Constructing a folding model for protein S6 guided by native fluctuations deduced from NMR structures

Heiko Lammert,[1] Jeffrey K. Noel,[1] Ellinor Haglund,[1] Alexander Schug,[2] and José N. Onuchic[1,a)]

[1] *Center for Theoretical Biological Physics and Department of Physics, Rice University, Houston, Texas 77005, USA*
[2] *Steinbuch Centre for Computing, Karlsruhe Institute of Technology, Karlsruhe, Germany*

(Received 7 August 2015; accepted 18 November 2015; published online 9 December 2015)

The diversity in a set of protein nuclear magnetic resonance (NMR) structures provides an estimate of native state fluctuations that can be used to refine and enrich structure-based protein models (SBMs). Dynamics are an essential part of a protein's functional native state. The dynamics in the native state are controlled by the same funneled energy landscape that guides the entire folding process. SBMs apply the principle of minimal frustration, drawn from energy landscape theory, to construct a funneled folding landscape for a given protein using only information from the native structure. On an energy landscape smoothed by evolution towards minimal frustration, geometrical constraints, imposed by the native structure, control the folding mechanism and shape the native dynamics revealed by the model. Native-state fluctuations can alternatively be estimated directly from the diversity in the set of NMR structures for a protein. Based on this information, we identify a highly flexible loop in the ribosomal protein S6 and modify the contact map in a SBM to accommodate the inferred dynamics. By taking into account the probable native state dynamics, the experimental transition state is recovered in the model, and the correct order of folding events is restored. Our study highlights how the shared energy landscape connects folding and function by showing that a better description of the native basin improves the prediction of the folding mechanism. © *2015 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4936881]

I. INTRODUCTION

Living organisms depend on the functioning of proteins in order to survive and reproduce. The activity of a protein relies on the properties of its native state, the state that is thermodynamically stable *in vivo*. A suitable description of the native state is thus essential for the understanding of protein function. Under native conditions in solution, most globular proteins adopt compact folded structures. For each protein, the specific native fold is encoded in its sequence, at least at a certain resolution that corresponds to elements of secondary structure and their relative tertiary arrangement. The folded protein chain in the native state, however, retains significant mobility, and protein function depends on the characteristic dynamics that are enabled by the native structure. A valid representation of the native state must, therefore, contain both the native structure and the native dynamics. Much insight into the dynamics can be derived from the structure in simulations with structure-based models (SBMs), guided by energy landscape theory.[1–6] Here, we study additional information about the dynamics that can be deduced directly from the experimental structures, and we complement the structures used to build a SBM with the inferred native state fluctuations.

The extent of dynamics involved in different protein functions varies from fast local fluctuations to global structural transitions, which may be facilitated by local cracking, or partial unfolding. All functional dynamics occur in the native region of the same energy landscape that also controls and guides the folding of the entire protein.[7,8] While the underlying interactions between the chemically diverse amino acid residues in the protein chain are complex, the overall shape of the energy landscape is determined by the need for reliable and robust folding and can be understood through energy landscape theory. Protein sequences have evolved to stabilize the native state relative to alternative compact structures via the principle of minimal frustration,[4] giving rise to a smooth, funneled energy landscape that guides the folding process down into the native state.

Structure-based models construct an ideally funneled energy landscape for the entire range of dynamics of a specific protein from the native structure alone. Effective interactions are assigned to the model based on assumptions drawn from energy landscape theory: The native state is stabilized by unfrustrated contact interactions, which are given uniform strength. Any non-native contacts remain unstabilized to eliminate energetic traps. Native contacts are identified based on the proximity of residues in the native structure. The contact potentials are soft and short-range, allowing contacts to break and the chain to unfold. Formation of native contacts is both the driving force of folding and a natural reaction coordinate for its description.[9,10] All dynamics of the model, including the folding mechanism, are determined by geometric constraints imposed by the structure of the native fold. As local unfolding is penalized by the loss of stabilizing contacts, the balance

[a)] Electronic mail: jonuchic@rice.edu

between contacts and loop entropy is a key factor in the dynamics.

SBMs have been used extensively to study protein folding and function. General trends[11,12] as well as the folding mechanisms of numerous individual proteins[13–18] and a wide range of functional dynamics[19–25] are indeed well described by SBM dynamics, controlled by the native geometry. Real protein sequences are, however, also subject to functional demands beyond foldability alone, and there are several well studied cases of sequence dependent folding mechanisms.[14,26–28] Generally, protein folding mechanisms are robust against perturbations, but certain folds with internal symmetries[29,30] or high helical content[31] are particularly susceptible to energetic effects.

The simplicity of SBMs with homogeneous interactions has its own value for investigations at a fundamental level, and in folding studies, it is often already instructive to detect discrepancies between experimental observations and the mechanism implied by the geometry alone.[14] But if the dynamics are shaped by a strongly non-uniform distribution of contact stabilities, or if non-native interactions become relevant, SBM potentials can be extended with non-uniform contact interactions. The strengths of individual contacts must then be chosen based on additional data about the system, like experimental ϕ-values.[32] Native state dynamics in SBMs have also been targeted, by fitting fluctuation amplitudes to data from explicit solvent simulations.[33] Fluctuations in the native basin are an interesting quantity to optimize because of the direct focus on the functional state. Furthermore, native state fluctuations are rapidly sampled in simulations compared to folding transitions, and experimental estimates are also readily available in conjunction with the structures that are necessary to create a SBM.

Native-state structures can be determined at a resolution that is suitable for the construction of SBMs either by X-ray diffraction or by nuclear magnetic resonance (NMR) in solution. Typically both methods are in agreement,[34] but in the case of the ribosomal protein S6, we find a significant discrepancy between the X-ray[35] and NMR[36] structures as well as between the dynamics of SBMs built around them.

X-ray structures may suggest an overly static view of the native state, even if crystallographic B-factors also provide a direct estimate of fluctuations, albeit in the unnatural crystal environment. The multiple structures in a NMR dataset give a more natural indication of the structural diversity in the native state, but due to the complex and indirect process of structure determination[37] from NMR constraints, it is not clear how well the native-state dynamics are represented by the resulting set of NMR structures. In the case of S6, none of the configurations in the set of NMR structures for a certain loop is compatible with the X-ray structure. The extent of variation between the NMR structures suggests that the loop may even remain unstructured in the folded protein.

The approach to infer dynamics from ensembles of alternative experimental structures has been taken before,[38] and the specific issue of building a SBM from a set of NMR structures has also been raised.[39] While the ensemble of configurations in a set of NMR structures has been described as tight,[40] the variance among the structures has already been linked to the solution dynamics,[39] and the shape of the implied profile of fluctuations was found to be robust.[41]

Our goal is to determine whether the diversity among the structures in a NMR dataset can directly serve as a useful measure of native state fluctuations in solution and to explore the relevance of accurate native state dynamics for folding simulations. To this end, we study a SBM for the ribosomal protein S6 with a contact map that is modified to accommodate the extent of diversity in the published set of NMR structures.

II. MODELS

S6 is a widely studied protein with a complex folding mechanism.[32,42–49] Several circular permutants (CPs) have been characterized in addition to the wild-type (WT). SBMs are well suited to describe the effects of circular permutation, which are due to changes in loop entropy,[50] and S6 thus offers multiple test cases in one system. Overlapping alternative folding nuclei in S6 provide corresponding alternative folding mechanisms for variants of S6. The outcome of simulations is expected to be parameter-dependent, offering a chance to differentiate the performance of alternative models. The behavior of S6 is further complicated by an experimentally noted correlation[42] between contact strengths and loop lengths, which is not represented in typical SBMs. S6 and its permutants have been studied before by theory and simulations,[32,43–47,51] and SBMs are capable of reproducing the general trends of the folding mechanism. A SBM with a heterogeneous contact map that was optimized towards quantitative agreement with experimental ϕ-values is also available for comparison.[32]

We use a model that is completely structure-based, both for all the backbone terms and for the tertiary contacts. All heavy atoms are explicitly included.[16] The contact map is determined via the shadow algorithm,[52] and contacts are represented by adaptable Gaussian contact functions.[53] Simulation input files are created with an automated online tool,[54] and all details of the setup are published.

Fig. 1 presents the structure of S6, the overlapping alternative folding nuclei, determined from experimental folding data, which have been identified with foldons,[55] and the corresponding subdivision of the contact map defined in order to analyse the simulations in these terms. The cartoon representation in Fig. 1(a) shows the crystal structure (Protein Data Bank (PDB) ID 1RIS, from t. thermophilus) for the native state of S6. Elements of secondary structure are distinguished by color. In the sketch of Fig. 1(b), strands and helices are grouped into foldons as identified by experiment. According to ϕ-value analysis of S6 and its circular permutants, the structure contains two competing potential folding nuclei or foldons. Each foldon consists of one α helix and two β strands; strand $\beta 1$ is part of both foldons. Foldon 1, shaded blue in the figure, adds helix $\alpha 1$ and strand $\beta 3$. Foldon 2, shaded red, contains helix $\alpha 2$, strand $\beta 4$, and the shared strand $\beta 1$. Strand $\beta 2$ is not part of either foldon. Arrows in the sketch symbolize contact interactions among elements of secondary structure, either inside of one foldon or with the rest of the structure. All contact interactions in the simulation model for S6 are detailed in the residue contact map, shown

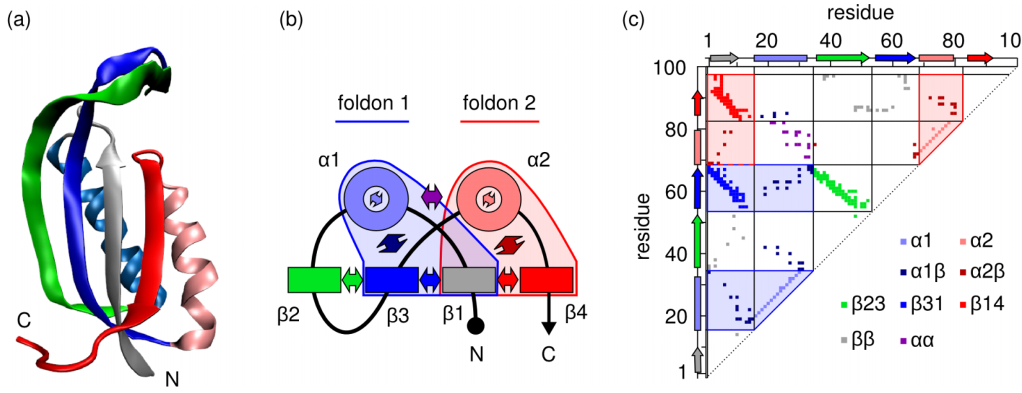

FIG. 1. Structure of S6 and its reported folding nuclei. (a) Cartoon of the native X-ray structure (PDB 1RIS). (b) Sketch of S6, seen from the top. Blue and red shadings indicate those elements of secondary structure, and the contacts among them, that according to experiment form competing potential folding nuclei, named foldon 1 and foldon 2. Strand $\beta 1$ is shared by both foldons. Strand $\beta 2$ is not part of either foldon. (c) Native residue contact map. Coloring indicates groups of contacts corresponding to elements of secondary structure or to groups of contacts between different elements. Blue and red shadings indicate those contacts that are unique to either foldon 1 or foldon 2, defined in (b).

as Fig. 1(c). With the sequence of S6 extended along both axes of the matrix, each dot marks a close contact between a pair of residues in the native structure, which is stabilized by a number of attractive contact interactions between atoms in the model. Colors and shading are the same as in the previous panels. For analysis of the folding mechanism in terms of contact formation, groups of contacts are introduced that are unique to either foldon. Specifically, foldon 1 is, therefore, described by the contacts in helix $\alpha 1$, between strands $\beta 1$ and $\beta 3$, and between these strands and the helix. Foldon 2 is characterized by the contacts in helix $\alpha 2$, between strands $\beta 1$ and $\beta 4$, and between these strands and helix $\alpha 2$. The contacts formed by strand $\beta 2$ with strand $\beta 3$ are not part of a foldon, and neither are the packing contacts between both helices. Also outside the foldons, there are contacts between either helix and strands that are not part of the same foldon, or finally contacts between non-adjacent strands in the β sheet. All these remaining contacts are collected into a third group.

Fig. 2(a) shows an overlay of the aligned backbone traces from the 20 structures in the NMR dataset (PDB 2KJV) for S6. Residues 1–97 are shown, which is the range also covered by the X-ray structure of S6. Substantial diversity between the NMR structures is apparent at the C-terminus and also in the configuration of loop-2/3, which is connecting strands $\beta 2$ and $\beta 3$ at one edge of the β sheet. These regions, highlighted by magenta color, also stand out in Fig. 2(b), where the average root mean square (RMS) deviation of the C_α positions in the 20 aligned NMR structures from their means is plotted as a function of sequence position. This quantity corresponds to the RMS fluctuation (RMSF) amplitude in a dynamical context. The intervals containing residues [44–56] and [93–97] are identified as regions with highly variable positions or variable

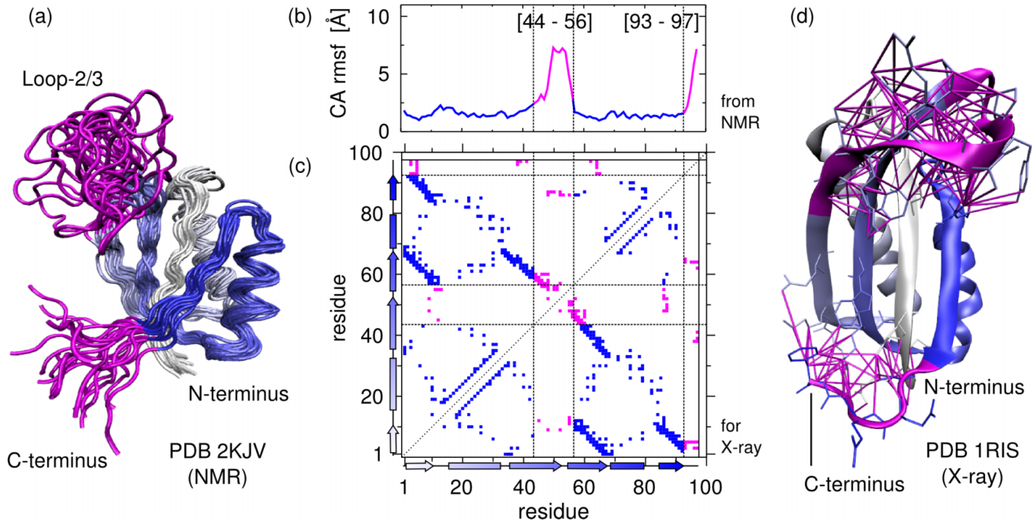

FIG. 2. Structural variation among the set of NMR structures for S6. (a) Overlay of aligned backbone traces for the 20 configurations contained in the NMR data set for S6, PDB 2KJV. Residues 1–97 are shown. The chain is colored white to blue from N to C-terminus. Regions with variable configurations are highlighted in magenta. (b) RMS deviations of the C_α positions from the mean of the aligned NMR structures, shown in panel (a), plotted as a function of residue number. Residues [44–56] and [93–97] are identified as regions with highly variable structure in the native state ensemble. (c) Residue contact map for the X-ray structure (PDB 1RIS), with contacts formed by residues in the variable regions, defined in (b), highlighted in magenta. (d) Cartoon of the X-ray structure, PDB 1RIS, containing residues 1–97, colored like in panel (a). Atomic contacts formed by residues in the highlighted variable regions are shown as magenta lines.

regions. Fig. 2(c) shows the residue contact map for S6 based on the X-ray structure. The variable regions, determined in Fig. 2(b) to contain residues with highly variable positions, are marked, and contacts that involve those residues are highlighted. All the atom-to-atom contacts that correspond to these residue contacts are drawn in Fig. 2(d). The X-ray structure (PDB 1RIS) of S6 is shown in cartoon representation, colored like in Fig. 2(a). The variable regions are again highlighted in purple. Residues in the variable regions and those residues that share contacts with them are also drawn explicitly in stick representations. Contacts between them are shown as purple lines. Out of the total of 282 atom-to-atom contacts identified for the X-ray structure, 57 are associated with the variable regions, compared to 225 among the other residues. The highlighted contacts form two localized clusters in the structure that correspond to the two selected regions. By one subset of 19 contacts, the C-terminus, which is free in the NMR structures, is attached across one end of the β sheet in the X-ray structure. The second subset of 38 contacts stabilizes the loop-2/3 in a configuration that is bent towards the β sheet and attaches it to the rest of the structure in this pose. Both the loop-2/3 and the C-terminus are thus being biased towards extreme positions compared to what the NMR dataset suggests as their respective ranges of motion. From this structural information, we create two SBMs. The first one, named *loop-fixed* or L_{fix}, is directly built from the X-ray structure alone. It stabilizes the loop-2/3 and the C-terminus in the crystal configuration by all the contacts derived from it. These 57 contacts, identified above, are omitted from the otherwise identical second model, which is named *loop-free* or L_{free}. It leaves the loop-2/3 and C-terminus free to move as suggested by the NMR structures.

III. RESULTS

The RMS fluctuations of the two models are compared in Fig. 3. In blue, the apparent C_α RMSF calculated from the variation among the NMR structures in set 2KJV is repeated from Fig. 2 for reference. The RMS fluctuations corresponding to the B-factors in the X-ray structure, 1RIS, are given in black. They share most features with the fluctuations derived from the NMR structures, with the exception of the prominent peak around loop-2/3 and the increase at the C-terminus, which are not present in the X-ray B-factors. The RMS fluctuations of the C_α atoms for the L_{fix} and L_{free} models are shown in purple and red, respectively. Outside the region of loop-2/3, both fluctuation profiles are similar, and both are in qualitative agreement with the profiles from NMR and X-ray data. But in the region of loop-2/3, only the L_{free} model shows a level of fluctuations that is compatible with the NMR data, while the smaller peak for the L_{fix} model follows the behavior of the B-factors from the crystal structure.

The folding behavior of the L_{fix} model is shown in Fig. 4. Folding occurs as a two-state process on the free energy landscape shown in Fig. 4(a). At the folding temperature, $T_f = 1.2$, the native state at high Q_{CA} is separated from the unfolded basin by a folding barrier of 6 $k_{\text{B}}T$.

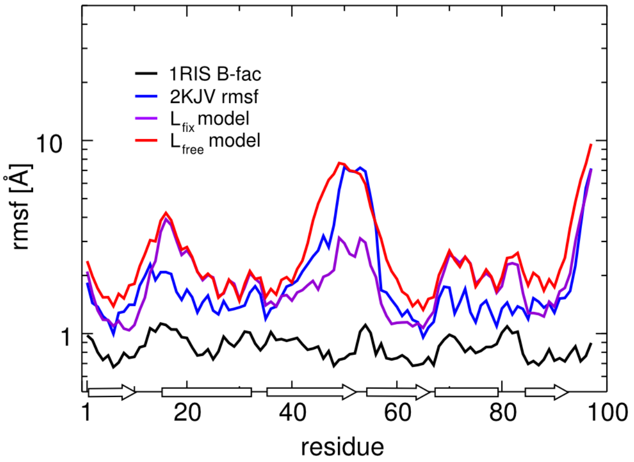

FIG. 3. RMS fluctuation profiles. The C_α RMSF derived from the variation between the NMR structures in set 2KJV is repeated from Fig. 2 in blue. The RMSF calculated from the X-ray B-factors in structure 1RIS, shown in black, shares most features except for the large peak around loop-2/3 and for the mobile C-terminus. RMSF profiles measured for the L_{fix} and L_{free} models are shown in purple and in red.

In order to characterize the folding mechanism of the L_{fix} model, the order of folding events is resolved in Fig. 4(b). Each curve follows the formation of either a single element of secondary structure or of the packing contacts among a group of such elements. Plotted is the average fraction of formed native contacts for each respective set of contacts as a function of the total fraction of formed native contacts Q_{CA}.

The three sets of sheet contacts between strands $\beta 2$ and $\beta 3$, between $\beta 3$ and $\beta 1$, and between $\beta 1$ and $\beta 4$, respectively, are being formed in sequence one after the other. First to form are the contacts between $\beta 2$ and $\beta 3$, which are already 60% formed ahead of the folding barrier. Contacts between $\beta 3$ and $\beta 1$ are being formed near the top of the barrier, accompanied by backtracking of contacts between $\beta 3$ and $\beta 2$. Contacts between $\beta 1$ and $\beta 4$ form last, after the barrier. Meanwhile, the two helices $\alpha 1$ and $\alpha 2$ and their packing contacts are being formed without a clear preference for either helix, and more gradually than the sheet contacts that shape the mechanism.

The folding progress in terms of the foldons, observed in experiment and defined in Fig. 1, is shown in Fig. 4(c). The blue and red curves show the degree of formation for the groups of those contacts that are characteristic for foldons 1 and 2, respectively. All remaining contacts are collected in a third group, shown in green. Formation of this group is initially preferred, driven by contacts between $\beta 2$ and $\beta 3$, until it backtracks in favor of foldon 1. From this point onwards, foldon 1 remains preferred over foldon 2 throughout the barrier region.

Fig. 5 shows the folding behavior of the L_{free} model. The free energy landscape for folding is plotted in Fig. 5(a) as a function of Q_{CA}, at the lower folding temperature of the L_{free} model of 1.14. Compared to the L_{fix} model, the barrier is broadened, and at 5.5 $k_{\text{B}}T$ also somewhat lowered. The folding mechanism is presented in Fig. 5(b) in terms of the folding progress of secondary structure elements and their packing contacts. The mechanism is again dominated by the behavior of the β sheet, while the helices and their packing contacts are again forming later and more gradually. Formation of the sheet now starts with the contacts between the central strands $\beta 3$

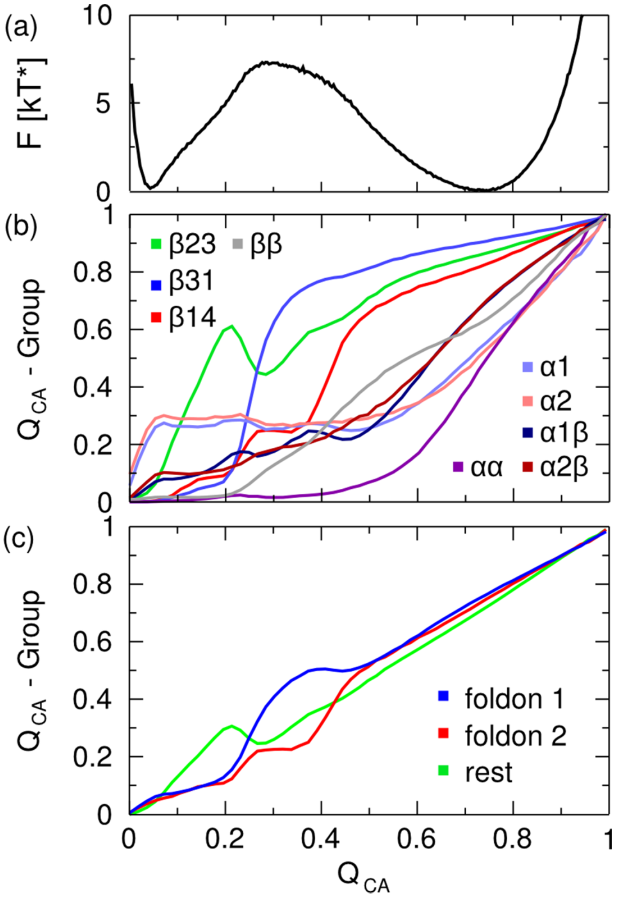

FIG. 4. Folding mechanism of S6 from simulations with the L_{fix} model. (a) Free energy landscape for the folding transition, plotted as a function of the fraction of formed native C_α contacts, Q_{CA}. (b) Average degree of contact formation, as a function of global folding progress, given by Q_{CA}, for the groups of contacts corresponding to elements of secondary structure, defined in Fig. 1(b). (c) Average degree of contact formation, as a function of Q_{CA}, shown for the exclusive components of the competing foldons 1 and 2, in blue and red, respectively. The rest, shown in green, consists of contacts that are either shared between both foldons or that are not part of either foldon.

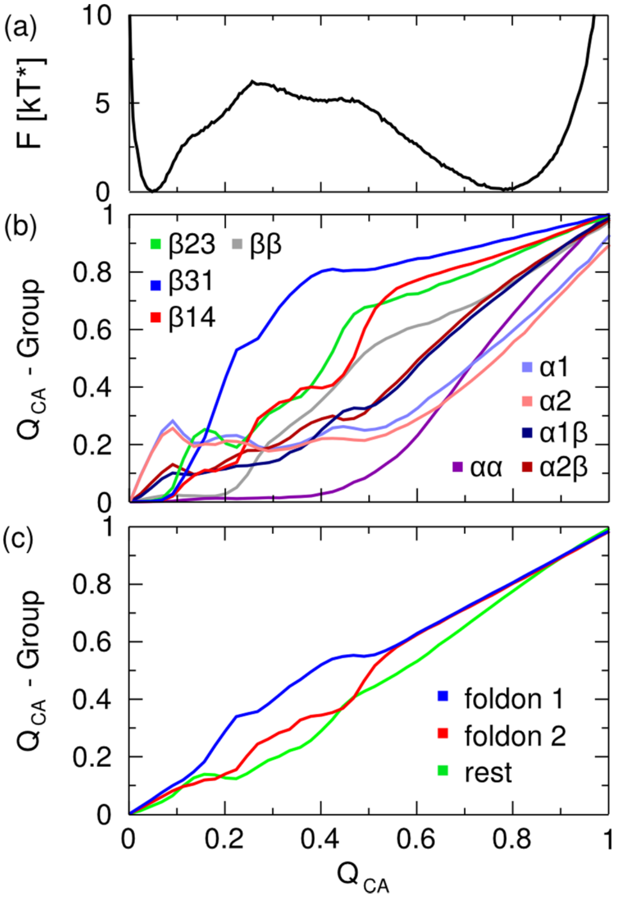

FIG. 5. Folding mechanism of the L_{free} model for S6. (a) Free energy landscape for the folding transition, plotted as a function of the fraction of formed native C_α contacts. (b) Average degree of contact formation, as a function of global folding progress, given by Q_{CA}, for the groups of contacts corresponding to elements of secondary structure, defined in Fig. 1(b). (c) Average degree of contact formation, as a function of Q_{CA}, shown for the exclusive components of the competing foldons 1 and 2, in blue and red, respectively. The rest, shown in green, consists of contacts that are either shared between both foldons or that are not part of either foldon.

and $\beta1$ well ahead of the barrier. Some contacts between $\beta2$ and $\beta3$ are also being formed early on, but, in contrast to the L_{fix} model, this region gains most of its contacts only past the top of the folding barrier, together with those joining $\beta1$ and $\beta4$. At this point, the helices also start to pack against the sheet and finally become more structured themselves. Compared to the distinct differences in folding progress between different strands in the β sheet, the two helices and their respective packing contacts are being formed nearly simultaneously. But compared to the L_{fix} model, the folding of the L_{free} model shows a small preference for contacts in helix $\alpha1$. The relative folding progress of the two entire foldons is compared in Fig. 5(c). The blue and red curves plot the fractions of formed native C_α contacts that are characteristic for foldons 1 and 2, respectively. The behavior of the remaining contacts in the system is given by the green curve. Contacts from foldon 1 are being formed preferentially throughout the entire barrier region. Only beyond the barrier, both foldons reach equal levels of contact formation and proceed to completion together. The remainder of the structure is less structured than either foldon at all points beyond $Q_{CA} = 0.2$. Crucially, the early formation of loop-2/3, which was characteristic for the folding of the L_{fix} model, is absent in the L_{free} system.

As a further test of the L_{free} model, the folding of its circular permutants is characterised in Fig. 6. The locations of the new termini for the set of five circular permutants that have been created and characterized experimentally for S6 are indicated in the sketch in Fig. 6(a). Permutants are named by the number of the residue before the incision, which becomes the new C-terminus of the permutant.

The observed folding mechanisms in simulations with the L_{free} model suggest a division of the five permutants into two groups, each defined by a shared characteristic folding mechanism. The respective mechanisms for these two groups of permutants are summarized in Figs. 6(b) and 6(c). Each plot shows the folding progress of the competing foldons 1 and 2 as a function of global folding progress measured by Q_{CA}. In each plot, the curves for the respective other group of permutants are shown in gray in the background for reference.

The folding mechanism for the first group of permutant systems, consisting of CP13, CP33, and CP54, shows a clear

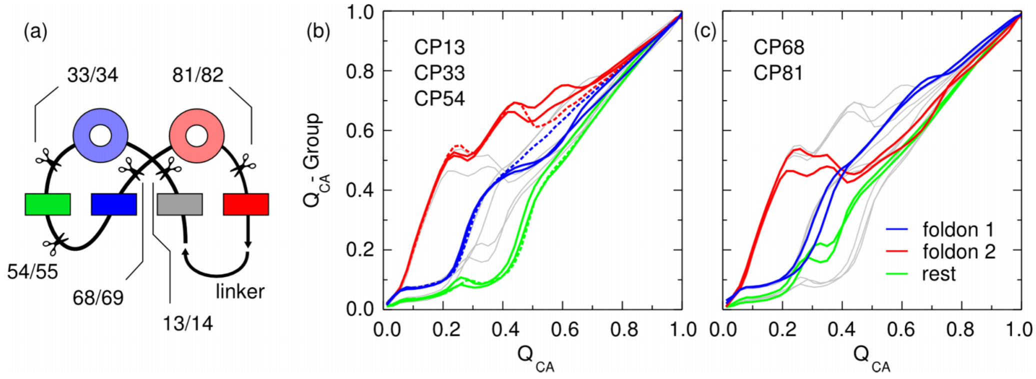

FIG. 6. Folding mechanisms of circular permutants. (a) Location of the termini for the 5 studied circular permutants, CP13, CP33, CP54, CP68, and CP81, marked on a sketch of the circularized protein S6. (b) Relative folding progress of foldons 1 and 2, and of the rest of the structure, for permutants CP13, CP33, and CP54 (dashed), which all follow a similar folding mechanism. The average fraction of formed native contacts from each group is plotted as a function of global folding progress, given by the total fraction of formed contacts, Q_{CA}. Data for foldons 1 and 2 are shown in blue and red, respectively, and for the rest in green. Traces for the remaining permutants, CP68 and CP81, are shown in gray in the background. (c) Relative folding progress of foldons 1 and 2 (blue and red), and of the rest of the structure (green), for permutants CP68 and CP81, which fold with a similar mechanism.

preference for foldon 2, which starts to form very early on and remains more highly structured than foldon 1 at all stages of folding. Contact formation in the rest of the structure always trails behind both foldons.

In the second group of permutant systems, made up of CP68 and CP81, the preference for foldon 2 is weakened, albeit still present. The early folding events up to $Q_{CA} = 0.2$ in this group are practically the same as for the other permutants, including the rapid initial structuring of foldon 2. But here the subsequent gain of contacts in foldon 2 is delayed until foldon 1 becomes the most highly formed part of the structure for the later stages of the folding process. The rest of the structure again remains less structured than either foldon.

An analogous partitioning of the permutants into the same two groups, defined by shared folding mechanisms, was also found in experiment. Notably for the first group, which shows a clear preference to form foldon 2, the simulations agree well with the experimental mechanism. The shift in the mechanism in favor of foldon 1 for the second group of permutants is more pronounced in experiment than in the simulations, but the trend is still correctly captured by the L_{free} model.

Experimental characterization of protein folding transition states relies primarily on ϕ-values, which are determined from mutation studies to quantify the extent of native-like interactions that a given residue forms in the transition state ensemble. An analysis of the observed folding mechanisms in terms of ϕ-values is available as supplementary material,[56] both for the permutant models and for the L_{fix} and L_{free} models of wild-type S6. ϕ-values can be approximated from simulation data by comparing the probability of contact formation for a given residue between the transition state and the folded state. In the supplementary material,[56] ϕ-values calculated in this way are compared to the experimental values. Fig. S1 shows ϕ-values for the L_{fix} and L_{free} models of wild-type S6 as well as their averages over elements of secondary structure. Fig. S2[56] shows ϕ-values for the circular permutants. The corresponding element averages are shown in Fig. S3.[56] Differences in ϕ-values upon circular permutation are shown for all permutants in Fig. S4, and their element averages are shown in Fig. S5.[56] It has been pointed out that the change in the experimental ϕ-values upon circular permutation shows a stronger correlation with loop length than the ϕ-values themselves.[48] This is expected, because wild-type and permutant share the same three dimensional native structure. When taking the difference between their ϕ-values, any interfering effects of shared structural features should cancel out, and only the effect of the changed loop lengths should remain.

For the same reason, the agreement with experimental ϕ-values improves for our simulations, when the difference upon circular permutation is considered instead of the raw ϕ-values. Furthermore, taking the average over elements of secondary structure helps to distinguish trends. The grouping of the circular permutants by mechanism, which was observed in Fig. 6 at the level of foldons, also shows up clearly in the averaged ϕ-values, which have similar structural resolution. But the experimental mechanism described above is also reflected qualitatively in the individual simulated ϕ-values. Notably the improved behavior of strand $\beta 2$ in the L_{free} model can be observed: the value for VAL37 is significantly reduced and thus brought into closer agreement with experiment for the L_{free} model. Overall, the ϕ-values from simulation and experiment show similar trends.

IV. DISCUSSION

Our results demonstrate the close connection between the dynamics in the native ensemble and the entire folding landscape of a protein. Because protein sequences have evolved to favor the native structure via the property of minimal frustration, structure-based folding models with idealized funnel landscapes can be constructed by assigning effective stabilizing interactions to features of the native structure. A model of S6 constructed with all the interactions derived from a static snapshot of the native structure could, however, not match the dynamics in the native ensemble, suggested by NMR, ϕ-value analysis, and protein engineering. Only a generalized SBM of S6, which only stabilized those interactions that are compatible with the experimental native state dynamics, reproduced their pattern correctly. Interestingly, this generalized model also showed significantly improved agreement with the overall experimental folding mechanism.

Experiments indicate that the strand $\beta2$ at the edge of the β sheet of S6 does not participate in the transition state. The loop, loop-2/3, which links it to the next strand, adopts no persistent structure even in the native state. Only one ϕ-value could be determined experimentally inside $\beta2$ to report on its role in the transition state. But the small changes in stability that hinder the analysis for other mutations in the element indirectly suggest a lack of consistent structure as well. The strongest experimental evidence that strand $\beta2$ does not participate in the transition state is provided by a construct that still folds like the native protein although the entire strand $\beta2$ has been removed.[49]

The behavior of strand $\beta2$ in the simulations is brought into agreement with experiment as soon as the dynamical behavior of loop-2/3 is allowed for in the model. The improvement can be clearly observed in the simulation data when the average behavior of entire elements of secondary structure is analyzed. Correctly delayed formation of contacts between $\beta2$ and $\beta3$ for the L_{free} model, built with awareness of native state dynamics, is contrasted with spurious early attachment of $\beta2$ in the L_{fix} model, based on the static native structure.

The diverse array of experimental data available for S6 has made it a prime target for previous theoretical protein folding studies. Earlier studies of S6[43,44] and its circular permutants[45,46] with SBMs could capture the general effects of circular permutation. Similar results were obtained with a computational model directly based on loop entropies.[51] Available ϕ-values have been used as constraints in some cases,[43,46] and one model has strengthened all long-range contacts to introduce the experimentally observed correlation between loop length and contact stability.[42] But all of these models have still predicted an unrealistically high degree of contact formation in the transition state for loop-2/3, driven by the small loss of loop entropy associated with forming these local contacts.[49]

A model that was optimized directly to reproduce the experimental rates and stabilities measured for various mutants has captured the late folding of strand $\beta2$.[32] The resulting optimized pattern of contact strengths contains the experimental correlation with loop lengths.[42] The same pattern has also been recreated from chemical considerations.[47]

The optimal model shares with our L_{free} model a destabilization of loop-2/3 as the most prominent feature of the contact map. The optimized model has additional features that are not present in the L_{free} contact map, where all retained contacts are treated as equal. Notably a difference in contact strength between the helices in the optimized contact map could lead to a more pronounced difference in folding between the two alternative nuclei.

The L_{free} model was designed to capture the most prominent features of the inferred native state fluctuations, guided by the notion that the diversity among a set of NMR structures reflects the diversity among the ensemble of structures found in solution.

As the amplitude of fluctuations varies strongly between regions of S6, a simple binary classification of residues and their corresponding contacts into groups with high and low mobility proved sufficient for the task. Reproducing finer details of the dynamics of S6 likely requires a continuous distribution of contact energies, similar to the one obtained for the optimized model.[32] Similarly, a binary classification of contacts may be insufficient to model proteins with less prominent native state fluctuations. Before detailed modifications of native contact strengths are attempted based on the properties of a set of NMR structures, the quantitative relation between such NMR structures and the native ensemble in solution should, however, be considered more carefully.

NMR also provides a wealth of methods for the direct investigation of dynamics,[57] but here, we are concerned with the common methodology for structure determination using solution NMR,[37] which is by now sufficiently standardized to invite automation.[58] Structure determination in this case is based on distance constraints derived from short-range interactions assigned to individual atom pairs. In combination with readily available knowledge about the local geometry, like bond lengths or dihedral preferences, these constraints are sufficient to generate candidate structures that are compatible with the NMR data. After scoring, a set of the best structures is selected to represent the range of solution structures.

Local variation among this set of structures at first indicates only a relative lack of constraints. Assuming that all accessible NMR constraints have been determined, this can however be taken as an indicator of dynamics, as sufficient local motion will make the interaction between nearby atoms unobservable. It has been pointed out that the RMS deviation over a set of NMR structures cannot be an exact estimate of the fluctuations in the native state because it neglects the proper Boltzmann weights of those structures, which must generally have different energies.[59] Estimated energies are determined as part of the data analysis, since empirical protein potentials are used both to generate candidate structures and in their scoring.

But without expensive computational choices for the treatment of long-range electrostatics and solvent effects, these energies can be only approximate. They are sufficient to signal steric clashes and other constraint violations but are unlikely to provide correct weights for different allowed structures. In any case, significant uncertainty about the implied dynamics is introduced by the small number of structures provided to characterize the NMR ensemble. Within the broad limits set by the low number of structures, it seems reasonable to forgo reweighting of the structures and directly interpret variation in the ensemble of NMR structures as fluctuations of the native state in solution. On the most qualitative level, the improvements achieved by the L_{free} model for S6 support this approximation. While ongoing efforts to incorporate NMR constraints directly into simulations[60,61] promise to be more rigorous, an approximate approach using just a set of published structures would be applicable to many existing NMR structures in the PDB, offering readily accessible improvements of structure-based folding models.

V. CONCLUSIONS

Native state fluctuations contain information about protein folding mechanisms that, in the case of S6, proved to be the key to a correct description of the folding transition state. The relevance of native state dynamics for protein folding behavior is confirmed by our results, and the connection is also supported by the common energy landscape shared by native state dynamics and late folding events.

ACKNOWLEDGMENTS

This work was supported by the Center for Theoretical Biological Physics sponsored by the NSF (under Grant Nos. PHY-1427654 and NSF-MCB-1214457), by the Cancer Prevention and Research Institute of Texas (CPRIT), and by the Welch Foundation (Grant No. C-1792). This work was supported in part by the Data Analysis and Visualization Cyberinfrastructure funded by the NSF under Grant No. OCI-0959097. A.S. acknowledges support from the Impuls- und Vernetzungsfond of the Helmholtz Association and a YIN international exchange grant.

[1] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, Science **254**, 1598 (1991).
[2] P. Leopold, M. Montal, and J. Onuchic, Proc. Natl. Acad. Sci. U. S. A. **89**, 8721 (1992).
[3] H. S. Chan and K. A. Dill, J. Phys. Chem. **100**, 9238 (1994).
[4] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins: Struct., Funct., Bioinf. **21**, 167 (1995).
[5] J. Kim, T. Keyes, and J. E. Straub, Phys. Rev. E **79**, 030902 (2009).
[6] J. N. Onuchic and P. G. Wolynes, Curr. Opin. Struct. Biol. **14**, 70 (2004).
[7] C.-J. Tsai, S. Kumar, B. Ma, and R. Nussinov, Protein Sci. **8**, 1181 (1999).
[8] M. S. Cheung, L. L. Chavez, and J. N. Onuchic, Polymer **45**, 547 (2004).
[9] S. S. Cho, Y. Levy, and P. G. Wolynes, Proc. Natl. Acad. Sci. U. S. A. **103**, 586 (2005).
[10] R. B. Best, G. Hummer, and W. A. Eaton, Proc. Natl. Acad. Sci. U. S. A. **110**, 17874 (2013).
[11] L. L. Chavez, J. N. Onuchic, and C. Clementi, J. Am. Chem. Soc. **126**, 8426 (2004).
[12] M. Kouza, M. S. Li, E. P. O'Brien, C.-K. Hu, and D. Thirumalai, J. Phys. Chem. A **110**, 671 (2006).
[13] C. Clementi, A. E. Garcia, and J. N. Onuchic, J. Mol. Biol. **326**, 933 (2003).
[14] S. Gosavi, P. C. Whitford, P. A. Jennings, and J. N. Onuchic, Proc. Natl. Acad. Sci. U. S. A. **105**, 10384 (2008).
[15] R. D. Hills and C. L. Brooks III, J. Mol. Biol. **382**, 485 (2008).
[16] P. C. Whitford, J. K. Noel, S. Gosavi, A. Schug, K. Sanbonmatsu, and J. N. Onuchic, Proteins: Struct., Funct., Bioinf. **75**, 430 (2009).
[17] J. I. Sulkowska, P. Sulkowski, and J. N. Onuchic, Proc. Natl. Acad. Sci. U. S. A. **106**, 3119 (2009).
[18] H. Lammert, J. K. Noel, and J. N. Onuchic, PLoS Comput. Biol. **8**, e1002776 (2012).
[19] Y. Levy, S. S. Cho, T. Shen, J. N. Onuchic, and P. G. Wolynes, Proc. Natl. Acad. Sci. U. S. A. **102**, 2373 (2005).
[20] A. Schug, P. C. Whitford, Y. Levy, and J. N. Onuchic, Proc. Natl. Acad. Sci. U. S. A. **104**, 17674 (2007).
[21] C. Hyeon and J. N. Onuchic, Proc. Natl. Acad. Sci. U. S. A. **104**, 17382 (2007).
[22] P. C. Whitford, P. Geggier, R. B. Altman, S. C. Blanchard, J. N. Onuchic, and K. Y. Sanbonmatsu, RNA **16**, 1196 (2010).
[23] R. Nechushtai, H. Lammert, D. Michaeli, Y. Eisenberg-Domovich, J. A. Zuris, M. A. Luca, D. T. Capraro, A. Fish, O. Shimshon, M. Roy, A. Schug, P. C. Whitford, O. Livnah, J. N. Onuchic, and P. A. Jennings, Proc. Natl. Acad. Sci. U. S. A. **108**, 2240 (2011).
[24] E. L. Baxter, P. A. Jennings, and J. N. Onuchic, Proc. Natl. Acad. Sci. U. S. A. **108**, 5266 (2011).
[25] X. Lin, N. Eddy, J. K. Noel, P. C. Whitford, Q. Wang, J. Ma, and J. N. Onuchic, Proc. Natl. Acad. Sci. U. S. A. **111**, 12049 (2014).
[26] N. Ferguson, A. P. Capaldi, R. James, C. Kleanthous, and S. E. Radford, J. Mol. Biol. **286**, 1597 (1999).
[27] L. Sutto, J. Lätzer, J. A. Hegler, D. Ferreiro, and P. G. Wolynes, Proc. Natl. Acad. Sci. U. S. A. **104**, 19825 (2007).
[28] B. G. Wensley, S. Batey, F. A. C. Bone, Z. M. Chan, N. R. Tumelty, A. Steward, L. G. Kwa, A. Borgia, and J. Clarke, Nature **463**, 685 (2010).
[29] D. K. Klimov and D. Thirumalai, J. Mol. Biol. **353**, 1171 (2005).
[30] J. K. Noel, A. Schug, W. Verma, A. Wenzel, A. E. Garcia, and J. N. Onuchic, J. Phys. Chem. B **116**, 6880 (2012).
[31] S. S. Cho, Y. Levy, and P. G. Wolynes, Proc. Natl. Acad. Sci. U. S. A. **106**, 434 (2009).
[32] S. Matysiak and C. Clementi, J. Mol. Biol. **343**, 235 (2004).
[33] W. Li, P. G. Wolynes, and S. Takada, Proc. Natl. Acad. Sci. U. S. A. **108**, 3504 (2011).
[34] G. Wagner, S. G. Hyberts, and T. F. Havel, Annu. Rev. Biophys. Biomol. Struct. **21**, 167 (1992).
[35] M. Lindahl, L. A. Svensson, A. Liljas, S. E. Sedelnikova, I. A. Eliseikina, N. P. Fomenkova, N. Nevskaya, S. V. Nikonov, M. B. Garber, T. A. Muranova, A. I. Rykonova, and R. Amons, EMBO J. **13**, 1249 (1994).
[36] A. Öhman, T. Öman, and M. Oliveberg, Protein Sci. **19**, 183 (2010).
[37] K. Wüthrich, *NMR of Proteins and Nucleic Acids* (John Wiley & Sons, New York, 1986).
[38] R. B. Best, K. Linddorf-Larsen, M. A. DePristo, and M. Vendruscolo, Proc. Natl. Acad. Sci. U. S. A. **103**, 10901 (2006).
[39] P. Jiang and U. H. E. Hansmann, J. Chem. Theory Comput. **8**, 2127 (2012).
[40] A. Perez, A. Roy, K. Kasavajhala, A. Wagaman, K. A. Dill, and J. L. MacCallum, Proteins: Struct., Funct., Bioinf. **82**, 2671 (2014).
[41] L.-W. Yang, E. Eyal, C. Chennubhotla, J. Jee, A. M. Groneborn, and I. Bahar, Structure **15**, 741 (2007).
[42] M. Lindberg, J. Tangrot, and M. Oliveberg, Nat. Struct. Biol. **9**, 818 (2002).
[43] I. A. Hubner, M. Oliveberg, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U. S. A. **101**, 8354 (2004).
[44] A. D. Stoycheva, C. L. Brooks III, and J. N. Onuchic, J. Mol. Biol. **340**, 571 (2004).
[45] J. Chen, J. Wang, and W. Wang, Proteins **57**, 153 (2004).
[46] I. A. Hubner, M. Lindberg, E. Haglund, M. Oliveberg, and E. I. Shakhnovich, J. Mol. Biol. **359**, 1075 (2006).
[47] L. Wu, J. Zhang, J. Wang, W. F. Li, and W. Wang, Phys. Rev. E **75**, 031914 (2007).
[48] E. Haglund, M. O. Lindberg, and M. Oliveberg, J. Biol. Chem **283**, 27904 (2008).
[49] E. Haglund, J. Danielsson, S. Kadhirvel, M. O. Lindberg, D. T. Logan, and M. Oliveberg, J. Biol. Chem. **287**, 2731 (2012).
[50] C. Clementi, P. A. Jennings, and J. N. Onuchic, J. Mol. Biol. **311**, 879 (2001).
[51] T. R. Weikl and K. A. Dill, J. Mol. Biol. **332**, 953 (2003).
[52] J. K. Noel, P. C. Whitford, and J. N. Onuchic, J. Phys. Chem. B **116**, 8692 (2012).
[53] H. Lammert, A. Schug, and J. N. Onuchic, Proteins **77**, 881 (2009).
[54] J. K. Noel, P. C. Whitford, K. Y. Sanbonmatsu, and J. N. Onuchic, Nucleic Acids Res. **38**, W657 (2010).
[55] M. Oliveberg and P. G. Wolynes, Q. Rev. Biophys. **38**, 245 (2005).
[56] See supplementary material at http://dx.doi.org/10.1063/1.4936881 for a comparison of simulated and experimental ϕ-values.
[57] I. R. Kleckner and M. P. Foster, Biochim. Biophys. Acta **1814**, 942 (2011).
[58] P. Guerry and T. Herrmann, Q. Rev. Biophys. **44**, 257 (2011).
[59] W. F. van Gunsteren, R. M. Brunne, P. Gros, R. C. van Schaik, C. A. Schiffer, and A. E. Torda, Methods Enzymol. **239**, 619 (1994).
[60] O. F. Lange, N. A. Lakomek, C. Fares, G. Schröder, K. F. A. Walter, S. Becker, J. Meiler, H. Grubmüller, C. Griesinger, and B. L. de Groot, Science **320**, 1471 (2008).
[61] R. W. Montalvao, A. De Simone, and M. Vendruscolo, J. Biomol. NMR **53**, 281 (2012).