Received: 21 June 2023 | Revised: 12 December 2023 | Accepted: 5 January 2024

DOI: 10.1002/pro.4905

RESEARCH ARTICLE

THE PROTEIN SOCIETY WILEY

Energy, water, and protein folding: A molecular dynamics-based quantitative inventory of molecular interactions and forces that make proteins stable

Juan José Galano-Frutos[1,2] | **Javier Sancho**[1,2,3]

[1]Biocomputation and Complex Systems Physics Institute (BIFI)-Joint Unit GBsC-CSIC, University of Zaragoza, Zaragoza, Spain

[2]Departamento de Bioquímica y Biología Molecular y Celular, Facultad de Ciencias, University of Zaragoza, Zaragoza, Spain

[3]Aragon Health Research Institute (IIS Aragón), Zaragoza, Spain

Correspondence
Javier Sancho, Biocomputation and Complex Systems Physics Institute (BIFI), 50018 Zaragoza, Spain
Email: jsancho@unizar.es

Funding information
Departamento de Educación, Cultura y Deporte, Gobierno de Aragón, Grant/Award Number: E45_23R; Ministerio de Ciencia e Innovación, Grant/Award Numbers: PDC2021-121341-I00, PID2019-107293GB-I00, PID2022-141068NB-I00

Review Editor: Shina Caroline Lynn Kamerlin

Abstract

Protein folding energetics can be determined experimentally on a case-by-case basis but it is not understood in sufficient detail to provide deep control in protein design. The fundamentals of protein stability have been outlined by calorimetry, protein engineering, and biophysical modeling, but these approaches still face great difficulty in elucidating the specific contributions of the intervening molecules and physical interactions. Recently, we have shown that the enthalpy and heat capacity changes associated to the protein folding reaction can be calculated within experimental error using molecular dynamics simulations of native protein structures and their corresponding unfolded ensembles. Analyzing in depth molecular dynamics simulations of four model proteins (CI2, barnase, SNase, and apoflavodoxin), we dissect here the energy contributions to ΔH (a key component of protein stability) made by the molecular players (polypeptide and solvent molecules) and physical interactions (electrostatic, van der Waals, and bonded) involved. Although the proteins analyzed differ in length, isoelectric point and fold class, their folding energetics is governed by the same quantitative pattern. Relative to the unfolded ensemble, the native conformations are enthalpically stabilized by comparable contributions from protein–protein and solvent–solvent interactions, and almost equally destabilized by interactions between protein and solvent molecules. The native protein surface seems to interact better with water than the unfolded one, but this is outweighed by the unfolded surface being larger. From the perspective of physical interactions, the native conformations are stabilized by van de Waals and Coulomb interactions and destabilized by conformational strain arising from bonded interactions. Also common to the four proteins, the sign of the heat capacity change is set by interactions between protein and solvent molecules or, from the alternative perspective, by Coulomb interactions.

KEYWORDS

Coulomb interactions, enthalpy change, heat capacity change, molecular dynamics simulations, protein folding, protein hydration, protein stability, protein thermodynamics, van der Waals interactions

1 | INTRODUCTION

Most proteins perform their biological functions once they have adopted a specific conformation from an enormity of available alternatives. Such a special conformation is known as the native state and constitutes the most stable spatial arrangement that polypeptide atoms can adopt at biologically relevant times when immersed, typically, in an aqueous solution (Anfinsen, 1973). Understanding the interactions that drive the transformation of the initially unfolded and highly hydrated polypeptide into the more stable native conformation remains a fundamental goal of Structural Biology (Chen et al., 2023; Dill & MacCallum, 2012; Moore et al., 2022). From a practical point of view, achieving a detailed, quantitative knowledge of the protein folding equilibrium could transform protein design (Huang et al., 2016) into a routine tool and exert a profound impact on a growing range of biotechnological processes, including the fabrication of biological drugs. It could also help to advance personalized medicine by putting the interpretation of genetic variants and their impact on human disease on firmer grounds (Stein et al., 2019). Heuristic approaches may offer useful alternatives to lack of quantitative knowledge, as recently illustrated in an important problem related to protein stability (Jumper et al., 2021), but there is no reason to slow down efforts toward understanding.

The stability of a folded protein is governed by the free energy difference (ΔG) of the equilibrium it establishes with a large ensemble of unfolded conformations. This free energy difference is the result of changes in the enthalpy (ΔH) and entropy (ΔS) of the system that occur when the protein folds, the temperature dependences of which are governed by the change in heat capacity (ΔCp). Both enthalpic and entropic changes need to be understood in detail, as both contribute significantly and in opposite directions to the change in free energy. Here, we will deal exclusively with dissecting the molecular and physical contributions to ΔH and ΔCp and will not attempt to dissect contributions to ΔS. Despite the long-standing efforts by the biophysical community to unravel the physics of protein stability, calculating ΔG from first principles is not yet possible. Although the experimental determination of ΔG is not usually complicated (Sancho, 2013), its interpretation in terms of specific atomic interactions is challenging (Lazaridis & Karplus, 2002; Prabhu & Sharp, 2005). Two key thermodynamic properties shaping ΔG and its temperature dependence—ΔH and ΔCp—can be measured directly by calorimetric experiments and are more amenable to interpretation (Gómez et al., 1995; Prabhu & Sharp, 2005). However, as it has been the case with purely biophysical approaches, protein engineering attempts to provide detailed knowledge of protein energetics at the molecular level have found great difficulty in assigning differences in stability, or in the other relevant thermodynamic quantities, to changes in specific physical interactions (e.g., Coulomb or van der Waals). This is because amino acid residue substitutions characteristically cause simultaneous energy changes of several types (Campos et al., 2005; Horovitz, 1996; Lazaridis & Karplus, 2002). More recently, molecular dynamics (MD) simulations have gained momentum in the study of protein energetics (Best, 2012; Bottaro & Lindorff-Larsen, 2018; Piana et al., 2012). Improvements in the accuracy of force fields and water models and sustained increases in computation power, enable better sampling of the folding equilibrium and make atomistic MD simulation ideally suited to provide an accurate dissection of protein folding energetics (Cui et al., 2021; Kamenik et al., 2020; Piana et al., 2020; Robustelli et al., 2018). Nevertheless, caution should be exercised in the choice of improved force fields as their performance on protein simulation is typically judged from geometric rather than energetic considerations (Chan-Yao-Chong et al., 2023; Huang et al., 2017; Robustelli et al., 2018; Zapletal et al., 2020).

So far, after intense scrutiny of natural proteins and engineered variants by experiment and simulation, no consensus has been reached on the relative contributions of protein–protein, protein–solvent, and solvent–solvent interactions to the enthalpy and heat capacity changes that make native proteins stable at physiological temperatures (Gómez et al., 1995; Lazaridis & Karplus, 2002; Prabhu & Sharp, 2005; Robertson & Murphy, 1997). The same uncertainty exists about the relative contribution of van der Waals and Coulomb interactions to these fundamental thermodynamic quantities (Lazaridis & Karplus, 2002; Newberry & Raines, 2019). To address this problem, we have recently described a method (Galano-Frutos & Sancho, 2019; Galano-Frutos et al., 2023) that uses MD simulations of native conformations and carefully generated unfolded ensembles (Estrada et al., 2009) to calculate ΔH and ΔCp values of folding in close agreement with the experimental ones (Galano-Frutos &

Sancho, 2019; Galano-Frutos et al., 2023). Those values, in combination with the experimentally determined mid denaturation temperature (T_m), yield the stability of the protein accurately (Galano-Frutos et al., 2023). The method circumvents the still limited ability of MD for simulating protein folding times, as well as the sampling problem posed by the overcompaction exerted on unfolded conformations by some force fields in long simulations (Best et al., 2014; Piana et al., 2014, 2015; Robustelli et al., 2018; Zerze et al., 2019). Here, we present a detailed energy dissection analysis, at different temperatures, of four model proteins whose thermodynamics have been successfully calculated. The energy patterns obtained for these proteins are recurrent, revealing the long-sought signs and magnitudes of the main contributions to protein folding energetics by the interacting molecules and the forces that are involved.

2 | RESULTS

2.1 | Accuracy of ΔH_{fol} and ΔCp_{fol} computed from atomistic MD simulations

Recently (Galano-Frutos & Sancho, 2019; Galano-Frutos et al., 2023), we have demonstrated that the changes in enthalpy and heat capacity associated with protein folding can be accurately calculated from short atomistic MD simulations of folded conformations and unfolded ensembles. The calculated ΔH_{fol} and ΔCp_{fol} values for the four proteins analyzed here (Table 1) agree well with the experimental ones (Table S1). Linear plots (calculated vs. experimental values) show correlations and slopes close to unity for both ΔH_{fol} ($R^2 = 0.99$, slope = 0.98) and ΔCp_{fol} ($R^2 = 0.99$, slope = 0.86) (Figure 1a,b). When those values were used in combination with experimental T_ms to calculate ΔG_{fol}, using the Gibbs–Helmholtz equation (Becktel & Schellman, 1987), the correlation with the experimental ones was also good ($R^2 = 0.83$, slope = 0.81; Figure 1c) (Galano-Frutos et al., 2023). We take this as strong evidence that the force field used in the simulations (Charmm22 with CMAP correction, version 2.0; Mackerell et al., 2004) describes the energetics of the folding equilibrium sufficiently well despite its known compacting effect on protein unfolded conformations (Galano-Frutos & Sancho, 2019; Piana et al., 2015; Robustelli et al., 2018). We have minimized such compacting effect by using short 2-ns simulations for sampling (Galano-Frutos & Sancho, 2019; Galano-Frutos et al., 2023). Each of the individual molecular or elementary energy differences contributing to ΔH_{fol}, and defined in Table 2, have been calculated by subtracting the corresponding time- and replica-averaged value obtained for the simulation boxes containing unfolded conformations from the value for the boxes containing folded conformations (Figure 2, left part). All the energy values obtained from the folded and unfolded boxes and the corresponding differences are shown in Tables S2–S6.

2.2 | Illustrative examples of lower accuracy provided by newer additive force fields and by a polarizable one

Since the CMAP correction for Charmm22—used in our recent work (Galano-Frutos & Sancho, 2019; Galano-Frutos et al., 2023)—was released in 2004, significant efforts have been done to improve force fields trying to reproduce better the geometry of unfolded conformations. One might expect that those improvements should retain the accurate energetic description of the folding reaction provided by some of the older force fields (Galano-Frutos & Sancho, 2019; Galano-Frutos et al., 2023). However, in our experience, the newer force fields specifically tuned to avoid compaction of unfolded conformations do not provide an accurate calculation of ΔH_{fol} by difference (Tables S7 and S8). We will illustrate this, using calculations of barnase (all performed at the temperature of 315 K). The Charmm22-CMAP force field in conjunction with the Tip3p water model (Jorgensen et al., 1983) reproduces well (−439 kJ/mol, see Table 1 and Table S8) the experimental barnase ΔH_{fol} at that temperature (−449.8 kJ/mol). However, we already described in previous work (Galano-Frutos & Sancho, 2019) that the Amber99SB-ILDN force field, released in 2010, also provided a reasonable value of barnase ΔH_{fol} (−365 kJ/mol), but that its newer version A99SB-disp (Robustelli et al., 2018), released in 2018 and specifically devised to improve the simulation of disordered protein states, did not (Table S8). This was the case whether using Tip3p (Galano-Frutos et al., 2023), the water model used in the successful simulations with Charmm22-CMAP and Amber99SB-ILDN, or Tip4p-d (Piana et al., 2015), a newer water model specifically devised in 2015 to provide a better solvation of the unfolded state with which A99SB-disp calculates an unphysical positive ΔH_{fol} of +335 kJ/mol for barnase (Table S8). Now, we have tested the performance of Charmm36m (Huang et al., 2017), a newer version of the Charmm force field family, released in 2017, that has also been shown to improve the description of unfolded ensembles. However, unlike the older Charmm22-CMAP, Charmm36m underestimate barnase ΔH_{fol}: −176 kJ/mol (Tables S7 and S8 for the raw data). Thus, the improvements introduced in the older Charmm and Amber versions seem to have worsened the description of the protein folding energetics. Another ongoing

TABLE 1 Contributions to ΔH and ΔCp of folding.

		Folding energy change (kJ/mol)				
		Temperature				
Protein	**Contributor[a]**	**295 K**	**315 K**	**335 K**	**ΔCp[b] (kJ/mol·K)**	**R^2**
Barnase[c]	ΔE_{PP}	−3109	−2978	−2850	6.48	1.00
	ΔE_{PN}	5261	4907	4584	−16.93	1.00
	ΔE_{NN}	−2481	−2362	−2238	6.08	1.00
	ΔE^{Coul}	−138	−216	−268	−3.23	0.99
	ΔE^{LJ}	−278	−278	−286	−0.18	0.66
	ΔE^{Bonded}	88	61	50	−0.94	0.95
	ΔE	−329	−433	−504	−4.38	0.99
	ΔE^{kin}	−3	−6	−5	−0.05	0.70
	Δ(pV)	0	0	0	NA	NA
	ΔH	−332	−439	−509	−4.40	0.99
		Folding energy change (kJ/mol)				
		Temperature				
Protein	**Contributor[a]**	**295 K**	**315 K**	**335 K**	**ΔCp[b] (kJ/mol·K)**	**R^2**
Barnase[d]	ΔE_{PP}	−3003	−2715	−2550	11.33	0.98
	ΔE_{PN}	4713	4181	3773	−23.50	0.99
	ΔE_{NN}	−2007	−1833	−1701	7.65	0.99
	ΔE^{Coul}	18	−66	−160	−4.45	1.00
	ΔE^{LJ}	−369	−346	−348	0.53	0.68
	ΔE^{Bonded}	54	47	31	−0.58	0.95
	ΔE	−297	−367	−478	−4.530	0.98
	ΔE^{kin}	2	2	−5	−0.18	0.75
	Δ(pV)	0	0	0	NA	NA
	ΔH	−295	−365	−483	−4.71	0.98
		Folding energy change (kJ/mol)				
		Temperature				
Protein	**Contributor[a]**	**307 K**	**317 K**	**327 K**	**ΔCp[b] (kJ/mol·K)**	**R^2**
SNase[c]	ΔE_{PP}	−4323	−4287	−4128	9.76	0.89
	ΔE_{PN}	7426	7260	6896	−26.51	0.96
	ΔE_{NN}	−3273	−3229	−3072	10.04	0.90
	ΔE^{Coul}	66	6	−17	−4.16	0.94
	ΔE^{LJ}	−403	−404	−399	0.18	0.61
	ΔE^{Bonded}	166	141	112	−2.73	1.00
	ΔE	−170	−257	−304	−6.70	0.97
	ΔE^{kin}	2	−1	−4	−0.30	1.00
	Δ(pV)	0	0	0	NA	NA
	ΔH	−168	−258	−308	−7.01	0.97

Protein	Contributor[a]	Folding energy change (kJ/mol) Temperature 305 K	320 K	335 K	ΔCp[b] (kJ/mol·K)	R^2
apoFld[c]	ΔE_{PP}	−4571	−4324	−4106	15.50	1.00
	ΔE_{PN}	8117	7439	6933	−39.47	0.99
	ΔE_{NN}	−3732	−3444	−3323	13.63	0.95
	ΔE^{Coul}	8	−90	−252	−8.67	0.98
	ΔE^{LJ}	−378	−384	−364	0.47	0.47
	ΔE^{Bonded}	185	145	119	−2.20	0.99
	ΔE	−186	−329	−496	−10.35	1.00
	ΔE^{kin}	−1	0	0	−0.05	1.00
	Δ(pV)	0	0	0	NA	NA
	ΔH	−187	−329	−496	−10.30	1.00
Protein	Contributor[a]	Folding energy change (kJ/mol) Temperature 320 K	335 K	350 K	ΔCp[b] (kJ/mol·K)	R^2
CI2[c]	ΔE_{PP}	−1481	−1421	−1379	3.40	0.99
	ΔE_{PN}	2381	2237	2146	−7.83	0.98
	ΔE_{NN}	−1061	−1008	−981	2.67	0.97
	ΔE^{Coul}	−62	−75	−98	−1.20	0.97
	ΔE^{LJ}	−115	−122	−121	−0.20	0.63
	ΔE^{Bonded}	17	6	5	−0.40	0.81
	ΔE	−161	−192	−214	1.77	0.99
	ΔE^{kin}	0	0	−1	−0.03	0.75
	Δ(pV)	0	0	0	NA	NA
	ΔH	−161	−192	−215	−1.80	1.00

[a]Equivalences: $\Delta E_{PP} = \Delta E_{PP}^{LJ} + \Delta E_{PP}^{Coul} + \Delta E^{Bonded}$; $\Delta E_{PN} = \Delta E_{PN}^{LJ} + \Delta E_{PN}^{Coul}$; $\Delta E_{NN} = \Delta E_{NN}^{LJ} + \Delta E_{NN}^{Coul} + \Delta E_{NN}^{Coul\_recip} + \Delta E_{NN}^{Disp\_Corr}$.
[b]Slope of corresponding energy change as a function of temperature.
[c]Data from simulations run with Charmm22-CMAP and Tip3p (Galano-Frutos et al., 2023).
[d]Data from simulations reported in Galano-Frutos and Sancho (2019) run with Amber99SB-ILDN, Tip3p and a smaller sampling of the folded and unfolded state (see Table S6).

development in protein simulation relates to the use of polarizable force fields aiming at describing nonbonded interactions more accurately. We have tested the performance of Drude (Lemkul et al., 2016), one polarizable force field that may be considered an evolution of recent Charmm versions, such as Charmm36, to which it was extensively compared before. Drude calculation of barnase ΔH_{fol} at −100 kJ/mol (Table S8) is also a clear underestimation of the experimental value. This sampling of two force fields tuned to improve the geometry of unfolded conformations and one representative of polarizable force fields suggests that while they may solve some limitations of previous force fields, they appear to have lost in the process the capacity to compute accurately the energetics of protein folding. It is thus suggested that further improvements of force fields intended to simulate proteins should not only focus on the correct geometrical description of unfolded ensembles but also on the accurate calculation of protein folding energetics. For what, trying to reproduce experimental ΔH_{fol} and ΔCp_{fol} values may be an appropriate goal.

2.3 | Contribution of intraproteic and solvation interactions to ΔH_{fol}

The protein folding enthalpy change is a consequence of the rearrangement of interactions between protein and solvent molecules. The relative contribution of changes

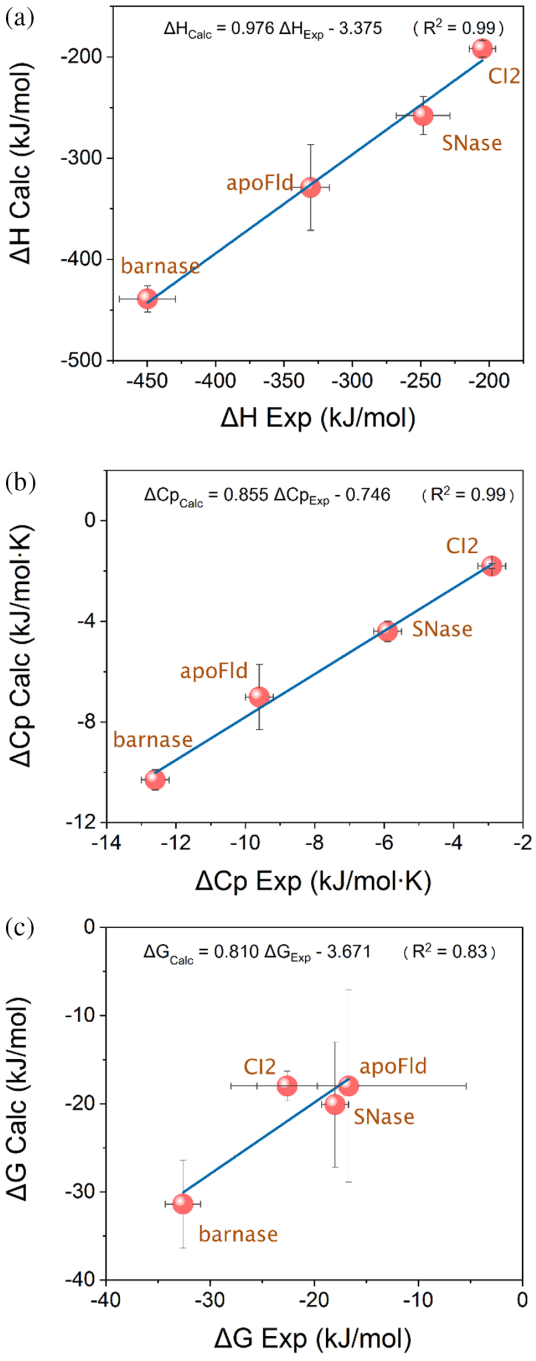

FIGURE 1 Linear correlation between calculated and experimentally determined ΔH_{fol} (a), ΔCp_{fol} (b), and ΔG_{fol} (c).

in intraproteic (ΔE_{PP}), intrasolvent (ΔE_{NN}), and protein-solvent (ΔE_{PN}) interactions to ΔH_{fol} has been determined first for barnase using the Charmm22-CMAP force field (Table 1). The enthalpic stabilization of the barnase native conformation at 42°C (−439 kJ/mol) is driven by the stabilizing contribution of intraproteic interactions ($\Delta E_{PP} = -2978$ kJ/mol) being slightly larger than the destabilizing contribution of solvation ($\Delta E_{PN} + \Delta E_{NN} = +2545$ kJ/mol). An additional minor contribution of −6 kJ/mol from the kinetic energy change completes the calculated enthalpy balance. This contribution probably comes from a non-perfect cancelation of the very large kinetic energies of the folded and unfolded simulation boxes (Table S2). The overall destabilizing effect of solvation is composed of two opposing contributions: a dominant, destabilizing contribution from protein–solvent interactions ($\Delta E_{PN} = +4907$ kJ/mol) and a smaller but significant stabilizing contribution of intrasolvent interactions ($\Delta E_{NN} = -2362$ kJ/mol; Figure 3a). Thus, as the barnase polypeptide folds establishing new internal stabilizing interactions and losing, to a much greater extent, pre-existing stabilizing interactions with solvent molecules, new stabilizing interactions appear between the solvent molecules surrounding the polypeptide. The energy change in internal solvent interactions is almost as large as that taking place in intraproteic interactions and makes, therefore, an important contribution to stabilizing the native conformation.

For the sake of comparison and to rule out that the energy pattern provided by Charmm22-CMAP might be force field specific rather than intrinsic to the protein simulated, we have performed the same dissection of barnase folding energetics using data from simulations performed with Amber99SB-ILDN, in otherwise identical conditions. Clearly, Amber99SB-ILDN force field provides the same pattern as Charmm22-CMAP. According to Amber99SB-ILDN, the contribution of intraproteic interactions is −2715 kJ/mol, and those of the two solvation terms are +4181 and −1833 kJ/mol, respectively (Table 1). Thus, two independent force fields, Charmm22-CMAP and Amber99SB-ILDN, agree in revealing an important stabilizing contribution of intrasolvent interactions to ΔH_{fol}.

To assess whether this pattern is specific to barnase or general to water-soluble globular proteins, MD simulations were performed in the same temperature range for SNase (at 44°C), apoFld (at 47°C), and CI2 (at 47°C) using Charmm22-CMAP (Table 1). The same quantitative pattern of stabilizing and destabilizing molecular interactions was found for all these proteins. Intrasolvent and intraproteic interactions are always stabilizing, the ratio ($\Delta E_{NN}/\Delta E_{PP}$) describing their relative contributions to ΔH_{fol} being 0.75 for SNase, 0.80 for apoFld, and 0.72 for CI2, very similar to the ratio of 0.79 found for barnase. For all four proteins, the stabilizing contributions are largely counterbalanced by destabilizing protein–solvent interactions. In absolute value, ΔE_{PN} equals 0.92 (barnase), 0.97 (SNase), 0.96 (apoFld) or 0.94 (CI2) times the stabilizing interactions combined ($\Delta E_{PP} + \Delta E_{NN}$) (Figure S1a). The averages of the relative molecular contributions to ΔH_{fol} in the four proteins are shown in Figure 2.

TABLE 2 Energy inventories of the Charmm22-CMAP and Amber99SB-ILDN force fields, thermodynamic equivalences, and energy grouping used to discuss the contribution of specific physical interactions (Bonded, Lennard-Jones or Coulomb) or specific molecular interactions: intraproteic (PP), protein-solvent (PN) or intrasolvent (NN) to protein folding ΔH and ΔC_p.

Energy terms of the force fields[a]	
Bonded terms[b]	
For Charmm22-CMAP	Bonds, U-B, Proper-Dih, CMAP-Dih, Improper-Dih
For Amber99SB-ILDN	Bonds, Angle, Proper-Dih, Improper-Dih
Nonbonded terms	
For both force fields	LJ-14, Coul-14, LJ-SR, Disp-corr, Coul-SR, Coul-recip
Energy grouping used in this article	
Bonded terms	
For Charmm22-CMAP	Bonded = U-B + Proper-Dih + CMAP-Dih + Improper-Dih
For Amber99SB-ILDN	Bonded = Angle + Proper-Dih + Improper-Dih
Nonbonded terms	
For both force fields	LJ = LJ-14 + LJ-SR + Disp-corr
	Coul = Coul-14 + Coul-SR + Coul-recip
Energy terms of the force fields in a secondary partition provided by Gromacs, based on interacting molecules[a,c]	
For both force fields	Coul-SR-P-P, LJ-SR-P-P, Coul-14-P-P, LJ-14-P-P, Coul-SR-P-N, LJ-SR-P-N, Coul-SR-N-N, LJ-SR-N-N
Energy grouping used in this article[c]	
$Coul_{PP}$ = Coul-SR-P-P + Coul-14-P-P	
LJ_{PP} = LJ-SR-P-P + LJ-14-P-P	
$E_{PP} = Coul_{PP} + LJ_{PP}$ + Bonded	
E_{PN} = Coul-SR-P-N + LJ-SR-P-N	
E_{NN} = Coul-SR-N-N + Coul-recip + LJ-SR-N-N + Disp-Corr	
Equivalences with thermodynamic terms	
Potential Energy (E) = Bonded + LJ + Coul = $E_{PP} + E_{PN} + E_{NN}$	
Enthalpy (H) = $E + E^{kin}$ + pV	

[a]Original terms of the force fields and of the energy partitions, as provided by Gromacs. "U-B" refers to the Urey-Bradley component (cross-term accounting for angle bending using 1,3 nonbonded interactions); "Coul" refers to Coulomb interactions; "LJ" refers to the 12-6 Lennard-Jones potential (van der Waals); "14" accounts for interactions between 1 and 4 pairs; "SR" accounts for short-range interactions; "Disp-corr" refers to long range dispersion corrections applied on energy and pressure (configured with Gromacs' mdp option EnerPres); "Coul-recip" refers to the reciprocal space contribution (long-range Coulomb interactions); "CMAP-Dih" refers to the correction maps added to the dihedrals in Charmm22-CMAP force field.
[b]Simulations carried out under constraints on all bonds.
[c]In the main text and in other tables, the hyphens in the P-N, P-N, and N-N force field terms are omitted for simplicity.

2.4 | Contribution of bonded, Coulomb, and Lennard-Jones interactions to ΔH_{fol}

The folding enthalpy change of a protein can be partitioned differently attending to the forces governing the reaction, rather to the participating molecular entities. At the indicated temperature of 42°C, and under the Charmm22-CMAP force field, the barnase enthalpy change ($\Delta H_{fol} = -439$ kJ/mol) arises from stabilizing energy contributions of Coulomb ($\Delta E^{Coul} = -216$ kJ/mol) and Lennard-Jones ($\Delta E^{LJ} = -278$ kJ/mol) interactions, which are partly opposed by moderately destabilizing bonded interactions ($\Delta E^{Bonded} = +61$ kJ/mol; Table 1 and Figure 3b). It seems that, as barnase folds, LJ interactions—and to a lower extent Coulomb interactions—strengthen at the expense of introducing some strain in the native conformation. The Amber99SB-ILDN force field reveals the same pattern for barnase folding as it also reports stabilizing contributions from Coulomb (−66 kJ/mol) and LJ (−346 kJ/mol) interactions and a moderate destabilizing contribution (+47 kJ/mol) from bonded ones (Table 1). An analysis of peptide and miniprotein (10–35 residues) folding pathways with Amber family force fields also showed a destabilizing contribution of bonded interactions (Shao et al., 2019).

To assess whether this is a general pattern characteristic of the protein folding reaction of globular proteins, the ΔH_{fol} values calculated for SNase, apoFld and CI2

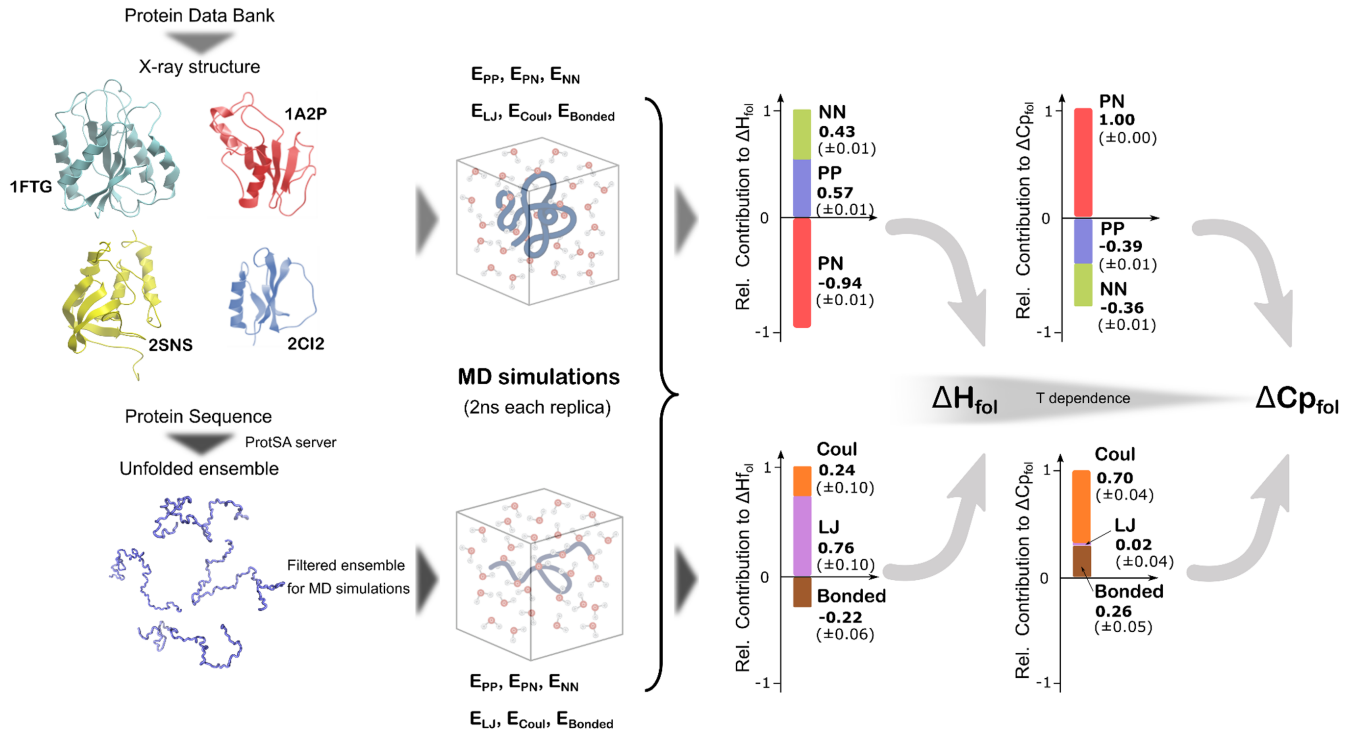

FIGURE 2 General scheme for the calculation of ΔH_{fold} and ΔCp_{fold} and their relative molecular and physical contributions. The scheme part at the left-hand side—including the simulation boxes—indicates the simulated proteins and summarizes the MD setup followed, which is fully described in Galano-Frutos et al. (2023). The bar graphs at the right-hand side show the averaged (standard errors between parentheses) relative molecular and physical contributions to ΔH_{fol} and ΔCp_{fol} obtained for the four proteins analyzed. Positive relative contributions to ΔH_{fol} indicate molecular or physical interactions that stabilize the folded conformation, while negative relative values indicate destabilizing contributions. In bars graphs showing relative contributions to ΔCp_{fol}, averaged positive values indicate either stabilizing contributions that become even more stabilizing with temperature (Coul) or destabilizing contributions that become less destabilizing with temperature (Bonded, PN), while negative values indicate stabilizing contributions that become less stabilizing with temperature (PP, NN). The reference value of 1 in the bars graphs represents the sum of the contributions whose sign aligns with that of ΔH_{fol} or ΔCp_{fol} positive contributions. The negligible (slightly positive) van der Waals (LJ) average contribution to ΔCp_{fol} is the only contribution whose sign varies among the proteins analyzed.

using Charmm22-CMAP (see above) has also been partitioned into Coulomb, LJ, and bonded contributions. As in the case of barnase, the major stabilization for all these proteins comes from LJ interactions, and there is a clear destabilizing contribution from bonded ones. The contribution of Coulomb interactions to ΔH_{fol} is more variable in the four proteins, as it is clearly stabilizing in barnase, apoFld, and CI2 but it is close to negligible in SNase at the temperature of comparison (Table 1 and Figure 3b). The averages of the relative physical contributions to ΔH_{fol} in the four proteins are shown in Figure 2.

2.5 | Brief description of contributions to barnase ΔH_{fol} from force fields that greatly underestimate the value or change the sign

For barnase, the different types of contributions to ΔH_{fol} have been additionally calculated from MD simulations done with newer force fields (Tables S7 and S8), tuned to better describe unfolded conformations. Although those force fields (Charmm36m [Huang et al., 2017] and Amber99SB-disp [Robustelli et al., 2018]) do not calculate well ΔH_{fol}, they provide a molecular pattern of interactions (i.e., ΔE_{PP}, ΔE_{NN}, and ΔE_{PN}) which is in qualitative agreement with Equations (1) and (2) (see Section 3). In addition, Charmm36m, which underestimates the value of ΔH_{fol} but correctly calculates its sign, also gives, for the contribution of forces, a pattern in qualitative agreement with that seen for the four proteins accurately calculated with Charmm22-CMAP. However, Amber99SB-disp, which calculates an unphysical positive value of ΔH_{fol}, attributes destabilizing contributions to all terms (i.e., Coulomb, LJ, and bonded). Finally, the polarizable Drude force field (Lemkul et al., 2016), which greatly underestimate ΔH_{fol}, also attributes a destabilizing contribution to LJ. In this respect, a recent benchmarking of the Drude force field on a non-protein system

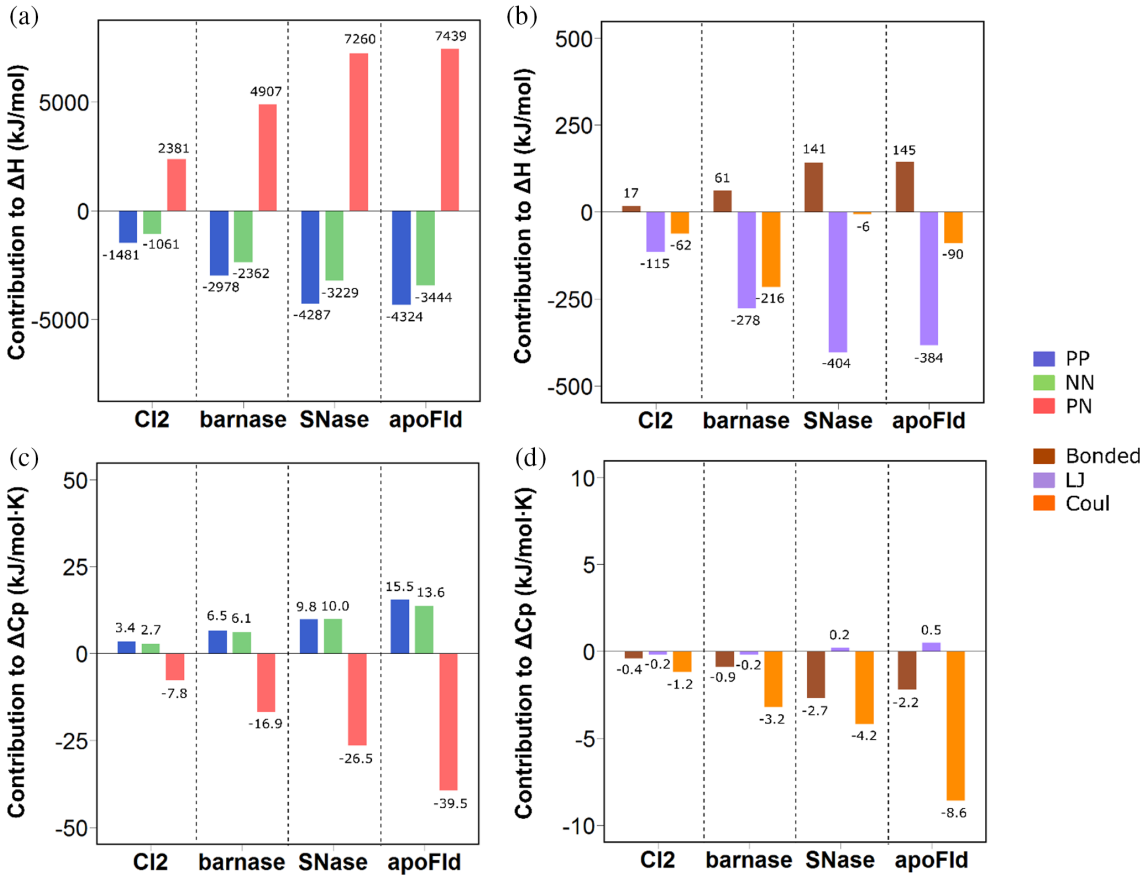

FIGURE 3 Contributions of molecular interactions and of physical interactions to ΔH_{fol} (a, b) and ΔCp_{fol} (c, d). "PP," "NN," and "PN" in panels a and c refer to intraproteic, intrasolvent, and protein-solvent energy contributions, respectively, whereas "Bonded," "LJ," and "Coul" and in panels b and d refers to bonded, Lennard-Jones (van der Waals) and Coulomb (electrostatics) contributions, respectively.

has also found that Drude overestimate electrostatic interactions at the expense of van der Waals, and it has concluded that "there is a misbalance of forces within the Drude FF, which likely requires a reweighting of the electrostatics or van der Waals interactions" (Winkler & Cheatham, 2023).

2.6 | Temperature dependence of intraproteic and solvation interactions: Contributions to ΔCp_{fol}

The folding enthalpies of barnase, SNase, apoFld, and CI2 have been calculated at three temperatures using Charmm22-CMAP, and their corresponding ΔE_{PP}, ΔE_{PN}, and ΔE_{NN} components have been obtained at each temperature. Then, their individual contributions to the heat capacity of folding (ΔCp_{fol}) have been calculated as the slopes of linear plots of energy change versus simulation temperature (e.g., ΔE_{PP} vs. T). These plots have shown excellent correlation coefficients (R^2) of 1.00 for barnase, 0.88–0.96 for SNase, 0.95–1.00 for apoFld, and 0.97–0.99 for CI2 (Table 1). In barnase, the contribution of intraproteic interactions to ΔCp_{fol} is positive ($\Delta Cp_{PP} = +6.48$ kJ/mol·K; Table 1), indicating that, with increasing temperature, the enthalpic stabilization they provide becomes less intense. On the other hand, the combined contribution to ΔCp_{fol} of the two solvation terms is negative ($\Delta Cp_{PN} + \Delta Cp_{NN} = -10.85$ kJ/mol·K; Table 1), meaning that the enthalpic destabilization due to solvation is also weakened as temperature increases. Incidentally, the two solvation terms make opposite contributions: that of protein–solvent interactions is negative and large ($\Delta Cp_{PN} = -16.93$ kJ/mol·K), while that of intrasolvent interactions is positive ($\Delta Cp_{NN} = +6.08$ kJ/mol·K; Table 1 and Figure 3c). Thus, as temperature increases, both the intraproteic and intrasolvent stabilizing interactions and the destabilizing protein–solvent interactions are weakened, but the latter are weakened to a greater extent. These effects taken together, the overall folding heat capacity becomes negative ($\Delta Cp_{fol} = -4.40$ kJ/mol·K; Table 1) and ΔH_{fol} stabilizes

the folded state more at higher temperatures than at lower ones. The same pattern appears when barnase is simulated using Amber 99SB-ILDN (Table 1). Likewise, analysis of the SNase, apoFld, and CI2 simulations performed with Charmm 22-CMAP (Table 1) indicates that the relative contributions of intraproteic, protein–solvent, and intrasolvent interactions to ΔCp_{fol} is essentially the same as that described for barnase (Figure 3c). For all four proteins, positive intraproteic and intrasolvent contributions of similar magnitude to ΔCp_{fol} are offset by a 1.29–1.36-fold negative contribution from protein–solvent interactions (Figure S1b). The averages of the relative molecular contributions to ΔCp_{fol} in the four proteins are shown in Figure 2.

2.7 | Temperature dependence of bonded, Coulomb, and Lennard-Jones interactions: Contributions to ΔCp_{fol}

The ΔCp_{fol} values calculated for the four proteins can be alternatively divided into specific contributions associated with physical interactions: LJ, Coulomb, and bonded. For all four proteins, linear plots of the folding energy change versus simulation temperature (e.g., ΔE_{Coul} vs. T) allow us to calculate the LJ, Coulomb, and bonded contributions to ΔCp_{fol}. The Coulomb and bonded plots show high correlation coefficients ($R^2 = 0.94$–0.99 for Coulomb, and 0.81–1.00 for bonded), whereas those of the LJ plots are lower (0.47–0.81). For all four proteins, the largest contribution to ΔCp_{fol} comes from Coulomb interactions, and there is also a significant contribution from bonded interactions (Figure 3d). In contrast, the LJ contributions to ΔCp_{fol} are consistently small in the four proteins ($\Delta Cp_{LJ} = -0.18$, $+0.18$, $+0.47$ and -0.20 kJ/mol·K; Table 1), indicating that the LJ balance of their folding equilibria is quite insensitive to temperature. The Coulomb and bonded contributions to ΔCp_{fol} are both negative, meaning that the stabilizing effect of Coulomb interactions increases with temperature while the destabilizing effect of bonded interactions decreases. Since the strong stabilizing effect exerted by LJ interactions on the native conformation (Figure 3b) appears to be rather insensitive to temperature (Figure 3d), the larger enthalpic stabilization of the native conformations of all these proteins at higher compared to lower temperatures does not arise from LJ interactions. Instead, it arises from the strengthening of stabilizing Coulomb interactions combined with the weakening of destabilizing bonded ones. The averages of the relative physical contributions to ΔCp_{fol} in the four proteins are shown in Figure 2.

2.8 | Correlation of ΔH_{fol} with protein length and changes in solvent exposure

There have been significant efforts to correlate folding enthalpy and heat capacity changes to simple properties of polypeptides such as protein size or solvent exposure (Gómez et al., 1995; Robertson & Murphy, 1997). Our analyses of barnase, SNase, apoFld, and CI2 MD simulations (Table 3) show (Figure 4a) that correlations between individual ΔE_{PP}, ΔE_{PN}, and ΔE_{NN} contributions to ΔH_{fol} and protein length (Table S1) are impressive ($R^2 = 0.97$–0.99), yet the correlation of ΔE (which combines the three energy terms) with protein length is poor ($R^2 = 0.16$). This can be so because, for each protein, the individual ΔE_{PP}, ΔE_{PN}, and ΔE_{NN} contributions are one order of magnitude higher than their added value (ΔE), and small errors in their calculated values can translate into a big error in ΔE. As ΔE only differs from ΔH_{fol} by a very small kinetic energy term, it follows that the correlation observed between calculated ΔH_{fol} and protein length is similarly poor ($R^2 = 0.16$; Figure 4e). This agrees with the fact that the experimental ΔH_{fol} values reported for these proteins (Table S1) do not correlate with protein length any better ($R^2 = 0.06$; Figure 4e). Considering the alternative partition of enthalpy changes attending to forces, both stabilizing LJ and destabilizing bonded interactions correlate with protein length ($R^2 = 0.94$ and 0.96, respectively) but Coulomb interactions do not ($R^2 = 0.04$; Figure 4c), which also explains from this angle the poor correlation of calculated ΔH_{fol} with protein length ($R^2 = 0.16$; Figure 4e). It has been pointed out that many quantities that reflect extensive properties should scale approximately linearly with the size of the protein (as it is seen in this section and in the following one) simply because the composition and packing of most proteins are similar (Lazaridis et al., 1995). In this context, a lack of correlation may be sometimes more indicative. For example, the lack of correlation between ΔE^{Coul} and protein length might suggest that hydrogen bonding is not a major contributor to ΔE^{Coul} (see Section 3).

On the other hand, the folding change in solvent-accessible surface area (ΔSASA), and its polar and apolar components ($\Delta SASA_{pol}$ and $\Delta SASA_{apol}$) have been calculated for each of these proteins (Table S1) (Estrada et al., 2009). Individual ΔE_{PP}, ΔE_{PN}, and ΔE_{NN} energy contributions to ΔE correlate best (Table 3) with $\Delta SASA_{pol}$ (R^2 from 0.97 to 0.99), then with ΔSASA (R^2 from 0.92 to 0.94; Figure 4b), and then with $\Delta SASA_{apol}$ (R^2 from 0.84 to 0.88). However, as seen above for the correlations with protein length, and likely for the same trivial reason of ΔE being very small compared to its components, ΔE does not correlate with either total,

TABLE 3 Correlation coefficients (R^2) between several protein properties and different contributions to ΔH and ΔCp of folding.[a]

	Property				Calculated contributions to ΔH							Calculated contributions to ΔCp						
					By partners			By force				By partners			By force			
	Length	ΔSASA	$\Delta SASA_{apol}$	$\Delta SASA_{pol}$	PP	PN	NN	Bonded	LJ	Coul	ΔH^{calc}	ΔH^{exp}	PP	PN	NN	Bonded	LJ	Coul	ΔCp^{calc}	ΔCp^{exp}
Length	-	-	-	-	0.97	0.99	0.98	0.96	0.94	0.04	0.16	0.06	0.90	0.94	0.97	0.83	0.83	0.80	0.95	0.97
ΔSASA	0.99	-	-	-	0.92	0.93	0.94	0.93	0.87	0.03	0.15	0.06	0.96	0.98	0.99	0.76	0.88	0.89	0.99	0.99
$\Delta SASA_{apol}$	0.94	0.98	-	-	0.84	0.86	0.88	0.85	0.78	0.01	0.17	0.08	0.99	0.99	0.98	0.63	0.89	0.95	0.99	0.98
$\Delta SASA_{pol}$	0.96	0.91	0.82	-	0.99	0.99	0.97	0.99	0.97	0.10	0.10	0.02	0.77	0.83	0.89	0.94	0.75	0.64	0.85	0.91

[a]Analysis carried out with energy data calculated from simulations of barnase, SNase, apoFld, and CI2. Correlation coefficients obtained from linear correlation with properties in the first column of the table.

polar or apolar changes in SASA (R^2 from 0.11 to 0.18). Consequently, calculated ΔH_{fol} values do not correlate either with any of the SASA changes (R^2 from 0.10 to 0.17) (Table 3 and Figure 4f). This agrees with the lack of correlation between SASA changes and experimental ΔH_{fol} values (R^2 from 0.02 to 0.08). For the alternative partition of enthalpy changes into elementary contributions, both the bonded and LJ terms correlate either strongly or moderately with ΔSASA ($R^2 = 0.93$ and 0.87, respectively; Figure 4d) and its polar and apolar components (R^2 values from 0.78 to 0.99; Table 3). In contrast, the Coulomb term does not correlate with the different ΔSASAs (R^2 from 0.01 to 0.10; Table 3 and Figure 4d).

The very high correlations found between protein length or changes in SASA and the individual energy terms (ΔE_{PP}, ΔE_{PN}, and ΔE_{NN}) totaling ΔE might appear to invite to calculate ΔH_{fol} from simple linear relationships. Unfortunately, as explained, once those individual terms are summed, the resulting ΔE values no longer correlate with either length or SASA changes. For the alternative energy partition, while LJ and bonded interactions correlate well with protein length and with the various changes in SASA considered, Coulomb interactions do not, which also explains the lack of correlation here observed between protein length or SASA changes and ΔH_{fol}. In practice, no accurate calculation of ΔH_{fol} from protein length or from changes in SASA seems possible. Therefore, intervals of confidence for the fittings described in this and in the following sections have not been calculated.

2.9 | Correlation of ΔCp_{fol} with protein length and changes in solvent exposure

The correlation between ΔCp_{fol} (and molecular or elementary contributions to it) and protein length has also been examined. Intraproteic (ΔCp_{PP}), protein–solvent (ΔCp_{PN}) or intrasolvent (ΔCp_{NN}) contributions to ΔCp_{fol} correlate well (R^2 from 0.90 to 0.97) with protein length (Table 3 and Figure 5a), and the combined calculated ΔCp_{fol} also do so ($R^2 = 0.95$; Table 3 and Figure 5e). In agreement with this, the correlation of experimental ΔCp_{fol} values with protein length is also high ($R^2 = 0.97$; Table 3 and Figure 5e). Considering the alternative energy partition, the contribution of bonded, LJ, and Coulomb interactions to ΔCp_{fol} all correlate moderately well with protein length (R^2 from 0.80 to 0.83; Table 3 and Figure 5c).

On the other hand, ΔCp_{PP}, ΔCp_{PN}, and ΔCp_{NN} contributions to ΔCp_{fol} show high correlations with total ΔSASA (R^2 from 0.96 to 0.99; Table 3 and Figure 5b), even better correlation with $\Delta SASA_{apol}$ (R^2 from 0.98 to

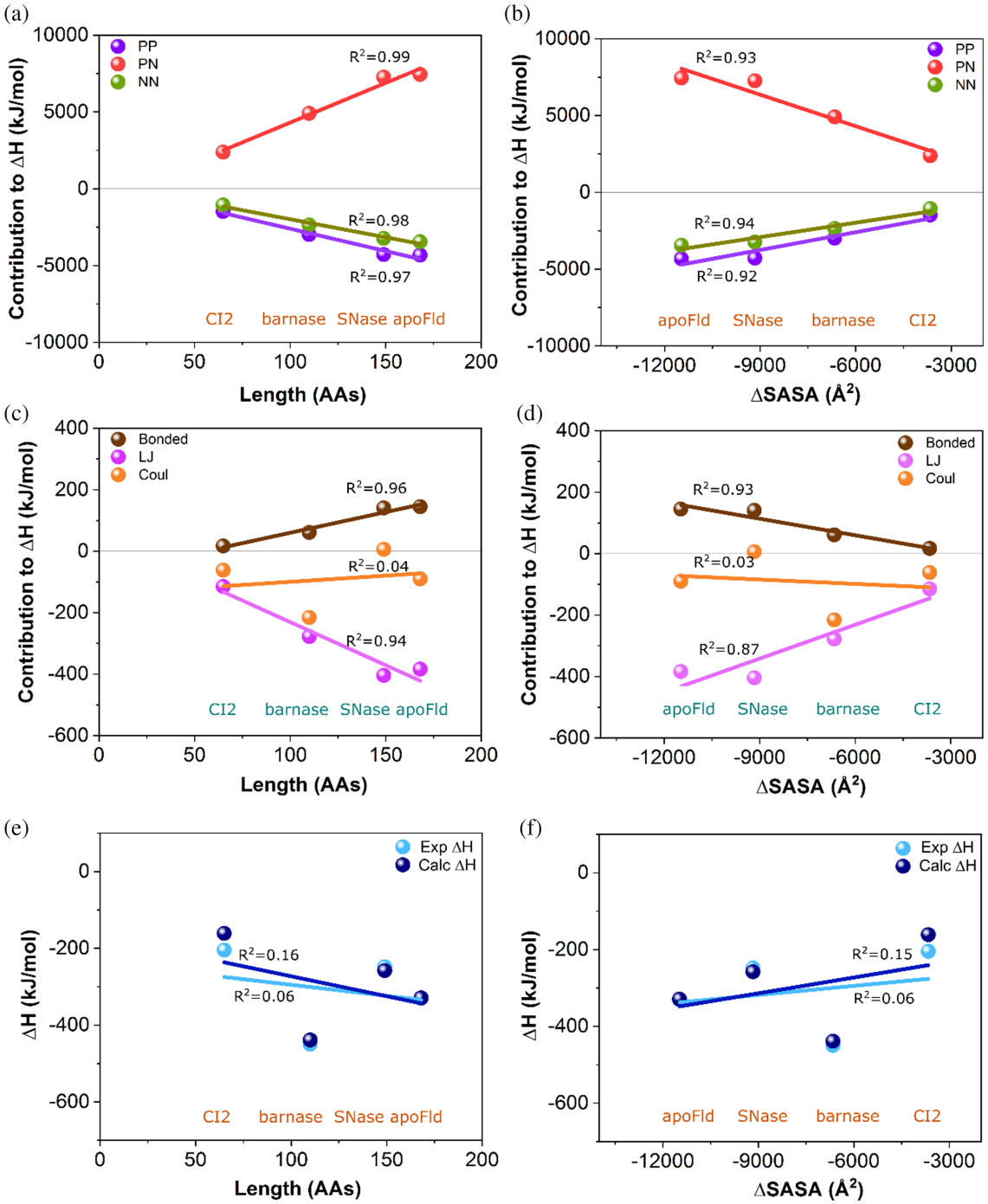

FIGURE 4 Linear correlation between molecular (a, b) and physical (c, d) contributions to ΔH_{fol} and protein length (a, c) or ΔSASA (b, d). Linear correlation between calculated and experimentally determined ΔH_{fol} and protein length (number of amino acid residues) (e) or ΔSASA (f). "PP," "NN," and "PN" in a and b refer to intraproteic, intrasolvent, and protein-solvent energy contributions, respectively, whereas "Bonded," "LJ," and "Coul" in c and d refer to bonded, Lennard-Jones (van der Waals) and Coulomb (electrostatics) contributions, respectively. Squared Pearson correlation coefficients are included close to each fitting line.

0.99; Table 3 and Figure S2a), and a clear but lower correlation with $\Delta SASA_{pol}$ (R^2 from 0.77 to 0.89; Table 3 and Figure S2c). For the alternative partition, the three contributions (ΔCp_{Bonded}, ΔCp_{LJ}, and ΔCp_{Coul}) correlate (R^2 from 0.63 to 0.95; Table 3) with total ΔSASA (Figure 5d) and apolar (Figure S2b) or polar SASA changes (Figure S2d). While the LJ contribution correlates similarly well with polar and apolar changes, the Coulomb one correlates better with $\Delta SASA_{apol}$ and the bonded contribution with $\Delta SASA_{pol}$. Overall, the calculated

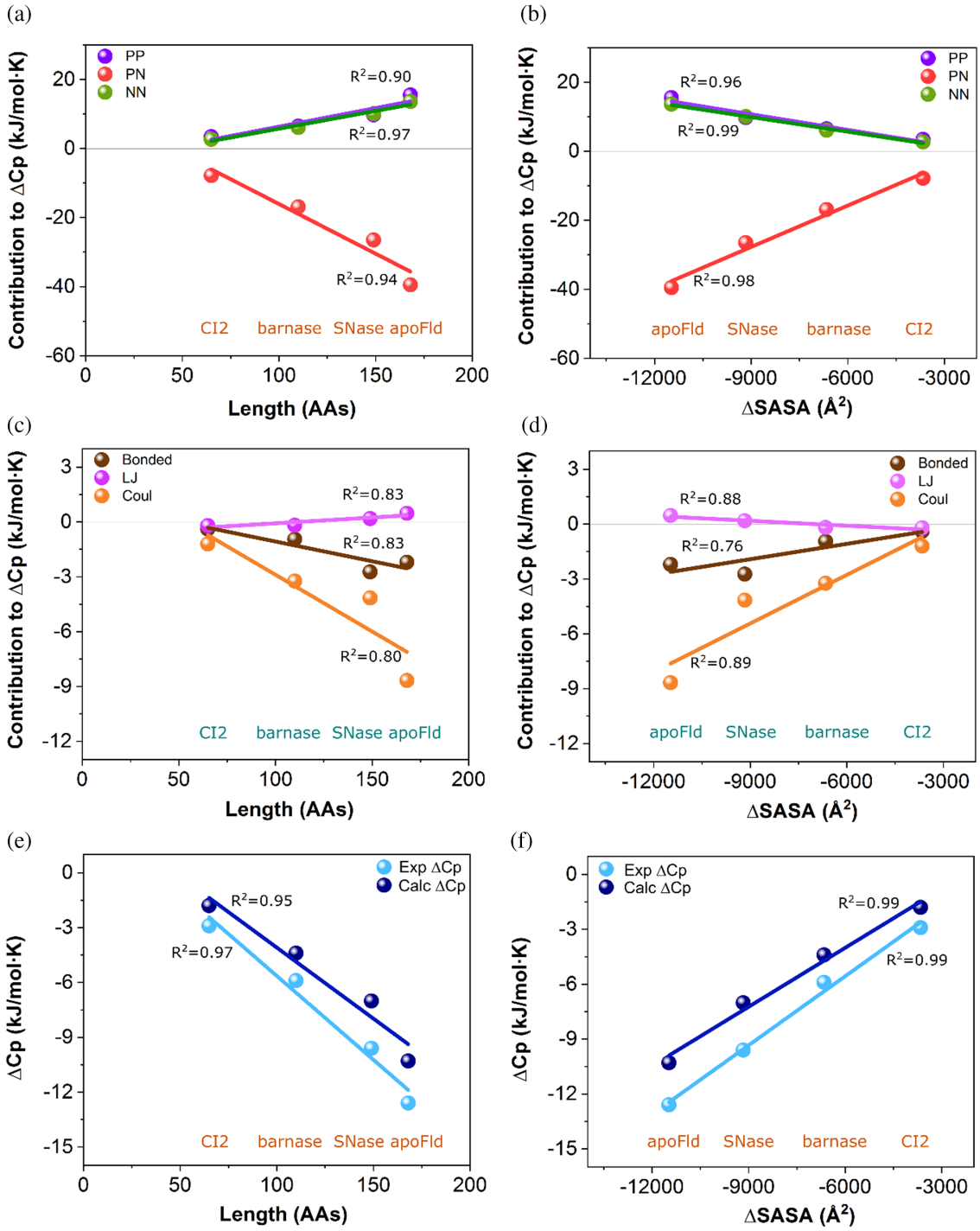

FIGURE 5 Linear correlation between molecular (a, b) and physical (c, d) contributions to ΔCp_{fol} and protein length (a, c) or ΔSASA (b, d). Linear correlation between calculated and experimentally determined ΔCp_{fol} and protein length (number of amino acids) (e) or ΔSASA (f). "PP," "NN," and "PN" in a and b refer to intraproteic, intrasolvent, and protein-solvent energy contributions, respectively, whereas "Bonded," "LJ," and "Coul" in c and d refers to bonded, Lennard-Jones (van der Waals) and Coulomb (electrostatics) contributions, respectively. Squared Pearson correlation coefficients are included close to each fitting line.

ΔCp_{fol} correlates best with either ΔSASA (Figure 5f) or its apolar component ($R^2 = 0.99$) and the experimental ΔCp_{fol} is also highly correlated with either of these two changes ($R^2 = 0.98$–0.99). Although, the ΔCp_{fol} could be calculated for these proteins with the linear equations (a): $\Delta Cp_{fol} = 3.53(\pm1.40) - 0.092(\pm0.011) \times \text{Length}$ ($R^2 = 0.97$; Figure 5e) or (b): $\Delta Cp_{fol} = 2.00(\pm0.59) + 0.0013(\pm0.0001) \times \Delta SASA$ ($R^2 = 0.99$; Figure 5f)—

where Length is the number of amino acid residues in the protein and ΔSASA is the change in total solvent exposure as calculated by ProtSA (Estrada et al., 2009)—the maximum prediction errors of these linear models for a 95% confidence level are high (4.9 kJ/mol·K over the length interval or 2.3 kJ/mol·K over the SASA interval). This precludes their predictive use to calculate ΔCp_{fol} from simple parameters.

3 | DISCUSSION

3.1 | The difficulty of dissecting protein folding energetics from experiments

Some basic facts about the protein folding equilibrium have long been known: (1) proteins that fold into compact native conformations tend to be stable at physiological temperatures and to unfold as temperature increases; (2) relative to unfolded ensembles, native conformations are stabilized at physiological temperatures by enthalpy and destabilized by entropy; (3) the folding enthalpy change is due to a combination of intraproteic interactions, protein–solvent interactions and interactions between solvent molecules; (4) the interactions that are established between protein and solvent atoms are driven primarily by electrostatic and van der Waals forces. However, despite this consolidated knowledge, the relative contribution of these different players to the stabilization of the native conformation remains controversial, since a myriad of concurrent interactions take place simultaneously in the protein folding equilibrium that cannot be measured independently.

Before protein engineering was developed and became widespread, extensive modeling efforts were undertaken to try to understand protein energetics by integrating available thermodynamic data for polar and apolar model compounds, analysis of the amino acid composition of proteins, and estimates of changes in SASA upon folding (as indicators of changes taking place in protein interactions with solvent) (Baldwin, 1986; Hilser et al., 1996; Makhatadze & Privalov, 1990; Murphy & Gill, 1991; Murphy et al., 1990; Privalov & Gill, 1988). Those studies were aided by the development of accurate calorimeters and led to attributing to hydration interactions a major contribution to ΔCp_{fol}. The rational for the proposal was that, in the unfolded state, interactions between solvent molecules and apolar groups are more intense than in the folded state, which would explain the sign of ΔCp_{fol} (Gómez et al., 1995; Madan & Sharp, 1996; Prabhu & Sharp, 2005; Privalov & Makhatadze, 1992; Robertson & Murphy, 1997). However, despite the extensive modeling carried out, the relative contributions to ΔCp_{fol} of protein–protein interactions and of the two solvation terms involved (protein–solvent interactions and intrasolvent interactions) has (Jumper et al., 2021) remained an open question (Prabhu & Sharp, 2005). Protein engineering attempted to clarify the problem by comparing the energetics of very similar protein variants, which often differed at a single amino acid residue. This approach was hampered by still existing limitations (Chin, 2017; Ravikumar et al., 2015) in the chemical changes that can be engineered and, more fundamentally, by the fact that removal or substitution of even small chemical groups present in a protein usually affects more than one type of interaction, as it has been discussed (Lazaridis & Karplus, 2002). Furthermore, protein engineering focused predominantly on determining free energy changes rather than enthalpy or heat capacity changes, which is not ideal for the purpose of performing a fine dissection of protein energetics. Ultimately, despite the large number of mutational experiments that have been carried out and that continue to feed the protein stability databases (Nikam et al., 2021; Xavier et al., 2021) used to train protein stability predictors (Dehouck et al., 2011; García-Cebollada et al., 2022; Liu & Kuhlman, 2006; Schymkowitz et al., 2005), the relative contribution of the different physical interactions to protein stability remains unclear.

More recently, protein simulation has finally opened a window into the detail. Atomistic MD simulations, the most popular and suitable simulation approach for studying protein stability and folding (Best, 2022; Ferina & Daggett, 2019; Lazaridis & Karplus, 2002; Lei et al., 2007; Lindorff-Larsen et al., 2011), makes it possible to calculate protein and solvent energetics using potential energy functions, known as force fields, whose individual terms are related to specific forces. Besides, MD programs can classify the quantified interactions into those involving only protein atoms, only solvent atoms, or atoms of both protein and solvent molecules. Usual force fields contain multiple parameters, obtained mainly from experimental data or quantum chemistry calculations on small molecules (Mackerell, 2004). Although it might be feared that such reductionist energy functions would not allow an accurate calculation of protein folding energetics, it has been recently shown that the folding enthalpy of polypeptides dissolved in water can indeed be accurately calculated using MD simulation (Galano-Frutos & Sancho, 2019; Galano-Frutos et al., 2023). For that purpose, the folded and unfolded states are first simulated independently at several temperatures. Then, the ΔH_{fol} is obtained by subtracting the time- and ensemble-averaged enthalpy of the unfolded state from that of the native conformation (Galano-Frutos & Sancho, 2019; Galano-Frutos et al., 2023), while ΔCp_{fol} is obtained as the slope

of the ΔH_{fol} temperature dependence. Accurate calculation of ΔH_{fol} and ΔCp_{fol} from MD simulation offers a long-sought opportunity to determine the contribution of different forces to protein folding energetics and to clarify the specific role played by protein and solvent interactions.

3.2 | Players and forces: A quantitative pattern of protein folding energetics

From a molecular perspective, the folding enthalpy change summarizes the rearrangement of protein internal (ΔE_{PP}) and solvent internal (ΔE_{NN}) interactions, as well as the rearrangement of the interactions established between protein and solvent molecules (ΔE_{PN}). In the four proteins dissected here, the folding reaction takes place amidst a strong cancelation of stabilizing and destabilizing contributions, which is quantitatively described by Equations (1) and (2). The native conformation is stabilized by the strengthening of both protein internal and solvent internal interactions, their relative contributions to ΔH_{fol} being similar and fairly constant:

$$\Delta E_{NN} = 0.78(\pm 0.01) \times \Delta E_{PP}. \quad (1)$$

Conversely, the native conformation is intensely destabilized by a weakening of protein–solvent interactions, which is almost as intense as the combination of the two stabilizing molecular contributions:

$$\Delta E_{PN} = -0.95(\pm 0.01) \times (\Delta E_{PP} + \Delta E_{NN}). \quad (2)$$

In other words, as the polypeptide folds (Figure 6a), the interactions it forms internally do not compensate for the interactions it loses with solvent molecules. However, the additional stabilizing interactions that form concomitantly between the solvent molecules drive the overall equilibrium in favor of the native conformation. In terms of energy, almost half of the protein–solvent interactions that are lost when the protein folds are compensated by the new interactions that are established between the solvent molecules.

From the perspective of forces, a clear pattern is also evident, in this case qualitative: (1) the major stabilizing contribution to ΔH_{fol} is made by van der Waals interactions; (2) electrostatic interactions also contribute to stabilize the native conformation but to a lesser extent; (3) the stabilizing van der Waals and electrostatic contributions are moderately opposed by destabilizing bonded contributions. As it appears (Figure 6c), protein folding in water is driven by a strengthening of van der Waals and electrostatic interactions, at the cost of introducing strain in the folded polypeptide. In agreement with this, avoiding local backbone strain has been reported to be critical for success in protein design (Baker, 2019).

Enthalpic contributions to protein stability were reported earlier by Karplus and coworkers (Lazaridis et al., 1995). Using data obtained from MD simulations in vacuum to describe the energetics of the polypeptide and an empirical model of the solvation enthalpy. Their work focused on discussing, in the new light offered by their MD simulations, the then available thermodynamics data and a variety of interpretations alternatively declaring that hydrogen bonding (Creighton, 1991), the hydrophobic effect (Dill, 1990) or both together (Privalov & Makhatadze, 1993) were the main contributors to protein stability. At the time, MD simulations such as those here analyzed were not possible due to limited computational capacity, and the simulations were done either in vacuum or using implicit solvent models, the unfolded state was represented by an extended chain and, to allow for a better comparison with alternative views proposed by Privalov and Makhatadze (Makhatadze & Privalov, 1995), ionizable residues were modeled as neutral. In this pioneering work and in a follow up by the same group (Lazaridis & Karplus, 2002) no ΔH_{fol} values were directly computed by difference using simulations of folded and unfolded polypeptides as done in the present work. Instead, the authors combined MD-based calculation of $\Delta H_{fol(in\text{-}vacuum)}$ for the polypeptide with estimations of the solvation enthalpy change. Those calculations led to unphysical positive ΔH_{fol} values for the proteins tested that, in absolute value, were one order of magnitude bigger than the experimental (negative) ones. The authors attributed such discrepancy to assumptions done in the calculation of the solvation enthalpy change from accessible surface area data and to having used a fully extended model for the denatured state. Our present study, which has been done using a fine representation of the unfolded ensemble, explicit solvent and charged residues where appropriate, cannot be directly compared with that of Karplus and coworkers due to significant methodological differences that translate in the calculation of energy terms that are not equivalent. Only one term, the contribution of bonded interactions to ΔH_{fol}, can be directly compared. According to Lazaridis and Karplus, ΔE^{Bonded} of folding is negative and small compared to other contributions, but they clearly pointed out that the actual sign could be positive if more accurate models of the denatured state were used. That is the case. Using realistic ensembles of unfolded conformations, we have determined that ΔE^{Bonded} is destabilizing and significant (Figures 2 and 3). On average, it counterbalances a 22% of the combined stabilizing van der Waals and Coulomb contribution to ΔH_{fol}, which indicates that folded

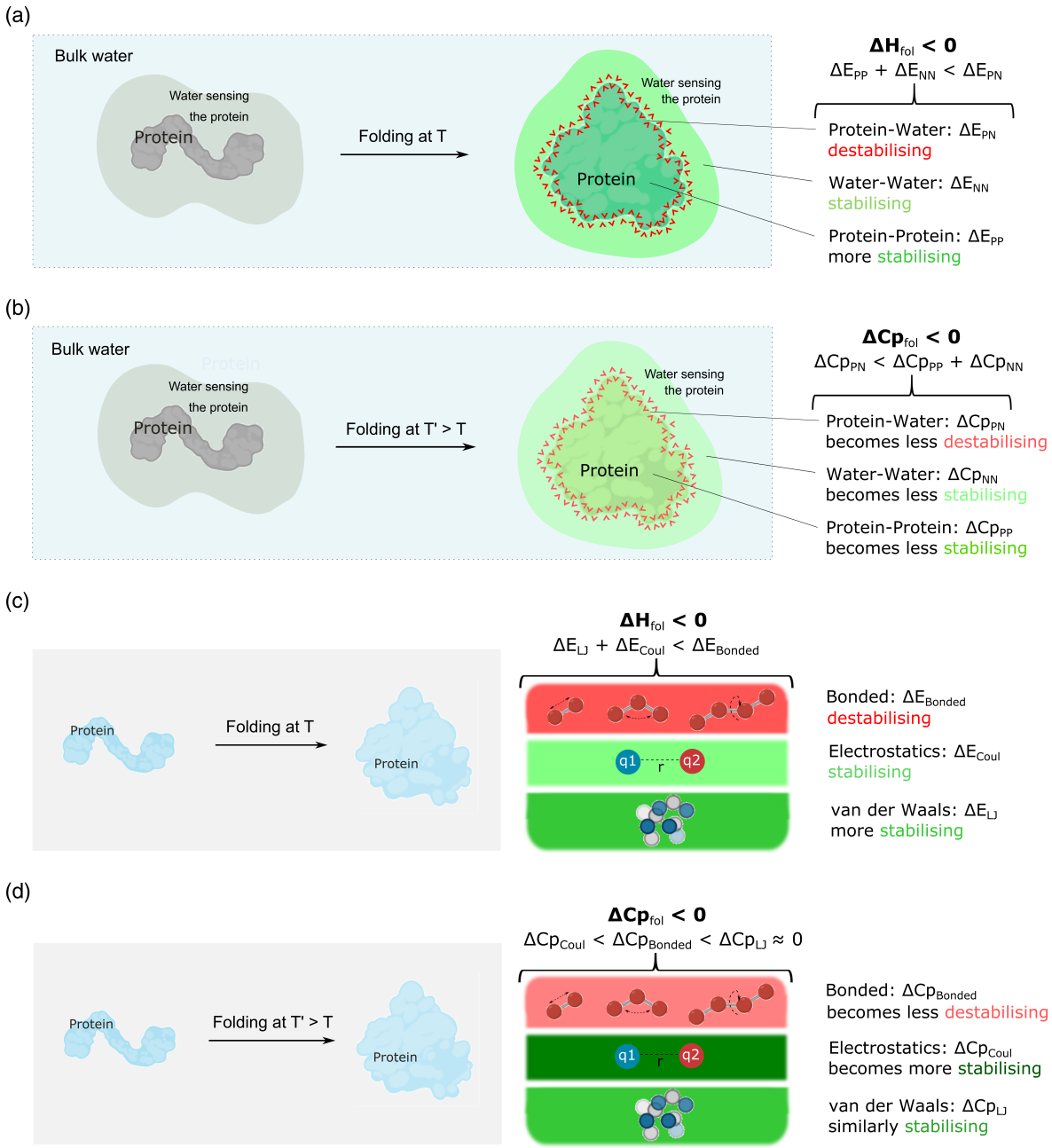

FIGURE 6 General pattern of protein folding energetics. Panels (a) and (b) illustrate the energetics of protein folding and its temperature dependence, focusing on inter and intramolecular interactions between protein and solvent (essentially water) atoms. The unfolded system at the left (protein and non-bulk surrounding water molecules sensing the protein) is taken as the energy reference and is depicted in gray. The folded system at the right is colored differently to distinguish interactions that stabilize (water–water and protein–protein interactions; green shades) or destabilize (protein–water interactions, red lines) the system and therefore the folded conformation relative to the unfolded one. The temperature effect on the different molecular interactions shaping ΔH_{fol} determines how much they contribute to ΔCp_{fol}. The temperature effects can be understood by comparing the comments at the right-hand side of panels (a) (lower temperature) and (b) (higher temperature). Relative values of the different molecular contributions to ΔH_{fol} and ΔCp_{fol} are indicated qualitatively. Panels (c) and (d) illustrate the energetics of protein folding, focusing on elementary contributions (bonded, van der Waals, and electrostatic) without differentiating between protein and solvent atoms. In these panels, the interactions stabilizing the folded state (van der Waals and electrostatic) are represented in green strips while the destabilizing contribution of bonded interactions arising from the protein covalent structure is indicated in the red strip on top of the green ones. The contribution of the different physical interactions to ΔCp_{fol} can be understood by comparing the comments at the right-hand side of panels (c) (lower temperature) and (d) (higher temperature). Relative values of the different contributions to ΔH_{fol} and ΔCp_{fol} are indicated qualitatively.

conformations are conformationally strained. Interestingly, Lazaridis and Karplus concluded that the stabilizing van der Waals contribution to $\Delta H_{\text{fol(in-vacuum)}}$ is consistently larger than that of Coulomb (0.7 vs. 0.3, for the proteins in their study). As explained, our partition of contributions by force does not refer to energy changes in the polypeptide but to changes in the whole system for which experimental data is available (polypeptide plus solvent). Nevertheless, we similarly see that the stabilizing contribution of van der Waals to ΔH_{fol} in the proteins we have analyzed is consistently larger than that of Coulomb (0.76 vs. 0.24). In both analyses, consistent patterns are observed in different protein sets for the relative contribution of van der Waals and Coulomb interactions to either $\Delta H_{\text{fol(in-vacuum)}}$ or to ΔH_{fol}. Reflecting on different proposals about the main contributions to protein stability (ΔG), Lazaridis and Karplus concluded from their study that hydrophobic interactions were the main contributors. Our analysis does not try to respond to the question of whether hydrogen bonds, the hydrophobic effect or both are the major contributors to protein stability because on the one hand, hydrogen bonds are just a part of the Coulomb interaction and, on the other hand, the hydrophobic effect refers to a complex phenomenon associated to entropy changes, which are not dealt with here. What our analysis makes clear from the perspective of forces is that the overall changes in van der Waals interactions (ΔE^{LJ}) are the dominant contributions to the enthalpic stabilization of folded proteins, followed by Coulomb interactions (ΔE^{Coul}) all together, and that their combined action on the system introduces strain in the folded state. A modest contribution of H-bonds to ΔE^{Coul} is indirectly suggested by the fact that, unlike ΔE^{LJ} and ΔE^{Bonded}, ΔE^{Coul} is not correlated with protein length ($R^2 = 0.004$; Table 3), whereas the number of H-bonds in the native conformation (barnase, 92; SNase, 127; CI2, 50 and apoFld, 152) is ($R^2 = 0.995$).

That van der Waals interactions (ΔE^{LJ}) are the major stabilizing contribution to the enthalpic stabilization of proteins might be interpreted as a suggestion that protein folding promotes a tighter packing of the system. Related to this, the packing density of folded proteins has been the object of substantial modeling over the years without a clear agreement having been reached on the extent to which folded proteins are more or similarly packed than amino acids in crystals or in solution. In fact, substantially different packing density has been reported for folded proteins depending on the method or the database used (Gaines et al., 2016; Harpaz et al., 1994; Richards, 1974). Whatever its exact value, the tight packing of folded proteins should certainly make an important contribution to reduce in vacuum the energy of the folded polypeptide relative to that of the unfolded one (in agreement with the calculations of $\Delta H_{\text{fol(in-vacuum)}}$ by Karplus and Lazaridis discussed above), but such energy difference—that may be similar to ΔE^{LJ}(PP)—is just one of the several components (PP, PN, and NN) of ΔE^{LJ}, and it does not reveal by itself the overall contribution (sign) of ΔE^{LJ} to protein stability. Experimentally, when a protein unfolds at atmospheric pressure, there appears to be a small reduction in the volume of the solution ($\Delta V_{\text{unf}} < 0.5\%$) rather than an increase (Brandts et al., 1970; Hawley, 1971; Roche & Royer, 2018; Zipp & Kauzmann, 1973). In our simulations, the change in solution volume upon folding for the four proteins simulated is always very small (Table S9). The experimental observations that the volume of protein solutions slightly increases upon folding means that the global packing density of solutions containing folded protein is lower, even if marginally lower, than that of solutions containing unfolded protein. Thus, interpretation of the negative ΔE^{LJ} of folding as arising from a tighter packing of the folded system would be misleading. As pointed out earlier: "the chemical environment of residues in solution and in protein interiors is very different, so the volumes they occupy in the two states cannot simply be related to energies" (Harpaz et al., 1994). Our calculations indicate that, globally, the van der Waals interactions among all protein and solvent atoms are stronger when the protein is folded, despite the slightly lower density experimentally observed for solutions containing folded protein.

That the native conformation is greatly destabilized by protein–solvent interactions is clearly indicated by the high, positive value of the ΔE_{PN} term, which is due to the higher stabilization of the unfolded conformation upon solvation, by both van der Waals and Coulomb interactions (Table S2). In the value of ΔE_{PN}, both extensive and intensive contributions can concur, as it seems to be the case. On average the interaction energy for the folded and unfolded conformations of the four proteins simulated are -1.7 ± 0.4 and -1.1 ± 0.2 kJ/mol·Å^2, respectively, referred to total SASA. Thus, one folded SASA unit appears to interact better with water than one unfolded SASA unit. Considering only the Coulomb contribution to E_{PN}, which is the larger one, and taking into account only the exposed polar surface area (SASA_{pol}), a similar energy bias is seen between the folded and the unfolded polypeptide: -3.8 ± 0.2 versus -2.8 ± 0.2 kJ/mol·Å^2, respectively. Therefore, ΔE_{PN} is destabilizing because a more intense interaction of individual water molecules with the folded surface is counteracted by the higher number of water-protein interactions in the unfolded conformation, due to its much larger surface. Why water interacts more strongly with the folded surface (even if only the Coulomb interaction and the polar surface are counted) may obey to different reasons. To

obtain some hints, we have done an exploratory analysis of the protein solvent interface. First, the residence time of individual water molecules at the protein interface (at ≤5.0 Å from the surface) has been determined for the folded and the unfolded conformations from a fitting of the fraction of initial waters still present in the shell versus time (Figure S3). The average diffusion constant for the folded and unfolded conformations of the four proteins are quite similar: -17.8 ± 2.0 versus -17.4 ± 1.3 ns^{-1} (Table S10) but the fraction of molecules still present at "infinite time" in folded and unfolded conformations differs, being 0.012 ± 0.005 and 0.005 ± 0.001, respectively. We interpret that those waters are likely tightly bound to the protein surface. However, although the fraction of waters bound to the native surfaces is higher, the actual numbers for a given protein are similar, at most differing by one. Second, we have calculated radial density functions of waters around the folded and unfolded proteins. Peaks at around 2.7 and 6.0 Å are found for both the folded and unfolded conformations that may correspond to the first and second solvation shells (Table S11 and Figures S4–S7). Depending on the protein atom type, the maximal water density of the first shell appears with slightly different radius (2.72 for O_w-O_p; 2.86 for O_w-N_p and 3.54 for O_w-C_p) but no significant differences are observed between the folded and unfolded conformations. Third, we have calculated the number of water molecules per SASA for the folded and unfolded conformations. For the four proteins, the folded conformation is surrounded by a higher number of water molecules per SASA unit: on average 0.034 ± 0.005 versus 0.027 ± 0.003, respectively. This higher density of water molecules around the folded surface is expected to contribute to the observed higher E_{PN} interaction energy per SASA unit of the folded state.

Finally, that the native conformation is stabilized by the balance of solvent–solvent interactions (ΔE_{NN}), which in our simulations are essentially water–water interactions, bears on the known stabilization of proteins exerted by cosolvents that increase the surface tension of water and on the destabilization exerted by denaturants that preferentially bind to the unfolded state (Timasheff, 2003). The simplest explanation of the negative value of ΔE_{NN} of folding is that, when the protein folds, the number of NN interactions increases in correspondence with the decrease in the number of PN interactions associated to the reduced SASA of the folded state. The interactions between water molecules are essentially coulombic and strong, and they are manifested in the high surface tension of water. Many solutes such as sugars, non-hydrophobic amino acids and certain salts are known to increase the surface tension of water. Those solutes are excluded from the protein surface and are not expected to greatly modify the indicated change in number of NN interactions that contributes to ΔE_{NN}. However, they will strengthen those interactions, increasing the magnitude of ΔE_{NN} and thus stabilizing the folded conformation. In contrast, denaturants such as urea of guanidine hydrochloride are known to preferentially bind to the denatured state and therefore their destabilizing effect on proteins can be explained in terms of PN interactions (i.e., they make ΔE_{PN} more positive, which means more destabilizing). Additionally, the interaction of denaturant molecules with the unfolded polypeptide might reduce the increase in water–water interactions upon folding thus making the stabilizing contribution of ΔE_{NN} smaller, which might also contribute to destabilize folded proteins.

As stated in the Introduction, we deal here exclusively with dissecting molecular and physical contributions to ΔH and ΔCp. From the accurately MD-calculated ΔH and ΔCp values (Galano-Frutos & Sancho, 2019; Galano-Frutos et al., 2023) that are here dissected, the ΔG (stabilities) of the corresponding proteins can be calculated when the T_m is known (Figure 1c). Deriving the values of ΔS from those of ΔH and ΔG would be trivial, but no insight would be obtained. We leave the task of dissecting the molecular contributions to ΔS to future work.

3.3 | Why ΔCp_{fol} is negative?

As indicated by the negative sign of ΔCp_{fol} (Prabhu & Sharp, 2005), raising the temperature increases the enthalpic stabilization of proteins. In the four dissected proteins, there is a consistent cancelation of molecular contributions to ΔCp_{fol}. On one hand, the contributions from intraproteic and intrasolvent interactions (ΔCp_{PP} and ΔCp_{NN}) are positive, meaning that, as the temperature is raised, ΔE_{PP} and ΔE_{NN} stabilize the native conformation less and less (Figure 6b). The relative impact on ΔCp_{fol} of these contributions can be quantitatively described as:

$$\Delta Cp_{NN} = 0.92(\pm 0.04) \times \Delta Cp_{PP}. \quad (3)$$

On the other hand, the positive contribution of intraproteic and intrasolvent interactions is more than offset by a larger negative contribution of ΔCp_{PN}. This negative contribution reflects that the weakening of the folded state by protein–solvent interactions, ΔE_{PN}, is markedly reduced at higher temperatures. As it appears (Figure 6b), increasing temperature weakens both the stabilizing and destabilizing molecular contributions to ΔH_{fol}, but it weakens the destabilizing ones more pronouncedly:

$$\Delta Cp_{PN} = -1.35(\pm 0.01) \times (\Delta Cp_{PP} + \Delta Cp_{NN}). \quad (4)$$

The overall consequence is that enthalpy stabilizes the folded state more at higher temperatures. The joint contribution of the two solvation terms to ΔCp_{fol} (negative ΔCp_{PN} + positive ΔCp_{NN}) is negative, but the sign of ΔCp_{fol} is determined specifically by protein–solvent interactions.

From the perspective of the acting forces (Figure 6d), the contributions of van der Waals, electrostatic and bonded interactions to ΔCp_{fol} are dissimilar. Van der Waals interactions show little temperature dependence and, therefore, their contribution to ΔCp_{fol} is very small. Clearly, the largest contribution to the negative sign of ΔCp_{fol} comes from electrostatic interactions, while bonded interactions also contribute significantly. Thus, raising the temperature markedly increases the stabilizing electrostatic interactions and decreases the conformational strain of the folded conformation. These two effects combine to increase the enthalpic stabilization of the native state at higher temperatures. The sign of the folding heat capacity is determined by electrostatic interactions.

The negative sign of ΔCp_{fol} has been traditionally attributed to desolvation of apolar groups upon folding, as their solvation is characterized by a positive ΔCp (Prabhu & Sharp, 2005). Mechanistically, this fact has been attributed to an increase of low-angle hydrogen bonds—which have a larger energy fluctuation and therefore higher Cp—among water molecules surrounding apolar compounds (Makhatadze & Privalov, 1995). We have calculated the angular distribution of all putative hydrogen bonds formed between water molecules within 5 Å of the protein surface in the folded and unfolded conformations (Figure S8) and compared it to that in bulk water. In both folded and unfolded conformations the probability of low-angle hydrogen bonds is higher than in the bulk, but no difference is observed between folded and unfolded proteins. However, as the number of concerned waters is bigger in the unfolded conformation, the indicated water ordering effect by the protein surface will contribute to the negative sign of ΔCp_{fol}. On the other hand, the relative contribution of protein–protein interactions and solvation interactions to ΔCp_{fol} has been largely debated (Prabhu & Sharp, 2005). Our analysis clearly indicates that intraproteic interactions contribute significantly to ΔCp_{fol}, albeit less, and in opposite direction, than the overall solvation interactions.

4 | CONCLUSIONS

Atomistic MD simulations of native globular proteins and their unfolded ensembles in explicit solvent reveal a consistent pattern of enthalpy and heat capacity changes that take place when a polypeptide folds. The native conformation is enthalpically stabilized by similarly intense contributions from internal protein–protein and solvent–solvent interactions, and destabilized by protein–solvent interactions, which are qualitatively weaker but quantitatively stronger in the unfolded conformation. Van de Waals interactions primarily, but also electrostatic interactions, stabilize the native conformation at the expense of introducing conformational strain. The sign of the heat capacity change is determined by the temperature dependence of protein–solvent interactions or, from the perspective of elementary contributions, by the temperature dependence of electrostatic interactions. Incidentally, changes in folding heat capacity but not in folding enthalpy appear to correlate well with protein length and with changes in protein solvent exposure due to folding. Here, we have shown how MD simulations can be used to calculate and dissect the enthalpy changes that contribute to protein stability. The information contained in atomistic MD simulations of proteins should also allow in the not-too-distant future the accurate calculation and dissection of the entropy changes, finally enabling the calculation and deep understanding of the free energy of folding. To help achieve this important goal, any improvement of existing force fields used for protein simulation should always include both structural and thermodynamic benchmarking.

5 | MATERIALS AND METHODS

5.1 | Proteins simulated and structural models used

Four proteins for which extensive reliable thermodynamic data of their folding/unfolding equilibria are available were simulated: ribonuclease from *Bacillus amiloliquefaciens* (barnase) (Hartley & Barker, 1972), *Staphylococcus aureus* nuclease (*SNase*) (Shortle, 1983) apoflavodoxin from *Anabaena* PCC 7119 (apoFld) (Fillat et al., 1990) and chymotrypsin inhibitor 2 from barley (CI2) (Svendsen et al., 1980). Their native ensembles were represented by X-ray structures (PDB IDs: 1A2P [Martin et al., 1999], 2SNS [Cotton et al., 1979], 1FTG [Genzor et al., 1996], and 2CI2 [McPhalen & James, 1987], respectively) with crystallographic resolutions of 1.5, 1.5, 2.0, and 2.0 Å, respectively. Their unfolded ensembles were obtained using the ProtSA server (Estrada et al., 2009), which generates from an input sequence a large ensemble of unfolded conformations consistent with NMR and SAXS properties of fully unfolded proteins. For each of the proteins indicated, initial unfolded ensembles of more than 2000 conformations

were generated, from which the most elongated structures (around 10%) were discarded. Then, 100 conformations were selected at random.

5.2 | MD simulation setup and sampling of folding energetics

The main analysis presented here is based on recently performed all-atom MD simulations (Galano-Frutos et al., 2023) of folded and unfolded conformations run with Charmm22-CMAP (Mackerell et al., 2004) force field and solvated in Tip3p (Jorgensen et al., 1983) explicit water. Results from additional simulations performed to test the most recent version of the Charmm force field series (Charmm36m [Huang et al., 2017]) and a polarizable force field (Drude [Lemkul et al., 2016]) are discussed, and comparisons are done with those from some of the force field/water model setups used in our previous work (Galano-Frutos & Sancho, 2019). A summary of the simulation setup and solvation conditions is given in Table S12. The Charmm22 force field with CMAP correction version 2.0 (Mackerell et al., 2004) (Charmm22-CMAP) was used because it results in folding enthalpies and heat capacities that match the experimental ones (Galano-Frutos & Sancho, 2019; Galano-Frutos et al., 2023). For each protein, short individual trajectories (2 ns) of 40 replicas of the folded conformation and 100 different unfolded conformations were obtained. Except for the simulations run with the Drude force field (see below) all were launched with the Gromacs 2020 package (Van Der Spoel et al., 2005). Each contribution from specific molecular or physical interactions to the energetics of each conformation was time-averaged. Then, the energetics of the folded state was obtained as the average of the 40 folded replicas and, likewise, the energetics of the unfolded state was obtained as the average of the 100 unfolded conformations. Averaging of short trajectories is an efficient strategy to ensure efficient sampling of the unfolded ensemble while avoiding protein compaction (Galano-Frutos & Sancho, 2019; Galano-Frutos et al., 2023).

From the simulations of the folded and unfolded conformations of a given protein, the ΔH of folding (ΔH_{fol}) was calculated (Galano-Frutos & Sancho, 2019; Galano-Frutos et al., 2023) as the averaged enthalpy of the 40 simulated boxes containing folded protein minus that of the 100 boxes containing unfolded conformations. In the simulations of barnase with the Drude force field, 20 folded and 40 unfolded boxes were used. For the ΔH_{fol} calculation by difference to be feasible, the boxes with folded protein must contain the same number of water molecules and counterions as the corresponding boxes with unfolded protein (Figure 2). Except for the Drude simulations of barnase (run only at 42°C), ΔH_{fol} was calculated at three temperatures and ΔCp_{fol} values were obtained as the slope of a representation of calculated ΔH_{fol} versus simulation temperature. Importantly, for all four proteins simulated under the Charmm22-CMAP/Tip3p setup, the calculated ΔH_{fol} and ΔCp_{fol} values agree with those determined experimentally, and the ΔG_{fol} values obtained from the calculated ΔH_{fol} and ΔCp_{fol} values plus the experimental melting temperature (Becktel & Schellman, 1987) also agrees with the experimentally determined conformational stabilities (Galano-Frutos et al., 2023) (see Figure 1).

Simulations of barnase to test the accuracy of the Charmm36m force field (with Tip3p [Jorgensen et al., 1983]) were setup with the same solvation conditions and MD input parameters reported in our previous work (Galano-Frutos et al., 2023). In the case of simulations run with the Drude polarizable force field (2019h master version for proteins) (Lemkul et al., 2016), they were launched with the NAMD3 simulation package (Phillips et al., 2005) because the method is still not fully or reliably implemented in Gromacs (according to the developers). The polarizable four-site water model SWM4–NDP (Lamoureux et al., 2006) was used. The numbers of water molecules and ions with which the simulated boxes were filled are indicated in Table S12 (volume averages of the folded and unfolded boxes are also indicated in Table S9). The isobaric-isothermal (NPT) ensemble was imposed to the simulations by applying a dual Langevin thermostat (Langevin dynamics) (Jiang et al., 2011) and a modified Nosé–Hoover barostat to control temperature and pressure fluctuations, respectively (Phillips et al., 2005). The rest of important input parameters used with NAMD3 for the Drude simulations appear described in Table S13.

Holonomic constraints were applied to all bonds across the MD simulations in all the systems analyzed. To evaluate the influence of applying bond constraints on the different energy terms analyzed, barnase was also simulated by imposing constraints only on bonds involving hydrogen atoms (see in next section). Finally, the values of the energy terms were printed out every 0.1 ps along the simulated trajectories in all the systems studied.

5.3 | Energy partitions provided by the force fields and terminology used for grouped terms

The Charmm22-CMAP and Charmm36m force fields use five energy terms to model the energy arising from the

covalent structure: bonds, angles with Urey-Bradley correction, proper dihedral, CMAP-dihedral correction, and improper dihedral. In this work, we report the energies corresponding to all those terms grouped in a single energy term named "Bonded" (Table 2). Since all bond distances were constrained in our simulations, the value of the bonds energy term did not change with time nor did it differ between the folded and unfolded conformations. To assess the influence that constraining bonds energy had on the values obtained for the other energy terms in the force field, we compared barnase simulations in which holonomic constraints were applied to all bonds (Table 1) with simulations where only bonds involving hydrogen atoms were constrained (Table S14). The comparison made it clear that, applying either of these two constraints configurations, the calculated contribution to folding enthalpy of the different bonded and nonbonded energy terms is not significantly altered (Table 1 and Table S14).

Charmm models the nonbonded interactions between protein atoms, between solvent molecules and between protein and solvent molecules by means of Lennard-Jones and Coulomb energy terms. These nonbonded interactions are computed distributed in six terms: LJ-14, Coul-14, LJ-SR, Coul-SR, Disp-corr, and Coul-recip (see detailed descriptions in Table 2). Here, we present them grouped in two terms: "LJ" (encompassing LJ-14, LJ-SR, and Disp-corr) and "Coul" (encompassing Coul-14, Coul-SR, and Coul-recip) (Table 2). The time-averaged enthalpy (H) of a simulated trajectory containing protein and explicit water is obtained by summing up the Bonded, LJ, and Coul terms (together making up the potential energy, E), the kinetic energy (E^{kin}), and the pV term (p = pressure, V = volume). The change in ΔH_{fol} can then be obtained as the average enthalpy corresponding to the boxes containing folded protein minus that of those containing unfolded protein (Galano-Frutos & Sancho, 2019; Galano-Frutos et al., 2023), provided that the number of water molecules and counterions is the same in all boxes.

Alternatively, the energy calculated from the simulations can be grouped differently to highlight the separate contribution of intraproteic (protein–protein), intrasolvent (solvent–solvent), and protein–solvent interactions. Indeed, the Gromacs' (Van Der Spoel et al., 2005) output provides the following terms: Coul-SR-PP, LJ-SR-PP, Coul-14-PP, LJ-14-PP, Coul-SR-PN, LJ-SR-PN, Coul-SR-NN, LJ-SR-NN, where PP refers to interactions between protein atoms, PN to interactions between protein and non-protein atoms (water molecules and ions, as the simulated proteins contain no cofactors), and NN to intrasolvent interactions. We report here the different PP terms grouped as "$Coul_{PP}$" (Coul-SR-PP + Coul-14-PP) and "LJ_{PP}" (LJ-SR-PP + LJ-14-PP). Thus, the potential energy due to intraproteic interactions (E_{PP}) is obtained as $Coul_{PP} + LJ_{PP}$ + Bonded, that corresponding to protein–solvent interactions (E_{PN}) is Coul-SR-PN + LJ-SR-PN, and that related to solvent–solvent interactions (E_{NN}) would be, in principle, Coul-SR-NN + LJ-SR-NN (Table 2). However, two additional terms, Disp-corr and Coul-recip, also contribute to the potential energy of the simulated box. These terms combine PP, PN, and NN contributions, but arise essentially from solvent–solvent interactions, which account for >98% of the pairwise interactions taking place in the simulated box. Thus, the contributions of the Disp-corr and Coul-recip terms have been added to those of Coul-SR-NN + LJ-SR-NN to jointly describe E_{NN}. It is important to note that ignoring the Disp-corr and Coul-recip terms in our analysis would make very little difference, since their contribution to ΔE_{NN}, the change in solvent–solvent interaction energy upon folding, is very small.

The energy terms in the Amber force fields are quite similar to those in Charmm, which facilitates pairwise comparisons of the values obtained for the same protein using either force field. The only difference is that Amber does not include CMAP dihedral correction, so it uses only four terms (bonds, angles, proper dihedral, and improper dihedral) to model the energy derived from the covalent structure. In the analysis of barnase simulations run with Amber, those four terms are reported grouped in a single energy term named "Bonded" (Table 2).

AUTHOR CONTRIBUTIONS

Juan José Galano-Frutos: Writing – review and editing; methodology; software; data curation; investigation; validation; formal analysis. **Javier Sancho:** Conceptualization; writing – original draft; writing – review and editing; project administration; resources; funding acquisition; supervision; data curation; investigation; validation; formal analysis.

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.

FUNDING INFORMATION

This work was supported by grants PID2019-107293GB-I00, PDC2021-121341-I00, and PID2022-141068NB-I00 (MICINN, Spain) and E45_23R (Gobierno de Aragón, Spain).

CONFLICT OF INTEREST STATEMENT

Juan José Galano-Frutos and Javier Sancho declare that they have no conflict of interest.

ORCID

Juan José Galano-Frutos https://orcid.org/0000-0002-1896-7805
Javier Sancho https://orcid.org/0000-0002-2879-9200

REFERENCES

Anfinsen CB. Principles that govern the folding of protein chains. Science. 1973;181:223–30.
Baker D. What has de novo protein design taught us about protein folding and biophysics? Protein Sci. 2019;28:678–83.
Baldwin RL. Temperature dependence of the hydrophobic interaction in protein folding. Proc Natl Acad Sci U S A. 1986;83:8069–72.
Becktel WJ, Schellman JA. Protein stability curves. Biopolymers. 1987;26:1859–77.
Best RB. Atomistic molecular simulations of protein folding. Curr Opin Struct Biol. 2012;22:52–61.
Best RB. Analysis of molecular dynamics simulations of protein folding. Methods Mol Biol. 2022;2376:317–29.
Best RB, Zheng W, Mittal J. Balanced protein-water interactions improve properties of disordered proteins and non-specific protein association. J Chem Theory Comput. 2014;10:5113–24.
Bottaro S, Lindorff-Larsen K. Biophysical experiments and biomolecular simulations: a perfect match? Science. 2018;361:355–60.
Brandts JF, Oliveira RJ, Westort C. Thermodynamics of protein denaturation. Effect of pressure on the denaturation of ribonuclease A. Biochemistry. 1970;9:1038–47.
Campos LA, Cuesta-López S, López-Llano J, Falo F, Sancho J. A double-deletion method to quantifying incremental binding energies in proteins from experiment: example of a destabilizing hydrogen bonding pair. Biophys J. 2005;88:1311–21.
Chan-Yao-Chong M, Chan J, Kono H. Benchmarking of force fields to characterize the intrinsically disordered R2-FUS-LC region. Sci Rep. 2023;131(13):1–12.
Chen SJ, Hassan M, Jernigan RL, Jia K, Kihara D, Kloczkowski A, et al. Protein folds vs. protein folding: differing questions, different challenges. Proc Natl Acad Sci U S A. 2023;120: e2214423119.
Chin JW. Expanding and reprogramming the genetic code. Nature. 2017;550:53–60.
Cotton FA, Hazen EE, Legg MJ. Staphylococcal nuclease: proposed mechanism of action based on structure of enzyme—thymidine 3′,5′-bisphosphate—calcium ion complex at 1.5-Å resolution. Proc Natl Acad Sci U S A. 1979;76:2551–5.
Creighton TE. Stability of folded conformations: current opinion in structural biology 1991, 1: 5–16. Curr Opin Struct Biol. 1991;1: 5–16.
Cui X, Liu H, Rehman AU, Chen HF. Extensive evaluation of environment-specific force field for ordered and disordered proteins. Phys Chem Chem Phys. 2021;23:12127–36.
Dehouck Y, Kwasigroch JM, Gilis D, Rooman M. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinform. 2011;12: 1–12.
Dill KA. Dominant forces in protein folding. Biochemistry. 1990;29: 7133–55.
Dill KA, MacCallum JL. The protein-folding problem, 50 years on. Science. 2012;338:1042–6.
Estrada J, Bernadó P, Blackledge M, Sancho J. ProtSA: a web application for calculating sequence specific protein solvent accessibilities in the unfolded ensemble. BMC Bioinform. 2009;10:1–8.
Ferina J, Daggett V. Visualizing protein folding and unfolding. J Mol Biol. 2019;431:1540–64.
Fillat MF, Edmondson DE, Gomez-Moreno C. Structural and chemical properties of a flavodoxin from Anabaena PCC 7119. Biochim Biophys Acta – Protein Struct Mol Enzymol. 1990;1040: 301–7.
Gaines JC, Smith WW, Regan L, O'Hern CS. Random close packing in protein cores. Phys Rev E. 2016;93:032415.
Galano-Frutos JJ, Nerín-Fonz F, Sancho J. Calculation of protein folding thermodynamics using molecular dynamics simulations. J Chem Inf Model. 2023;63:24:7791–806.
Galano-Frutos JJ, 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;59:4350–60.
García-Cebollada H, López A, Sancho J. Protposer: the web server that readily proposes protein stabilizing mutations with high PPV. Comput Struct Biotechnol J. 2022;20:2415–33.
Genzor CG, Perales-Alcon A, Sancho J, Romero A. Closure of a tyrosine/tryptophan aromatic gate leads to a compact fold in apo flavodoxin. Nat Struct Biol. 1996;3:329–32.
Gómez J, Hilser VJ, Xie D, Freire E. The heat capacity of proteins. Proteins Struct Funct Bioinform. 1995;22:404–12.
Harpaz Y, Gerstein M, Chothia C. Volume changes on protein folding. Structure. 1994;2:641–9.
Hartley RW, Barker EA. Amino-acid sequence of extracellular ribonuclease (Barnase) of Bacillus amyloliquefaciens. Nat New Biol. 1972;235:15–6.
Hawley SA. Reversible pressure-temperature denaturation of chymotrypsinogen. Biochemistry. 1971;10:2436–42.
Hilser VJ, Gómez J, Freire E. The enthalpy change in protein folding and binding: refinement of parameters for structure-based calculations. Proteins. 1996;26:123–33.
Horovitz A. Double-mutant cycles: a powerful tool for analyzing protein structure and function. Fold Des. 1996;1:R121–6.
Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, De Groot BL, et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins HHS public access. Nat Methods. 2017;14:71–3.
Huang PS, Boyken SE, Baker D. The coming of age of de novo protein design. Nature. 2016;537:320–7.
Jiang W, Hardy DJ, Phillips JC, MacKerell AD, Schulten K, Roux B. High-performance scalable molecular dynamics simulations of a polarizable force field based on classical Drude oscillators in NAMD. J Phys Chem Lett. 2011;2:87–92.
Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. J Chem Phys. 1983;79:926–35.
Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9.

Kamenik AS, Handle PH, Hofer F, Kahler U, Kraml J, Liedl KR. Polarizable and non-polarizable force fields: protein folding, unfolding, and misfolding. J Chem Phys. 2020;153:185102.

Lamoureux G, Harder E, Vorobyov IV, Roux B, MacKerell AD. A polarizable model of water for molecular dynamics simulations of biomolecules. Chem Phys Lett. 2006;418:245–9.

Lazaridis T, Archontis G, Karplus M. Enthalpic contribution to protein stability: insights from atom-based calculations and statistical mechanics. Adv Protein Chem. 1995;47:231–306.

Lazaridis T, Karplus M. Thermodynamics of protein folding: a microscopic view. Biophys Chem. 2002;100:367–95.

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;104:4925–30.

Lemkul JA, Huang J, Roux B, Mackerell AD. An empirical polarizable force field based on the classical drude oscillator model: development history and recent applications. Chem Rev. 2016; 116:4983–5013.

Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. How fast-folding proteins fold. Science. 2011;334:517–20.

Liu Y, Kuhlman B. RosettaDesign server for protein design. Nucleic Acids Res. 2006;34:W235–8.

Mackerell AD. Empirical force fields for biological macromolecules: overview and issues. J Comput Chem. 2004;25:1584–604.

Mackerell AD, Feig M, Brooks CL. 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;25:1400–15.

Madan B, Sharp K. Heat capacity changes accompanying hydrophobic and ionic solvation: a Monte Carlo and Random Network Model Study. J Phys Chem. 1996;100:7713–21.

Makhatadze GI, Privalov PL. Heat capacity of proteins: I. Partial molar heat capacity of individual amino acid residues in aqueous solution: hydration effect. J Mol Biol. 1990;213:375–84.

Makhatadze GI, Privalov PL. Energetics of protein structure. Adv Protein Chem. 1995;47:307–425.

Martin C, Richard V, Salem M, Hartley R, Mauguen Y. Refinement and structural analysis of barnase at 1.5 A resolution. Acta Crystallogr D Biol Crystallogr. 1999;55:386–98.

McPhalen CA, James MN. Crystal and molecular structure of the serine proteinase inhibitor CI-2 from barley seeds. Biochemistry. 1987;26:261–9.

Moore PB, Hendrickson WA, Henderson R, Brunger AT. The protein-folding problem: not yet solved. Science. 2022;375:507.

Murphy KP, Gill SJ. Solid model compounds and the thermodynamics of protein unfolding. J Mol Biol. 1991;222:699–709.

Murphy KP, Privalov PL, Gill SJ. Common features of protein unfolding and dissolution of hydrophobic compounds. Science. 1990;247:559–61.

Newberry RW, Raines RT. Secondary forces in protein folding. ACS Chem Biol. 2019;14:1677–86.

Nikam R, Kulandaisamy A, Harini K, Sharma D, Michael Gromiha M. ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. Nucleic Acids Res. 2021;49:D420–4.

Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, et al. Scalable molecular dynamics with NAMD. J Comput Chem. 2005;26:1781–802.

Piana S, Donchev AG, Robustelli P, Shaw DE. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. J Phys Chem B. 2015;119:5113–23.

Piana S, Klepeis JL, Shaw DE. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. Curr Opin Struct Biol. 2014;24:98–105.

Piana S, Lindorff-Larsen K, Shaw DE. Protein folding kinetics and thermodynamics from atomistic simulation. Proc Natl Acad Sci U S A. 2012;109:17845–50.

Piana S, Robustelli P, Tan D, Chen S, Shaw DE. Development of a force field for the simulation of single-chain proteins and protein-protein complexes. J Chem Theory Comput. 2020;16: 2494–507.

Prabhu NV, Sharp KA. Heat capacity in proteins. Annu Rev Phys Chem. 2005;56:521–48.

Privalov PL, Gill SJ. Stability of protein structure and hydrophobic interaction. Adv Protein Chem. 1988;39:191–234.

Privalov PL, Makhatadze GI. Contribution of hydration and noncovalent interactions to the heat capacity effect on protein unfolding. J Mol Biol. 1992;224:715–23.

Privalov PL, Makhatadze GI. Contribution of hydration to protein folding thermodynamics: II. The entropy and Gibbs energy of hydration. J Mol Biol. 1993;232:660–79.

Ravikumar Y, Nadarajan SP, Hyeon Yoo T, Lee CS, Yun H. Unnatural amino acid mutagenesis-based enzyme engineering. Trends Biotechnol. 2015;33:462–70.

Richards FM. The interpretation of protein structures: Total volume, group volume distributions and packing density. J Mol Biol. 1974;82:1–14.

Robertson AD, Murphy KP. Protein structure and the energetics of protein stability. Chem Rev. 1997;97:1251–67.

Robustelli P, Piana S, Shaw DE. Developing a molecular dynamics force field for both folded and disordered protein states. Proc Natl Acad Sci U S A. 2018;115:E4758–66.

Roche J, Royer CA. Lessons from pressure denaturation of proteins. J R Soc Interface. 2018;15:20180244.

Sancho J. The stability of 2-state, 3-state and more-state proteins from simple spectroscopic techniques... plus the structure of the equilibrium intermediates at the same time. Arch Biochem Biophys. 2013;531:4–13.

Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic Acids Res. 2005;33:W382–8.

Shao Q, Yang L, Zhu W. Selective enhanced sampling in dihedral energy facilitates overcoming the dihedral energy increase in protein folding and accelerates the searching for protein native structure. Phys Chem Chem Phys. 2019;21:10423–35.

Shortle D. A genetic system for analysis of staphylococcal nuclease. Gene. 1983;22:181–9.

Stein A, Fowler DM, Hartmann-Petersen R, Lindorff-Larsen K. Biophysical and mechanistic models for disease-causing protein variants. Trends Biochem Sci. 2019;44:575–88.

Svendsen I, Martin B, Jonassen I. Characteristics of Hiproly barley III. Amino acid sequences of two lysine-rich proteins. Carlsberg Res Commun. 1980;45:79–85.

Timasheff SN. The control of protein stability and association by weak interactions with water: how do solvents affect these processes? Annu Rev Biophys Biomol Struct. 2003;22:67–97.

Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: fast, flexible, and free. J Comput Chem. 2005;26:1701–18.

Winkler L, Cheatham TE. Benchmarking the Drude polarizable force field using the r(GACC) Tetranucleotide. J Chem Inf Model. 2023;63:2505–11.

Xavier JS, Nguyen TB, Karmarkar M, Portelli S, Rezende PM, Velloso JPL, et al. ThermoMutDB: a thermodynamic database for missense mutations. Nucleic Acids Res. 2021;49: D475–9.

Zapletal V, Mládek A, Melková K, Louša P, Nomilner E, Jaseňáková Z, et al. Choice of force field for proteins containing structured and intrinsically disordered regions. Biophys J. 2020; 118:1621–33.

Zerze GH, Zheng W, Best RB, Mittal J. Evolution of all-atom protein force fields to improve local and global properties. J Phys Chem Lett. 2019;10:2227–34.

Zipp A, Kauzmann W. Pressure denaturation of metmyoglobin. Biochemistry. 1973;12:4217–28.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Galano-Frutos JJ, Sancho J. Energy, water, and protein folding: A molecular dynamics-based quantitative inventory of molecular interactions and forces that make proteins stable. Protein Science. 2024;33(2):e4905. https://doi.org/10.1002/pro.4905