diameter cutoff used to filter too-elongated unfolded structures obtained from ProtSA[25] (left, see also Figure S5), and temperatures selected for the MD-based calculation (Charmm22-CMAP) of apoFld thermodynamics. b-d) Global stability curves ($\Delta G_{unf}(T) = \Delta G_{unf(F\text{-}to\text{-}I)}(T) + \Delta G_{unf(I\text{-}to\text{-}U)}(T)$), thermograms (Excess Cp + $\sum \chi_i \times \Delta Cp_i$ vs T), and protein molar fractions (χ_i) vs T plot (*in silico* vs experimental), respectively. Inset in b depicts linear plots of calculated ΔH_{unf} from the MD simulations vs T, with the fitted equation (the slope being ΔCp_{unf}) obtained. The color coding is indicated in the legends of the panels.

temperature. For each protein, three not-distant temperatures spanning 30−40 degrees have been selected so that the temperature range covered contains the experimental T_m of the simulated protein. The three calculated ΔH_{unf} values have been represented as a function of simulation temperature, and the ΔCp_{unf} has been calculated as the slope of a linear fit. For three-state proteins (e.g., apoFld), $\Delta Cp_{unf(F\text{-}to\text{-}I)}$ and $\Delta Cp_{unf(I\text{-}to\text{-}U)}$ have been obtained as the temperature dependence of the calculated enthalpy changes of the corresponding unfolding transition, assuming a linear dependency of ΔH_{unf} with temperature (i.e., a temperature independent ΔCp_{unf}) is a good and common approximation for performing short extrapolations. However, ΔCp_{unf} is temperature dependent.[51,52] To assess whether assuming a constant ΔCp_{unf} affects ΔH_{unf} extrapolation to T_m, we have additionally calculated barnase ΔH_{unf} at six temperatures spanning 100 °C and compared the calculated ΔCp_{unf} and ΔH_{unf} extrapolated to T_m with those obtained as indicated above.

The calculation of the protein stability curves (ΔG_{unf} as a function of temperature) has been done through the Gibbs−Helmholtz equation[26] (eq 1)

$$\Delta G(T) = \Delta H_{T_m} \times (1 - T/T_m) - \Delta Cp \times [T_m - T + T \times \ln(T/T_m)] \quad (1)$$

introducing the calculated ΔH_{unf} and ΔCp_{unf} values and the reported experimental T_m.

2.2. Specific MD Simulation Workflow for Calculation of Unfolding Energetics (ΔH_{unf}, ΔCp_{unf}, and ΔG_{unf}) in Holoproteins. In the case of holoproteins (noncovalent complexes of apoprotein and cofactor; e.g. holoFld), the ensemble-averaged enthalpy of the folded (bound) state, $\langle H_{holo(F)} \rangle$, has been obtained from simulations (40 replicas) each consisting of one molecule of holoFld solvated with water molecules and ions, as needed (Figure 1 and Figure 4a). Similarly, the energetics of the unfolded (unbound) state has been modeled from simulations (100 replicates) in which one unfolded protein molecule generated with ProtSA[25] and one cofactor molecule (placed at a minimum distance of 3 nm from the protein) have been put together in a box, where they have been solvated in the same way (Figure 4a). The ensemble-averaged enthalpy of such boxes, $\langle H_{apo(U)+cofactor} \rangle$, has been obtained following the averaging scheme of the general workflow. Then, the unfolding enthalpy change has been calculated as $\Delta H_{unf} = \langle H_{apo(U)+cofactor} \rangle - \langle H_{holo(F)} \rangle$. As required for this enthalpy change calculation by difference, the number of water molecules and ions in the box containing unfolded protein and cofactor must equal those in the box containing folded holoprotein (Table S2). The simulations have also been performed at three different temperatures, and the unfolding

https://doi.org/10.1021/acs.jcim.3c01107
J. Chem. Inf. Model. 2023, 63, 7791−7806

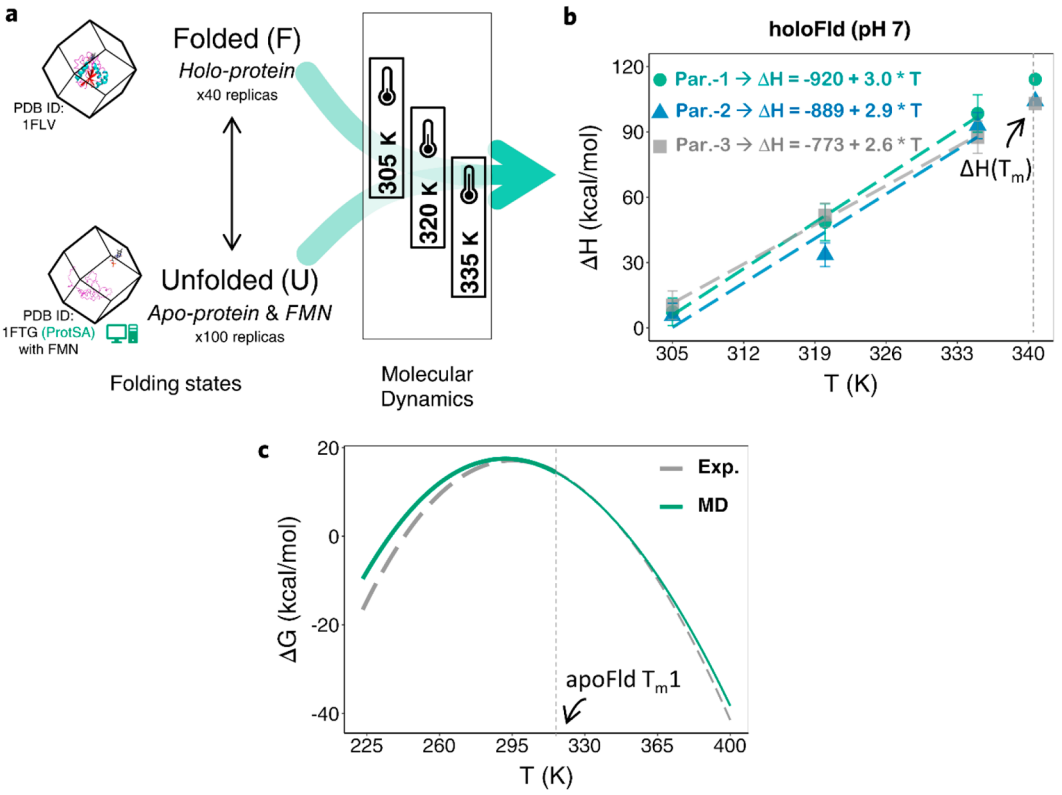


Figure 4. Simplified MD-based scheme and comparison with experimental results for a holoprotein example: holoFld. a) Protein and cofactor models placed in the simulation boxes, folding states, number of structures (unfolded) and replicas (folded) simulated, and temperatures selected for the MD-based calculation (Charmm22-CMAP) of holoFld thermodynamics. b) Calculated ΔH_{unf} vs T linear plots, with the fitted equations (slopes are the respective ΔCp_{unf}) obtained for the three FMN parametrizations tested. Extrapolated ΔH_{unf} values at T_m (340.7 K) are indicated over the vertical dashed line at this temperature. c) Stability curves ($\Delta G_{unf}(T)$) (*in silico* vs experimental) obtained from SI eq 5. Curves appear depicted with finer lines beyond the first T_m of the apoprotein (316.2 K, Table 1, vertical dashed line) to indicate that in this region the ΔG_{unf} values calculated are not reliable. This is so because the van't Hoff approximation to model the temperature dependence of the binding constant[53] should work fine as long as the conformation of the protein binding site does not change significantly. However, this will not be the case at temperatures where the apoprotein begins to unfold, and we consider the stability curve of the holoprotein (panel c) to be not reliable beyond the first melting temperature (T_m1) of the apoprotein (316.2 K in the case of apoFld). The fact that at 298.15 K the calculated stability of HoloFld (17.3 ± 2.6 kcal/mol) agrees within error with the stability measured from experimental thermal unfolding curves (19.0 ± 0.9 kcal/mol)[54] seems to validate the accuracy of the profiles in the range of temperatures below the apoprotein T_m1. Similar to the case of apoproteins, the ΔH_{unf} and ΔCp_{unf} values calculated for the holoprotein can be combined with the experimental T_m to obtain the protein stability curves (ΔG_{unf} as a function of temperature). However, as the conformational stability of holoproteins is cofactor concentration dependent, a modified Gibbs−Helmholtz equation that takes into account the binding energetics (SI eq 5, see details in SI Methods) has been derived to calculate the conformational stability as a function of temperature and concentration of free cofactor.

ΔCp_{unf} has been obtained as the slope of a ΔH_{unf} versus temperature plot (Figure 4a-b).

2.3. Target Proteins and Case Studies. *2.3.1. Barnase from B. amyloliquefaciens and Nuclease from S. aureus.* 110-Residue barnase[55−59] and 149-residue nuclease[60−63] (C-terminal fragment) are well characterized proteins with a two-state equilibrium, as summarized in previous work.[24] Here, the two-state unfolding energetics of WT barnase and nuclease was determined using the present computational approach. In addition, the reported effect of pH on nuclease stability has been addressed (see Table 1).

2.3.2. CI2 from Barley Seeds. CI2[37,38] is a small, 84-residue, globular serine proteinase inhibitory protein extensively studied and reported to fold in a two-state manner as well as to display a two-state thermal unfolding equilibrium.[64−66] Its 19-residue N-terminal tail is completely unstructured.[67,68] We have focused here on a truncated form of CI2 lacking the unstructured N-terminal tail because the structure of the full-length protein is not available and because it has been shown that the tail does not contribute to the protein stability.[64,65] The truncated WT CI2 variant has been modeled at a solvating condition equivalent to pH 3.0 under which experimental energetics is available.[65] Due to the significantly different thermodynamics quantities reported for WT CI2 at pH 6.3[64,66] compared to those at pH 3.0 (see Table 1), we have also modeled WT CI2 at pH 6.3 in order to evaluate the sensibility of the method to solvent effects. On the other hand, the CI2 variant Ile76Ala which, relative to WT in identical solvent, shows a significantly lower unfolding enthalpy change and a large destabilization[65] (Table 1), has been selected to evaluate the feasibility of the approach to calculate the effect of single amino acid replacements on protein stability.

2.3.3. Phage T4 Lysozyme. T4 endolysin (lysozyme)[39] is a two-domain, 164-residue globular protein that has also been the subject of extensive study and widely used to investigate the role of hydrophobic interactions in protein structural stabilization.[69−71] Over 500 X-ray structures of T4 lysozyme have been obtained under a variety of experimental conditions

https://doi.org/10.1021/acs.jcim.3c01107
J. Chem. Inf. Model. 2023, 63, 7791−7806

Table 1. Experimental Thermal Unfolding Data

Protein	Method	Ref.	pH & IS (mM)[a]	T_m(s) (K)[b,c]	ΔH_{unf} (kcal/mol)[c,d]	ΔCp_{unf} (kcal/mol·K)[c]	ΔG^0_{unf} (kcal/mol)[c,e]
barnase	DSC	86	3.8 & 0.4[f]	321.0 ± 0.1	107.0 ± 7.0	0.9 ± 0.1	6.9 ± 0.7
		59	4.0 & 0.2[g]	321.7[g]	106.4[g]	1.4[g]	6.6
		87	4.0 & 8[h]	323.5[g]	125.0 ± 6.3	1.6 ± 0.2	8.2 ± 0.9
		88	4.4 & 6[i]	324.5[g]	126.7[g]	1.6[g]	8.5[g]
		89	4.4 & 6[i]	324.5 ± 0.0	128.6 ± 1.5	1.6 ± 0.2	8.7 ± 0.1
	(Ave ± SE)		**~4.1 & 4.0**	**323.0 ± 0.7**	**118.7 ± 4.9**	**1.4 ± 0.1**	**7.8 ± 0.4**
nuclease[j]	DSC	60	7.0 & 141[k]	327.3 ± 0.5	77.4 ± 7.7	2.2 ± 0.5	4.0 ± 3.0
			5.0 & 141[l]	323.8 ± 0.5	73.1 ± 7.3	2.2 ± 0.5	3.6 ± 2.7
			4.1 & 0.4[m]	316.4 ± 0.5	49.2 ± 4.9	1.2 ± 0.5	2.6 ± 1.5
		61	7.0 & 151[n]	326.5 ± 0.2	86.8 ± 1.4	2.4 ± 0.4	4.6 ± 1.4
			4.9 & 112[o]	321.5 ± 0.3	73.0 ± 6.6	2.4 ± 0.4	3.4 ± 1.9
			4.1 & 103[o]	314.4 ± 0.3	59.0 ± 5.3	1.2 ± 0.4	3.1 ± 1.0
	(Ave ± SE)		**pH 7.0[p]**	**326.9 ± 0.4**	**82.1 ± 4.7**	**2.3 ± 0.1**	**4.3 ± 0.3**
			pH 5.0[p]	**322.7 ± 1.1**	**73.1 ± 0.1**	**2.3 ± 0.1**	**3.5 ± 0.1**
			pH 4.1[p]	**315.4 ± 1.0**	**54.1 ± 4.9**	**1.2 ± 0.0**	**2.9 ± 0.3**
WT CI2[q]	DSC	65	3.0 & 1.9[f]	337.2 ± 0.2	61.0 ± 2.3	0.7[g]	5.4 ± 0.7
	Spectroscopy	64,66	6.3 & 33[r]	358.7[g,s]	78.4 ± 1.4[t]	0.8 ± 0.1[t]	7.2 ± 0.4[u]
Ile76Ala CI2[q]	DSC	65	3.0 & 1.9[f]	311.5[g]	30.2[g]	0.7[g]	1.1 ± 0.3
WT lysozyme	DSC	73	2.4 & 34[v]	317.4 ± 0.6[w]	109.0 ± 4.2[w]	2.3 ± 0.2[w]	5.3 ± 0.8
		75	2.4 & 34[v]	317.8 ± 0.2[x]	104.3 ± 2.2[x]	2.5 ± 0.0[x]	4.9 ± 0.6
	(Ave ± SE)		**2.4 & 34**	**317.6 ± 0.2**	**106.6 ± 2.3**	**2.4 ± 0.1**	**5.1 ± 0.6**
Ile3Glu lysozyme	DSC	75	2.4 & 34[v]	312.9 ± 0.3[x]	105.1 ± 3.0[x]	2.5 ± 0.0[x]	4.1 ± 0.8
pseudo-WT lysozyme[y]	DSC	77	3.0 & 3.8[z]	327.1 ± 0.5	127.4 ± 12.7	1.6 ± 0.2[§]	9.3 ± 1.5
		77	3.7 & 0.9[z]	337.6 ± 0.5	138.9 ± 13.9	1.6 ± 0.2[§]	12.5 ± 1.1
apoFld	Spectroscopy	42	7.0 & 17[‡]	316.5 ± 0.3	28.1 ± 0.6	1.5 ± 0.3	0.8 ± 0.7
				329.6 ± 0.1	54.9 ± 1.2	1.6 ± 0.5	2.8 ± 1.1
	Spectroscopy	43	7.0 & 17[‡]	316.0 ± 0.3	34.7 ± 2.0		
				328.3 ± 0.4	55.6 ± 2.0		
	Spectroscopy	83	7.0 & 17[‡]	317.3 ± 0.2	33.9 ± 0.1		
				329.0 ± 0.1	52.7 ± 0.6		
	Spectroscopy	90	7.0 & 17[‡]	317.3 ± 0.2	34.1 ± 0.1		
				329.0 ± 0.1	52.9 ± 0.6		
	Spectroscopy	50	7.0 & 17[‡]	317.2 ± 0.3	32.2 ± 0.1		
				328.9 ± 0.1	51.0 ± 1.1		
	DSC	50	7.0 & 17[‡]	317.1 ± 0.6	32.7 ± 0.4	1.2 ± 0.1	1.3 ± 0.5
				329.4 ± 0.1	66.0 ± 0.2	1.5 ± 0.1	4.0 ± 0.3
	DSC	42	7.0 & 17[‡]	311.8 ± 2.5	28.0 ± 3.7	≤1.5 ± <0.3	≥0.8 ± 4.2
				326.5 ± 1.2	56.0 ± 2.5	≥1.0 ± <0.3	≤3.6 ± 1.9
	(Ave ± SE)		**7.0 & 17**	**316.2 ± 0.8**	**32.0 ± 1.1**	**1.4 ± 0.3**	**1.1 ± 1.4**
				328.7 ± 0.4	**55.6 ± 2.0**	**1.6 ± 0.3**	**2.9 ± 1.3**
holoFld	DSC	84	7.0 & 17[‡]	340.7 ± 0.1	101.9 ± 0.6	3.5 ± 0.6[ϕ]	17.1 ± 2.7[‡]

[a]Experimental pH and ionic strength (IS) conditions. IS reported or calculated according to buffer, concentration, and pH reported. [b]Mid-denaturation temperature (T_m) reported or calculated from a reported empirical equation. [c]For three-state apoFld, two values are shown. The first one corresponds to the Native-to-Intermediate transition, and the second one corresponds to the Intermediate-to-Unfolded transition. [d]Enthalpy change upon thermal unfolding (ΔH_{unf}) either reported or calculated from a given empirical equation at T_m. [e]Standard (298.15 K) conformational stability (ΔG^0_{unf}) obtained from the Gibbs−Helmholtz equation[26] (eq 1), except otherwise noted. When more than one experimental data are reported, the ΔG^0_{unf} values shown in the "Ave ± SE" row are the average among those values (Ave), and the standard error is obtained by dividing the standard deviation (SD) between the square root of the number of data ($SD/\sqrt{n}$) (it is not the value calculated through the Gibbs−Helmholtz equation and its propagated associate error). For nuclease, values are calculated at 293.15 K, as experimental data appear reported at that temperature. [f]10 mM glycine hydrochloride. IS calculated from the Henderson-Haselbach equation and the Glycine pK_a values of 2.37 and 9.78.[91] [g]No error reported. [h]50 mM sodium acetate. [i]20 mM sodium acetate. [j]The modeled nuclease is the 149-residue C-ter fragment of the protein. [k]20 mM sodium phosphate, 100 mM NaCl, 1 mM EDTA. [l]20 mM sodium acetate, 100 mM NaCl, 1 mM EDTA. [m]20 mM glycine hydrochloride. The influence of salt concentration (between 0 and 800 mM) on measurements seems negligible (see Figure 1c of the reference paper).[60] [n]25 mM sodium phosphate, 100 mM NaCl. [o]20 mM sodium acetate, 100 mM NaCl. [p]As measurements of nuclease unfolding thermodynamics are independent of IS[60] and this parameter largely varied in the experiments reported, the buffer IS is not taken into account in the modeling of this protein. [q]Truncated wild-type CI2 and Ile76Ala variant lacking the first 19 amino acid residues. [r]50 mM MES, as reported by Jackson et al.[64] [s]T_m

https://doi.org/10.1021/acs.jcim.3c01107
J. Chem. Inf. Model. 2023, 63, 7791−7806

Table 1. continued

reported by Tan et al.[66] [t]Obtained by extrapolating at T_m after doing a ΔH_{unf} vs. T_m fitting with reported data,[64] the slope being ΔCp_{unf}. [u]Value extrapolated to [GdnHCl] = 0 M from thermal denaturation data.[64] [v]20 mM potassium phosphate, 25 mM KCl, 0.5 mM dithiothreitol. [w]Values obtained from the reported empirical equations $T_m = 9.63 + 14.41 \times pH$ and $\Delta H_{unf} = 5.97 + 2.33 \times T$. ΔCp_{unf} is the slope of this fitting equation. [x]Values obtained from the reported empirical equations $T_m = 9.13 + 14.81 \times pH$ and $\Delta H_{unf} = -10.51\ (\pm 0.83) + 2.57\ (\pm 0.02) \times T$ for the wild-type protein, $T_m = -0.62\ (\pm 0.13) + 16.84\ (\pm 0.05) \times pH$ and $\Delta H_{unf} = 5.22\ (\pm 1.14) + 2.51\ (\pm 0.03) \times T$ (T in Celsius degrees) for the Ile3Glu variant. ΔCp_{unf} is the slope of the ΔH_{unf} vs. T fitting line. [y]Lysozyme variant where residues 54 and 97 appear replaced by a threonine and an alanine, respectively. [z]20 mM glycine hydrochloride. IS calculated from the Henderson-Haselbach equation and the glycine pK_a values of 2.37 and 9.78.[91] [§]Value obtained from the ΔH_{unf} vs. T linear fitting plot in Figure 6a of the reference paper.[77] [♯]50 mM MOPS at 298.15 K. [‡]Standard Gibbs free-energy of unfolding ([FMN] = 1 M) obtained from SI eq 5 (includes the correction of temperature and ligand concentration, see the SI Methods). For the calculation of this stability, the average ($K_b = 3.61(\pm 1.4) \times 10^9$ M) of binding constants reported for FMN,[54,80,84] as well as the enthalpy ($\Delta H_{bind} = -11.0 \pm 0.2$ kcal/mol) and heat capacity changes ($\Delta Cp_{bind} = -0.6 \pm 0.02$ kcal/mol·K) upon binding,[80] was used. As additional data, a standard Gibbs free-energy change of 19.0 ± 0.9 kcal/mol has been reported by Campos and co-workers.[54] [Φ]ΔCp_{unf} value estimated as follows: sum of ΔCp_{unf} of the two partial unfolding steps of apoFld (1.4 ± 0.3 and 1.6 ± 0.3 kcal/mol·K) plus the ΔCp of binding reported for FMN (−0.6 ± 0.02 kcal/mol·K).[80]

(buffer, pH, ionic strength), including those of an engineered pseudolysozyme (see below) and many variants thereof.[46] WT lysozyme carries two cysteine residues at positions 54 and 97. To ease experimental work on the protein, a Cys54Thr/Cys97Ala variant (termed pseudo-WT lysozyme) has often been studied. WT and pseudo-WT lysozymes[72] slightly differ in structure and thermodynamics[73−77] (Table 1). For the sake of testing the method, the energetics of these two lysozyme variants has been calculated. Besides, the energetics of the nonpseudolysozyme variant, Ile3Glu,[75] has been addressed as a further attempt to capture the effect of single amino acid replacements, and the pseudo-WT lysozyme[77] has been simulated in different solvent conditions (different pH values) to assess, as with nuclease and CI2, whether the method can capture pH-related effects on protein stability (Table 1).

2.3.4. Anabaena PCC 7119 Flavodoxin (Fld). Fld[40,78] is a 169-residue protein that carries electrons from photosystem I to ferredoxin-NADP+ reductase.[79] Fld capability to transfer electrons is conferred by the presence of a molecule of noncovalently bound FMN cofactor. Reversible removal of the cofactor from the holoprotein (holoFld) leads to the apo form (apoFld). Fld has been widely studied to investigate protein/cofactor interactions,[80,81] as well as non-native protein conformations.[42,50,82−84] While apoFld thermal unfolding equilibrium is three-state,[41−43] binding of FMN greatly stabilizes the complex so that holoFld unfolds following a two-state mechanism.[54,84] A detailed picture of Fld folding and binding thermodynamics is available.[41−43,50,54,80,83−85] The reasonably high enthalpy and heat capacity changes (Table 1) of the two apoFld unfolding transitions, namely folded-to-intermediate (F-to-I) and intermediate-to-unfolded (I-to-U), together with the availability of a representative structure of the intermediate conformation[50] have made us select this protein to test the simulation approach on the calculation of unfolding energetics in three-state proteins.

Structure Models (PDB Files) and Coverage. The starting structures used to simulate the folded state of the proteins analyzed have been those with the highest resolution available in the RCSB Protein Data Bank[46,47] at the time of writing this manuscript, namely the following: 1A2P (1.5 Å resolution)[58] for barnase, 2SNS (1.5 Å)[92] for nuclease (C-ter fragment), 2CI2 (2.0 Å)[93] for CI2 (truncated form), 6LZM (1.8 Å)[72] for lysozyme, 1L63 (1.75 Å)[94] for pseudolysozyme, 1FTG (2.0 Å)[95] for apoFld, and 1FLV (2.0 Å)[96] for holoFld. On the other hand, the thermal unfolding intermediate state of apoFld has been represented by 2KQU,[50] a 20-model NMR ensemble of the Phe99Asn mutant previously shown to constitute a reliable representation of this state.[50,82,97] According to the reference sequences in UniProt,[98] the structural coverage of the solved sequences is 3-110 (barnase), 83-231 (nuclease C-terminal fragment), 20-84 (WT CI2 and Ile76Ala mutant), 1-162 (WT lysozyme and Ile3Glu mutant), 1-162 (pseudo-WT lysozyme), and 3-170 (apo and holoFld).

2.4. Solvation Conditions and MD Simulation General Details. Solvation conditions on the simulated proteins (i.e., protonation states and the number of ions added) have been selected in each case to reproduce the experimental pH and ionic strength (IS) under which the experimental thermodynamics measurements were performed (see detailed information in SI Methods and Table S2). Box dimensions have been adopted from the diameter of the most elongated structure in the unfolded ensemble sampled for a given protein, plus a minimum distance of 1 nm from protein atoms to the simulation box edges. The MD simulation setup has been similar to that previously described[24] (details are also given in Table S1). All the systems have been simulated with the force field Charmm22 with CMAP correction (version 2.0)[15] and the explicit water model Tip3p:[27] the most accurate force field/water model combination reported in previous work.[24] The Amber99SB-ILDN[16] force field has been tested again, combined with Tip3p, by modeling the apoFld unfolding thermodynamics. MD simulations have been run and analyzed with Gromacs 2020.[99] Setting short 2 ns productive trajectories in the workflow[24] circumvents the known issue of structure overcompaction in long simulations[14,24] for force fields like Charmm22-CMAP[15] and Amber99SB-ILDN.[16] In addition, the simulations performed have been tested for protein overcompaction through the analysis of the evolution of the radius of gyration (Rg) along the trajectories (Table S3). Results of this analysis have confirmed that no significant protein compaction occurs over the trajectories of the systems simulated (Table S3). The mutant variants tested (of CI2 and lysozyme) have been modeled by replacing the wild-type residue by the new one, using the mutator tool of Chimera (v.1.15),[100] as no solved structures were available. No clashes have been observed in the final mutant structures of the lowest energy obtained after accommodating the new residues, which have been taken as the starting structures in simulations of their folded states. In the case of the apoFld intermediate state, the representative model used (see below) has been mutated back to the wild-type sequence (Chimera v.1.15)[100] in order to keep the same amino acid sequence as that of the other structural models used in simulations of apo and holoFld. No clashes have been

https://doi.org/10.1021/acs.jcim.3c01107
J. Chem. Inf. Model. 2023, 63, 7791−7806

Table 2. Calculated Thermal Unfolding Energetics from MD Simulations

Protein (condition & temperatures)[a]	Force Field[b]	ΔH_{unf} (kcal/mol)[c,d]	ΔCp_{unf} (kcal/mol·K)[d,e]	ΔG^0_{unf} (kcal/mol)[d,f]
barnase (pH: ~4.1, T: 295, 315, 335 K)	Charmm22-CMAP	110.4 ± 3.1	1.0 ± 0.1	7.5 ± 1.2
nuclease (pH: 7.0, T: 307, 317, 327 K)		75.1 ± 4.5	1.7 ± 0.3	4.8 ± 1.7
nuclease (pH: 5.0, T: 307, 317, 327 K)		71.0 ± 4.5	1.5 ± 0.4	4.4 ± 2.8
nuclease (pH: 4.1, T: 307, 317, 327 K)		86.3 ± 4.8	1.7 ± 0.1	4.8 ± 2.2
WT CI2 (pH: 3.0, T: 320, 335, 350 K)		46.1 ± 1.9	0.4 ± 0.0	4.3 ± 0.4
WT CI2 (pH: 6.3, T: 335, 350, 365 K)		57.1 ± 1.9	0.5 ± 0.1	6.9 ± 0.6
Ile76Ala CI2 (pH: 3.0, T: 320, 335, 350 K)		27.7 ± 1.7	0.5 ± 0.0	1.0 ± 0.2
WT lysozyme (pH: 2.4, T: 305, 320, 335 K)		184.6 ± 2.9	1.3 ± 0.2	10.6 ± 0.6
Ile3Glu lysozyme (pH: 2.4, T: 305, 320, 335 K)		174.6 ± 2.9	1.5 ± 0.1	7.8 ± 0.6
pseudo-WT lysozyme (pH: 3.0, T: 305, 320, 335 K)		205.1 ± 1.0	1.6 ± 0.2	16.2 ± 0.7
pseudo-WT lysozyme (pH: 3.7, T: 305, 320, 335 K)		227.4 ± 1.2	1.7 ± 0.3	22.5 ± 2.0
apoFld (pH: 7.0, T: 305, 320, 335 K)	Charmm22-CMAP	35.6 ± 6.0 48.1 ± 4.1 (83.7 ± 10.1)	1.5 ± 0.1 1.0 ± 0.0 (2.5 ± 0.1)	1.3 ± 1.7 3.0 ± 0.9 (4.3 ± 2.6)
	Amber99SB-ILDN	49.5 ± 4.2 18.3 ± 3.3 (67.8 ± 7.5)	1.4 ± 0.1 1.1 ± 0.1 (2.5 ± 0.2)	2.1 ± 1.5 0.1 ± 0.9 (2.2 ± 2.4)
holoFld (pH: 7.0, T: 305, 320, 335 K)	Charmm22-CMAP/Par.-1	114.2 ± 7.8	3.0 ± 0.2	17.3 ± 2.6[g]
	Charmm22-CMAP/Par.-2	104.2 ± 5.8	2.9 ± 0.6	
	Charmm22-CMAP/Par.-3	103.0 ± 6.5	2.6 ± 0.1	

[a] ΔH_{unf} is calculated at the three indicated temperatures. ΔCp_{unf} obtained as the slope of a ΔH_{unf} vs. T linear plot. [b] Force fields tested for the calculation, and FMN parametrizations used in holoFld systems (see Methods). The water model used is always Tip3p, as described in Methods. [c] Calculated enthalpy change upon thermal unfolding (ΔH_{unf}) at T_m (see values in Table 1), obtained by extrapolation. Given errors are standard error (SE) obtained as the sum of the SE from folded simulations (40 replicas) plus the SE from unfolded simulations (100 replicas) (see Table S4). [d] For three-state apoFld, three calculated ΔH values are shown. The upper one corresponds to the enthalpy change of the Native-to-Intermediate transition; the intermediate one corresponds to the enthalpy change of the Intermediate-to-Unfolded transition; and the lower one (between parentheses) corresponds to the total ΔH_{unf}, obtained by adding up the values calculated for each transition. Likewise, in the ΔCp_{unf} and ΔG^0_{unf} columns, the three values indicated correspond (from top to bottom) to the Native-to-Intermediate, Intermediate-to-Unfolded, and global (Native-to-Unfolded) heat capacity or Gibbs free-energy changes, respectively. [e] Calculated ΔCp_{unf} obtained as the slope of a ΔH_{unf} vs. T linear plot. Fitting errors are given as SE. [f] Unfolding Gibbs free-energy changes at 298.15 K calculated using the Gibbs−Helmholtz equation[26] (or SI eq 5 for HoloFld; see SI Methods). For nuclease, the temperature of reference used, 293.15 K, is the one at which most of the experimental data are reported (Table 1). Given errors are SE obtained by error propagation through the Gibbs−Helmholtz equation[26] (or SI eq 5 for HoloFld). [g] Standard Gibbs free-energy (at 1 M FMN) calculated through SI eq 5 (SI Methods and the footnote *bb* in Table 1).

observed after this replacement either. Crystal waters and any other nonprotein molecule have been removed from the PDB structural models chosen (see below).

2.5. FMN Parametrization. Three different parametrizations of the FMN molecule (charge −2) have been tested. Namely, 'Par.-1' has been obtained ad hoc, assisted by the AmberTools20 package[101] and the Gaussian 09 program;[102] 'Par.-2' is that reported by Schulten et al.;[103] and 'Par.-3' has been obtained through the SwissParam server.[104] FMN coordinates have been extracted from the crystal structure of holoFld (PDB ID: 1FLV[96]). For ad-hoc 'Par.-1', partial atomic charges have been modeled with Gaussian 09 (HF/6-31G*) and then fitted through the RESP method[105,106] (with Antechamber),[101,107] and finally, parameters have been obtained from the General Amber Force Field (GAFF,[108] Antechamber[101,107]). FMN coordinates have been uploaded to SwissParam[104] ('Par.-3') in mol2 format after adding hydrogen atoms. Except for van der Waals parameters, which have been taken from the closest atom type in Charmm2, parameters and

https://doi.org/10.1021/acs.jcim.3c01107
J. Chem. Inf. Model. 2023, 63, 7791−7806

charges with this server derive from the Merck Molecular Force Field (MMFF).[104]

2.6. Increased Sampling for Higher Precision. Individual enthalpies (H^i_F, H^i_U, or H^i_I) of the simulated systems (i.e., boxes containing one protein molecule, several ions, and thousands of water molecules) can mount to 10^5 (negative values) or even higher (see Table S4). These big figures are owed to the large number of water molecules present in the large simulation boxes required to solvate the unfolded conformations. In general, the larger the protein, the larger the negative enthalpy of the simulated box. Therefore, the calculation of unfolding thermodynamics by difference requires a high precision (a low standard error in the calculation) to be able to assess the accuracy of the approach (the difference between experimental and calculated results). Since the enthalpy change of a partial thermal unfolding step of a protein (e.g., the apoFld F-to-I or I-to-U transitions) can be significantly lower than the global enthalpy changes modeled before[24] (for barnase and nuclease, see Table 1), a higher precision (standard error $\leq$ 10) than that previously achieved[24] has been here guaranteed a priori by running a higher number of replicas. For each system (i.e., folded or unfolded), the minimum sample size (40 and 100, respectively) necessary to meet such precision has been estimated as reported.[24]

3. RESULTS

3.1. Energetics of Two-State Proteins: Barnase, Nuclease, CI2, and Lysozyme. The equilibrium thermal unfolding of barnase, nuclease, CI2, and lysozyme has been described to be two-state. Accordingly, we have calculated their unfolding energetics: ΔH_{unf} (at T_m), ΔCp_{unf}, and ΔG^0_{unf} (at 25.0 °C or, for nuclease, at 20.0 °C) using the general workflow described in Methods (see Figure 1) where the number of simulated replicas of the folded state and simulated structures in the unfolded ensemble has been increased relative to its initial formulation.[24] All calculated and experimentally determined ΔH_{unf}, ΔCp_{unf}, and ΔG^0_{unf} values will be reported in kcal/mol, kcal/mol·K, and kcal/mol units, respectively. For simplicity, the units are omitted in this Results section.

Barnase has been simulated (Figure 2a) at pH ∼ 4.1 (Table 1 and Table S2) under solvating conditions similar to those reported in experimental measurements. In previous modeling,[24] a reasonable agreement was found between experimental and calculated data. Here, the calculated values of ΔH_{unf}, ΔCp_{unf}, and ΔG^0_{unf} obtained with a larger conformational sampling (110.4 ± 3.1, 1.0 ± 0.1, and 7.5 ± 1.2, respectively, Table 2) agree very well with the averaged experimentally determined energetics (118.7 ± 4.9, 1.4 ± 0.1, and 7.8 ± 0.4, Table 1). Due to this fine agreement, the experimental and calculated temperature dependencies of ΔG_{unf} (stability curve, Figure 2b), (thermogram, Figure 2c), and state fractions (Figure 2d) nearly coincide. The agreement between experimental and calculated magnitudes is better than that obtained with a smaller sampling (92.3 ± 5.7, 0.9 ± 0.1, and 6.5 ± 0.8, respectively) in the previous calculation.[24]

Alternatively, barnase ΔCp_{unf} has been calculated from a linear fit of not just 3 but 6 ΔH_{unf} values newly obtained from MD simulations spanning 100 °C (from 275 to 375 K). The value and error obtained for ΔCp_{unf} are the same (1.0 ± 0.1), and the calculated ΔH_{unf} at T_m is 100.1 ± 2.2, which is close to the value of 110.4 ± 3.1 previously obtained. Considering the two calculations as independent experiments and using only the data obtained in the common temperature interval, the average values and standard errors obtained for ΔCp_{unf} and ΔH_{unf} at T_m are 1.1 ± 0.1 and 106.3 ± 4.0, respectively. The standard errors obtained are only slightly bigger than those reported in Table 2, obtained from a single calculation using ΔH_{unf} at three temperatures. On the other hand, we have noticed that the ΔH_{unf} versus T plot spanning 100 °C shows a slight departure from linearity (Figure S6) as expected if ΔCp_{unf} is not constant.[51,52] Because the experimental information on the temperature dependence of ΔCp_{unf} is lacking for most of the proteins analyzed here, both the calculated and experimental stability curves displayed in Figures 2−4 and Figures S1−S4 are obtained from eq 1 or SI eq 5 (Figure 4), using constant ΔCp_{unf} values, either experimental or calculated.

Nuclease unfolding thermodynamic data are available over a range of pH (from 3 to 8.5) and solvating conditions.[60,61] WT nuclease has been simulated (Figure S1a) at three pH values: 7.0, 5.0, and 4.1 (see solvating conditions and protonation states in Table 1 and Table S2). At pH 7.0, the calculated ΔH_{unf}, ΔCp_{unf}, and ΔG^0_{unf} values (75.1 ± 4.5, 1.7 ± 0.3, and 4.8 ± 1.7, respectively, Table 2) match very well the averaged experimental ones (82.1 ± 4.7, 2.3 ± 0.3, and 4.3 ± 0.3, Table 1). This excellent agreement is reflected, as seen for barnase, in a fine correspondence between the experimental and calculated temperature dependences of the Gibbs free-energy difference, thermogram, and molar fractions (Figure S1b-d). The second solvating condition simulated for nuclease reproduces a protonation scheme previously used,[24] corresponding to pH 5.0. Under this condition, our calculated energetics (ΔH_{unf} = 71.0 ± 4.5, ΔCp_{unf} = 1.5 ± 0.4, and ΔG^0_{unf} = 4.4 ± 2.8, Table 2) matches fairly well the experimental values (73.1 ± 0.1, 2.3 ± 0.1, and 3.5 ± 0.1, respectively, Table 1 and Figure S1e-f). The application here of a more exhaustive sampling yields results for nuclease that are as accurate as those obtained for this protein with a smaller sampling in previous work (ΔH_{unf} = 76.0 ± 8.1, ΔCp_{unf} = 1.8 ± 0.1, and ΔG^0_{unf} = 4.6 ± 1.4).[24] Nuclease stability is thus accurately calculated in the pH range 5.0−7.0. At lower pH (pH 4.1), however, the method overestimates ΔH_{unf} and ΔCp_{unf}, which leads to a less accurate calculated stability (4.8 ± 2.2, Table 2) compared to the experimental value (2.9 ± 0.3, see Table 1 and Figure S1g-h).

Thermodynamic data for chymotrypsin inhibitor 2 (WT truncated form, see Methods) and for a broad set of point mutants analyzed under different solvation conditions (varying in pH and ionic strength) are available[64−66] (Table 1). Here, WT CI2 has been simulated (Figure S2a) at two pH conditions for which reliable experimental data are reported (Table 1 and Table S2). At pH 3.0, the calculated ΔH_{unf} and ΔCp_{unf} values (46.1 ± 1.9 and 0.4 ± 0.03, respectively) are a bit lower than the corresponding experimental values (61.0 ± 2.3 and 0.72). Notwithstanding, the calculated ΔG^0_{unf} at this pH (4.3 ± 0.4) virtually agrees within error of the experimental stability (5.4 ± 0.7). At pH 6.3, CI2 is more stable than at pH 3.0, as the experimental ΔH_{unf} and ΔCp_{unf} values (78.4 ± 0.7 and 0.8 ± 0.1, respectively) combine to a higher conformational stability (ΔG^0_{unf} = 7.2 ± 0.4). The higher experimental ΔH_{unf} and ΔCp_{unf} values at pH 6.3 relative to pH 3.0 are captured by our simulations (calculated values at pH 6.3:57.1 ± 0.5 and 0.5 ± 0.07), and so is the increase in conformational stability (calculated value at pH 6.3: 6.9 ± 0.6). We have also assessed the capability of the simulation approach to detect changes in stability associated with point mutations. For that,

https://doi.org/10.1021/acs.jcim.3c01107
J. Chem. Inf. Model. 2023, 63, 7791−7806

we have computed the energetics of the Ile76Ala CI2 variant at pH 3.0 and compared it to that of WT CI2 at the same pH. Substitution of the bulky WT isoleucine residue by alanine creates a cavity that severely destabilizes the folded structure of the mutant. The reduced stability of Ile76Ala CI2 compared to WT is evidenced in its experimental unfolding energetics (ΔH_{unf} = 30.2, ΔCp_{unf} = 0.7, and ΔG^0_{unf} = 1.1 ± 0.3, Table 1), which is accurately obtained from our simulations (27.7 ± 1.7, 0.5 ± 0.01, and 1.0 ± 0.2, respectively, Table 2). Thus, the simulation workflow allows capture of the experimental observations that 1) WT CI2 is stabilized by raising the pH from 3.0 to 6.3 (experimental $\Delta\Delta G_{unf(pH3\rightarrow pH6.3)}$ = +1.8 ± 1.1; calculated value = +2.5 ± 1.2) and 2) WT CI2 is severely destabilized by replacing Ile76 by Ala (experimental $\Delta\Delta G^0_{unf(WT\rightarrow I76A)}$ = −4.3 ± 1.0; calculated value = −3.3 ± 0.6). Experimental and calculated stability curves, thermograms, and state fractions of WT (pH 3.0), WT (pH 6.3), and Ile76Ala CI2 mutant (pH 3.0) are compared in Figure S2b-h. A good agreement between calculated and experimental data can be observed, which is particularly remarkable for the Ile76Ala CI2 variant (Figure S2g-h).

The thermal stability of WT lysozyme and many variants thereof have been reported.[73−76] Lysozyme has been simulated here (Figure S3a) at pH 2.4 (WT and Ile3Glu mutant) and at pH 3.0 and 3.7 (pseudo-WT; Figure S4a). The experimental ΔCp_{unf} is accurately calculated for the pseudo-WT but underestimated for the WT. For the four simulated lysozyme variants or pH conditions (Table 1), the calculated ΔH_{unf} values (Table 2) clearly overestimate the corresponding experimental ones (Table 1). As a consequence, the stabilities calculated also overestimate the experimental values, and the stability temperature dependencies (Figures S3b-f and S4b-f) do not match the calculated ones. Thus, the actual lysozyme stabilities are not correctly calculated. Possible reasons for this are indicated in the Discussion section. Still, both the lower stability of the Ile3Glu mutant relative to WT at pH 2.4 ($\Delta\Delta G^0_{unf(WT\rightarrow Ile3Glu)}$ = −1.0 ± 1.4) and the higher stability of pseudo-WT at pH 3.7 compared to pH 3.0 ($\Delta\Delta G_{unf(pH3.0\rightarrow pH3.7)}$ = +3.2 ± 2.6) are qualitatively captured (−2.8 ± 1.2 and +6.3 ± 2.7, respectively).

3.2. Energetics of a Three-State Protein: apoFld.

ApoFld thermal unfolding equilibrium is three-state, with a well-defined intermediate accumulating at equilibrium with the folded and unfolded conformations. For this protein, the unfolding enthalpy changes of the sequential partial unfolding equilibria (F-to-I and I-to-U) have been separately calculated using the general workflow (Figure 1). Structures or ensembles (see Methods) representing the three states involved in the transitions have been simulated (Figure 3a). The results show that the calculated enthalpy changes of the two unfolding transitions, $\Delta H_{unf(F\text{-}to\text{-}I)}$ = 35.6 ± 6.0 and $\Delta H_{unf(I\text{-}to\text{-}U)}$ = 48.1 ± 4.1 (Table 2), are in excellent agreement with the corresponding experimental enthalpies of 32.0 ± 1.1 and 55.6 ± 2.0 (Table 1). The heat capacity changes calculated for each partial unfolding step, $\Delta Cp_{unf(F\text{-}to\text{-}I)}$ = 1.5 ± 0.1 and $\Delta Cp_{unf(I\text{-}to\text{-}U)}$ = 1.0 ± 0.0, respectively (2.5 ± 0.1 for the global transition, Table 2), are also in fair agreement with the experimental values of 1.35 ± 0.3 and 1.55 ± 0.3, respectively (2.9 ± 0.6 for the global transition, Table 1). From these calculated data and the corresponding experimental T_ms (Table 1), the Gibbs free-energy changes of the individual apoFld unfolding transitions are calculated at 25.0 °C using the Gibbs−Helmholtz equation[26] (eq 1), and the global apoFld stability is then obtained as the sum of the individual free-energy changes. A fine correspondence between the calculated stability values, $\Delta G^0_{unf(F\text{-}to\text{-}I)}$ = 1.3 ± 1.7, $\Delta G^0_{unf(I\text{-}to\text{-}U)}$ = 3.0 ± 0.9, and $\Delta G^0_{unf(F\text{-}to\text{-}U)}$ = 4.3 ± 2.6 (Table 2), and the corresponding experimental ones, 1.1 ± 1.4, 2.9 ± 1.3, and 4.0 ± 2.7, is observed. The outstanding correspondence between calculated and experimentally determined apoFld thermal unfolding thermodynamics is also observed in the compared stability curves, thermograms, and folded/intermediate/unfolded state fractions depicted in Figure 3b-d.

An otherwise identical calculation of apoFld thermal unfolding thermodynamics has been carried out using the Amber99SB-ILDN force field instead of Charmm22-CMAP. Although accurate heat capacity changes have been calculated with Amber99SB-ILDN for the two equilibria (1.4 ± 0.1 and 1.1 ± 0.1, respectively, Table 2), the calculated enthalpy changes (Table 2) do not agree well with the experimental values (Table 1), which results in less accurate calculations of the individual Gibbs free-energy changes (Table 2) compared to those obtained with Charmm22-CMAP. For barnase and nuclease, the better agreement of Charmm22-CMAP thermodynamics calculations with experimental values compared to calculations with Amber99SB-ILDN was already reported.[24]

3.3. Energetics of a Holoprotein: holoFld.

The calculation of the thermal unfolding energetics of a holoprotein (a protein carrying a noncovalently bound cofactor) has been performed as described in Methods and illustrated in Figure 4a. To model holoFld energetics, three different FMN parametrizations have been tested (see Methods). ΔH_{unf} calculated for holoFld with any of them (ranging from 103.0 ± 6.5 to 114.2 ± 7.8, Table 2) is in fair agreement with the experimental value reported by Lamazares and co-workers[84] from DSC measurements (101.9 ± 0.6, Table 1).

holoFld ΔCp_{unf} has not been reported, but an estimation can be done by adding the reported value for FMN dissociation ($\Delta Cp_{diss} = -\Delta Cp_{bind}$ = 0.6 ± 0.0)[80] to the apoFld ΔCp_{unf} (2.9 ± 0.6, Table 1). Thus, the holoFld ΔCp_{unf} is estimated to be 3.5 ± 0.6. Our calculated holoFld ΔCp_{unf} values (reported in Table 2 and depicted as the slope of fitting lines in Figure 4b) indicate that ΔCp_{unf} obtained with either FMN Par.-1 or FMN Par.-2 (3.0 ± 0.2 and 2.9 ± 0.6, respectively) agrees within experimental error, and that obtained with FMN Par.-3 (2.6 ± 0.1) while lower is still above the value previously calculated for apoFld (2.5 ± 0.1, Table 2), in agreement with the observed positive value of ΔCp_{diss}.

The stability of holoFld at 25.0 °C is obtained through SI eq 5 (see derivation in SI Methods). To the apoprotein Gibbs free-energy, SI eq 5 applies a correction due to the ligand concentration and incorporates the van't Hoff approximation[53] to account for the temperature dependence of the binding constant. Thus, SI eq 5 is not based on the thermodynamics derived from the holoFld simulations but on those of the apoprotein ($\Delta H_{apo(unf)}$, $\Delta Cp_{apo(unf)}$) plus the cofactor energetics. Using SI eq 5, the ΔG^0_{unf} value calculated (17.3 ± 2.6, Table 2) is in close agreement with the experimental value (17.1 ± 2.7, Table 1) similarly obtained with SI eq 5 using experimental $\Delta H_{apo(unf)}$ and $\Delta Cp_{apo(unf)}$ data. Importantly, the calculated ΔG^0_{unf} also matches, within error, the experimental stability of holoFld directly obtained from thermal unfolding curves (19.0 ± 0.9).[54]

https://doi.org/10.1021/acs.jcim.3c01107
J. Chem. Inf. Model. 2023, 63, 7791−7806

4. DISCUSSION

The devised MD simulation workflow allows for the calculation of ΔH_{unf}, ΔCp_{unf}, and ΔG_{unf}, i.e., three of the main thermodynamic magnitudes governing the stability of proteins. The overall accuracy of the method can be assessed from lineal plots of calculated versus experimentally determined values of each of those magnitudes.

The primary figure calculated is the unfolding enthalpy change (ΔH_{unf}) of the proteins investigated. With the exception of lysozyme (simulated in four conditions) and nuclease (when simulated at low pH, pH 4.1), which are clear outliers, the linear plot (Figure 5a) can be fitted to a straight line with an ordinate close to zero (−2.8), slope close to unity (0.95), and a correlation of $R^2 = 0.93$. The fitting includes the data from ten simulated systems (barnase, nuclease at two pH values, and two partial unfolding equilibria, as well as the whole transition of three-state apoFld, CI2 at two pH values plus one mutant, and holoFld) spanning a range of ΔH_{unf} values from 30 to 120 kcal/mol. It is thus clear that ΔH_{unf} can be accurately calculated by using this approach.

The second figure is the unfolding heat capacity change (ΔCp_{unf}), which is also captured for the 10 protein systems well fitted in Figure 5a. The four lysozyme systems simulated (WT, a variant of WT, and a pseudo-WT variant at two pHs), as well as nuclease at pH 4.1, fit worse than the other 10 systems (Figure 5b). Albeit their calculated ΔCpunf values do not differ too much from their experimental ones, they have been treated as outliers for consistency. The linear fit with data from the other 10 simulated systems yields a straight line with an ordinate close to zero (−0.15), slope close to unity (0.85), and a correlation of $R^2 = 0.94$, indicating that the change in heat capacity of unfolding can be also calculated in an accurate manner. The range of ΔCp_{unf} values spanned in the plot goes from 0.6 to 3.5 kcal/mol·K.

The third figure is the unfolding Gibbs free-energy change (ΔG_{unf}), i.e., the conformational stability of the protein. To derive it, the workflow combines the calculated enthalpy and heat capacity changes with experimental values of melting temperatures, using the Gibbs−Helmholtz equation (eq 1) for apoproteins, or an analogous equation (SI eq 5) for holoproteins. As expected, in the linear plot of calculated versus experimentally determined stabilities (Figure 5c) lysozyme yields outliers, as the high enthalpy changes calculated for this protein system are carried over in the calculation of the stability. Although nuclease at pH 4.1 is not a clear outlier in the stability representation, it has been kept as such for consistency. The fitting of the calculated and experimental values for the other 10 systems simulated gives rise once again to a straight line with close to zero intercept (0.10), close to unity slope (0.99), and a high correlation of $R^2 = 0.99$. It seems thus that protein conformational stability can be accurately calculated from first-principles using the described simulation workflow. The range of Gibbs free-energies spanned in the plot goes from 1 to 17 kcal/mol.

The MD simulation workflow accurately calculates the protein changes in enthalpy, heat capacity, and Gibbs free-energy upon unfolding and can also be used to compare the stability of a protein under different pH values or to compare the stability of a wild-type protein with that of its mutants. According to our literature search, no similar approach for the calculation of protein folding energetics has been described, which precludes a direct comparison of our approach with

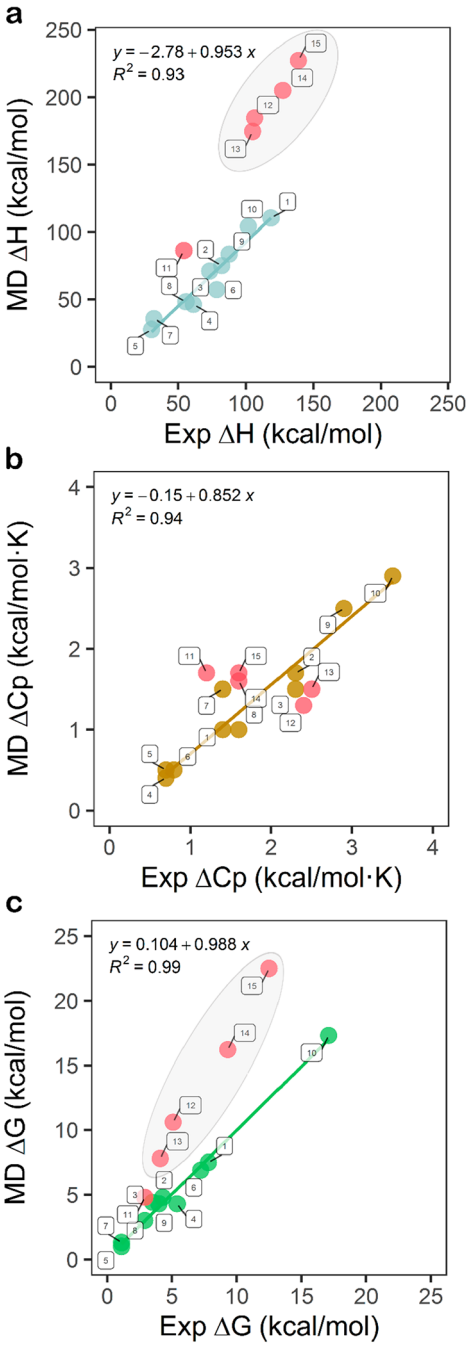


Figure 5. Global assessment of the approach for calculation of unfolding thermodynamics with Charmm22-CMAP/Tip3p. a) Scatter plot of MD-calculated vs experimental ΔH_{unf} for the set of proteins simulated (including different solvating conditions and variants). The linear fit shown in this panel (also in panels b and c) was performed over the following ten systems: barnase at pH ∼ 4.1 (dot number 1 in legend), nuclease at pH 7.0 (2) and pH 5.0 (3), WT CI2 at pH 3.0 (4), Ile76Ala CI2 at pH 3.0 (5), WT CI2 at pH 6.3 (6), apoFld(F-to-I) (7), apoFld(I-to-U) (8), apoFld(F-to-U) (9), and holoFld(FMN Par.-2) (10). The fitting equation and the square Pearson correlation coefficient are given. b) Scatter plot and linear fit of MD-calculated vs experimental ΔCp_{unf}. c) Scatter plot and linear fitting of MD-calculated vs experimental protein stability (ΔG^0_{unf} at 298.15 K for all proteins except for nuclease that is compared at 293.15 K). Experimental values (*x*-axis) are the averages (or individual value in some cases) of data obtained from the literature, as summarized in Table 1, while calculated values are those presented in Table 2. Red

https://doi.org/10.1021/acs.jcim.3c01107
J. Chem. Inf. Model. 2023, 63, 7791−7806

Figure 5. continued

circles represent outliers (or cases treated as such, see the Results and the Discussion sections) not considered in the linear fitting, namely the following: nuclease at pH 4.1 (dot number 11 in legend), WT lysozyme at pH 2.4 (12), Ile3Glu lysozyme at pH 2.4 (13), pseudo-WT lysozyme at pH 3.0 (14), and pseudo-WT lysozyme at pH 3.7 (15). In panels a and c, the 4 outliers of lysozyme and pseudolysozyme systems are enclosed in a semitransparent gray oval to visualize them as similar systems whose enthalpy change upon unfolding (ΔH_{unf}) and protein stability (ΔG_{unf}) are all overestimated by our simulations. Out of the three setups tested for holoFld, the results obtained with FMN parametrization 2 (the most accurate one, see Tables 1 and 2) are depicted.

other methods. The systems successfully calculated here contain representatives of the main protein classes (mainly alpha, mainly beta, and alpha beta),[109] with sequences ranging from 84 to 169 residues, and isoelectric points from 4.0 to 8.9. They include proteins that undergo two- or three-state thermal unfolding as well as proteins that do or do not carry a tightly bound cofactor. Altogether, these proteins offer a fair representation of natively folded proteins, for which the unfolding process leads to fully unfolded conformations. Detailed thermodynamic studies on much larger proteins are scarce, and the approach has not been tested on large proteins. We foresee no reasons why the energetics of larger proteins cannot be calculated with similar accuracy using sufficient sampling, provided that they adopt fully unfolded conformations after heating. Full unfolding of the denatured state is a requisite, as it is necessary to be able to build realistic models of the unfolded ensemble using ProtSA.[25]

For one of the proteins simulated, lysozyme, the calculations have consistently led to overestimated ΔH_{unf} values, which has translated to overestimated stability. In principle, the method could have failed for this protein due to insufficient quality of the models used to represent its folded and unfolded conformations. This is unlikely, however, as the folded structures have been solved in a highly experience lab,[110] and they get good marks (not shown) when subjected to quality control with the MolProbity server.[111] On the other hand, the model of the unfolded ensemble generated by ProtSA[25] would be wrong if the lysozyme unfolded state were compact, but we have found no reports pointing to that. A different possible reason for the inaccurate lysozyme calculation may be small inaccuracies in force field parameters. Although the same force field has been used in lysozyme and in the successfully calculated proteins, it should be noticed that force field parameters are globally optimized, and optimal individual performance from each parameter cannot be taken for granted. In this respect, of all the systems simulated here, lysozyme stands out as the one containing the highest net (positive) charge (Table S2), only paralleled by the high net (positive) charge of nuclease under the simulation condition of pH 4.1, where inaccurate results have also been obtained. It is thus possible that the discrepancy between calculated and experimental lysozyme unfolding magnitudes is related to insufficient tuning of Coulombic treatment by the Charmm22-CMAP force field[15] for lysine and arginine protonated side chains. Alternatively, or in addition to this, some uncertainty in the protonation state of lysozyme carboxyl groups at the acidic pH of the simulations could contribute to inaccuracy. Whatever the reason, the poorer performance of the method on lysozyme suggests that it should be used with caution when highly positively charged proteins are simulated at acidic pH values. As proteins are rarely studied experimentally under basic pH conditions, we have not tested the performance of the method at high pH values.

Although the described approach is based on a specific force field and water model, it suggests that current force fields are already close to capturing the complexity of the protein folding energetics. We hope that our results will encourage further improvement of the force fields and water models. Toward that goal, the described methodology constitutes an effective and efficient way to assess the ability of a given force field to replicate the changes in energy that govern protein equilibria.

5. CONCLUSIONS

The energetics (folding ΔH and ΔCp) of two- and three-state proteins (with or without bound cofactors) can be accurately computed using conventional force fields and water models by sampling the unfolded ensemble energy with many short MD simulations of conformationally diverse starting structures. If the melting temperature of the simulated protein is known, the stability curve providing the value of ΔG as a function of the temperature can also be obtained. Besides, smaller stability differences ($\Delta\Delta G$) due to differences in solution conditions (e.g., differences in pH value) or caused by point mutations can be semiquantitatively obtained. However, the combination of force field and water model used here (which is nevertheless better than other combinations based on force fields specifically tuned to avoid overcompaction) overestimates ΔH in the case of highly charged proteins if they are simulated at low pH. We propose that the thermodynamic approach described here for calculating protein energetics from MD simulations can be of help to force field developers to fine-tune force fields and water models, which, until now, have paid great attention to reproducing geometric and dynamical features of proteins but little attention to reproducing the energy changes governing protein equilibria.

■ ASSOCIATED CONTENT

Data Availability Statement

The files used/necessary for the calculations done in this work using Molecular Dynamics simulations can be downloaded from https://zenodo.org/record/8165111.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.3c01107.

Additional methods, MD simulations setup, and experimental details. Data of calculated thermodynamics for additional protein systems simulated compared with their experimental values, including stability curves, thermograms, and protein molar fractions plots (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Javier Sancho − *Department of Biochemistry, Molecular and Cell Biology, Faculty of Science, University of Zaragoza, 50009 Zaragoza, Spain; Biocomputation and Complex Systems Physics Institute (BIFI), Joint Unit GBs-CSIC, University of Zaragoza, 50018 Zaragoza, Spain; Aragon Health Research Institute (IIS Aragón), 50009 Zaragoza, Spain;* orcid.org/0000-0002-2879-9200; Email: jsancho@unizar.es

https://doi.org/10.1021/acs.jcim.3c01107
J. Chem. Inf. Model. 2023, 63, 7791−7806

Authors

Juan J. Galano-Frutos − *Department of Biochemistry, Molecular and Cell Biology, Faculty of Science, University of Zaragoza, 50009 Zaragoza, Spain; Biocomputation and Complex Systems Physics Institute (BIFI), Joint Unit GBs-CSIC, University of Zaragoza, 50018 Zaragoza, Spain;* orcid.org/0000-0002-1896-7805

Francho Nerín-Fonz − *Department of Biochemistry, Molecular and Cell Biology, Faculty of Science, University of Zaragoza, 50009 Zaragoza, Spain;* Present Address: Biomedical Research Foundation, Academy of Athens, 11527 Athens, Greece; orcid.org/0000-0002-5548-7608

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c01107

Author Contributions

J.S. conceived and directed the investigation. J.J.G-F. and F.N.-F. carried out and analyzed the Molecular Dynamics simulations. J.J.G-F. and J.S. analyzed data and wrote the manuscript. All authors have given approval to the final version of the manuscript.

Funding

This work was supported by grants PID2019-107293GB-I00, PID2022-141068NB-I00, and PDC2021-121341-I00 (MICINN, Spain) and E45_20R (Gobierno de Aragón, Spain).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank the Biocomputation and Complex Systems Physics Institute (BIFI) of the University of Zaragoza and the Red Española de Supercomputación (RES) for computing facilities granted to perform Molecular Dynamics simulations. We thank Ritwik Maity for help with the cover figure.

■ ABBREVIATIONS

AI, Artificial Intelligence; DSC, Differential Scanning Calorimetry; FMN, Flavin Mononucleotide; IS, Ionic Strength; LEM, Linear Extrapolation Method; MD, Molecular Dynamics; NMR, Nuclear Magnetic Resonance; SAXS, Small-Angle X-ray Scattering; NPT, Isothermal−isobaric ensemble in MD simulations; NVT, Canonical ensemble in MD simulations; PBC, Periodic Boundary Conditions; PME, Particle Mesh Ewald; Rg, Radius of gyration; SE, Standard Error

■ REFERENCES

(1) Berg, J. M.; Tymoczko, J. L.; Stryer, L. *Protein Structure and Function*; W. H. Freeman: 2002.

(2) Dill, K. A.; MacCallum, J. L. The Protein-Folding Problem, 50 Years On. *Science.* **2012**, *338* (6110), 1042−1046.

(3) Goldenzweig, A.; Fleishman, S. J. Principles of Protein Stability and Their Application in Computational Design. *Annu. Rev. Biochem.* **2018**, *87* (1), 105−129.

(4) Lindorff-Larsen, K.; Trbovic, N.; Maragakis, P.; Piana, S.; Shaw, D. E. Structure and Dynamics of an Unfolded Protein Examined by Molecular Dynamics Simulation. *J. Am. Chem. Soc.* **2012**, *134* (8), 3787−3791.

(5) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How Fast-Folding Proteins Fold. *Science.* **2011**, *334* (6055), 517−520.

(6) Sedov, I. A.; Magsumov, T. I. Molecular Dynamics Study of Unfolding of Lysozyme in Water and Its Mixtures with Dimethyl Sulfoxide. *J. Mol. Graph. Model.* **2017**, *76*, 466−474.

(7) Gsponer, J.; Caflisch, A. Molecular Dynamics Simulations of Protein Folding from the Transition State. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (10), 6719−6724.

(8) Jiang, F.; Wu, Y. D. Folding of Fourteen Small Proteins with a Residue-Specific Force Field and Replica-Exchange Molecular Dynamics. *J. Am. Chem. Soc.* **2014**, *136* (27), 9536−9539.

(9) Daggett, V.; Levitt, M. Protein Unfolding Pathways Explored through Molecular Dynamics Simulations. *J. Mol. Biol.* **1993**, *232* (2), 600−619.

(10) Miao, Y.; Feixas, F.; Eun, C.; McCammon, J. A. Accelerated Molecular Dynamics Simulations of Protein Folding. *J. Comput. Chem.* **2015**, *36* (20), 1536−1549.

(11) Lei, H.; Wu, C.; Liu, H.; Duan, Y. Folding Free-Energy Landscape of Villin Headpiece Subdomain from Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (12), 4925−4930.

(12) Best, R. B. Atomistic Molecular Simulations of Protein Folding. *Curr. Opin. Struct. Biol.* **2012**, *22* (1), 52−61.

(13) Freddolino, P. L.; Harrison, C. B.; Liu, Y.; Schulten, K. Challenges in Protein-Folding Simulations. *Nat. Phys.* **2010**, *6* (10), 751−758.

(14) Robustelli, P.; Piana, S.; Shaw, D. E. Developing a Molecular Dynamics Force Field for Both Folded and Disordered Protein States. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (21), E4758−E4766.

(15) Mackerell, A. D.; Feig, M.; Brooks, C. L. Extending the Treatment of Backbone Energetics in Protein Force Fields: Limitations of Gas-Phase Quantum Mechanics in Reproducing Protein Conformational Distributions in Molecular Dynamics Simulation. *J. Comput. Chem.* **2004**, *25* (11), 1400−1415.

(16) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber Ff99SB Protein Force Field. *Proteins Struct. Funct. Bioinforma.* **2010**, *78* (8), 1950−1958.

(17) Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Systematic Validation of Protein Force Fields against Experimental Data. *PLoS One* **2012**, *7* (2), e32131.

(18) Ben-Naim, A. The Rise and Fall of the Hydrophobic Effect in Protein Folding and Protein-Protein Association, and Molecular Recognition. *Open J. Biophys.* **2011**, *01* (01), 1−7.

(19) Baker, D. What Has de Novo Protein Design Taught Us about Protein Folding and Biophysics? *Protein Sci.* **2019**, *28* (4), 678.

(20) Bender, B. J.; Gahbauer, S.; Luttens, A.; Lyu, J.; Webb, C. M.; Stein, R. M.; Fink, E. A.; Balius, T. E.; Carlsson, J.; Irwin, J. J.; Shoichet, B. K. A Practical Guide to Large-Scale Docking. *Nat. Protoc.* **2021**, *16* (10), 4799.

(21) Galano-Frutos, J. J.; Garciá-Cebollada, H.; Sancho, J. Molecular Dynamics Simulations for Genetic Interpretation in Protein Coding Regions: Where We Are, Where to Go and When. *Brief. Bioinform.* **2021**, *22* (1), 3−19.

(22) Castro, K. M.; Scheck, A.; Xiao, S.; Correia, B. E. Computational Design of Vaccine Immunogens. *Curr. Opin. Biotechnol.* **2022**, *78*, No. 102821.

(23) Zielinski, D. C.; Patel, A.; Palsson, B. O. The Expanding Computational Toolbox for Engineering Microbial Phenotypes at the Genome Scale. *Microorganisms* **2020**, *8* (12), 2050.

(24) Galano-Frutos, J. J.; Sancho, J. Accurate Calculation of Barnase and SNase Folding Energetics Using Short Molecular Dynamics Simulations and an Atomistic Model of the Unfolded Ensemble: Evaluation of Force Fields and Water Models. *J. Chem. Inf. Model.* **2019**, *59* (10), 4350−4360.

(25) Estrada, J.; Bernadó, P.; Blackledge, M.; Sancho, J. ProtSA: A Web Application for Calculating Sequence Specific Protein Solvent Accessibilities in the Unfolded Ensemble. *BMC Bioinformatics* **2009**, *10* (1), 104.

(26) Becktel, W. J.; Schellman, J. A. Protein Stability Curves. *Biopolymers* **1987**, *26* (11), 1859−1877.

(27) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926−935.

(28) Piana, S.; Robustelli, P.; Tan, D.; Chen, S.; Shaw, D. E. Development of a Force Field for the Simulation of Single-Chain Proteins and Protein-Protein Complexes. *J. Chem. Theory Comput.* **2020**, *16* (4), 2494−2507.
(29) Mahoney, M. W.; Jorgensen, W. L. A Five-Site Model for Liquid Water and the Reproduction of the Density Anomaly by Rigid, Nonpolarizable Potential Functions. *J. Chem. Phys.* **2000**, *112* (20), 8910−8922.
(30) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. *Interaction Models for Water in Relation to Protein Hydration*; Springer: Dordrecht, 1981; pp 331−342.
(31) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The Missing Term in Effective Pair Potentials. *J. Phys. Chem.* **1987**, *91* (24), 6269−6271.
(32) Deller, M. C.; Kong, L.; Rupp, B. Protein Stability: A Crystallographer's Perspective. *Acta Crystallogr. F Struct. Biol. Commun.* **2016**, *72*, 72−95.
(33) Hartley, R. W.; Barker, E. A. Amino-Acid Sequence of Extracellular Ribonuclease (Barnase) of Bacillus Amyloliquefaciens. *Nat. New Biol.* **1972**, *235* (53), 15−16.
(34) Paddon, C. J.; Hartley, R. W. Cloning, Sequencing and Transcription of an Inactivated Copy of Bacillus Amyloliquefaciens Extracellular Ribonuclease (Barnase). *Gene* **1985**, *40* (2−3), 231−239.
(35) Cone, J. L.; Cusumano, C. L.; Taniuchi, H.; Anfinsen, C. B. Staphylococcal Nuclease (Foggi Strain). II. The Amino Acid Sequence. *J. Biol. Chem.* **1971**, *246* (10), 3103−3110.
(36) Davis, A.; Moore, I. B.; Parker, D. S.; Taniuchi, H. Nuclease B: A Possible Precursor of Nuclease A, an Extracellular Nuclease of Staphylococcus Aureus. *J. Biol. Chem.* **1977**, *252* (18), 6544−6553.
(37) Svendsen, I.; Martin, B.; Jonassen, I. Characteristics of Hiproly Barley III. Amino Acid Sequences of Two Lysine-Rich Proteins. *Carlsberg Res. Commun.* **1980**, *45* (2), 79−85.
(38) Williamson, M. S.; Forde, J.; Buxton, B.; Kreis, M. Nucleotide Sequence of Barley Chymotrypsin Inhibitor-2 (CI-2) and Its Expression in Normal and High-lysine Barley. *Eur. J. Biochem.* **1987**, *165* (1), 99−106.
(39) Tsugita, A.; Inouye, M.; et al. Purification of Bacteriophage T4 Lysozyme. *J. Biol. Chem.* **1968**, *243* (2), 391−397.
(40) Fillat, M. F.; Edmondson, D. E.; Gomez-Moreno, C. Structural and Chemical Properties of a Flavodoxin from Anabaena PCC 7119. *Biochim. Biophys. Acta (BBA)/Protein Struct. Mol.* **1990**, *1040* (2), 301−307.
(41) Irún, M. P.; Maldonado, S.; Sancho, J. Stabilization of Apoflavodoxin by Replacing Hydrogen-Bonded Charged Asp or Glu Residues by the Neutral Isosteric Asn or Gln. *Protein Eng.* **2001**, *14* (3), 173−181.
(42) Irún, M. P.; Garcia-Mira, M. M.; Sanchez-Ruiz, J. M.; Sancho, J. Native Hydrogen Bonds in a Molten Globule: The Apoflavodoxin Thermal Intermediate. *J. Mol. Biol.* **2001**, *306* (4), 877−888.
(43) Lamazares, E.; Clemente, I.; Bueno, M.; Velázquez-Campoy, A.; Sancho, J. Rational Stabilization of Complex Proteins: A Divide and Combine Approach. *Sci. Rep.* **2015**, *5* (1), 9129.
(44) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nat. 2021 5967873* **2021**, *596* (7873), 583−589.
(45) Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; Žídek, A.; Green, T.; Tunyasuvunakool, K.; Petersen, S.; Jumper, J.; Clancy, E.; Green, R.; Vora, A.; Lutfi, M.; Figurnov, M.; Cowie, A.; Hobbs, N.; Kohli, P.; Kleywegt, G.; Birney, E.; Hassabis, D.; Velankar, S. AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models. *Nucleic Acids Res.* **2022**, *50*, D439.
(46) Burley, S. K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chen, L.; Crichlow, G. V.; Christie, C. H.; Dalenberg, K.; Di Costanzo, L.; Duarte, J. M.; Dutta, S.; Feng, Z.; Ganesan, S.; Goodsell, D. S.; Ghosh, S.; Green, R. K.; Guranovic, V.; Guzenko, D.; Hudson, B. P.; Lawson, C. L.; Liang, Y.; Lowe, R.; Namkoong, H.; Peisach, E.; Persikova, I.; Randle, C.; Rose, A.; Rose, Y.; Sali, A.; Segura, J.; Sekharan, M.; Shao, C.; Tao, Y. P.; Voigt, M.; Westbrook, J. D.; Young, J. Y.; Zardecki, C.; Zhuravleva, M. RCSB Protein Data Bank: Powerful New Tools for Exploring 3D Structures of Biological Macromolecules for Basic and Applied Research and Education in Fundamental Biology, Biomedicine, Biotechnology, Bioengineering and Energy Sciences. *Nucleic Acids Res.* **2021**, *49* (D1), D437−D451.
(47) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−242.
(48) Ozenne, V.; Bauer, F.; Salmon, L.; Huang, J. R.; Jensen, M. R.; Segard, S.; Bernadó, P.; Charavay, C.; Blackledge, M. Flexible-Meccano: A Tool for the Generation of Explicit Ensemble Descriptions of Intrinsically Disordered Proteins and Their Associated Experimental Observables. *Bioinformatics* **2012**, *28* (11), 1463−1470.
(49) Eyal, E.; Najmanovich, R.; Mcconkey, B. J.; Edelman, M.; Sobolev, V. Importance of Solvent Accessibility and Contact Surfaces in Modeling Side-Chain Conformations in Proteins. *J. Comput. Chem.* **2004**, *25* (5), 712−724.
(50) Ayuso-Tejedor, S.; Angarica, V. E.; Bueno, M.; Campos, L. A.; Abián, O.; Bernadó, P.; Sancho, J.; Jiménez, M. A. Design and Structure of an Equilibrium Protein Folding Intermediate: A Hint into Dynamical Regions of Proteins. *J. Mol. Biol.* **2010**, *400* (4), 922−934.
(51) Makhatadze, G. I.; Kim, K. -S; Woodward, C.; Privalov, P. L. Thermodynamics of Bpti Folding. *Protein Sci.* **1993**, *2* (12), 2028−2036.
(52) Privalov, P. L.; Makhatadze, G. I. Heat Capacity of Proteins. II. Partial Molar Heat Capacity of the Unfolded Polypeptide Chain of Proteins: Protein Unfolding Effects. *J. Mol. Biol.* **1990**, *213* (2), 385−391.
(53) Fukada, H.; Sturtevant, J. M.; Quiocho, F. A. Thermodynamics of the Binding of L-Arabinose and of D-Galactose to the L-Arabinose-Binding Protein of Escherichia Coli. *J. Biol. Chem.* **1983**, *258* (21), 13193−13198.
(54) Campos, L. A.; Sancho, J. Native-Specific Stabilization of Flavodoxin by the FMN Cofactor: Structural and Thermodynamical Explanation. *Proteins Struct. Funct. Genet.* **2006**, *63* (3), 581−594.
(55) Serrano, L.; Kellis, J. T.; Cann, P.; Matouschek, A.; Fersht, A. R. The Folding of an Enzyme. II. Substructure of Barnase and the Contribution of Different Interactions to Protein Stability. *J. Mol. Biol.* **1992**, *224* (3), 783−804.
(56) Serrano, L.; Matouschek, A.; Fersht, A. R. The Folding of an Enzyme. VI. The Folding Pathway of Barnase: Comparison with Theoretical Models. *J. Mol. Biol.* **1992**, *224* (3), 847−859.
(57) Oliveberg, M.; Vuilleumier, S.; Fersht, A. R. Thermodynamic Study of the Acid Denaturation of Barnase and Its Dependence on Ionic Strength: Evidence for Residual Electrostatic Interactions in the Acid/ Thermally Denatured State. *Biochemistry* **1994**, *33* (29), 8826−8832.
(58) Martin, C.; Richard, V.; Salem, M.; Hartley, R.; Mauguen, Y. Refinement and Structural Analysis of Barnase at 1.5 Å Resolution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **1999**, *55* (2), 386−398.
(59) Griko, Y. V.; Makhatadze, G. I.; Privalov, P. L.; Hartley, R. W. Thermodynamics of Barnase Unfolding. *Protein Sci.* **1994**, *3* (4), 669−676.
(60) Carra, J. H.; Anderson, E. A.; Privalov, P. L. Thermodynamics of Staphylococcal Nuclease Denaturation. I. The Acid-denatured State. *Protein Sci.* **1994**, *3* (6), 944−951.
(61) 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* (13), 4761−4768.

https://doi.org/10.1021/acs.jcim.3c01107
J. Chem. Inf. Model. 2023, 63, 7791−7806

(62) Shortle, D.; Meeker, A. K. Mutant Forms of Staphylococcal Nuclease with Altered Patterns of Guanidine Hydrochloride and Urea Denaturation. *Proteins Struct. Funct. Bioinforma.* **1986**, *1* (1), 81−89.
(63) Eftink, M. R.; Ghiron, C. A.; Kautz, R. A.; Fox, R. O. Fluorescence and Conformational Stability Studies of Staphylococcus Nuclease and Its Mutants, Including the Less Stable Nuclease-Concanavalin a Hybrids. *Biochemistry* **1991**, *30* (5), 1193−1199.
(64) Jackson, S. E.; Fersht, A. R. Folding of Chymotrypsin Inhibitor 2.1. Evidence for a Two-State Transition. *Biochemistry* **1991**, *30* (43), 10428−10435.
(65) Jackson, S. E.; Moracci, M.; ElMasry, N.; Johnson, C. M.; Fersht, A. R. Effect of Cavity-Creating Mutations in the Hydrophobic Core of Chymotrypsin Inhibitor 2. *Biochemistry* **1993**, *32* (42), 11259−11269.
(66) Tan, Y. J.; Oliveberg, M.; Davis, B.; Fersht, A. R. Perturbed PKA-Values in the Denatured States of Proteins. *J. Mol. Biol.* **1995**, *254* (5), 980−992.
(67) Kjær, M.; Ludvigsen, S.; Sørensen, O. W.; Denys, L. A.; Kindtler, J.; Poulsen, F. M. Sequence Specific Assignment of the Proton Nuclear Magnetic Resonance Spectrum of Barley Serine Proteinase Inhibitor 2. *Carlsberg Res. Commun.* **1987**, *52* (5), 327−354.
(68) Kjær, M.; Poulsen, F. M. Secondary Structure of Barley Serine Proteinase Inhibitor 2 Determined by Proton Nuclear Magnetic Resonance Spectroscopy. *Carlsberg Res. Commun.* **1987**, *52* (5), 355−362.
(69) Klemm, J. D.; Wozniak, J. A.; Alber, T.; Goldenberg, D. P. Correlation between Mutational Destabilization of Phage T4 Lysozyme and Increased Unfolding Rates. *Biochemistry* **1991**, *30* (2), 589−594.
(70) Hawkes, R.; Grutter, M. G.; Schellman, J. Thermodynamic Stability and Point Mutations of Bacteriophage T4 Lysozyme. *J. Mol. Biol.* **1984**, *175* (2), 195−212.
(71) Matthews, B. W. Genetic and Structural Analysis of the Protein Stability Problem. *Biochemistry* **1987**, *26* (22), 6885−6888.
(72) Bell, J. A.; Wilson, K. P.; Zhang, X. -J; Faber, H. R.; Nicholson, H.; Matthews, B. W. Comparison of the Crystal Structure of Bacteriophage T4 Lysozyme at Low, Medium, and High Ionic Strengths. *Proteins Struct. Funct. Bioinforma.* **1991**, *10* (1), 10−21.
(73) Kitamura, S.; Sturtevant, J. M. A Scanning Calorimetric Study of the Thermal Denaturation of the Lysozyme of Phage T4 and the Arg 96 →His Mutant Form Thereof. *Biochemistry* **1989**, *28* (9), 3788−3792.
(74) Matsumura, M.; Becktel, W. J.; Matthews, B. W. Hydrophobic Stabilization in T4 Lysozyme Determined Directly by Multiple Substitutions of Ile 3. *Nature* **1988**, *334* (6181), 406−410.
(75) Ladbury, J. E.; Sturtevant, J. M.; Hu, C. Q. A Differential Scanning Calorimetric Study of the Thermal Unfolding of Mutant Forms of Phage T4 Lysozyme. *Biochemistry* **1992**, *31* (44), 10699−10702.
(76) Connelly, P.; Ghosaini, L.; Hu, C. Q.; Kitamura, S.; Tanaka, A.; Sturtevant, J. M. A Differential Scanning Calorimetric Study of the Thermal Unfolding of Seven Mutant Forms of Phage T4 Lysozyme. *Biochemistry* **1991**, *30* (7), 1887−1891.
(77) Carra, J. H.; Murphy, E. C.; Privalov, P. L. Thermodynamic Effects of Mutations on the Denaturation of T4 Lysozyme. *Biophys. J.* **1996**, *71* (4), 1994−2001.
(78) Sancho, J. Flavodoxins: Sequence, Folding, Binding, Function and Beyond. *Cell. Mol. Life Sci. C* **2006**, *63* (7), 855−864.
(79) Casaus, J. L.; Navarro, J. A.; Hervás, M.; Lostao, A.; De La Rosa, M. A.; Gómez-Moreno, C.; Sancho, J.; Medina, M. Anabaena Sp. PCC 7119 Flavodoxin as Electron Carrier from Photosystem I to Ferredoxin-NADP + Reductase. Role of Trp 57 and Tyr 94. *J. Biol. Chem.* **2002**, *277* (25), 22338−22344.
(80) Lostao, A.; El Harrous, M.; Daoudi, F.; Romero, A.; Parody-Morreale, A.; Sancho, J. Dissecting the Energetics of the Apoflavodoxin-FMN Complex. *J. Biol. Chem.* **2000**, *275* (13), 9518−9526.
(81) Lostao, A.; Daoudi, F.; Irún, M. P.; Ramón, Á.; Fernández-Cabrera, C.; Romero, A.; Sancho, J. How FMN Binds to Anabaena Apoflavodoxin: A Hydrophobic Encounter at an Open Binding Site. *J. Biol. Chem.* **2003**, *278* (26), 24053−24061.
(82) García-Fandiño, R.; Bernadó, P.; Ayuso-Tejedor, S.; Sancho, J.; Orozco, M. Defining the Nature of Thermal Intermediate in 3 State Folding Proteins: Apoflavodoxin, a Study Case. *PLoS Comput. Biol.* **2012**, *8* (8), No. e1002647.
(83) Campos, L. A.; Bueno, M.; Lopez-Llano, J.; Jiménez, M. Á.; Sancho, J. Structure of Stable Protein Folding Intermediates by Equilibrium φ-Analysis: The Apoflavodoxin Thermal Intermediate. *J. Mol. Biol.* **2004**, *344* (1), 239−255.
(84) Lamazares, E.; Vega, S.; Ferreira, P.; Medina, M.; Galano-Frutos, J. J.; Martínez-Júlvez, M.; Velázquez-Campoy, A.; Sancho, J. Direct Examination of the Relevance for Folding, Binding and Electron Transfer of a Conserved Protein Folding Intermediate. *Phys. Chem. Chem. Phys.* **2017**, *19* (29), 19021−19031.
(85) Lostao, A.; Gómez-Moreno, C.; Mayhew, S. G.; Sancho, J. Differential Stabilization of the Three FMN Redox Forms by Tyrosine 94 and Tryptophan 57 in Flavodoxin from Anabaena and Its Influence on the Redox Potentials. *Biochemistry* **1997**, *36* (47), 14334−14344.
(86) Makarov, A. A.; Protasevich, I. I.; Kuznetsova, N. V.; Fedorov, B. B.; Korolev, S. V.; Struminskaya, N. K.; Bazhulina, N. P.; Leshchinskaya, I. B.; Hartley, R. W.; Kirpichnikov, M. P.; Yakovlev, G. I.; Esipova, N. G. Comparative Study of Thermostability and Structure of Close Homologues - Bamase and Binase. *J. Biomol. Struct. Dyn.* **1993**, *10* (6), 1047.
(87) Martínez, J. C.; Filimonov, V. V.; Mateo, P. L.; Schreiber, G.; Fersht, A. R. A Calorimetric Study of the Thermal Stability of Barstar and Its Interaction with Barnase. *Biochemistry* **1995**, *34* (15), 5224−5233.
(88) Vuilleumier, S.; Fersht, A. R. Insertion in Barnase of a Loop Sequence from Ribonuclease T1: Investigating Sequence and Structure Alignments by Protein Engineering. *Eur. J. Biochem.* **1994**, *221* (3), 1003−1012.
(89) Matouschek, A.; Matthews, J. M.; Johnson, C. M.; Fersht, A. R. Extrapolation to Water of Kinetic and Equilibrium Data for the Unfolding of Barnase in Urea Solutions. *Protein Eng. Des. Sel.* **1994**, *7* (9), 1089−1095.
(90) Bueno, M.; Campos, L. A.; Estrada, J.; Sancho, J. Energetics of Aliphatic Deletions in Protein Cores. *Protein Sci.* **2006**, *15* (8), 1858−1872.
(91) Budavari, S. *The Merck Index: An Encyclopedia of Chemicals, Drugs, and Biologicals*, 12th ed.; Merck and Co., Inc.: Whitehouse Station, NJ, 1996.
(92) Legg, M. J.; Cotton, F. A.; Hazen, E. E., Jr. RCSB PDB - 2SNS: Staphylococcal Nuclease. Proposed Mechanism of Action Based on Structure of Enzyme-Thymidine 3(Prime),5(Prime)-Biphosphate-Calcium Ion Complex at 1.5-Angstroms Resolution. *Protein Data Bank*; 1982.
(93) McPhalen, C. A.; James, M. N. Crystal and Molecular Structure of the Serine Proteinase Inhibitor CI-2 from Barley Seeds. *Biochemistry* **1987**, *26* (1), 261−269.
(94) Nicholson, H.; Anderson, D. E.; Dao-pin, S.; Matthews, B. W. Analysis of the Interaction between Charged Side Chains and the α-Helix Dipole Using Designed Thermostable Mutants of Phage T4 Lysozyme. *Biochemistry* **1991**, *30* (41), 9816−9828.
(95) Genzor, C. G.; Perales-Alcón, A.; Sancho, J.; Romero, A. Closure of a Tyrosine/Tryptophan Aromatic Gate Leads to a Compact Fold in Apo Flavodoxin. *Nat. Struct. Mol. Biol.* **1996**, *3* (4), 329−332.
(96) Rao, S. T.; Shaffie, F.; Yu, C.; Satyshur, K. A.; Stockman, B. J.; Markley, J. L.; Sundaralingam, M. Structure of the Oxidized Long-chain Flavodoxin from Anabaena 7120 at 2 å Resolution. *Protein Sci.* **1992**, *1* (11), 1413−1427.
(97) Ayuso-Tejedor, S.; García-Fandiño, R.; Orozco, M.; Sancho, J.; Bernadó, P. Structural Analysis of an Equilibrium Folding Intermediate in the Apoflavodoxin Native Ensemble by Small-Angle X-Ray Scattering. *J. Mol. Biol.* **2011**, *406* (4), 604−619.

https://doi.org/10.1021/acs.jcim.3c01107
J. Chem. Inf. Model. 2023, 63, 7791−7806

(98) UniProt Consortium. UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* **2019**, *47* (D1), D506−D515.
(99) 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.
(100) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25* (13), 1605−1612.
(101) Case, D. A.; Aktulga, H. M.; Belfon, K.; Ben-Shalom, I. Y.; Berryman, J. T.; Brozell, S. R.; Cerutti, D. S.; Cheatham, T. E., III; Cisneros, G. A.; Cruzeiro, V. W. D.; Darden, T. A.; Forouzesh, N.; Giambaşu, G.; Giese, T.; Gilson, M. K.; Gohlke, H.; Goetz, A. W.; Harris, J.; Izad, S.; Izmailov, S. A.; Kasavajhala, K.; Kaymak, M. C.; King, E.; Kovalenko, A.; Kurtzman, T.; Lee, T. 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.; Sagu, C.; Schott-Verdugo, S.; Shajan, A.; Shen, J.; Simmerling, C. L.; Skrynnikov, N. R.; Smith, J.; Swails, J.; Walker, R. C.; Wang, J.; Wang, J.; Wei, H.; Wu, X.; Wu, Y.; Xiong, Y.; Xue, Y.; York, D. M.; Zhao, S.; Zhu, Q.; Kollman, P. A. *AMBER 2020*; University of California: San Francisco, 2018.
(102) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egid, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; J. A. Montgomery, J.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Keith, T.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 09*, Revision A.02; Wallingford, CT, 2016.
(103) Freddolino, P. L.; Gardner, K. H.; Schulten, K. Signaling Mechanisms of LOV Domains: New Insights from Molecular Dynamics Studies. *Photochem. Photobiol. Sci.* **2013**, *12* (7), 1158−1170.
(104) Zoete, V.; Cuendet, M. A.; Grosdidier, A.; Michielin, O. SwissParam: A Fast Force Field Generation Tool for Small Organic Molecules. *J. Comput. Chem.* **2011**, *32* (11), 2359−2368.
(105) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem.* **1993**, *97* (40), 10269−10280.
(106) Cieplak, P.; Cornell, W. D.; Bayly, C.; Kollman, P. A. Application of the Multimolecule and Multiconformational RESP Methodology to Biopolymers: Charge Derivation for DNA, RNA, and Proteins. *J. Comput. Chem.* **1995**, *16* (11), 1357−1377.
(107) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graph. Model.* **2006**, *25* (2), 247−260.
(108) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25* (9), 1157−1174.
(109) Sillitoe, I.; Bordin, N.; Dawson, N.; Waman, V. P.; Ashford, P.; Scholes, H. M.; Pang, C. S. M.; Woodridge, L.; Rauer, C.; Sen, N.; Abbasian, M.; Le Cornu, S.; Lam, S. D.; Berka, K.; Varekova, I. H.; Svobodova, R.; Lees, J.; Orengo, C. A. CATH: Increased Structural Coverage of Functional Space. *Nucleic Acids Res.* **2021**, *49* (D1), D266.
(110) Baase, W. A.; Liu, L.; Tronrud, D. E.; Matthews, B. W. Lessons from the Lysozyme of Phage T4. *Protein Sci.* **2010**, *19* (4), 631.
(111) 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, 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* (1), 293.

https://doi.org/10.1021/acs.jcim.3c01107
J. Chem. Inf. Model. 2023, 63, 7791−7806