Received: 18 July 2022 | Revised: 7 September 2022 | Accepted: 9 September 2022

DOI: 10.1002/pro.4445

FULL-LENGTH PAPER

THE PROTEIN SOCIETY WILEY

A look at the face of the molten globule: Structural model of the *Helicobacter pylori* apoflavodoxin ensemble at acidic pH

Juan José Galano-Frutos[1,2] | **Renzo Torreblanca**[1,2] | **Helena García-Cebollada**[1,2] | **Javier Sancho**[1,2,3]

[1]Departamento de Bioquímica y Biología Molecular y Celular, Facultad de Ciencias, Universidad de Zaragoza, Zaragoza, Spain

[2]Biocomputation and Complex Systems Physics Institute (BIFI)-GBsC-CSIC Joint Unit, Universidad de Zaragoza, Zaragoza, Spain

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

Correspondence

Javier Sancho, Departamento de Bioquímica y Biología Molecular y Celular, Facultad de Ciencias, Universidad de Zaragoza, Zaragoza 50009, Spain.
Email: jsancho@unizar.es

Funding information

Ministerio de Ciencia e Innovación, Grant/Award Numbers: PDC2021-121341-I00, PID2019-107293GB-I00; Gobierno de Aragón, Grant/Award Number: E45_20R; MICINN

Review Editor: Nir Ben-Tal

Abstract

Molten globule (MG) is the name given to a compact, non-native conformation of proteins that has stimulated the imagination and work in the protein folding field for more than 40 years. The MG has been proposed to play a central role in the folding reaction and in important cell functions, and to be related to the onset of misfolding diseases. Due to its inherent intractability to high-resolution studies, atomistic structural models have not yet been obtained. We present here an integrative atomistic model of the MG formed at acidic pH by the apoflavodoxin from the human pathogen *Helicobacter pylori*. This MG has been previously shown to exhibit the archetypical expansion, spectroscopic and thermodynamic features of a molten conformation. To obtain the model, we have analyzed the stability of wild-type and 55 apoflavodoxin mutants to derive experimental equilibrium Φ values that have been used in biased molecular dynamics simulations to convert the native conformation into an MG ensemble. The ensemble has been refined to reproduce the experimental hydrodynamic radius and circular dichroism (CD) spectrum. The refined ensemble, deposited in PDB-Dev, successfully explains the characteristic ^{1}H-nuclear magnetic resonance (NMR) and near-UV CD spectral features of the MG as well as its solvent-accessible surface area (SASA) change upon unfolding. This integrative model of an MG will help to understand the energetics and roles of these elusive conformations in protein folding and misfolding. Interestingly, the apoflavodoxin MG is structurally unrelated to previously described partly unfolded conformations of this protein, exemplifying that equilibrium MGs need not to reflect the properties of kinetic intermediates.

KEYWORDS

apoflavodoxin, biased MD simulations, integrative protein model, intrinsically disordered protein, molten globule, protein folding landscape

Abbreviations: BMD, biased molecular dynamics; CD, circular dichroism; IDP, intrinsically disordered protein; LEM, linear extrapolation method; MD, molecular dynamics; MG, molten globule; NMR, nuclear magnetic resonance; RC, reaction coordinate; SASA, solvent-accessible surface area; SAXS, small-angle X-ray scattering; WT, wild type.

1 | INTRODUCTION

The molten globule (MG)[1–4] is a partially unfolded conformation adopted by many natively folded proteins either transiently during the folding reaction or at equilibrium in solution conditions, where the native conformation is weakened (e.g., acidic pH or moderate denaturant concentration). Besides, many intrinsically disordered proteins (IDPs) adopt MG-like conformations at physiological conditions.[5,6] Over the years, the MGs of many proteins have been characterized using a combination of spectroscopic, hydrodynamic, thermodynamic and simulation techniques.[7–17] The consensus view of the MG describes a compact conformation rich in secondary structure but lacking the precise and persistent tertiary interactions that, in natively folded proteins, bring together distant parts of the polypeptide. In MGs, apolar residues are more exposed to the solvent than in native proteins, which explains both their convenient detection by apolar fluorescent probes and their fastidious tendency to aggregate. Although dynamic information and structural detail continues to be obtained for some classical MG-forming proteins,[18] their high conformational heterogeneity has so far precluded their crystallization and has significantly hampered nuclear magnetic resonance (NMR) studies. Despite this lack of structural information, MGs have been attributed a central role in the protein folding landscape,[4,19–29] have been shown to be involved in important cell functions,[30–34] and have been associated with human diseases.[35–41] Yet, recent reviews[18,42] stress the facts that quantitative descriptions of MG folding energetics are scarce, and that arriving to atomistic models of MGs structures remains an unaccomplished challenge.

The flavodoxin (Fld) from *Helicobacter pylori* (*Hp*), a pathogenic bacteria responsible for gastritis and gastric ulcers,[43] is an essential protein[44,45] which is being targeted for the development of *Hp*-specific antimicrobials.[46–48] The crystal structures of the holo[44] and apo[49] forms are known, and both forms have been extensively characterized in terms of protein stability and cofactor binding. In particular, the apo form[50,51] is natively folded between pH 5–9 but, as previously observed for other flavodoxins (e.g., that from *Anabaena* PCC7119),[52] it adopts an MG conformation at acidic pH. The earlier discovered MG of *Anabaena* PCC7119 apoFld aggregates[52] and only limited structural information could be obtained of it from the study of a monomeric truncated variant,[53,54] using equilibrium Φ-analysis.[55] These early studies pointed to an overall and marked debilitation of the native interactions. In advantageous contrast, the MG formed by *Hp* apoFld at acidic pH is monomeric, it has been extensively characterized,[50,51] displays a simple two-state unfolding equilibrium and constitutes a much more promising model to try and obtain detailed structural information. As for other MGs previously studied, efforts to crystallize the *Hp* apoFld MG have failed (not shown), and its ^{1}H-NMR spectrum shows typical signal broadening and low resolution.[51] However, taking advantage of the easy engineering, expression and purification of *Hp* Fld, and of the monomeric nature, simple unfolding equilibrium, and high solubility of its acidic MG, we apply here a combination of protein engineering, equilibrium Φ-analysis, and molecular simulation that allows us to propose a conformational ensemble consistent with all the available observables.[50,51] To that end, the conformational stability of wild-type (WT) *Hp* apoFld and 55 purposely designed variants has been determined in solution conditions, in which virtually all protein molecules populate either the native or the MG conformation. Then, the Φ values obtained from equilibrium Φ-analysis[53,55] have been used in biased molecular dynamics (BMD) simulations[56,57] to transform the native structure of the protein into an initial ensemble of MG conformations, which has been refined to be consistent with the known hydrodynamic radius and far UV circular dichroism (CD) spectrum of the MG, as well as with the diverse spectroscopic and thermodynamic data so far reported.[50,51] Classical unbiased molecular dynamics (MD) simulations at pH 2 have also been performed to explore the feasibility of alternatively deriving the MG structure from relaxation of the native conformation dominant at pH 7.

2 | RESULTS

2.1 | Stability of native and MG apoflavodoxin variants

The conformational stability of WT *Hp* apoFld and 55 point mutants (Table 1) has been determined at either pH 7 or pH 2 by urea denaturation, following the change in emission fluorescence in the near-UV. The unfolding curves recorded for the native conformations at pH 7 are reported in Figure S1 and the calculated stability of the variants in Table S1. The mutants were designed to break interactions present in the native state and, as expected, most of them are less stable than the WT protein. Only two mutants, I49A and Q61A, are significantly more stable. In I49A, the mutation removes a side chain in van der Waals contact with W155, one of the two tryptophan residues in the protein. In the absence of repacking, this mutation would have created a cavity of 66 Å^3, which could have destabilized the protein by around 3 kcal/mol.[58] However, the unfolding curve shows the fluorescence emission of this mutant is quenched relative to that

TABLE 1 Structural and thermodynamic description of mutations implemented in *Hp* apoFld

							ΔΔG (kcal/mol)[b]		Φ value[c]
Residue	Structure element	ΔSASA (Å^2)[a]	H-bond/salt bridge partner residue (atom)	van der Waals contacts (<4 Å)	Contacting elements	Mutation	pH 2 (MG)	pH 7 (native)	
I6	β1	0.0		A19/S23	α1	I6A	−0.23 ± 0.06	0.26 ± 0.02	[d]
F7	β1	0.0		Q41/F42/F45	α2/loop → β3	F7A	−0.23 ± 0.05	0.27 ± 0.02	[d]
T10	Loop → α1	28.9	G13(N)	F8	β1	T10A	−0.30 ± 0.05	−0.02 ± 0.02	[d]
D11	Loop → α1	97.3	Q61(NE2)	Q61	α3	D11A	0.46 ± 0.07	−0.05 ± 0.11	[e]
S12	Loop → α1	109.5	N14(N)		α1	S12A	−0.10 ± 0.04	0.08 ± 0.02	[d]
E16	α1	54.6		F8	β1	E16A	0.30 ± 0.05	−0.10 ± 0.09	[d]
I18	α1	1.1		E141/T148	Loop → α5/α5	I18A	0.45 ± 0.06	2.42 ± 0.03	0.18 ± 0.02
A19	α1	0.0		I6/V31	β1/β2	A19G	−0.24 ± 0.03	5.14 ± 0.03	[e]
E20	α1	50.4		V31	β2	E20A	0.43 ± 0.04	0.35 ± 0.04	1.23 ± 0.18
K21	α1	54.2	D149(DO1)		α5	K21A	−0.91 ± 0.07	2.07 ± 0.02	[e]
S23	α1	19.6		A29/V31	β2	S23A	−0.74 ± 0.05	−0.12 ± 0.05	[e]
A25	α1	69.5		V156	α5	A25G	0.67 ± 0.04	1.38 ± 0.07	0.48 ± 0.04
I26	α1	23.5		I4/F163	β1/α5	I26V	0.28 ± 0.06	1.57 ± 0.11	0.18 ± 0.04
E30	β2	34.0	G2(N)		Loop → s1	E30A	0.19 ± 0.08	1.23 ± 0.20	0.15 ± 0.07
S38	α2	60.2	Q41(N)	E40/Q41	α2	S38A	0.29 ± 0.06	2.13 ± 0.13	0.14 ± 0.03
E40	α2	100.4		S38/K39	α2	E40A	−0.03 ± 0.04	0.08 ± 0.11	[d]

(Continues)

TABLE 1 (Continued)

Residue	Structure element	ΔSASA (Å^2)[a]	H-bond/salt bridge partner residue (atom)	van der Waals contacts (<4 Å)	Contacting elements	Mutation	ΔΔG (kcal/mol)[b]		Φ value[c]
							pH 2 (MG)	pH 7 (native)	
N43	α2	70.6	D75(OD1)/K79(NZ)	K39/F42	α2/α3/β4	N43A	−0.28 ± 0.06	0.84 ± 0.12	[e]
F45	Loop → β3	4.6		F7/E30/S44	β1/β2/loop → s3	F45A	−0.22 ± 0.04	0.12 ± 0.08	[d]
K47	β3	27.7		T46/T80/F163/A164	β3/β4/loop → C-ter	K47A	0.26 ± 0.08	1.54 ± 0.05	0.17 ± 0.05
I49	β3	0.0		I22/V84/W155/F163	α1/β4/α5	I49A	−0.26 ± 0.07	−0.91 ± 0.09	0.29 ± 0.08
L50	β3	0.0		F42/V48/W64/F76/I81/L83/I103	α2/β3/α3/β4/α4	L50A	−0.17 ± 0.05	0.71 ± 0.02	e
T54	β3	28.3	L86(O)/G99(N)	G58/D59/A97/E98	β4/α4	T54A	−0.12 ± 0.04	2.32 ± 0.09	0
L60	Loop → α3	5.2		A52/W64/L68/I103	β3/α3/α4	L60A	−0.05 ± 0.03	2.46 ± 0.04	0
Q61	α3	1.7	D11(OD1)/D63(N/OD2)/W64(N)	P53	Loop → α1/α3/β3	Q61A	−0.42 ± 0.04	−0.82 ± 0.03	0.51 ± 0.05
D63	α3	44.4	Q61(NE2)	A35	α3/β2	D63A	−0.17 ± 0.08	1.63 ± 0.18	0
L68	α3	10.7		L60/F76/K106	Loop → α3/α3/α4	L68A	0.71 ± 0.06	3.19 ± 0.13	0.22 ± 0.02
T70	α3	92.0		L71	α3	T70A	−0.25 ± 0.04	0.87 ± 0.32	[e]
D75	α3	14.4	N43(ND2)/E72(N)		α2/α3	D75A	0.05 ± 0.06	3.63 ± 0.11	0
I81	β4	0.0		L50/F76/K79/L83	β3/α3	I81A	−0.17 ± 0.04	0.41 ± 0.03	[e]
V84	β4	0.0		I22/I49/V51/W155	α1/β3/α5	V84A	0.16 ± 0.03	4.33 ± 0.05	0
L86	β4	10.3		N14	α1	L86A	0.16 ± 0.03	3.83 ± 0.09	0
D88	Loop → α4	25.5	T91(OG1,N)/N143(ND2)/D90(N)	D142	Loop → α4/loop → α5	D88A	−0.35 ± 0.04	1.12 ± 0.12	[e]

TABLE 1 (Continued)

Residue	Structure element	ΔSASA (Å^2)[a]	H-bond/salt bridge partner residue (atom)	van der Waals contacts (<4 Å)	Contacting elements	Mutation	ΔΔG (kcal/mol)[b]		Φ value[c]
							pH 2 (MG)	pH 7 (native)	
Q89	Loop → α4	0.6	H122(N)/D140(OD2)	D90/Y121/N143/Q144	Loop → α4/ loop → α5	Q89A	−0.11 ± 0.06	1.20 ± 0.03	0
S93	Loop → α4	36.3	E124(N)	F123/E124	Loop → α5	S93A	−0.25 ± 0.06	0.63 ± 0.05	[e]
T95	Loop → α4	18.9		Y92 /E98	Loop → α4	T95A	−0.04 ± 0.06	1.50 ± 0.08	0
T95	Loop → α4	18.9		Y92 /E98		T95S	−0.15 ± 0.04	−0.07 ± 0.06	[d]
E98	Loop → α4	16.1	K127(N)	T95	Loop → α4/loop → α5	E98A	−0.14 ± 0.04	1.31 ± 0.13	0
I100	α4	0.2		V135	β6c	I100A	0.47 ± 0.04	4.01 ± 0.08	0.12 ± 0.01
Y104	α4	16.1	E130(OE1)	K108	α4/β6b	Y104F	−0.29 ± 0.04	1.59 ± 0.25	[e]
E105	α4	63.2	K108(NZ)	F101	α4	E105A	0.07 ± 0.05	0.48 ± 0.34	0.15 ± 0.15[f]
Q115	Loop → α5	37.1	V112(N)	K133	β5/β6c	Q115A	−0.14 ± 0.04	1.26 ± 0.21	0
T116	β6a	20.8		W155	α5	T116A	0.01 ± 0.05	2.72 ± 0.18	0
Y121	Loop → α5	16.6	V138(C)	F96/Y121/F134/Q144/R151	Loop → α4/loop → α5/ s6c/α5	Y121F	−0.17 ± 0.05	2.26 ± 0.06	0
S126	Loop → α5	17.7	A128(N)		β6c	S126A	−0.11 ± 0.06	3.30 ± 0.02	0
V129	β6b	47.1		F134	β6c	V129A	0.23 ± 0.12	1.87 ± 0.15	0.12 ± 0.06
E141	Loop → α5	41.7	T148(OG1)	I18/I139/D145	α1/loop → α5	E141A	−0.08 ± 0.05	1.80 ± 0.10	0
Q144	Loop → α5	12.7	R151(NH1,NH2)/G120(O)	Q89/Y121/H122	Loop → α4/loop → α5/ α5	Q144A	−0.41 ± 0.05	2.23 ± 0.15	[e]
T148	α5	2.2	E141(OE1)	I18/I139/E141/I152	α1/loop → α5	T148A	0.22 ± 0.07	3.50 ± 0.04	0

(Continues)

TABLE 1 (Continued)

							$\Delta\Delta G$ (kcal/mol)[b]		Φ value[c]
Residue	Structure element	ΔSASA (Å^2)[a]	H-bond/salt bridge partner residue (atom)	van der Waals contacts (<4 Å)	Contacting elements	Mutation	pH 2 (MG)	pH 7 (native)	
T148	α5	2.2	E141(OE1)	I18/I139/E141/I152		T148S	−0.04 ± 0.05	−0.26 ± 0.05	[d]
T148	α5	2.2	E141(OE1)	I18/I139/E141/I152		T148V	−0.07 ± 0.05	2.08 ± 0.33	0
D149	α5	84.8	K21(NZ)		α1	D149A	−0.04 ± 0.04	1.30 ± 0.10	0
R151	α5	16.4	OE1(Q144)	Y121/I139/Q144/L147	Loop/loop → α5	R151A	−0.72 ± 0.07	3.24 ± 0.15	[e]
I152	α5	5.8		T148	α5	I152A	−0.05 ± 0.04	0.75 ± 0.07	0
V156	α5	10.1		A25/I26	α1	V156A	0.09 ± 0.04	3.05 ± 0.11	0
V159	α5	1.3		W155/F163	α5	V159A	0.01 ± 0.05	4.55 ± 0.03	0

Abbreviations: MG, molten globule; SASA, solvent-accessible surface area.

[a] ΔSASA difference calculated with the ProtSA server[59] as the SASA in the folded structure minus the mean SASA in the unfolded ensemble.

[b] $\Delta\Delta G$ errors calculated as $\delta\Delta\Delta G = \sqrt{(\delta\Delta G_{\mathrm{wt}})^2 + (\delta\Delta G_{\mathrm{mut}})^2}$.

[c] Φ values calculated as $\Delta\Delta G^{\mathrm{MG}}/\Delta\Delta G^{\mathrm{native}}$. Errors calculated as: $\delta(\Phi) = \Phi \times \sqrt{\left(\frac{\delta\Delta\Delta G^{\mathrm{MG}}}{\Delta\Delta G^{\mathrm{MG}}}\right)^2 + \left(\frac{\delta\Delta\Delta G^{\mathrm{native}}}{\Delta\Delta G^{\mathrm{native}}}\right)^2}$. Thirteen very small negative values, that is, $-0.11 \le \Phi \le 0$, have been rounded to $\Phi = 0$.

[d] Non-reliable Φ value due to small $\Delta\Delta G$ native.

[e] Φ-value outside the 0–1 range, indicative of formation of non-native interactions in MG.

[f] Non-reliable Φ value due to its high relative error.

of WT, suggesting that side chain repacking has occurred near W155, which might be related to the stabilization observed. In Q61A, the WT glutamine, fully buried according to ProtSA,[59] is replaced by an alanine, aiming to create a destabilizing cavity. However, replacement of buried glutamines by apolar residues has been previously associated with thermostabilisation,[60] which in this mutation appears to be the dominant effect.

The unfolding curves recorded for the MG conformations at pH 2 are reported in Figure S2 Importantly, no protein aggregation was observed for any of the mutants at the denaturant concentrations used to record the unfolding curves. The calculated stability of the variants is reported in Table S1. For all mutants, the MG is significantly less stable than the corresponding native conformation. On the other hand, while at pH 7 most mutants are less stable than native WT, at pH 2 there is a more even distribution of stabilities around that of MG WT. Three mutants (K21A, S23A, and R151A) stand out as more stable than WT. In two of them (K21A and R151A) basic residues that in the native structure are located in an α-helix close to salt bridges (E20-K24 and E150-K154, respectively) are replaced by alanines. Assuming some helical conformation remains in the concerned segments of the MG, disruption of the salt bridges at pH 2 due to protonation of the acidic residues involved will allow the basic residues K24 and K154 to exert repulsion on K21 and R151, respectively. The K21A and R151A mutants would be more stable than WT at pH 2 because they would not experience the indicated repulsions. The higher stability of the third mutant, S23A, could arise from a combination of the generally superior α-helix stabilizing character of alanine residues[61,62] and the specific destabilizing low solvent accessibility of Ser23, as per ProtSA[59] calculation.

2.2 | MG Apoflavodoxin in Φ values

Obtaining structural information of partially unfolded conformations of proteins is challenging. A useful approach is Φ-analysis,[63] which was developed to characterize transition states of protein folding and has been applied to the characterization of equilibrium intermediates.[53,55] The method determines the extent to which an interaction established by a residue in a folded conformation of known three-dimensional structure remains formed in an alternative conformation in equilibrium with the folded one. Using site-directed mutagenesis, a probe residue is replaced by a shorter one (often alanine) incapable of establishing the interaction probed. Then, the changes in the folding free energies of the two conformations analyzed are determined and a Φ value is calculated. We have used the data in Table S1 to calculate, through Equation (1), equilibrium Φ values (Table 1) probing the integrity in the MG of each of the interactions broken in the native structure by the mutations described in Table 1.

$$\Phi = \frac{\Delta\Delta G^{\mathrm{MG}}_{\mathrm{wt-mut}}}{\Delta\Delta G^{\mathrm{native}}_{\mathrm{wt-mut}}} \qquad (1)$$

In Equation (1), $\Delta\Delta G^{\mathrm{MG}}_{\mathrm{wt-mut}}$ is the unfolding free energy difference of WT MG minus that of a mutant MG, and $\Delta\Delta G^{\mathrm{native}}_{\mathrm{wt-mut}}$ is the unfolding free energy difference of the WT native conformation minus that of the same mutant. A Φ value of 1 indicates that the interaction probed by the mutation is fully retained in MG (i.e., the mutated side chain is in a native-like region of MG), while a Φ value of 0 indicates the interaction is totally lost (i.e., the mutated side chain is in a completely unfolded region of MG). Fractional Φ values between 0 and 1 can be interpreted as indicating that the interaction is lost in a 1-Φ fraction of molecules, or that it is debilitated in all the molecules to the fraction of its initial value indicated by the Φ value. As previously explained,[55] we favor the latter interpretation for highly structured intermediates, but the former one seems more realistic for ensembles with significant conformational heterogeneity such as the *Hp* apoFld MG. On the other hand, Φ values outside the [0–1] interval are indicative of the substitution in the MG of native interactions by non-native ones.

Ten mutations produced small changes $|\Delta\Delta G_{\mathrm{wt-mut}}|$ (≤0.3 kcal/mol) in both the stability of the native and MG conformations (I6A, F7A, T10A, S12A, E16A, E40A, F45A, T95S, T148S) or displayed a particularly high relative error (E105A). The Φ values calculated for those 10 mutations are not reliable and will not be further considered. Thirteen mutations yielded reliable Φ values that fell outside the canonical 0–1 interval, and they are indicative of formation of non-native interactions in the MG. Eleven of them (A19G, K21A, N43A, L50A, T70A, I81A, D88A, S93A, Y104F, Q144A, R151A) destabilized significantly the native conformation but stabilized the MG by at least 0.15 kcal/mol leading to negative Φ values. Two of them hardly modified the stability of native apoflavodoxin but significantly stabilized (D11A) or destabilized (S23A) the MG leading to very large Φ values. The additional 32 mutations gave canonical Φ values between 0 and 1. They include 13 very small negative values, that is, $|\Phi| \leq 0.11$, that will be considered to indicate full disruption of the probed interaction, that is, $\Phi = 0$. The 32 canonical Φ values, the location of the corresponding probe residues in the secondary structure elements as well as in the tridimensional structure of the native conformation, and the distribution of changes in free energy of unfolding $\Delta\Delta G_{\mathrm{wt-mut}}$ in both the native and MG *Hp* apoFld conformations are shown in Figure 1a–c. A detailed analysis of the likely structural meaning of the 32 canonical and 13 non-canonical but reliable Φ values is presented in the *SI Qualitative Structure of the Hp apoFld MG* section, from which a preliminary schematic structure of the MG can be proposed (Figure 1d,e). Essentially, the MG appears to retain an organization in three layers, with the packing of the α1α5 layer onto the central β-sheet being mostly native-like (although much debilitated) and that of the α2α3α4 layer mostly non-native. The loops, including the FMN binding

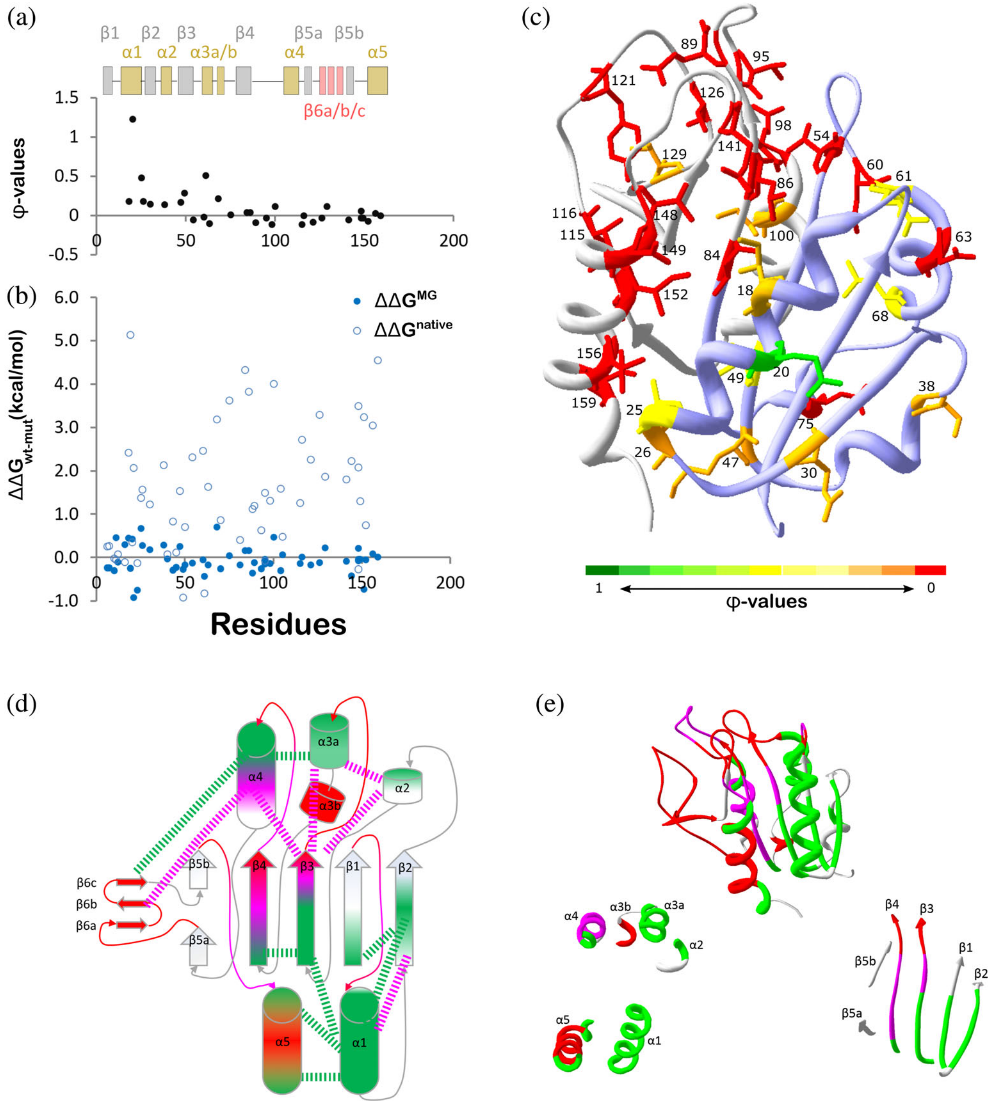

FIGURE 1 Distribution of stability effects and Φ values along the sequence and structure of *Hp* apoFld, and schematic structure of the molten globule (MG) obtained from equilibrium Φ-analysis. (a) Distribution of canonical Φ values (filled black circles), excluding non-reliable values and those indicative of non-native interactions. Very small negative Φ values, assimilated to Φ = 0 are also shown. A linear representation of the main elements of secondary structure along the sequence is shown at the top of the panel. (b) Distribution of changes in free energy of unfolding (wild-type [WT] minus mutant) in the *Hp* apoFld MG (filled blue circles), and in the native conformation (open circles). (c) Ribbon drawing of the apoflavodoxin native structure displaying the color-coded Φ values of the different residues probed as indicated by the color scale at the bottom (red: native interactions totally lost; green: native interactions totally retained). For non-probed residues, the ribbon color is pale blue for the N-terminal half of the protein (residues 1–82) and gray for the C-terminal half (residues 83–164). (d) Schematic structure of the MG showing integrity and presence of native-like interactions (debilitated) in the elements of secondary structure and loops: persistence of debilitated native-like interactions (green), complete disruption of native interactions (red), formation of non-native interactions (purple), regions not probed (white). (e) Ribbon drawing showing the native structure color-coded as in (d) (top): the two helical layers (bottom left) and the central β-sheet (bottom right) are represented separately.

loops and the extra loop of long-chain flavodoxins, appear to be either disordered or mostly engaged in non-native interactions. Thus, equilibrium Φ-analysis reveals persistence of secondary structural elements in the MG, which agrees with the reported preservation[50,51] of a far-UV CD spectrum with a significant helical content, albeit much reduced compared to the native conformation. This schematic structure is illustrative of the overall MG topology but it does not constitute the atomistic model needed to perform a quantitative comparison with the reported MG observables. To be able to do that comparison, the 32 experimentally determined canonical Φ values have been used to drive the transformation of the apoflavodoxin native structure into an ensemble of conformations representative of the MG.

2.3 | Biased MD simulations transforming the native conformation into an MG ensemble

BMD simulations[56,57,64,65] have been performed to progressively drive the starting *Hp* apoFld conformation (PDB ID 2BMV)[49] towards satisfying the experimental Φ values. Along the biased dynamics, residue Φ values have been computed as fractions of native contacts.[57] Three simulation sets (named *side chain heavy atom*, *whole side chain* and *all-atom*) of 21 independent replicas each have generated reaction coordinates leading the computed Φ values to closely match those determined experimentally (Figure S3a1–a3 and Table S2). The radius of gyration and RMSD plots (Figures S3b1–b3 and S3c1–c3, respectively) for the metatrajectories built for each set combining the last nanosecond (ns) of the simulated replicas are indicative of the conformational heterogeneity in the generated structures. The correspondence between experimental and computed Φ values from a representative replica of the *all-atom* simulation set is shown in Figure 2, where the stability of the calculated Φ values during the last ns of the trajectory is illustrated. For subsequent analyses, each replica has been represented by the centroid structure obtained from the last ns of the corresponding trajectory (see *SI Materials and Methods*). In this way, ensembles encompassing the 21 representative structures of each of the 3 simulation sets have been constructed. To assess whether those ensembles capture well the experimentally determined geometric and spectroscopic properties of the MG, we have computed on them a number of experimentally reported features. They include the far-UV CD spectrum, the hydrodynamic

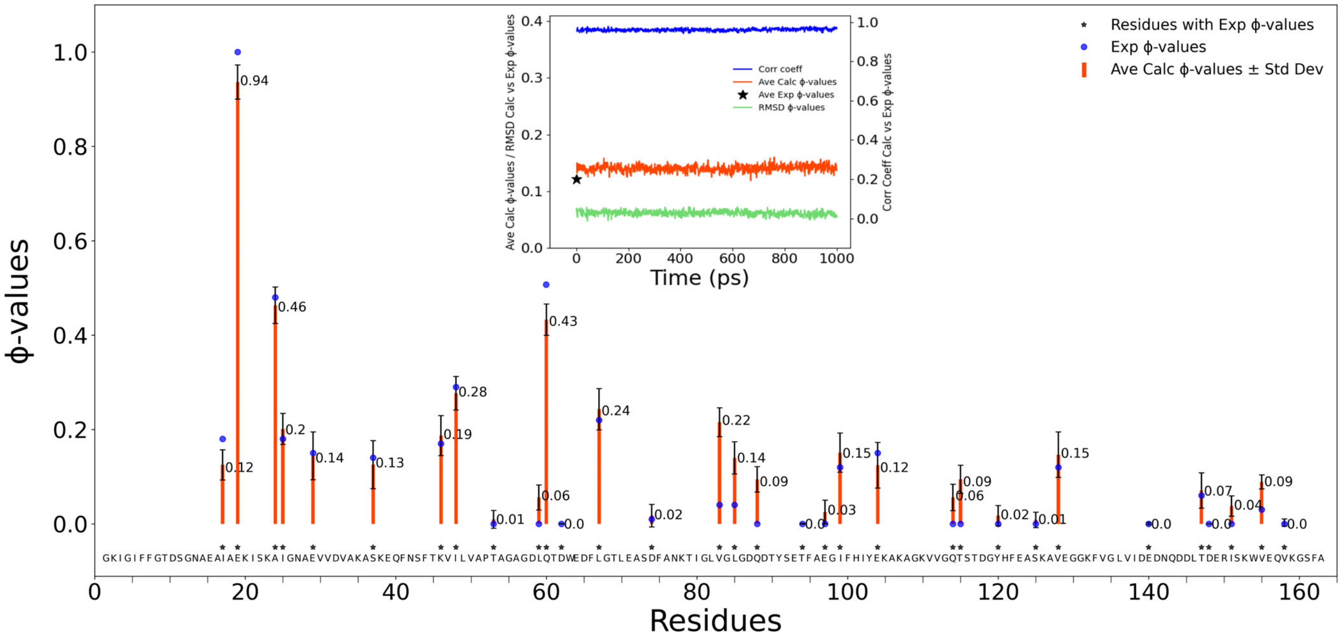

FIGURE 2 Comparison of calculated Φ values with experimentally determined ones. Experimental versus calculated comparison of Φ values for residues used as restraints in the biased molecular dynamics (BMD) simulations is shown for one replica of the *all-atom* simulations set. The sequence of the protein is shown at the bottom, with the mutated residues indicated with small asterisks. Orange bars represent the averages (with their *SD*) along the last nanosecond (ns) simulated. Blue circles indicate the experimental Φ values (see Table 1). The inset depicts the average of calculated Φ values compared with the average of the experimental ones (big asterisk), the RMSD of calculated versus experimental Φ values, and the correlation between them along the last ns of the BMD trajectory. These plots reflect the stability of the calculated Φ values and the good matching between them and the experimental ones.

radius, the change in SASA upon unfolding, and the fluctuations near aromatic or near aliphatic residues that may contribute, respectively, to the disappearance of native fingerprints in the near UV-CD and in the ^{1}H-NMR spectra.[66,67] For all the above properties, the *all-atom* ensemble has provided the best matching between calculated and observed properties, so it will be described and discussed in more detail. Data obtained for the two other ensembles (*whole side chain* and *side chain heavy atom*) are shown in Tables S2–S8, and in Figures S3–S5, either separately or together with data from the *all-atom* ensemble.

In the *all-atom* ensemble, the native helices and strands appear significantly shortened, albeit to different extents (Figure 3a). The central β-sheet remains formed but it shows a rather homogeneous debilitation, which is more evident towards the protein C-terminus. There, the three short strands (β6a–c) of the long loop that intercalates in strand β5, plus the second segment of this strand (β5b) are more debilitated than preceding strands β1–β5a. On the other hand, the helical layer formed by helices α2, α3, and α4 is quite homogeneously destabilized, with the individual helices retaining 32–40% of their initial residues (Figure 3a). Helix α3 encompasses two helical stretches (α3a and α3b) in the native protein, of which α3b appears unfolded in the ensemble. In contrast with the previous two layers, the helical layer comprising helices α1 and α5 is inhomogeneously destabilized, with a strong persistence of most of helix α1 (90%) and a severe destabilization of α5 (15% persistence). On average, around 40% of the residues populating either helices or strands in the native apoflavodoxin conformation are still populating the original secondary structure elements in the *all-atom* MG ensemble (Figure 3a).

This ensemble has been refined to fit the far-UV CD experimental spectrum of the MG which, together with the spectrum of the native conformation, is shown in Figure 3b,c. The shapes of native and MG experimental spectra are dissimilar, possibly due to aromatic side chain contributions to the native spectrum in the 205–215 nm region,[68] which would be absent in the MG.[50,51] In any case, the CD signals of either spectrum at 222 nm, essentially due to helical contributions, clearly indicate that the MG helical content is significantly lower than that of the native conformation. To assess whether the *all-atom* ensemble captures the observed far UV CD spectrum of the MG, we have calculated, using the PDBMD2CD server,[69] the CD spectra of the 21 ensemble conformations. On the other hand, to eliminate potential contributions of aromatic residues to the shape of the native spectrum, a theoretical spectrum of the native conformation has also been calculated from the X-ray structure in the same way. The calculated spectra of the native conformation and of the *all-atom* ensemble conformations are compared in Figure 3b,c with the corresponding experimental spectra. The MG reduced secondary structure content (see assignments given by DSSP[70] in Table S3) is clearly captured by the 21 conformations of the *all-atom* ensemble, all of which display a lower helical content than the native conformation. As the secondary structure content varies notably among the 21 conformations, we have used the RMSD between the experimental and the calculated CD spectra as a criterion to identify the 10 best-matching conformations. The RMSD of the 10 selected structures range from 0.23 to 0.5 $\Delta\varepsilon$ units, while those of the 10-best matching ones from the alternative BMD ensembles (*whole side chain* and *side chain heavy atom*) are higher (0.51–0.74 $\Delta\varepsilon$ units and 0.37–0.78 $\Delta\varepsilon$ units, respectively, Figure S5a,b). The 10 best conformations so identified in the *all-atom* ensemble constitute the *all-atom-10* ensemble (Figure 3c). This ensemble, as will be shown, explains well the properties of the MG.

MGs are expanded conformations compared to native ones[71] and the hydrodynamic radius (R_h) is a suitable property to assess molecular expansion. In previous work, the R_h of native and MG *Hp* apoFld were determined,[51] confirming that this MG is an expanded conformation. To evaluate whether the *all-atom-10* ensemble can reproduce the reported expansion, the self-diffusion coefficient of the 10 structures, as well as that of the native conformation (PDB ID 2BMV), have been calculated with the program HYDROPRO v.10[72,73] (Table S4) and used to obtain the corresponding R_h, as described.[51] Comparison of the R_h calculated for the native structure (20.4 Å) with the experimental one (18.8 Å) suggests HYDROPRO v.10 calculations slightly overestimate the hydrodynamic radii. Therefore, the representativeness of the *all-atom-10* ensemble can be best assessed comparing the increase in R_h associated with the transformation of the native conformation into the MG, rather than directly comparing the actual R_h values. The experimental data indicate that the hydrodynamic radius of native *Hp* apoFld increases by 9.6% in the MG, which compares well with the calculated average increase of 8.4 ± 0.4% obtained for the *all-atom-10* ensemble (Table S4). The two alternative BMD ensembles considered show worse agreements (*whole side chain*, 6.9 ± 0.4% R_h increase; *side chain heavy atom*, 6.7 ± 0.3% R_h increase) as they consist of more compact conformations (Figure S3b1–b3).

On the other hand, the study of *Hp* apoFld stability as a function of pH using urea denaturation and the linear extrapolation method (LEM)[74] revealed a large difference between the *m* value (proportionality constant between change in unfolding free energy and denaturant

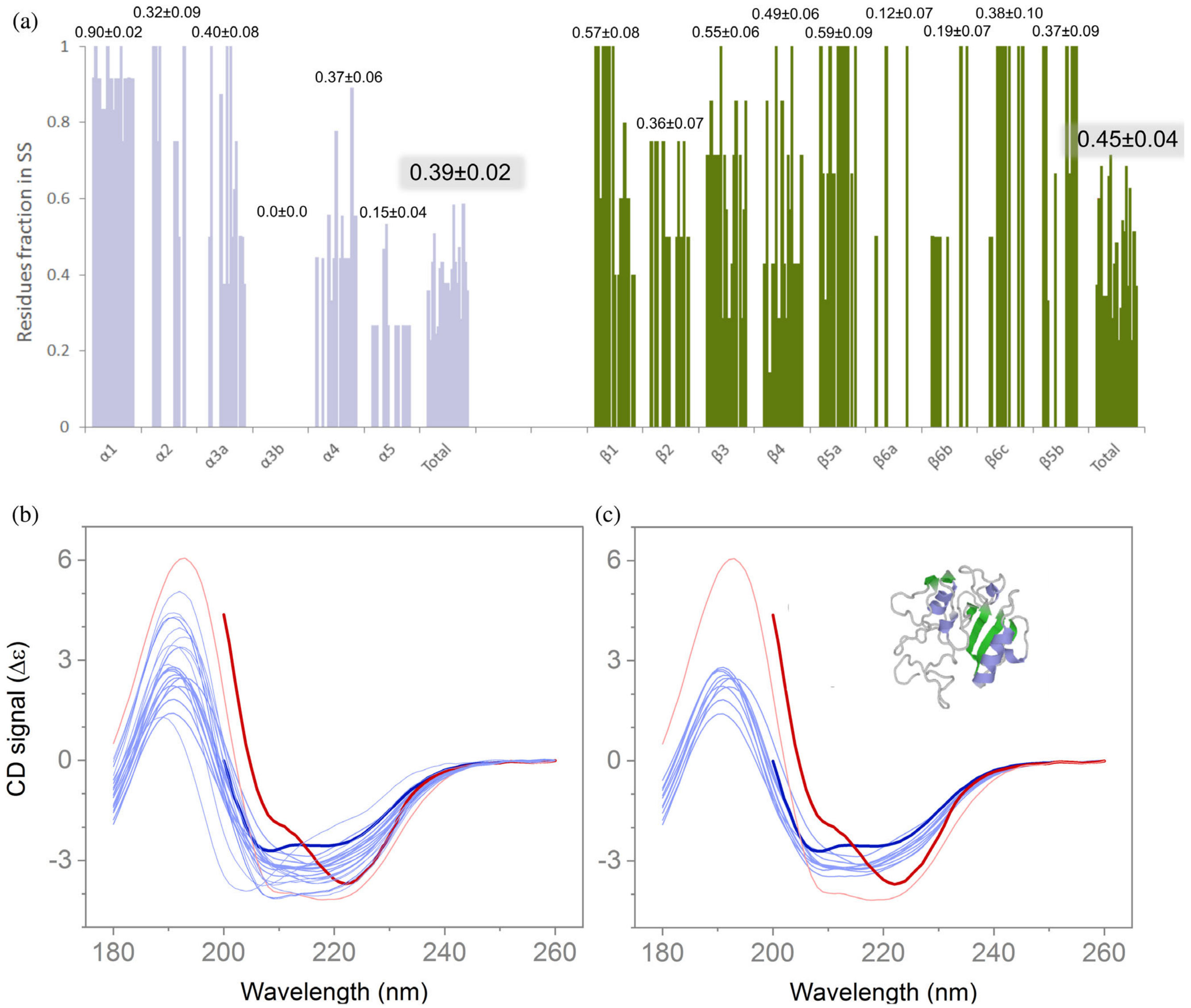

FIGURE 3 Secondary structure and predicted versus experimental far-UV circular dichroism (CD) spectra of the *all-atom* molten globule (MG) ensemble. (a) Persistence of native secondary structure in the biased molecular dynamics (BMD) *all-atom* MG ensemble depicted as fractions of residues in secondary structure elements of the native structure (PDB ID 2BMV[49] minimized) that remain forming the same elements in the MG ensemble. Fractions are shown correlatively for each of the 21 conformations in the ensemble as thin bars (pale blue for α-helices; pale green for β-strands) over the specified secondary structure element. The average fraction in the ensemble (±*SE*) for each secondary structure element is given on top of it. Overall α-helix and β-strand average fractions are highlighted with a gray box background. Secondary structure assignments were made by the DSSP program.[70] (b) Far-UV CD profiles predicted by the PDBMD2CD server[69] for the 21 conformations in the BMD *all-atom* MG ensemble (thin light blue lines) compared to the *Hp* apoFld experimental spectrum at pH 2 (thick blue line).[51] The calculated far-UV CD profile of the minimized *Hp* apoFld structure (native, thin red line) and the experimental spectrum of the native structure at pH 7 (thick red line)[51] are also depicted for comparison. As the experimental spectrum of the native conformation at pH 7 may contain contributions from aromatic residues that alter the shape of the spectrum, the spectrum calculated for the minimized native conformation, rather than the experimental native spectrum, should be taken as the reference to compare with the MG data. (c) Same comparison as in (b) but showing only the 10 predicted CD profiles (*all-atom-10* ensemble, thin light blue lines) best matching the experimental spectrum (thick blue line). The best matching conformation of this sub-ensemble is shown in cartoon representation. RMSDs (in Δ*ε* units) for the 10 best-matching predicted CD profiles (i.e., those with the lowest RMSD versus the experimental one of the *Hp* apoFld MG) range from 0.23 to 0.50.

concentration) for the native protein (m = 2,150 ± 50 cal/(mol × M)) and that for the MG (m = 1,330 ± 40 cal/(mol × M)).[51] As m values are linearly related to the increase in solvent exposure (ΔSASA) that takes place as the protein unfolds,[75] the agreement of the *all-atom-10* and the other MG ensembles with the experimental m values can be evaluated by comparing ΔSASAs derived from m values with ΔSASAs directly calculated[59] on the structures. Considering that the linear relationship obtained by Myers et al. ($m = 368 + 0.11 \times \Delta$SASA, with $R = 0.90$, m in cal/(mol × M) and ΔSASA in Å^2) was based on unrealistic fully-extended unfolded conformations,[75] we have first computed for the same protein dataset realistic values of unfolded SASAs using the ProtSA server.[59] Then, we have obtained an equivalent upgraded relationship: $m = 362(\pm 137) + 0.16 (\pm 0.01) \times \Delta$SASA ($R = 0.90$; see details in Table S9 and Figure S6a), from which the ΔSASA of native *Hp* apoFld has been calculated at 11,203 (±2,205) Å^2 and that of the MG at 6,065 (±1,669) Å^2. The ΔSASA computed with ProtSA for the native conformation, 10,482 Å^2 (Table S5), agrees within experimental error with that derived from the m value. Out of the different MG ensembles analyzed, the averaged ΔSASA computed for the *all-atom-10* MG ensemble (8,652 Å^2) is the one closer to that derived from the experimental m value, albeit still a bit higher than its upper limit of 7,734 Å^2. We have extended this ΔSASA analysis to separately consider polar and apolar exposure changes in the unfolding. Using the more accurate exposures obtained with ProtSA for the unfolded ensembles of the proteins included in the fit,[75] new linear relationships between either polar or nonpolar ΔSASA and experimental m values have been obtained (with R =0.96 or 0.92, respectively, see Table S9 and Figure S6b,c), which are analogous to those in Myers' paper.[70] For the native conformation, both the polar (3,227 ± 635 Å^2) and nonpolar (7,218 ± 1,420 Å^2) ΔSASAs calculated from the m values using the updated relationships agree with the ΔSASAs computed from the PDB file (2BMV): 2,977 ± 82 and 7,506 ± 81 Å^2, respectively. For the *all-atom-10* ensemble, both the polar (2,856 ± 112 Å^2) and apolar (5,796 ± 248 Å^2) ΔSASA computed from the *all-atom-10* ensemble (Table S5) are a bit higher than those obtained from the m value (1,743 ± 480 and 4,045 ± 1,114 Å^2, respectively).

The near-UV CD spectra of some proteins contain peaks characteristic of aromatic residues in asymmetric environments.[76] The spectrum of native *Hp* apoFld displays two such distinct peaks centered at 287 and 295 nm.[51] The peaks disappear in the MG spectrum indicating that some tyrosine or, more likely considering the wavelengths, tryptophan residues are sensing a less defined environment in the MG than in the native conformation.[67] To assess whether the *all-atom-10* ensemble captures the structural basis of this spectroscopic observation we have computed the root mean square fluctuations (RMSFs) of the three tyrosine (Tyr92, Tyr104, and Tyr121) and two tryptophan (Trp64 and Trp155) residues in native *Hp* apoFld and in the ensemble (Table S6). In the native state both Trp64 and Trp155 display low RMSFs (0.69 and 0.80, respectively), with Tyr104 fluctuating more (0.90) and Tyr92 (1.44) and Tyr121 (1.78) showing the greater fluctuations. In the *all-atom-10* ensemble, the RMSFs of the tyrosines are not much higher than in the native structure, but those of the tryptophans are greatly increased (1.34 and 1.95, respectively) so that the five aromatic residues show comparable RMSFs. The large increase in RMSF associated with the tryptophans in the *all-atom-10* ensemble is consistent with the loss of the native 287 and 295 nm peaks in the near-UV CD spectrum of the MG, and points to those residues as likely responsible for the indicated native peaks.

The ^{1}H-NMR monodimensional spectrum of native *Hp* apoFld contains several high field resonances that disappear in the spectrum of the MG.[51] Such resonances usually arise in proteins from persistent interactions between methyl groups in aliphatic residues and aromatic side chains, and are indicative of a well-defined tertiary structure.[66,77] We have assessed the persistence of aromatic/methyl interactions in native *Hp* apoFld and in the *all-atom-10* ensemble by determining the fraction of frames where a given (Phe, Tyr, Trp)/methyl interaction takes place along the different simulations (using a distance cutoff of 4.5 Å, Table S7). A total of 976 pairwise interactions between 11 Phe, 3 Tyr, and 2 Trp aromatic residues on the one hand and 18 Ala, 11 Ile, 8 Leu, 11 Thr, and 13 Val methyl-containing residues on the other have been analyzed. While interactions involving Phe or Tyr residues tend to be similarly persistent in the native structure and in the *all-atom-10* ensemble, those established by the two tryptophan side chains markedly decrease in the ensemble. This is especially noticeable in the case of Trp/(Ile, Val) contacts (Table S8), and remarkable for the contacts between Trp155 and residues Ile22, Ile49, Leu137, and Val159 which give rise to a strongly persistent cluster in the native conformation (93–98% persistence) that is hardly detected (0–12% persistence) in the ensemble. Loss of persistent interactions between methyl groups in aliphatic residues and tryptophan side chains is thus evident in the *all-atom-10* ensemble, which is consistent with the reported absence of the native high field methyl-resonances in the MG ^{1}H-NMR spectrum.[51]

2.4 | Attempts to define an MG ensemble using unbiased MD simulations to follow relaxation of the native conformation into the MG

Classical unbiased MD simulations are increasingly used to explore protein conformational landscapes and to derive the structure of protein variants from the corresponding WT structures.[78–81] A frequent way to achieve this is to model an amino acid substitution in the WT structure and simulate the created variant, hoping to capture a structural relaxation that leads to and reveals its actual conformation. To explore the feasibility of using MD simulations to likewise unveil the *Hp* apoFld MG conformation from relaxation of the native structure dominant at pH 7, we have performed atomistic MD simulations after having set the charges of the ionizable residues to their expected values at pH 2 (MG). Two force field/water model combinations have been tested. The first one, Charmm22 with CMAP correction/Tip3, has been previously reported to capture well protein folding energetics using short simulations but also to artifactually compact unfolded conformations if the simulations are prolonged.[79] The second one, Amber99SB-disp/Tip4p-d-disp, does not capture equally well the folding energetics[79] but does not appear to overcompact unstructured conformations.[79,82] The protocol used to try to select representative MG ensembles from relaxation MD simulations carried out with either combination of force field and water model is described in Figure 4. For the first combination, five replicas of the native conformation at pH 7 (a total of 15 μs) and five replicas of the native conformation transferred to pH 2 (MG conditions; a total of 19 μs) were simulated (Figure 4a). Likewise, for the second combination, five replicas for native (a total of 10 μs) and five replicas for MG conditions (a total of 10 μs) were simulated (Figure 4b). For each group of replicas, full metatrajectories were used to try to obtain representative MG ensembles.

A first attempt to obtain the ensembles was done by performing 2D-RMSD-based clustering (see *SI Materials and Methods*). The clustering aimed to identify exclusive (not present in the native metatrajectory), recurrent (repeatedly appearing) or sustained (0.5 or more microseconds of continuous presence) conformational clusters. Three such clusters were identified in the Charmm22/CMAP metatrajectory (Figure S8a), but none in the Amber99SB-disp one (Figure S7b). The three clusters identified were refined (see scheme in Figure 4a) to retain only conformations exhibiting experimental versus calculated Φ values correlations ≥0.5 (76 conformations were extracted, Table S10). Regardless of cluster of origin, the helical content in the filtered conformations was unrealistically high (>40%, see Table S11) compared to that in the MG (Figure S8a–c) and even to that in native *Hp* apoFld (Table S11). Consistent with this high helical content, the average increase in R_h relative to that of the native conformation (7.0 ± 0.2%) is clearly below the experimental observation (9.6%, Table S12), and the calculated average ΔSASA of 9,105 (±33) Å^2 (Table S5) is above that calculated from the urea m value (6,065 ± 1,669). Related to the lack of high-field resonances in the MG ^{1}H-NMR spectrum, these ensembles capture that the interactions between tryptophan side chains and methyl groups from Ile and Val residues are less persistent than in the native conformation (particularly evident in clusters C5 and C58, see Table S13). However, they do not capture a clear increase in the fluctuations of the tryptophan residues (see RMSF values in Table S14) that could explain the observed loss of the native near-UV CD peaks in the MG spectrum. Overall, these ensembles constitute poor recapitulations of the known structural and spectroscopic features of the *Hp* apoFld MG.

As the dominant clusters identified in the Charmm22/CMAP metatrajectory (Figure S7a) did not lead to a realistic MG ensemble, and as no dominant clusters were identified in the Amber99SB-disp metatrajectory (Figure S7b), a second more stringent selection was done to try to identify representative MG ensembles by focusing on the properties of the individual conformations. Thus, from the Charmm22/CMAP pH 2 metatrajectory, a reduced 10-conformation ensemble (*Charmm-10*, Table S10) was obtained encompassing the conformations exhibiting ≥0.6 experimental versus calculated Φ correlations (Figure S9a) and the lowest experimental versus calculated RMSD of CD spectra (Figure S8d). This *Charmm-10* ensemble exhibits an average helical content of 41% (Table S11) that, albeit a bit lower than that in the persistent clusters analyzed above, still exceeds the content calculated for the native conformation (33%, Table S11). The fractions of native residues initially present in the native α-helices and β-strands that are retained in the corresponding structural elements present in this ensemble (Figure S10a) are quite high and, in addition, longer, and new helices are formed in a non-transient manner (not shown) in these simulations. On the other hand, the R_h increase in this ensemble (7.6 ± 0.6%) is clearly below the experimental observation (Table S12). The *Charmm-10* ensemble, as those obtained above from cluster analysis of the same trajectories, also constitutes a poor representation of the MG. It is possible that the high helical content of the conformations in the Charmm22/CMAP metatrajectory is related to the reported fact that this force field tends to overcompact (partially) unfolded conformations.[73,76]

Finally, to try to extract a representative MG ensemble from the Amber99SB-disp simulations, 69 conformations (*Amber-69* ensemble) exhibiting Φ correlations (experimental vs. calculated) ≥0.6 (Table S10 and Figures 4b and S9b) were identified, from which those 10 exhibiting the higher correlations between their calculated far-UV CD spectra and the experimental one (lowest RMSD) were selected (Table S10 and Figure S11). This MG ensemble (*Amber-10*) exhibits an average helical

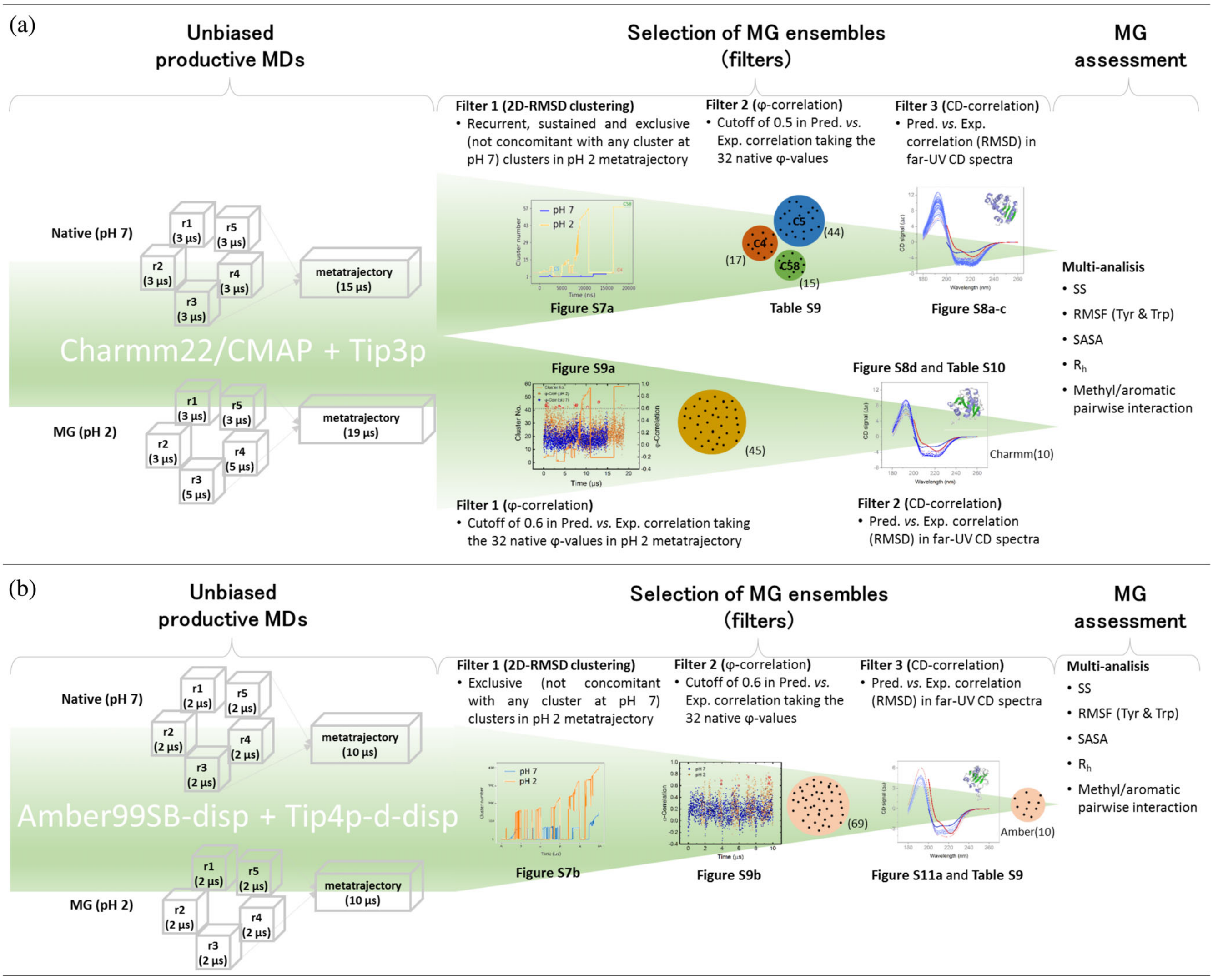
(a)
Unbiased productive MDs
Selection of MG ensembles (filters)
MG assessment
Filter 1 (2D-RMSD clustering)
• Recurrent, sustained and exclusive (not concomitant with any cluster at pH 7) clusters in pH 2 metatrajectory
Filter 2 (φ-correlation)
• Cutoff of 0.5 in Pred. vs. Exp. correlation taking the 32 native φ-values
Filter 3 (CD-correlation)
• Pred. vs. Exp. correlation (RMSD) in far-UV CD spectra
Native (pH 7)
r1 (3 µs)
r2 (3 µs)
r3 (3 µs)
r4 (3 µs)
r5 (3 µs)
metatrajectory (15 µs)
Charmm22/CMAP + Tip3p
MG (pH 2)
r1 (3 µs)
r2 (3 µs)
r3 (5 µs)
r4 (5 µs)
r5 (3 µs)
metatrajectory (19 µs)
pH 7
pH 2
Figure S7a
C5
C4
C58
(44)
(17)
(15)
Table S9
Figure S8a-c
Multi-analisis
• SS
• RMSF (Tyr & Trp)
• SASA
• R$_h$
• Methyl/aromatic pairwise interaction
Figure S9a
(45)
Figure S8d and Table S10
Charmm(10)
Filter 1 (φ-correlation)
• Cutoff of 0.6 in Pred. vs. Exp. correlation taking the 32 native φ-values in pH 2 metatrajectory
Filter 2 (CD-correlation)
• Pred. vs. Exp. correlation (RMSD) in far-UV CD spectra
(b)
Unbiased productive MDs
Selection of MG ensembles (filters)
MG assessment
Filter 1 (2D-RMSD clustering)
• Exclusive (not concomitant with any cluster at pH 7) clusters in pH 2 metatrajectory
Filter 2 (φ-correlation)
• Cutoff of 0.6 in Pred. vs. Exp. correlation taking the 32 native φ-values
Filter 3 (CD-correlation)
• Pred. vs. Exp. correlation (RMSD) in far-UV CD spectra
Multi-analisis
• SS
• RMSF (Tyr & Trp)
• SASA
• R$_h$
• Methyl/aromatic pairwise interaction
Native (pH 7)
r1 (2 µs)
r2 (2 µs)
r3 (2 µs)
r4 (2 µs)
r5 (2 µs)
metatrajectory (10 µs)
Amber99SB-disp + Tip4p-d-disp
MG (pH 2)
r1 (2 µs)
r2 (2 µs)
r3 (2 µs)
r4 (2 µs)
r5 (2 µs)
metatrajectory (10 µs)
pH 7
pH 2
Figure S7b
Figure S9b
(69)
Figure S11a and Table S9
Amber(10)

FIGURE 4 General workflow used to filter molten globule (MG) ensembles from classical unbiased molecular dynamics (MD) simulations. (a) Two-path scheme used for collecting and evaluating MG ensembles from simulations run with the Charmm22/CMAP force field. The number of simulated replicas (cubic boxes), the simulation time run for each one, and the total time in the metatrajectories built from pH 7 and pH 2 simulations are indicated at the left-hand side block of the scheme (unbiased productive MDs, also in (b)). The first path (top) followed to filter MG ensembles consisted of a sequence of three filtering steps including 2D-RMSD-based clustering, Φ values correlation and far-UV circular dichroism (CD) comparative analyses. The second path (bottom) discards the results of the clustering and performs filtering through Φ values correlation and far-UV CD comparative analyses. A 10-conformation set (*Charmm-10*) resulted from this path, which was subsequently analyzed in order to assess the representativeness of the selected ensemble (MG assessment). (b) One-path route followed to select and evaluate MG ensembles from simulations run with the Amber99SB-disp force field. Three filtering steps, 2D-RMSD-based clustering, Φ-correlation and far-UV CD comparative analyses, were implemented which led to a 10-conformation ensemble (*Amber-10*) that was assessed through the indicated analyses (MG assessment). The figures and tables associated with the filtering carried out in (a) or (b) appear indicated. Φ-correlation thresholds are also indicated in the filters descriptions.

content of 25 (±2.5) % (Table S11), which is below that of the native conformation (33%) but still doubles the amount calculated for the *all-atom-10* BMD ensemble (Table S3). Besides, the fractions of retained residues in α-helices and β-strands (Figure S10b) are still high compared to those calculated for the BMD *all-atom-10* ensemble and hardly match the qualitative MG model obtained from the Φ-analysis alone (Figure 1d). The calculated average R_h increase in the *Amber-10* ensemble (11.0 ± 1.1%, Table S15) is above the experimental one (9.6%). This force field, that was designed to avoid over-compacting (partially) unfolded conformations, appears to allow for an excessive expansion in this particular case. Average SASAs for the *Amber-10* ensemble are shown in Table S5. Its average unfolding ΔSASA of 8,192 (±336) $Å^2$ is slightly larger than the upper limit calculated from

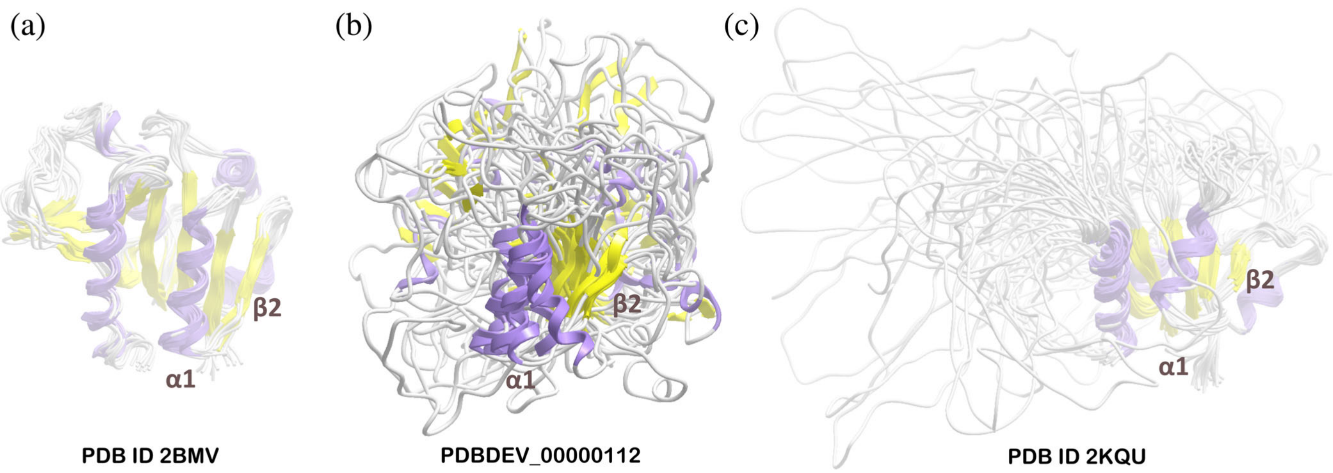

FIGURE 5 Molten globule (MG) ensemble of *Hp* apoFld. Comparison with the native conformation and with the equilibrium intermediate ensemble of the homologous apoFld from Anabaena PCC7119. (a) Ten-conformer ensemble of native *Hp* apoFld, PDB ID 2BMV.[49] The conformers were selected (every 0.1 ns) from the final nanosecond of the equilibration step run before the biased molecular dynamics (BMD) productive simulations. (b) Integrative model of the *Hp* apoFld molten globule (PDBDEV_00000112) consisting of the 10 structures (*all-atom-10*; selected from the BMD *all-atom* ensemble) that best match the far-UV circular dichroism (CD) spectrum, R_h, and other structural and spectroscopic experimental features of the MG conformation at pH 2. Given the structural variability of the 10 conformers selected and for greater visual clarity, the structural alignment was performed on the first 50 amino acid residues of the sequence, as the N-terminal region of the MG appears to display a more homogeneous structure than the C-terminal one. (c) Experimental ensemble (solution NMR, 20 conformers) of the *Anabaena* PCC7119 apoFld equilibrium intermediate accumulating in thermal unfolding (PDB ID 2KQU).[87] To facilitate visual comparison of the three structural ensembles depicted (native (a), MG (b), and thermal intermediate (c)), the same secondary structure elements α1 and β2 are indicated in the corresponding panel.

the m value (7,734 Å^2). The ensemble cannot offer a reason for the observed loss of native near-UV CD signals in the MG because the RMSFs of the tryptophan and tyrosine residues are similar in either conformation (Table S16). However, the ensemble does capture the debilitation of Trp/methyl interactions (Table S17) associated with the disappearance of high field peaks in the ^{1}H-NMR spectra. Thus, while the *Amber-10* ensemble offers a better representation of the MG than the *Charmm-10* ensemble, both are clearly surpassed by the *all-atom-10* ensemble obtained from BMD simulations. The 10 conformations of the *all-atom-10* ensemble are shown superimposed in Figure 5b, where they are compared with other apoflavodoxin conformations.

3 | DISCUSSION

3.1 | General destabilization of native interactions in the MG

Comparison of the WT *Hp* apoFld conformational stability at pH 7 and pH 2 in buffers of similar ionic strength indicates the MG is far less stable than the native conformation (0.9 vs. 7.9 kcal/mol, Table S1), pointing to a severe weakening of native interactions in the MG that is not compensated by formation of similarly stabilizing non-native ones. As the only chemical difference between the pH 7 and pH 2 polypeptides is the protonation state of some ionizable residues, the stability difference between the native and MG conformations has to be related to electrostatic interactions. *Hp* apoFld is a highly acidic protein with an excess of acidic residues over basic ones (28 vs. 17). This electrostatic imbalance, common to other apoflavodoxins, has provided a means to stabilize the native conformation by replacing acidic residues by neutral or basic ones.[83] On the other hand, it determines that, as the pH is lowered from 7 to 5, the electrostatic repulsion is reduced and the stability of the native conformation increases. For *Anabaena* PCC7119 apoflavodoxin, further reduction of pH to the isoelectric point (4.2) drastically lowers its aqueous solubility,[52] but at lower pHs, for example, at pH 2, the protein aqueous solubility is recovered and a soluble MG conformation emerges. Unfortunately, this *Anabaena* MG is not amenable to thorough thermodynamic characterization due to its oligomeric nature.[52] In contrast, *Hp* apoFld, whose stability as a function of pH follows a very similar trend, adopts at pH 2 a monomeric MG conformation,[49] the stability of which can be determined accurately by chemical denaturation and LEM analysis.[84,85] The electrostatic imbalance of apoflavodoxins below the isoelectric point is characterized by a smaller net charge than at neutral pH, but also by an absence of stabilizing charge/charge interactions, as all the acidic residues are presumably neutralized. The low conformational stability of the *Hp* apoFld MG (Table S1, Figures S1 and S2) indicates that this particular combination of less electrostatic repulsions with less attractive interactions arising at pH 2 is highly detrimental for the stability of the native conformation, which relaxes to the partially unfolded MG, itself not very stable.

To probe the structure of the MG by equilibrium Φ-analysis, 55 mutations consisting in shortenings of WT residues (Table 1) have been designed as cleanly and specifically as possible. For those mutations to be useful to perform a Φ-analysis, large stability changes relative to native WT are desired. The mutations selected do their job (Figure 1) as their average destabilization of the native conformation is large (1.6 ± 1.4 kcal/mol). At the same time, they confirm the MG is severely destabilized in a rather general manner, as their average effect on MG stability is small (0.0 ± 0.3 kcal/mol). Only five mutations destabilize the MG by more than 0.4 kcal/mol and only five others increase the stability by at least that same amount. Out of these 10 mutations exerting a strong effect on MG stability, as many as 8 are located in the N-terminal half of the sequence (Figure 1). From the analysis of the 55 variants, 45 reliable Φ values have been calculated, which fall in two categories: 32 mutations displaying canonical Φ values from 0 to 1 that report on the debilitation of native interactions in MG, and 13 mutations displaying either negative Φ values (11 cases) or very large ones (2 cases), which are indicative of the engagement of the mutated side chains in non-native interactions in the MG. In general, the canonical Φ values obtained along the structure are quite low (0.1 ± 0.1, mean ± *SD*, or 0.32 ± 0.04 if the zero Φ values are excluded). The only five Φ values >0.2 concentrate in the N-terminal half of the sequence (Figure 1a,b).

A detailed structural interpretation of the 32 canonical and 13 noncanonical but reliable Φ values is presented in the *Supplementary Information*, where a schematic structure of the MG is proposed. The structure (Figure 1d) captures the greater persistence of native-like but highly debilitated interactions in the N-terminal half of the MG and suggests that the three layers (α/β/α) of the native conformation persist in a greatly debilitated form that is stabilized by a combination of native and non-native interactions. To obtain an atomistic structural model of the MG ensemble that can be validated against the structural and spectroscopic experimental properties, both biased and classical MD simulations have been performed.

3.2 | Structural model of the MG

BMD simulations have been performed using a reaction coordinate based on the multiple experimental Φ values, which has transformed native *Hp* apoFld into an ensemble of conformations representative of the MG. On the other hand, unbiased microsecond-long explicit-solvent MD simulations have been run to allow native *Hp* apoFld (PBD ID 2BMV) to evolve after having been "transferred" to pH 2 conditions. In either approach, observable properties (far-UV CD spectrum, hydrodynamic radius, change in SASA upon unfolding, and fluctuations near aromatic or aliphatic residues) have been calculated for the ensembles obtained, which have been compared with the corresponding experimental values. The best correspondence between experimentally determined observables and calculated values or features in the modeled ensembles has been obtained using the BMD approach termed *all-atom*, wherein the computed Φ values used to bias native *Hp* apoFld account for all atoms of the concerned residues (see *SI Materials and Methods*). An almost perfect agreement between experimental and calculated Φ values has been obtained for the last ns of the 21 *all-atom* replica trajectories (Figure 2), from which an ensemble representative of the MG has been obtained and further refined to encompass the 10 structures that best match the experimental far-UV CD spectrum (*all-atom-10* ensemble). The CD signal at 222 nm calculated for this refined ensemble (Figure 3d) is 1.11 times that of the experimental MG spectrum, which is similar to the 1.16 ratio obtained for the calculated and experimental CD signals of the native conformation at the same wavelength. The expansion of the *Hp* apoFld MG (R_h 9.6% larger than that of the native conformation) is reasonably well captured by the ensemble (8.4 ± 0.3% larger R_h, see Table S4). This MG expansion is also reflected in the experimental unfolding ΔSASA (6,065 ± 1,665 Å^2), which is much lower than that of the native conformation (11,203 ± 2,205 Å^2). The calculated ΔSASA of the refined ensemble is a bit higher (8,652 ± 298 Å^2) than the experimental ΔSASA, which may indicate the ensemble is a bit more compact than the actual MG. The *all-atom* ensemble is also consistent with the experimental observation of loss of the native near-UV CD peaks at 287 and 295 nm in the MG. Moreover, it is also consistent with the disappearance of the native ^{1}H-NMR high field methyl resonances.[71] Detailed structural analysis of the ensemble points to a higher fluctuation of Trp64 and Trp155 (Table S6) and to loss of persistent Trp/(Ile, Val) contacts (Table S7) —particularly those of the cluster formed between Trp155 and residues Ile22, Ile49, Leu137, and Val159— as involved in the observed spectral changes. Overall, based on the good correspondence between the average values of the observables calculated from the conformations of this ensemble and the corresponding experimentally determined values of the MG, we propose the *all-atom-10* ensemble (Figure 5b) as a structural model of the *Hp* apoFld MG. The coordinates of the 10 structures constituting the ensemble have been deposited in the PDB-Dev database[86] with accession code PDBDEV_00000112.

Attempts have also been carried out to arrive to a realistic MG ensemble using classical, unbiased MD simulations. The rationale for this is that, upon transfer to pH 2 simulation conditions, the native structure might relax in a reasonable simulation time into the MG conformation, driven by force fields like those used for simulating native structures. However, simulation of *Hp* apoFld at pH 2 using Charmm22/CMAP has consistently led to overcompacted conformations with unrealistically high helical content (Figure S9). This may be related to previous observations that this force field tends to overcompact unfolded conformations.[79,82] It seems, therefore, that simulation of the expanded MG with this force field also suffers from this caveat. On the other hand, similar unbiased MD simulation using AmberSB99-disp (a force field designed to avoid the compaction of unfolded structures)[82] has not led to an overcompacted ensemble but rather to an over expanded one (Table S15), which still displays an elevated helical content (Table S11 and Figure S10b). Moreover, while the calculated average ΔSASA obtained is not far from the experimental one (Table S5) and the debilitation of methyl/Trp interactions is captured (Table S17), the fluctuation of tryptophans in the ensemble is similar to that observed in the native structure (Table S16), which does not explain the loss of near-UV CD signal in the MG. Whether the differences between calculated observables for the Charmm22/CMAP- or AmberSB99-disp-derived ensembles and the experimental MG values should be attributed to insufficient simulation time or to intrinsic limitations of these force fields for simulating partially unfolded conformations is not clear at present. Whatever the case, biased MD simulations based on experimental Φ values have provided a satisfactory ensemble (Figure 5b) that recapitulates the known properties of the *Hp* apoFld MG. It is possible that the approach developed here could also be used to analyze the conformational ensembles of the MG-like IDPs and IDP regions,[5,6] constituting up to 40% of the human proteome.

3.3 | The MG and the other non-native conformations in the apoflavodoxin landscape

Flavodoxins have been used over the years to investigate protein conformational plasticity and a number of non-native conformations (i.e., MGs, equilibrium intermediates, kinetic intermediates, and transition states of folding) have been found and described in different detail, mainly in the flavodoxins from *Hp*, *Anabaena* PCC7119,[88] and *Azotobacter vinelandii*.[89] Such wealth of non-native conformations, which is not exclusive of the flavodoxin family, raises questions such as whether equilibrium intermediates correspond with the intermediates transiently found during the folding reaction or, more specifically, whether equilibrium MGs and kinetic MGs are the same thing, as it has been proposed for other proteins.[22,23] The ensemble model of the equilibrium *Hp* apoFld MG here obtained can shed some light on this important question, so we will briefly compare its tridimensional structure with what is structurally known of other non-native apoflavodoxin conformations.

The best known non-native apoflavodoxin conformation is the equilibrium intermediate accumulating in the thermal unfolding of *Anabaena* apoFld (Figure 5c), which has been characterized in great detail by equilibrium Φ-analysis,[55] NMR and SAXS,[90] and classical MD simulations.[91] In this intermediate, which contains a large natively folded moiety exhibiting native-like energetics, the three characteristic flavodoxin layers (αβα) are formed, all secondary structure elements but β5b are present, and the precise side chain packing in the hydrophobic clusters established at the layers interfaces are essentially preserved. Analogous conformations have been described in equilibrium studies for *Hp* (the lower temperature equilibrium intermediate in Reference 92) and for *A. vinelandii* (PUF2 and PUF3 in Reference 93) apoflavodoxins. The MG here studied (Figure 5b) greatly differs in both extent of native structure and energetics from the *Anabaena* thermal intermediate (Figure 5c). The two conformations only appear to share a tendency to loop disordering and to a greater debilitation of persisting β strands at their C-termini and of persisting α-helices at their N-termini. Limited data available on a more thermostable *Hp* thermal intermediate[92] and on additional *A vinelandii* PUFs[93] also suggest they are not related to the MG.

In the folding reaction of the *Anabaena* and *A. vinelandii* apoflavodoxins transient kinetic intermediates have been detected whose similarities and differences have been discussed.[88,89] The *A. vinelandii* off-pathway kinetic intermediate has been identified with an intermediate that accumulates in the Gu-HCl chemical unfolding of the native protein.[94] Extensive characterization indicates that this intermediate lacks a central β-sheet and its core is formed by non-native packing of helical regions,[95] which is most dissimilar to the MG here studied. Finally, for *Anabaena* apoFld, the transition state of the folding reaction has been characterized in some detail by classic kinetic Φ-analysis and a schematic structure has been proposed.[96] The transition state displays relatively high Φ values and its schematic structure describes a folding nucleus formed by packing of helices α3, α4, and α5 onto strands β3 and β4, which does not correspond either with the schematic structure of the MG (Figure 1d).

Thus, the tridimensional structure of the *Hp* apoFld MG here obtained appears to differ from the native structure (Figure 5a) and from any other partially unfolded conformation (either thermal intermediate, kinetic intermediate or transition state) previously characterized in structurally related long-chain flavodoxins. Overall, the structural information available suggests that, when exposed to similar solution conditions, structurally related apoflavodoxins might populate analogous non-native conformations, but it also indicates that those non-native conformations do not need to resemble one another. As this may also be the case for other proteins, the hope that the structure and energetics of MG-like kinetic intermediates can be inferred in a general way from the analysis of MGs at equilibrium seems questionable, as it may depend on the protein studied and cannot be anticipated. The apoflavodoxin case strongly argues that structural information of the highest resolution possible has to be obtained and compared before any two partially unfolded conformations of a protein arising upon dissimilar perturbations of the energy landscape can be claimed to be equivalent.

4 | CONCLUSIONS

At present, BMD simulation guided by experimentally determined Φ values seems a valuable approach to obtain realistic ensembles with atomic resolution representing elusive protein conformations such as equilibrium MGs. Classical unbiased relaxation MD simulations may not be equally successful for this task yet, but sustained progress in force fields, water models, and simulation techniques might soon fill the gap. The atomistic ensemble model of an MG presented here may help to better understand the energetics of this long puzzling protein conformation. The methodology developed could be used to analyze the structure of IDPs that exhibit MG characteristics.

5 | MATERIALS AND METHODS

5.1 | Experimental part

To probe the 164-residue *Hp* apoFld MG structure, 52 residues have been mutated to shorter ones, usually alanines (Table 1). Where appropriate or needed, other shortenings were done (e.g., Tyr to Phe, Ile to Val, or Ala to Gly) that in no case introduced side chain branching not present in the WT residue. For two residues (Thr95 and Thr148), two and three alternative shortenings, respectively, were introduced. A total of 55 mutants have been analyzed. In each mutant, hydrogen bonds and/or van der Waals interactions formed between different secondary structure elements packed in the native apoflavodoxin structure have been removed. The materials used and procedures followed for the site-directed mutagenesis, protein expression and purification,[97,98] as well as for the measurements of the conformational stability of WT and mutant apoflavodoxins at pH 2 and at pH 7 are detailed in the *SI Materials and Methods* (Supplementary Information).

5.2 | Modeling of *Hp* apoFld MG and analyses

The HQBM module with the reaction coordinate RC3[56,57] implemented in the CHARMM package (v.44b2)[64,65] was used to perform the BMD simulations, whereas for the unbiased (all-atom explicit-solvent) MD simulations the GROMACS package (v.2018)[99] was used under the force-field/water-model pairs of either the Charmm22 with CMAP correction (version 2.0)[100] plus Tip3p[101] or the Amber99SB-disp plus Tip4p-d-disp.[82] The remaining relevant details on the implementation of these two modeling approaches and the analyses performed on the simulated trajectories and the selected ensembles of conformations are also summarized in the *SI Materials and Methods* (Supplementary Information).

AUTHOR CONTRIBUTIONS

Juan José Galano-Frutos: Conceptualization (supporting); data curation (lead); formal analysis (equal); investigation (equal); methodology (equal); writing – original draft (equal); writing – review and editing (supporting). **Renzo Torreblanca:** Data curation (equal); formal analysis (supporting); methodology (equal); writing – original draft (supporting). **Helena García-Cebollada:** Conceptualization (supporting); data curation (equal); formal analysis (equal); investigation (equal); methodology (equal); writing – original draft (equal). **Javier Sancho:** Conceptualization (lead); data curation (equal); formal analysis (equal); funding acquisition (lead); investigation (equal); writing – original draft (equal); writing – review and editing (lead).

ACKNOWLEDGMENTS

We acknowledge financial support from MICINN, Spain, (PID2019-107293GB-I00 and PDC2021-121341-I00 grants) and Gobierno de Aragón, Spain, (E45_20R). H. G.-C. is the recipient of an FPU16/04232 doctoral contract from MCINN.

CONFLICT OF INTEREST

The authors declare no conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

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

REFERENCES

1. Ohgushi M, Wada A. 'Molten-globule state': A compact form of globular proteins with mobile side-chains. FEBS Lett. 1983;164:21–24.
2. Ptitsyn OB, Dolgikh DA, Gilmanshin RI, Shakhnovich EI, Finkelshtein AV. Fluctuating state of the protein globule. Mol Biol (Mosk). 1983;17:569–576.
3. Dolgikh DA, Gilmanshin RI, Brazhnikov EV, et al. α-Lactalbumin: Compact state with fluctuating tertiary structure? FEBS Lett. 1981;136:311–315.
4. Baldwin RL, Rose GD. Molten globules, entropy-driven conformational change and protein folding. Curr Opin Struct Biol. 2013;23:4–10.
5. Uversky VN. Biophysical methods to investigate intrinsically disordered proteins: Avoiding an "elephant and blind men" situation. In: Felli IC, Pierattelli R, editors. Advances in experimental medicine and biology. Volume 870. New York LLC: Springer, 2015; p. 215–260.
6. Oldfield CJ, Dunker AK. Intrinsically disordered proteins and intrinsically disordered protein regions. Annu Rev Biochem. 2014;83:553–584.
7. Daggett V, Levitt M. A model of the molten globule state from molecular dynamics simulations. Proc Natl Acad Sci U S A. 1992;89:5142–5146.
8. Kjaergaard M, Teilum K, Poulsen FM. Conformational selection in the molten globule state of the nuclear coactivator binding domain of CBP. Proc Natl Acad Sci U S A. 2010;107:12535–12540.
9. Bhattacharjee N, Rani P, Biswas P. Capturing molten globule state of α-lactalbumin through constant pH molecular dynamics simulations. J Chem Phys. 2013;138:095101.
10. Shimizu M, Kajikawa Y, Kuwajima K, Dobson CM, Okamoto Y. Determination of the structural ensemble of the molten globule state of a protein by computer simulations. Proteins. 2019;87:635–645.
11. Rösner HI, Redfield C. The human α-lactalbumin molten globule: Comparison of structural preferences at pH 2 and pH 7. J Mol Biol. 2009;394:351–362.
12. Naganathan AN, Orozco M. The native ensemble and folding of a protein molten-globule: Functional consequence of downhill folding. J Am Chem Soc. 2011;133:12154–12161.
13. Nakamura S, Seki Y, Katoh E, Kidokoro S. Thermodynamic and structural properties of the acid molten globule state of horse cytochrome c. Biochemistry. 2011;50:3116–3126.
14. Naiyer A, Hassan MI, Islam A, Sundd M, Ahmad F. Structural characterization of MG and pre-MG states of proteins by MD simulations, NMR, and other techniques. J Biomol Struct Dyn. 2015;33:2267–2284.
15. Marion J, Trovaslet M, Martinez N, et al. Pressure-induced molten globule state of human acetylcholinesterase: Structural and dynamical changes monitored by neutron scattering. Phys Chem Chem Phys. 2015;17:3157–3163.
16. Elms PJ, Chodera JD, Bustamante C, Marqusee S. The molten globule state is unusually deformable under mechanical force. Proc Natl Acad Sci U S A. 2012;109:3796–3801.
17. Lindhoud S, Pirchi M, Westphal AH, Haran G, Van Mierlo CPM. Gradual folding of an off-pathway molten globule detected at the single-molecule level. J Mol Biol. 2015;427:3148–3157.
18. Bychkova VE, Semisotnov GV, Balobanov VA, Finkelstein AV. The molten globule concept: 45 years later. Biochemistry. 2018;83:S33-S47.
19. Kuwajima K. The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure. Proteins Struct Funct Bioinform. 1989;6:87–103.
20. Ptitsyn OB, Pain RH, Semisotnov GV, Zerovnik E, Razgulyaev OI. Evidence for a molten globule state as a general intermediate in protein folding. FEBS Lett. 1990;262:20–24.
21. Bhattacharyya S, Varadarajan R. Packing in molten globules and native states. Curr Opin Struct Biol. 2013;23:11–21.
22. Eliezer D, Jennings PA, Wright PE, Doniach S, Hodgson KO, Tsuruta H. The radius of gyration of an apomyoglobin folding intermediate. Science. 1995;270:487–488.
23. Jennings PA, Wright PE. Formation of a molten globule intermediate early in the kinetic folding pathway of apomyoglobin. Science. 1993;262:892–896.
24. Roder H, Elöve GA, Englander SW. Structural characterization of folding intermediates in cytochrome c by H-exchange labelling and proton NMR. Nature. 1988;335:700–704.
25. Kuwajima K. The molten globule state of α-lactalbumin. FASEB J. 1996;10:102–109.
26. Colón W, Roder H. Kinetic intermediates in the formation of the cytochrome c molten globule. Nat Struct Biol. 1996;3:1019–1025.
27. Zhou B, Tian K, Jing G. An in vitro peptide folding model suggests the presence of the molten globule state during nascent peptide folding. Protein Eng. 2000;13:35–39.
28. Bai P, Peng Z. Cooperative folding of the isolated alpha-helical domain of hen egg-white lysozyme. J Mol Biol. 2001;314:321–329.
29. Arai M, Ito K, Inobe T, et al. Fast compaction of α-lactalbumin during folding studied by stopped-flow X-ray scattering. J Mol Biol. 2002;321:121–132.
30. Bychkova VE, Pain RH, Ptitsyn OB. The 'molten globule' state is involved in the translocation of proteins across membranes? FEBS Lett. 1988;238:231–234.
31. van der Goot FG, González-Mañas JM, Lakey JH, Pattus F. A "molten-globule" membrane-insertion intermediate of the pore-forming domain of colicin A. Nature. 1991;354:408–410.

32. Uversky VN, Narizhneva NV, Ivanova TV, Kirkitadze MD, Tomashevski AY. Ligand-free form of human α-fetoprotein: Evidence for the molten globule state. FEBS Lett. 1997;410: 280–284.
33. Cai S, Singh BR. Role of the disulfide cleavage induced molten globule state of type A botulinum neurotoxin in its endopeptidase activity. Biochemistry. 2001;40:15327–15333.
34. Benke S, Roderer D, Wunderlich B, Nettels D, Glockshuber R, Schuler B. The assembly dynamics of the cytolytic pore toxin ClyA. Nat Commun. 2015;6:1–15.
35. Dobson CM. Protein folding and misfolding. Nat. 2003;426: 884–890.
36. Thomas PJ, Qu BH, Pedersen PL. Defective protein folding as a basis of human disease. Trends Biochem Sci. 1995;20: 456–459.
37. Morrow JA, Hatters DM, Lu B, et al. Apolipoprotein E4 forms a molten globule: A potential basis for its association with disease*. J Biol Chem. 2002;277:50380–50385.
38. Santucci R, Sinibaldi F, Fiorucci L. Protein folding, unfolding and misfolding: Role played by intermediate States. Mini Rev Med Chem. 2008;8:57–62.
39. Hatters DM, Peters-Libeu CA, Weisgraber KH. Apolipoprotein E structure: Insights into function. Trends Biochem Sci. 2006;31:445–454.
40. Svensson M, Sabharwal H, Håkansson A, et al. Molecular characterization of α–lactalbumin folding variants that induce apoptosis in tumor cells*. J Biol Chem. 1999;274:6388–6396.
41. Bychkova VE, Ptitsyn OB. Folding intermediates are involved in genetic diseases? FEBS Lett. 1995;359:6–8.
42. Judy E, Kishore N. A look back at the molten globule state of proteins: Thermodynamic aspects. Biophys Rev. 2019;11: 365–375.
43. Marshall B, Warren JR. Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. Lancet. 1984;323:1311–1315.
44. Freigang J, Diederichs K, Schäfer KP, Welte W, Paul R. Crystal structure of oxidized flavodoxin, an essential protein in *Helicobacter pylori*. Protein Sci. 2002;11:253–261.
45. Salillas S, Sancho J. Flavodoxins as novel therapeutic targets against *Helicobacter pylori* and other gastric pathogens. Int J Mol Sci. 2020;21:1881.
46. Galano JJ, Alías M, Pérez R, Velázquez-Campoy A, Hoffman PS, Sancho J. Improved flavodoxin inhibitors with potential therapeutic effects against *Helicobacter pylori* infection. J Med Chem. 2013;56:6248–6258.
47. Cremades N, Velázquez-Campoy A, Martínez-Júlvez M, et al. Discovery of specific flavodoxin inhibitors as potential therapeutic agents against *Helicobacter pylori* infection. ACS Chem Biol. 2009;4:928–938.
48. Salillas S, Alías M, Michel V, et al. Design, synthesis, and efficacy testing of nitroethylene- and 7-nitrobenzoxadiazol-based flavodoxin inhibitors against *Helicobacter pylori* drug-resistant clinical strains and in *Helicobacter pylori*-infected mice. J Med Chem. 2019;62:6102–6115.
49. Martínez-Júlvez M, Cremades N, Bueno M, et al. Common conformational changes in flavodoxins induced by FMN and anion binding: The structure of *Helicobacter pylori* apoflavodoxin. Proteins Struct Funct Bioinform. 2007;69:581–594.
50. Cremades N, Sancho J. Molten globule and native state ensemble of *Helicobacter pylori* flavodoxin: Can crowding, osmolytes or cofactors stabilize the native conformation relative to the molten globule? Biophys J. 2008;95:1913–1927.
51. Cremades N, Bueno M, Neira JL, Velázquez-Campoy A, Sancho J. Conformational stability of *Helicobacter pylori* flavodoxin: Fit to function at pH 5. J Biol Chem. 2008;283:2883–2895.
52. Genzor CG, Beldarraín A, Gómez-Moreno C, López-Lacomba JL, Cortijo M, Sancho J. Conformational stability of apoflavodoxin. Protein Sci. 1996;5:1376–1388.
53. López-Llano J, Campos LA, Bueno M, Sancho J. Equilibrium Φ-analysis of a molten globule: The 1-149 apoflavodoxin fragment. J Mol Biol. 2006;356:354–366.
54. Maldonado S, Jiménez MÁ, Langdon GM, Sancho J. Cooperative stabilization of a molten globule apoflavodoxin fragment. Biochemistry. 1998;37:10589–10596.
55. Campos LA, Bueno M, Lopez-Llano J, Jiménez MÁ, Sancho J. Structure of stable protein folding intermediates by equilibrium ϕ-analysis: The apoflavodoxin thermal intermediate. J Mol Biol. 2004;344:239–255.
56. Paci E, Karplus M. Forced unfolding of fibronectin type 3 modules: An analysis by biased molecular dynamics simulations. J Mol Biol. 1999;288:441–459.
57. Paci E, Vendruscolo M, Dobson CM, Karplus M. Determination of a transition state at atomic resolution from protein engineering data. J Mol Biol. 2002;324:151–163.
58. Bueno M, Campos LA, Estrada J, Sancho J. Energetics of aliphatic deletions in protein cores. Protein Sci. 2006;15:1858–1872.
59. 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.
60. Ayuso-Tejedor S, Abián O, Sancho J. Underexposed polar residues and protein stabilization. Protein Eng Des Sel. 2011;24: 171–177.
61. Serrano L, Sancho J, Hirshberg M, Fersht AR. α-Helix stability in proteins: I. Empirical correlations concerning substitution of side-chains at the N and C-caps and the replacement of alanine by glycine or serine at solvent-exposed surfaces. J Mol Biol. 1992;227:544–559.
62. López-Llano J, Campos LA, Sancho J. Alpha-helix stabilization by alanine relative to glycine: Roles of polar and apolar solvent exposures and of backbone entropy. Proteins. 2006;64: 769–778.
63. Fersht AR, Matouschek A, Serrano L. The folding of an enzyme: I. Theory of protein engineering analysis of stability and pathway of protein folding. J Mol Biol. 1992;224:771–782.
64. Brooks BR, Brooks CL III, MacKerell AD Jr, et al. CHARMM: The biomolecular simulation program. J Comput Chem. 2009; 30:1545–1614.
65. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem. 1983;4:187–217.
66. Perkins SJ, Wüthrich K. Ring current effects in the conformation dependent NMR chemical shifts of aliphatic protons in

the basic pancreatic trypsin inhibitor. Biochim Biophys Acta—Protein Struct. 1979;576:409–423.

67. Kelly S, Price N. The use of circular dichroism in the investigation of protein structure and function. Curr Protein Pept Sci. 2000;1:349–384.

68. Vuilleumier S, Sancho J, Loewenthal R, Fersht AR. Circular dichroism studies of barnase and its mutants: Characterization of the contribution of aromatic side chains. Biochemistry. 1993;32:10303–10313.

69. Drew ED, Janes RW. PDBMD2CD: Providing predicted protein circular dichroism spectra from multiple molecular dynamics-generated protein structures. Nucleic Acids Res. 2020;48:W17–W24.

70. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22:2577–2637.

71. Kataoka M, Kuwajima K, Tokunaga F, Goto Y. Structural characterization of the molten globule of α-lactalbumin by solution X-ray scattering. Protein Sci. 1997;6:422–430.

72. García De La Torre J, Huertas ML, Carrasco B. Calculation of hydrodynamic properties of globular proteins from their atomic-level structure. Biophys J. 2000;78:719–730.

73. Ortega A, Amorós D, García de la Torre J. Prediction of hydrodynamic and other solution properties of rigid proteins from atomic- and residue-level models. Biophys J. 2011;101: 892–898.

74. Greene RF, Pace CN. Urea and guanidine hydrochloride denaturation of ribonuclease, lysozyme, α-chymotrypsin, and β-lactoglobulin. J Biol Chem. 1974;249:5388–5393.

75. Myers JK, Pace CN, Scholtz JM. Denaturant m values and heat capacity changes: Relation to changes in accessible surface areas of protein unfolding. Protein Sci. 1995;4:2138–2148.

76. Li Z, Hirst JD. Quantitative first principles calculations of protein circular dichroism in the near-ultraviolet. Chem Sci. 2017;8:4318–4333.

77. Redfield C. NMR studies of partially folded molten-globule states. Methods Mol Biol. 2004;278:233–254.

78. Galano-Frutos JJ, Garciá-Cebollada H, Sancho J. Molecular dynamics simulations for genetic interpretation in protein coding regions: Where we are, where to go and when. Brief Bioinform. 2021;22:3–19.

79. 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;59:4350–4360.

80. Hospital A, Goñi JR, Orozco M, Gelpí JL. Molecular dynamics simulations: Advances and applications. Adv Appl Bioinform Chem. 2015;8:37.

81. Kumari I, Sandhu P, Ahmed M, Akhter Y. Molecular dynamics simulations, challenges and opportunities: A biologist's prospective. Curr Protein Pept Sci. 2017;18:1163-1179.

82. 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–E4766.

83. Campos LA, Garcia-Mira MM, Godoy-Ruiz R, Sanchez-Ruiz JM, Sancho J. Do proteins always benefit from a stability increase? Relevant and residual stabilisation in a three-state protein by charge optimisation. J Mol Biol. 2004;344:223–237.

84. Pace CN. Determination and analysis of urea and guanidine hydrochloride denaturation curves. Methods Enzymol. 1986; 131:266–280.

85. Santoro MM, Bolen DW. Unfolding free energy changes determined by the linear extrapolation method. 1. Unfolding of phenylmethanesulfonyl alpha-chymotrypsin using different denaturants. Biochemistry. 1988;27:8063–8068.

86. Vallat B, Webb B, Fayazi M, et al. New system for archiving integrative structures. Acta Crystallogr Sect D: Struct Biol. 2021;77:1486–1496.

87. Ayuso-Tejedor S, Angarica VE, Bueno M, et al. Design and structure of an equilibrium protein folding intermediate: A hint into dynamical regions of proteins. J Mol Biol. 2010;400:922–934.

88. Sancho J. Flavodoxins: Sequence, folding, binding, function and beyond. Cell Mol Life Sci. 2006;63:855–864.

89. Houwman JA, van Mierlo CPM. Folding of proteins with a flavodoxin-like architecture. FEBS J. 2017;284:3145–3167.

90. Ayuso-Tejedor S, García-Fandiño R, Orozco M, Sancho J, Bernadó P. Structural analysis of an equilibrium folding intermediate in the apoflavodoxin native ensemble by small-angle X-ray scattering. J Mol Biol. 2011;406:604–619.

91. García-Fandiño R, Bernadó P, Ayuso-Tejedor S, Sancho J, Orozco M. Defining the nature of thermal intermediate in 3 state folding proteins: Apoflavodoxin, a study case. PLoS Comput Biol. 2012;8:e1002647.

92. Cremades N, Velazquez-Campoy A, Freire E, Sancho J. The flavodoxin from *Helicobacter pylori*: Structural determinants of thermostability and FMN cofactor binding. Biochemistry. 2008;47:627–639.

93. Bollen YJM, Kamphuis MB, van Mierlo CPM. The folding energy landscape of apoflavodoxin is rugged: Hydrogen exchange reveals nonproductive misfolded intermediates. Proc Natl Acad Sci U S A. 2006;103:4095–4100.

94. Bollen YJM, Sánchez IE, van Mierlo CPM. Formation of on- and off-pathway intermediates in the folding kinetics of *Azotobacter vinelandii* apoflavodoxin. Biochemistry. 2004;43: 10475–10489.

95. Nabuurs SM, van Mierlo CPM. Interrupted hydrogen/deuterium exchange reveals the stable core of the remarkably helical molten globule of α-β parallel protein flavodoxin. J Biol Chem. 2010;285:4165–4172.

96. Bueno M, Ayuso-Tejedor S, Sancho J. Do proteins with similar folds have similar transition state structures? A diffuse transition state of the 169 residue apoflavodoxin. J Mol Biol. 2006;359:813–824.

97. Cremades N, Bueno M, Toja M, Sancho J. Towards a new therapeutic target: *Helicobacter pylori* flavodoxin. Biophys Chem. 2005;115:267–276.

98. Tollin G, Edmondson DE. Flavoprotein chemistry. III. Flavine protein interactions and the redox properties of the Shethna flavoprotein. Biochemistry. 1971;10:133–145.

99. van der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: Fast, flexible, and free. J Comput Chem. 2005;26:1701–1718.

100. 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 simulations. J Comput Chem. 2004;25:1400–1415.

101. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. J Chem Phys. 1998;79:926–935.

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, Torreblanca R, García-Cebollada H, Sancho J. A look at the face of the molten globule: Structural model of the *Helicobacter pylori* apoflavodoxin ensemble at acidic pH. Protein Science. 2022; 31(11):e4445. https://doi.org/10.1002/pro.4445