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,*}

¹Departamento de Bioquímica y Biología Molecular y Celular, Facultad de Ciencias, Universidad de Zaragoza, Zaragoza 50009, Spain; ²Biocomputation and Complex Systems Physics Institute (BIFI)-GBsC-CSIC Joint Unit, Universidad de Zaragoza, Zaragoza 50018, Spain; ³Aragon Health Research Institute (IIS Aragón), Zaragoza 50009, Spain.

KEYWORDS: Molten globule, Integrative protein model, Biased MD simulations, Apoflavodoxin, Protein folding landscape, Intrinsically disordered protein.

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 WT and 55 apoflavodoxin mutants to derive experimental equilibrium Φ values that have been refined to reproduce the experimental hydrodynamic radius and circular dichroism (CD) spectrum. The refined ensemble, deposited in PDB-Dev, successfully explains the characteristic ¹H-NMR and near-UV CD spectral features of the MG as well as its unfolding Δ SASA. This integrative model of a 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.

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 molten globule-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 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 to 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 holo⁴⁴ and apo⁴⁹ 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 a MG conformation at acidic pH. The earlier discovered MG of Anabaena PCC7119 apoFld aggregates⁵² and only limited structural information could be obtained of it from the study of a monomeric truncated variant,^{53,54} using equilibrium Φ -analysis⁵⁵. 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 ¹H-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 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.

RESULTS

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 Å³, 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 of WT, suggesting that side chain repacking has occurred near W155, which might be related to the stabilization observed. In Q61A, the wild-type glutamine, fully buried according to ProtSA,59 is replaced by an alanine, aiming to create a destabilising cavity. However, replacement of buried glutamines by apolar residues has been previously associated to 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 ProtSA⁵⁹ calculation.

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 analysed are determined and a Φ value is calculated. We have used the data in Table S1 to calculate, through Eq. 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.

$$\Phi = \frac{\Delta \Delta G_{wt-mut}^{MG}}{\Delta \Delta G_{wt-mut}^{native}} \tag{1}$$

In Eq. 1, $\Delta\Delta G_{wt-mut}^{MG}$ is the unfolding free energy difference of WT MG minus that of a mutant MG, and $\Delta\Delta G_{wt-mut}^{native}$ 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 nativelike 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 favour 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.

Residue	Structure Element	∆SASA (Ų)ª	H-bond/Salt Bridge Partner Residue (atom)	van der Waals Contacts (<4 Å)	Contacting Elements	Mutation	$\Delta\Delta \mathbf{G} \ (\mathbf{kcal/mol})^{\mathbf{b}}$		Φ value ^c
							рН 2	pH 7	
							(MG)	(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	$0.46{\pm}0.07$	-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	$0.30{\pm}0.05$	-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	$-0.12{\pm}0.05$	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	$0.19{\pm}0.08$	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
N43	α2	70.6	D75(OD1)/K79(NZ)	K39/F42	$\alpha 2/\alpha 3/\beta 4$	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	$\beta 3/\beta 4/loop \rightarrow C$ -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	$-0.91{\pm}0.09$	0.29 ± 0.08
L50	β3	0.0		F42/V48/W64/F76/I81/L83/I103	$\alpha 2/\beta 3/\alpha 3/\beta 4/\alpha 4$	L50A	-0.17 ± 0.05	0.71 ± 0.02	e
T54	β3	28.3	L86(O)/G99(N)	G58/D59/A97/E98	$\beta 4/\alpha 4$	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	$loop \rightarrow \alpha 1/\alpha 3/\beta 3$	Q61A	-0.42 ± 0.04	$-0.82{\pm}0.03$	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	$loop \rightarrow \alpha 3/\alpha 3/\alpha 4$	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)		$\alpha 2/\alpha 3$	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. Structural and thermodynamic description of mutations implemented in *Hp* apoFld.

Q89	loop→α4	0.6	H122(N)/D140(OD2)	D90/Y121/N143/Q144	$loop \rightarrow \alpha 4/loop \rightarrow \alpha 5$	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	$0.15{\pm}0.15^{ m f}$
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	β6Ь	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
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

^a Solvent accessible surface area difference (Δ SASA) calculated with the ProtSA server⁵⁹ as the SASA in the folded structure minus the mean SASA in the unfolded ensemble.

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

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

 d Non reliable Φ value due to small $\Delta\!\Delta G$ native.

 $^{\rm e}$ Φ -value outside the 0-1 range, indicative of formation of non-native interactions in MG.

 $^{\rm f}$ Non reliable Φ value due to its high relative error.



Figure 1. Distribution of stability effects and Φ values along the sequence and structure of Hp apoFld, and schematic structure of the 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 $\Phi = 0$ 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 (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 colour-coded Φ values of the different residues probed as indicated by the colour scale at the bottom (red: native interactions totally lost; green: native interactions totally retained). For non-probed residues, the ribbon colour is pale blue for the N-terminal half of the protein (residues 1–82) and grey 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.

Ten mutations produced small changes $|\Delta\Delta G_{wt-mut}| \le 0.3 \text{ kcal/mol})$ 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 ten 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 thirty-two mutations gave canonical Φ values between 0 and 1. They include thirteen very small negative values, i.e. $|\Phi| \le 0.11$, that will be considered to indicate full disruption of the probed interaction, i.e. $\Phi = 0$. The thirty-two 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 $\Delta\Delta G_{wt-mut}$ 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 thirty-two canonical and thirteen 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 and 1e). Essentially, the MG appears to retain an organization in three layers, with the packing of the $\alpha 1 \alpha 5$ layer onto the central β -sheet being mostly native-like (although much debilitated) and that of the $\alpha 2\alpha 3\alpha 4$ layer mostly non-native. The loops, including the FMN binding 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.

Biased MD Simulations Transforming the Native Conformation into a MG Ensemble. BMD simulations^{56,57,64,65} have been performed to progressively drive the starting Hp apoFld conformation (PDB ID 2BMV)⁴⁹ toward 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 atoms, 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 (Figure 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 three 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 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 ¹H-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 atoms) 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 (β 6a-c) of the long loop that intercalates in strand β 5, plus the second segment of this strand (β 5b) are more debilitated than preceding strands β 1- β 5a. On the other hand, the helical layer formed by helices $\alpha 2$, $\alpha 3$ and $\alpha 4$ is quite homogeneously destabilised, with the individual helices retaining 32-40 % of their initial residues (Figure 3a). Helix α 3 encompasses two helical stretches (α 3a and α 3b) in the native protein, of which α 3b appears unfolded in the ensemble. In contrast with the previous two layers, the helical layer comprising helices $\alpha 1$ and $\alpha 5$ is inhomogeneously destabilised, with a strong persistence of most of helix $\alpha 1$ (90 %) and a severe destabilisation of $\alpha 5$ (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 and 3c. 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 and 3c with the corresponding experimental spectra. The MG reduced secondary structure content (see assignments given by DSSP⁷⁰ 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 $\Delta\epsilon$ units, while those of the 10-best matching ones from the alternative BMD ensembles (*whole side chain* and *side chain heavy atoms*) are higher (0.51-0.74 $\Delta\epsilon$ units and $0.37-0.78 \Delta \varepsilon$ units, respectively, Figure S5a and S5b). 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.



Figure 2. Comparison of calculated Φ values with experimentally determined ones. Experimental vs. calculated comparison of Φ values for residues used as restraints in the BMD simulations is shown for one replica of the *all-atom* set of simulations. 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 vs. experimental Φ values, and the correlation between them along the last no of the BMD trajectory. These plots reflect the stability of the calculated Φ values and the good matching between them and the experimental ones.

MGs are expanded conformations compared to native ones⁷¹ 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 to the transformation of the native conformation into the MG, rather than directly comparing the actual Rh values. The experimental data indicates 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 atoms*, 6.7 ± 0.3 % R_h increase) as they consist of more compact conformations (Figure S3b1-b3).



Figure 3. Secondary structure and predicted vs. experimental far-UV CD spectra of the *all-atom* MG ensembles. (a) Persistence of native secondary structure in the BMD *all-atom* MG ensemble depicted as fractions of residues in secondary structure elements of the native structure (PDB ID 2BMV⁴⁹ 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 server⁶⁹ 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 $\Delta \epsilon$ units) for the 10 best-matching predicted CD profiles (i.e. those with the lowest RMSD vs. the experimental one of the *Hp* apoFld MG) range from 0.23 to 0.50.

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 concentration) for the native protein (m = 2,150 ± 50 cal/(mol×M)) and that for the MG (m =

 $1,330 \pm 40 \text{ cal/(mol \times M)}$).⁵¹ As m values are linearly related to the increase in solvent exposure (Δ SASA) 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 Δ SASAs derived from m values with Δ SASAs directly calculated⁵⁹ on the structures. Considering

that the linear relationship obtained by Myers et al. (m = 368 +0.11× Δ SASA, with R = 0.90, m in cal/(mol×M) and Δ SASA in Å²) 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 Δ SASA of native Hp apoFld has been calculated at 11,203 ($\pm 2,205$) Å² and that of the MG at 6.065 (± 1.669) Å². The Δ SASA computed with ProtSA for the native conformation, 10,482 Å2 (Table S5), agrees within experimental error with that derived from the m value. Out of the different MG ensembles analyzed, the averaged Δ SASA 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 Å². We have extended this Δ SASA 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 Δ SASA and experimental m values have been obtained (with R = 0.96 or 0.92, respectively, see Table S9 and Figure S6b and S6c) which are analogous to those in Myers' paper (70). For the native conformation, both the polar (3.227 ± 635) Å²) and non-polar (7,218 \pm 1,420 Å²) Δ SASAs calculated from the m values using the updated relationships agree with the Δ SASAs computed from the PDB file (2BMV): 2,977 ± 82 and 7,506 \pm 81 Å², respectively. For the *all-atom-10* ensemble, both the polar $(2,856 \pm 112 \text{ Å}^2)$ and apolar $(5,796 \pm 248 \text{ Å}^2)$ Δ SASA computed from the *all-atom-10* ensemble (Table S5) are a bit higher than those obtained from the m value (1,743 \pm 480 and 4,045 \pm 1,114 Å², 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 (RMSF) of the three tyrosine (Tyr92, Tyr104 and Tvr121) 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 to 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 ¹H-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 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 ¹H-NMR spectrum.⁵¹

Attempts to Define a 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 wild type structures.78-81 A frequent way to achieve this is to model an amino acid substitution in the wildtype 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 energetics⁷⁹ but does not appear to over-compact 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 $10 \ \mu s$) 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.



Figure 4. General workflow used to filter MG ensembles from classical unbiased 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 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 to the filtering carried out in (a) or (b) appear indicated. Φ -correlation thresholds are also indicated in the filters descriptions.

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 μ s 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 Δ SASA of 9,105 (±33) Å² (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 ¹H-NMR spectrum, these ensembles capture that the interactions between tryptophan side chains and methyl groups from IIe 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.

dominant clusters identified in As the 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 10conformation ensemble (Charmm-10, Table S10) was obtained encompassing the conformations exhibiting ≥ 0.6 experimental vs. calculated Φ correlations (Figure S9 a) and the lowest experimental vs. 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 analysed 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 \pm 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 S9 b) 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 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 Rh 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 overcompacting (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 Δ SASA of 8,192 (± 336) Å² is slightly larger than the upper limit calculated from 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 to the disappearance of high field peaks in the ¹H-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.

DISCUSSION

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 versus 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 versus 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, e.g. 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 wild-type 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 \pm 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 \pm 0.3 kcal/mol). Only 5 mutations destabilize the MG by more than 0.4 kcal/mol and only 5 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 and 1b).

A detailed structural interpretation of the 32 canonical and 13 non-canonical 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 ($\alpha/\beta/\alpha$) 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.

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 us-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 modelled 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 (Rh 9.6 % larger than that of the native conformation) is reasonably well captured by the ensemble (8.4 \pm 0.3 % larger $R_h,$ see Table S4). This MG expansion is also reflected in the experimental unfolding Δ SASA (6,065 ± 1,665 Å²), which is much lower

than that of the native conformation $(11,203 \pm 2,205 \text{ Å}^2)$. The calculated Δ SASA of the refined ensemble is a bit higher $(8,652 \pm 298 \text{ Å}^2)$ than the experimental Δ SASA, 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 ¹H-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 database⁸⁶ 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 over compacted 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 Δ SASA obtained is not far from the experimental one (Table 2) 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 AmberSB99disp-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 analyse the conformational ensembles of the molten globule-like IDPs and IDP regions,^{5,6} constituting up to 40 % of the human proteome.

Figure 5. 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 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 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 α 1 and β 2 are indicated in the corresponding panel.

The Molten Globule 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. molten globules, equilibrium intermediates, kinetic intermediates and transition states of folding) have been found and described in different detail, mainly in the flavodoxins from Hp, Anabaena PCC711988 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 ($\alpha\beta\alpha$) are formed, all secondary structure elements but β 5b 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 ⁹²) and for *A. vinelandii* (PUF2 and PUF3 in 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 intermediate⁹² and on additional *A vinelandii* PUFs⁹³ 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 $\alpha 3$, $\alpha 4$ and $\alpha 5$ onto strands β 3 and β 4, 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 nonnative 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.

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 a MG presented here may help to better understand the energetics of this long puzzling protein conformation. The methodology developed could be used to analyse the structure of IDPs that exhibit molten globule characteristics.

MATERIALS AND METHODS

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 wild-type residue. For two residues (Thr95 and Thr148), two and three alternative shortenings, respectively, were introduced. A total of 55 mutants have been analysed. 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 wild-type and mutant apoflavodoxins at pH 2 and at pH 7 are detailed in the Supplementary Materials and Methods (Supplementary Information).

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 Tip3p¹⁰¹ 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 summarised in the **Supplementary Materials and Methods** (Supplementary Information).

ASSOCIATED CONTENT

Supplementary Information. Qualitative structural description of the MG; experimental details, including mutagenesis, protein expression and purification, urea denaturation and data fitting for Hp apoFld and the mutants; modeling details and analysis of simulated trajectories and selected MG ensembles. This material is available free of charge via the Internet at <u>http://pubs.acs.org</u>. (PDF)

AUTHOR INFORMATION

Corresponding Author

* Javier Sancho - Email: jsancho@unizar.es.

Notes

The authors declare no competing financial interests.

ACKNOWLEDGMENT

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.

ABBREVIATIONS

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

REFERENCES

1. Ohgushi M, Wada A (1983) 'Molten-globule state': a compact form of globular proteins with mobile side-chains. FEBS Lett. 164:21–24.

2. Ptitsyn OB, Dolgikh DA, Gilmanshin RI, Shakhnovich EI, Finkelshtein A V. (1983) Fluctuating state of the protein globule. Mol. Biol. (Mosk). 17:569–576.

3. Dolgikh DA, Gilmanshin RI, Brazhnikov E V., Bychkova VE, Semisotnov G V., Venyaminov SY, Ptitsyn OB (1981) α -lactalbumin: compact state with fluctuating tertiary structure? FEBS Lett. 136:311–315.

4. Baldwin RL, Rose GD (2013) Molten globules, entropy-driven conformational change and protein folding. Curr. Opin. Struct. Biol. 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. Vol. 870. Springer New York LLC; 2015. pp. 215–260.

6. Oldfield CJ, Dunker AK (2014) Intrinsically disordered proteins and intrinsically disordered protein regions. Annu. Rev. Biochem. 83:553-584.

7. Daggett V, Levitt M (1992) A model of the molten globule state from molecular dynamics simulations. Proc. Natl. Acad. Sci. U. S. A. 89:5142–5146.

8. Kjaergaard M, Teilum K, Poulsen FM (2010) Conformational selection in the molten globule state of the nuclear coactivator binding domain of CBP. Proc. Natl. Acad. Sci. U. S. A. 107:12535.

9. Bhattacharjee N, Rani P, Biswas P (2013) Capturing molten globule state of α -lactalbumin through constant pH molecular dynamics simulations. J. Chem. Phys. 138:095101.

10. Shimizu M, Kajikawa Y, Kuwajima K, Dobson CM, Okamoto Y (2019) Determination of the structural ensemble of the molten globule state of a protein by computer simulations. Proteins 87:635–645.

11. Rösner HI, Redfield C (2009) The Human α -Lactalbumin Molten Globule: Comparison of Structural Preferences at pH 2 and pH 7. J. Mol. Biol. 394:351.

12. Naganathan AN, Orozco M (2011) The Native Ensemble and Folding of a Protein Molten-Globule: Functional Consequence of Downhill Folding. J. Am. Chem. Soc. 133:12154–12161.

13. Nakamura S, Seki Y, Katoh E, Kidokoro S (2011) Thermodynamic and Structural Properties of the Acid Molten Globule State of Horse

Cytochrome c. Biochemistry 50:3116-3126.

14. Naiyer A, Hassan MI, Islam A, Sundd M, Ahmad F (2015) Structural characterization of MG and pre-MG states of proteins by MD simulations, NMR, and other techniques. J. Biomol. Struct. Dyn. 33:2267–2284.

15. Marion J, Trovaslet M, Martinez N, Masson P, Schweins R, Nachon F, Trapp M, Peters J (2015) Pressure-induced molten globule state of human acetylcholinesterase: structural and dynamical changes monitored by neutron scattering. Phys. Chem. Chem. Phys. 17:3157–3163.

16. Elms PJ, Chodera JD, Bustamante C, Marqusee S (2012) The molten globule state is unusually deformable under mechanical force. Proc. Natl. Acad. Sci. 109:3796–3801.

17. Lindhoud S, Pirchi M, Westphal AH, Haran G, Van Mierlo CPM (2015) Gradual Folding of an Off-Pathway Molten Globule Detected at the Single-Molecule Level. J. Mol. Biol. 427:3148–3157.

18. Bychkova VE, Semisotnov G V., Balobanov VA, Finkelstein A V. (2018) The Molten Globule Concept: 45 Years Later. Biochem. 83.

19. Kuwajima K (1989) The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure. Proteins Struct. Funct. Bioinforma. 6:87-103.

20. Ptitsyn OB, Pain RH, Semisotnov GV, Zerovnik E, Razgulyaev OI (1990) Evidence for a molten globule state as a general intermediate in protein folding. FEBS Lett. 262:20–24.

21. Bhattacharyya S, Varadarajan R (2013) Packing in molten globules and native states. Curr. Opin. Struct. Biol. 23:11–21.

22. Eliezer D, Jennings PA, Wright PE, Doniach S, Hodgson KO, Tsuruta H (1995) The radius of gyration of an apomyoglobin folding intermediate. Science (80-.). 270:487–488.

23. Jennings PA, Wright PE (1993) Formation of a Molten Globule Intermediate Early in the Kinetic Folding Pathway of Apomyoglobin. Science (80-.). 262:892–896.

24. Roder H, Elöve GA, Englander SW (1988) Structural characterization of folding intermediates in cytochrome c by H-exchange labelling and proton NMR. Nature 335:700.

25. Kuwajima K (1996) The molten globule state of α -lactalbumin. FASEB J. 10:102–109.

26. Colón W, Roder H (1996) Kinetic intermediates in the formation of the cytochrome c molten globule. Nat. Struct. Biol. 3:1019–1025.

27. Zhou B, Tian K, Jing G (2000) An in vitro peptide folding model suggests the presence of the molten globule state during nascent peptide folding. Protein Eng. 13:35–39.

28. Bai P, Peng Z (2001) Cooperative folding of the isolated alpha-helical domain of hen egg-white lysozyme. J. Mol. Biol. 314:321–329.

29. Arai M, Ito K, Inobe T, Nakao M, Maki K, Kamagata K, Kihara H, Amemiya Y, Kuwajima K (2002) Fast Compaction of α -Lactalbumin During Folding Studied by Stopped-flow X-ray Scattering. J. Mol. Biol. 321:121–132.

30. Bychkova VE, Pain RH, Ptitsyn OB (1988) The 'molten globule' state is involved in the translocation of proteins across membranes? FEBS Lett. 238:231–234.

31. van der Goot FG, González-Mañas JM, Lakey JH, Pattus F (1991) A "molten-globule" membrane-insertion intermediate of the pore-forming domain of colicin A. Nature 354:408–410.

32. Uversky VN, Narizhneva N V, Ivanova T V, Kirkitadze MD, Tomashevski AY (1997) Ligand-free form of human α -fetoprotein: evidence for the molten globule state. FEBS Lett. 410:280–284.

33. Cai S, Singh BR (2001) Role of the Disulfide Cleavage Induced Molten Globule State of Type A Botulinum Neurotoxin in Its Endopeptidase Activity[†]. Biochemistry 40:15327–15333.

34. Benke S, Roderer D, Wunderlich B, Nettels D, Glockshuber R, Schuler B (2015) The assembly dynamics of the cytolytic pore toxin ClyA. Nat. Commun. 6:1–15.

35. Dobson CM (2003) Protein folding and misfolding. Nat. 2003 4266968 426:884–890.

36. Thomas PJ, Qu BH, Pedersen PL (1995) Defective protein folding as a basis of human disease. Trends Biochem. Sci. 20:456–459.

37. Morrow JA, Hatters DM, Lu B, Höchtl P, Oberg KA, Rupp B, Weisgraber KH (2002) Apolipoprotein E4 Forms a Molten Globule: A

Potential Basis for its Association with Disease *. J. Biol. Chem. 277:50380-50385.

38. Santucci R, Sinibaldi F, Fiorucci L (2008) Protein folding, unfolding and misfolding: role played by intermediate States. Mini Rev. Med. Chem. 8:57–62.

39. Hatters DM, Peters-Libeu CA, Weisgraber KH (2006) Apolipoprotein E structure: insights into function. Trends Biochem. Sci. 31:445–454.

40. Svensson M, Sabharwal H, Håkansson A, Mossberg A-K, Lipniunas P, Leffler H, Svanborg C, Linse S (1999) Molecular Characterization of α -Lactalbumin Folding Variants That Induce Apoptosis in Tumor Cells *. J. Biol. Chem. 274:6388–6396.

41. Bychkova VE, Ptitsyn OB (1995) Folding intermediates are involved in genetic diseases? FEBS Lett. 359:6–8.

42. Judy E, Kishore N (2019) A look back at the molten globule state of proteins: thermodynamic aspects. Biophys. Rev. 11:365–375.

43. Marshall B, Warren JR (1984) Unidentified Curved Bacilli in the Stomach of Patients with Gastritis and Peptic Ulceration. Lancet 323:1311–1315.

44. Freigang J, Diederichs K, Schäfer KP, Welte W, Paul R (2002) Crystal structure of oxidized flavodoxin, an essential protein in Helicobacter pylori. Protein Sci. 11:253.

45. Salillas S, Sancho J (2020) Flavodoxins as Novel Therapeutic Targets against Helicobacter pylori and Other Gastric Pathogens. Int. J. Mol. Sci. 21.

46. Galano JJ, Alías M, Pérez R, Velázquez-Campoy A, Hoffman PS, Sancho J (2013) Improved Flavodoxin Inhibitors with Potential Therapeutic Effects against Helicobacter pylori Infection. J. Med. Chem. 56:6248–6258.

47. Cremades N, Velázquez-Campoy A, Martínez-Júlvez M, Neira JL, Pérez-Dorado I, Hermoso J, Jiménez P, Lanas A, Hoffman PS, Sancho J (2009) Discovery of Specific Flavodoxin Inhibitors as Potential Therapeutic Agents against Helicobacter pylori Infection. ACS Chem. Biol. 4:928–938.

48. Salillas S, Alías M, Michel V, Mahía A, Lucía A, Rodrigues L, Bueno J, Galano-Frutos JJ, Reuse H De, Velázquez-Campoy A, et al. (2019) 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. 62:6102–6115.

49. Martínez-Júlvez M, Cremades N, Bueno M, Pérez-Dorado I, Maya C, Cuesta-López S, Prada D, Falo F, Hermoso JA, Sancho J (2007) Common conformational changes in flavodoxins induced by FMN and anion binding: The structure of Helicobacter pylori apoflavodoxin. Proteins Struct. Funct. Bioinforma. 69:581–594.

50. Cremades N, Sancho J (2008) 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. 95:1913–1927.

51. Cremades N, Bueno M, Neira JL, Velázquez-Campoy A, Sancho J (2008) Conformational Stability of Helicobacter pylori Flavodoxin: Fit to Function at pH 5. J. Biol. Chem. 283:2883–2895.

52. Genzor CG, Beldarraín A, Gómez-Moreno C, López-Lacomba JL, Cortijo M, Sancho J (1996) Conformational stability of apoflavodoxin. Protein Sci. 5:1376.

53. López-Llano J, Campos LA, Bueno M, Sancho J (2006) Equilibrium Φ -Analysis of a Molten Globule: The 1-149 Apoflavodoxin Fragment. J. Mol. Biol. 356:354–366.

54. Maldonado S, Jiménez MÁ, Langdon GM, Sancho J (1998) Cooperative Stabilization of a Molten Globule Apoflavodoxin Fragment. Biochemistry 37:10589–10596.

55. Campos LA, Bueno M, Lopez-Llano J, Jiménez MÁ, Sancho J (2004) Structure of Stable Protein Folding Intermediates by Equilibrium φ-Analysis: The Apoflavodoxin Thermal Intermediate. J. Mol. Biol. 344:239–255.

56. Paci E, Karplus M (1999) Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations. J. Mol. Biol. 288:441-459.

57. Paci E, Vendruscolo M, Dobson CM, Karplus M (2002) Determination of a Transition State at Atomic Resolution from Protein

Engineering Data. J. Mol. Biol. 324:151-163.

58. Bueno M, Campos LA, Estrada J, Sancho J (2006) Energetics of aliphatic deletions in protein cores. Protein Sci. 15:1858.

59. Estrada J, Bernadó P, Blackledge M, Sancho J (2009) ProtSA: a web application for calculating sequence specific protein solvent accessibilities in the unfolded ensemble. BMC Bioinforma. 2009 10:1–8.

60. Ayuso-Tejedor S, Abián O, Sancho J (2011) Underexposed polar residues and protein stabilization. Protein Eng. Des. Sel. 24:171–177.

61. Serrano L, Sancho J, Hirshberg M, Fersht AR (1992) 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. 227:544–559.

62. López-Llano J, Campos LA, Sancho J (2006) Alpha-helix stabilization by alanine relative to glycine: roles of polar and apolar solvent exposures and of backbone entropy. Proteins 64:769–778.

63. Fersht AR, Matouschek A, Serrano L (1992) The folding of an enzyme: I. Theory of protein engineering analysis of stability and pathway of protein folding. J. Mol. Biol. 224:771–782.

64. Brooks BR, Brooks CL, III, MacKerell AD, Jr., Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, et al. (2009) CHARMM: The Biomolecular Simulation Program. J. Comput. Chem. 30:1545.

65. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem. 4:187–217.

66. Perkins SJ, Wüthrich K (1979) Ring current effects in the conformation dependent NMR chemical shifts of aliphatic protons in the basic pancreatic trypsin inhibitor. Biochim. Biophys. Acta - Protein Struct. 576:409–423.

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

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

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

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

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

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

73. Ortega A, Amorós D, García de la Torre J (2011) Prediction of Hydrodynamic and Other Solution Properties of Rigid Proteins from Atomic- and Residue-Level Models. Biophys. J. 101:892.

74. Greene RF, Pace CN (1974) Urea and Guanidine Hydrochloride Denaturation of Ribonuclease, Lysozyme, α -Chymotrypsin, and β -Lactoglobulin. J. Biol. Chem. 249:5388–5393.

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

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

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

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

79. Galano-Frutos JJ, Sancho J (2019) 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. 59:4350–4360. Hospital A, Goñi JR, Orozco M, Gelpí JL (2015) Molecular dynamics simulations: advances and applications. Adv. Appl. Bioinform. Chem. 8:37.

81. Kumari I, Sandhu P, Ahmed M, Akhter Y (2017) Molecular Dynamics Simulations, Challenges and Opportunities: A Biologist's Prospective. Curr. Protein Pept. Sci. 18.

82. Robustelli P, Piana S, Shaw DE (2018) Developing a molecular dynamics force field for both folded and disordered protein states. Proc. Natl. Acad. Sci. 115:E4758–E4766.

83. Campos LA, Garcia-Mira MM, Godoy-Ruiz R, Sanchez-Ruiz JM, Sancho J (2004) Do Proteins Always Benefit from a Stability Increase? Relevant and Residual Stabilisation in a Three-state Protein by Charge Optimisation. J. Mol. Biol. 344:223–237.

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

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

86. Vallat B, Webb B, Fayazi M, Voinea S, Tangmunarunkit H, Ganesan SJ, Lawson CL, Westbrook JD, Kesselman C, Sali A, et al. (2021) New system for archiving integrative structures. Acta Crystallogr. Sect. D, Struct. Biol. 77:1486–1496.

87. Ayuso-Tejedor S, Angarica VE, Bueno M, Campos LA, Abián O, Bernadó P, Sancho J, Jiménez MA (2010) Design and Structure of an Equilibrium Protein Folding Intermediate: A Hint into Dynamical Regions of Proteins. J. Mol. Biol. 400:922–934.

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

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

90. Ayuso-Tejedor S, García-Fandiño R, Orozco M, Sancho J, Bernadó P (2011) Structural Analysis of an Equilibrium Folding Intermediate in the Apoflavodoxin Native Ensemble by Small-Angle X-ray Scattering. J. Mol. Biol. 406:604–619.

91. García-Fandiño R, Bernadó P, Ayuso-Tejedor S, Sancho J, Orozco M (2012) Defining the Nature of Thermal Intermediate in 3 State Folding Proteins: Apoflavodoxin, a Study Case. PLOS Comput. Biol. 8:e1002647.

92. Cremades N, Velazquez-Campoy A, Freire E, Sancho J (2008) The flavodoxin from Helicobacter pylori: Structural Determinants of Thermostability and FMN Cofactor Binding. Biochemistry 47:627.

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

94. Bollen YJM, Sánchez IE, van Mierlo CPM (2004) Formation of Onand Off-Pathway Intermediates in the Folding Kinetics of Azotobacter vinelandii Apoflavodoxin. Biochemistry 43:10475–10489.

95. Nabuurs SM, Mierlo CPM van (2010) Interrupted Hydrogen/Deuterium Exchange Reveals the Stable Core of the Remarkably Helical Molten Globule of α - β Parallel Protein Flavodoxin. J. Biol. Chem. 285:4165.

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

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

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

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

100. Mackerell AD, Feig M, Brooks CL (2004) 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. 25:1400–1415.

water. J. Chem. Phys. 79:926.