Biophysical Journal, Volume *111*

Supplemental Information

Structure-Based Prediction of Protein-Folding Transition Paths

William M. Jacobs and Eugene I. Shakhnovich

SUPPLEMENTARY INFORMATION FOR "STRUCTURE-BASED PREDICTION OF PROTEIN-FOLDING TRANSITION PATHS"

S1. CONTACT-GRAPH MODEL

A. Allowed microstates

In this section, we describe the allowed microstates of the contact-graph model using the language of graph theory. The microstate that corresponds to the completely folded polymer, i.e., the configuration in which all possible contacts are formed, is denoted by the graph G. The vertices of this graph correspond to residues, while the edges indicate native contacts. The vertices $\{u\}$ are labeled by their positions on the polymer backbone, i.e., $u \in \{1, \ldots, L\}$, where L is the total number of residues in the chain. We denote the set of all allowed subgraphs by $\{g\}$ and the set of edges in a microstate g by $\mathcal{E}(g)$. Because g is a subgraph of G, every edge in $\mathcal{E}(g)$ is also an edge in $\mathcal{E}(G)$. Only residues that form one or more contacts are represented by vertices in g; this set of vertices is denoted by the set $\mathcal{V}(g)$. The set of connected components of g is $\mathcal{C}(g)$, and the edge and vertex sets of a connected component $c \in \mathcal{C}(g)$ are $\mathcal{E}(g, c)$ and $\mathcal{V}(g, c)$, respectively.

For each microstate, the associated graph of native contacts can be decomposed into a disjoint set of connected components (maximal subgraphs in which all pairs of vertices are connected by paths through the subgraph). As described in the main text, the fact that the residues occupy non-overlapping finite volumes implies that many contacts must be correlated. These correlations place restrictions on the combinations of contacts that can be simultaneously formed. In the generation of a contact-graph model from a crystal structure, we have ignored contacts between residues that are separated by less than one Kuhn segment, b, in the polymer sequence; for consistency, we must therefore consider contacts involving sequences of residues that are shorter than one Kuhn segment to be correlated as well. Consequently, we restrict the set of allowed microstates to those subgraphs that satisfy the following two rules:

1. Every connected component, $c \in \mathcal{C}(g)$, must be an induced subgraph of G. This means that every edge (u, v) in the connected component c must appear in the subgraph g if the vertices u and v are adjacent in the supergraph G.

2. Assume that two vertices $v' > u'$ belong to the same connected component c and are separated by at most b residues in the sequence, i.e., $v' - u' \leq b$. Then every intervening vertex u, i.e., $u' < u < v'$, must also be included in the connected component c if an edge exists between u and any vertex v in c.

B. Loop entropy

In Eq. (2), we define a loop to be any contiguous sequence of non-interacting residues, with the exception of 'bridge' segments (residues that, if removed, would break a polymer configuration given by a specific microstate into two non-interacting pieces). For example, the microstate shown on the right in Figure 1c contains two loops, 4–5–6–7 and 18, and one bridge segment, 11–12–13, where the residues are labeled starting from 1 at the top right of the figure. In Eq. (2), $r(l)$ is the end-to-end distance of loop l, expressed as a dimensionless multiple of the covalent backbone bond length; $r = 0$ if the residues at the loop ends form a native contact.

C. Native-contact energies

It is important to note that the native-contact energies are themselves free energies, since they depend on the average potential energy between two amino acids as well as solvent effects. Here we assume that these attractive interactions are short-ranged and discrete, i.e., a contact is either completely formed or not present. In a real polymer, there are likely to be other random interactions between residues. Such nonspecific interactions contribute to the average energy of the ensemble of random coil configurations, which is taken to be the reference state for all free-energy calculations. Consequently, the attractive interactions that are associated with specific contacts are, more precisely, associated with the *differences* between the specific contact free energies and the average interaction energy between any pair of residues in the chain. We assume that only these free-energy differences determine the folding pathways of the polymer.

The two-parameter empirical potential introduced in the Materials and Methods was manually tuned to achieve good agreement with the experimental ϕ-values for protein G (1igd). We verified that our values for the two adjustable parameters, α_{helix} and α_{hb}, also result in close to optimal agreement with the experimental ϕ-values for the α/β proteins 1k53, 1ubq and 2ci2.

S2. MONTE CARLO FREE-ENERGY CALCULATIONS

We compute free energies in this model using Monte Carlo integration. This application of the Monte Carlo method is not a conventional simulation, as the sequence of microstates generated by our algorithm does not correspond to a physical folding trajectory. Instead, the approach used here is simply an efficient means to integrate over the set of microstates with the same topological configuration. (For a related application of this Monte Carlo technique, see Ref. 3.) To do so, we first construct a Markov Chain to sample from the space of allowed subgraphs $\{g\}$. We then use the Wang–Landau method (8) to calculate F_i, the free energy of all microstates in topological configuration i. Finally, we compute the contact and vertex probabilities $\langle \mathbf{1}_{uv} \rangle_i$ and $\langle \mathbf{1}_u \rangle_i$. Below, we first describe the construction of the Markov Chain and then provide details of these algorithms.

A. Monte Carlo acceptance probabilities

In order to calculate equilibrium properties of the contact-graph model, the underlying Markov Chain must obey detailed balance. That is, the probability of making forward and reverse moves between two subgraphs g and g' must be equal. To do so, we propose transitions between microstates (which obey the two rules given in Sec. S1 A) with uniform probability and then correct for this bias by calculating the ratio of the generation probabilities between forward and backward moves, $\alpha(g \to g')/\alpha(g' \to g)$.

Assuming a single connected component (i.e., a single structured region) c, we implement moves that add or remove individual vertices. The set of vertices that are adjacent to c in the supergraph G but are not in $\mathcal{V}(g)$ is denoted by $\mathcal{A}(g, c)$. We choose one vertex u from $\mathcal{A}(g, c)$ with uniform probability and form all edges $(u, v) \in \mathcal{E}(G)$ between u and the existing vertices $v \in \mathcal{V}(g, c)$. With the addition of these edges, we denote the new graph as g' and the updated connected component as c'.

For the reverse move, we must avoid breaking c' into two or more disconnected subgraphs. Consequently, we must be careful not to remove any vertex that is an articulation point of c'. The set of such points is denoted by $\mathcal{B}(g', c')$. We therefore select one vertex with uniform probability from the set $\mathcal{V}(g', c') \setminus \mathcal{B}(g', c')$. For this move, we only consider connected components that are larger than a dyad. The ratio of the forward to reverse generation probabilities is

$$\frac{\alpha_{\mathrm{N}+}(g, c \to g', c')}{\alpha_{\mathrm{N}-}(g', c' \to g, c)} = \frac{|\mathcal{A}(g, c)|}{\mathbf{1}\left[|\mathcal{V}(g', c')| > 2\right] |\mathcal{V}(g', c') \setminus \mathcal{B}(g', c')|}. \tag{S1}$$

To ensure ergodicity and to improve sampling efficiency, we implement a super-detailed balance sampling scheme (9) for vertex additions and removals. If a move $g, c \to g', c'$ results in a subgraph that violates rule 2 in Sec. S1 A, we immediately attempt another move of the same type, starting from the new subgraph g' using the updated connected component c'. This process is repeated until the resulting subgraph, $g^{(n)}$, satisfies rule 2. The total probability of following this path from g to $g^{(n)}$ is the product of the generation probabilities at each step, $\alpha(g \to g^{(1)}) \times \alpha(g^{(1)} \to g^{(2)}) \times \cdots \times \alpha(g^{(n-1)} \to g^{(n)})$. The ratio of generation probabilities depends on the total probability of following this forward path and the total probability of traversing the path in reverse, following precisely the same sequence of steps:

$$\frac{\alpha_{\mathrm{N}+}^{(n)}}{\alpha_{\mathrm{N}-}^{(n)}} = \frac{\prod_{i=1}^{n} \alpha_{\mathrm{N}+}(g^{(i-1)}, c^{(i-1)} \to g^{(i)}, c^{(i)})}{\prod_{i=0}^{n-1} \alpha_{\mathrm{N}-}(g^{(n-i)}, c^{(n-i)} \to g^{(n-i-1)}, c^{(n-i-1)})}, \tag{S2}$$

where each step is indexed by i and $g^{(0)} \equiv g$. If at any step on the forward move we find that $|\mathcal{A}(g^{(i)}, c^{(i)})| = 0$, then the entire move is rejected. In order to obey detailed balance, vertex additions and removals are attempted with equal probability at every Monte Carlo step.

B. Wang–Landau sampling

Wang–Landau sampling (8) provides an efficient algorithm for calculating the free-energy difference between two disjoint sets of microstates. Here we implement the variant of this algorithm described in Ref. 10. In essence, the Wang–Landau algorithm calculates an equilibrium free-energy landscape stochastically by continually updating an estimate of the free energy, F_t, as the Monte Carlo calculation samples from the space of allowed subgraphs. At every step, the underlying Monte Carlo algorithm uses F_t to bias the acceptance probabilities of individual moves.

For these calculations, we use an order parameter to measure progress toward the completely folded microstate. Excluding the effects of the backbone connectivity, which are entirely contained in $\Delta S_l(g)$, the entropic contribution to the free energy in Eq. (1) is proportional to

$$X(g) \equiv \sum_{c \in \mathcal{C}(g)} \left[|\mathcal{V}(g, c)| - 1\right] = N(g) - C(g), \tag{S3}$$

where $N(g)$ is the total number of interacting residues and $C(g) \equiv |\mathcal{C}(g)|$ is the number of connected components of the microstate g. Like the commonly used fraction of native contacts, Q (1), the order parameter X characterizes the similarity between any given microstate and the native configuration. However, X is preferable for analyzing a discrete model, since it measures the degree of assembly of the independent monomers as opposed to the (likely correlated) interactions among them. Since our calculations only consider the largest structured region, $C(g) = 1$ for all topological configurations except $\varnothing$, in which case $C(g) = 0$.

To perform free-energy calculations for a specific topological configuration i, we first find the subgraph of G that contains the maximum number of compatible contacts. (We find the maximal subgraph containing all possible edges from all substructures in topological configuration i, without including edges from substructures that are not represented in configuration i.) The free energy of this microstate serves as the reference state for the Wang–Landau calculation, $F[i, \max_i(X)]$. We then apply the algorithm described in Ref. 10 using the following acceptance probabilities for proposed moves $g \to g'$:

$$p_{\rm acc}(g \to g') = \min\left\{1,\, \frac{\alpha(g' \to g)}{\alpha(g \to g')}\, e^{-\left[F(g') - F(g)\right]/k_{\rm B}T} \right. \qquad \text{(S4)}$$
$$\left. \times\, e^{\left[F_t\left[i, X(g')\right] - F_t\left[i, X(g)\right]\right]/k_{\rm B}T} \right\}.$$

The Wang–Landau algorithm breaks detailed balance, since the bias changes as a function of the Monte Carlo ‘time,’ t. However, the amount by which $F_t(i,X)$ is updated between Monte Carlo moves is gradually decreased as the algorithm runs such that the estimated $F_t(i,X)$ converges to the equilibrium free-energy landscape. The total free energy of each topological configuration is then $F_i = -k_{\rm B}T \ln \sum_X \exp[-F_t(i,X)/k_{\rm B}T]$. For proteins with ~ 60 residues, sufficiently converged results for all topological configurations can typically be obtained in a few minutes on a single processor.

C. Calculation of ensemble averages

Once the Wang–Landau sampling is complete, we use $F_t(i,X)$ as a biasing potential to accelerate the calculation of equilibrium averages via standard Metropolis Monte Carlo sampling. If the free-energy differences between adjacent coarse-grained states have converged to within $\sim 1\ k_{\rm B}T$, then biased Metropolis Monte Carlo sampling will visit all coarse-grained states with roughly equal frequency. This means that the Metropolis algorithm can provide a direct verification of the convergence of the Wang–Landau sampling.

We use Metropolis Monte Carlo sampling to compute the equilibrium contact probability, $\langle \mathbf{1}_{uv} \rangle$, and vertex probability, $\langle \mathbf{1}_u \rangle$, within each topological configuration i. We calculate the probability that the contact (u,v) or the vertex u appears in the set of visited microstates,

$$\langle \mathbf{1}_{uv} \rangle_i \simeq \frac{\sum_X \sum_{\{y\}_X} \mathbf{1}_{uv}(g_y)\, e^{-F(i,X)/k_{\rm B}T}}{\sum_X \sum_{\{y\}_X} e^{-F(i,X)/k_{\rm B}T}}, \qquad \text{(S5)}$$

$$\langle \mathbf{1}_{u} \rangle_i \simeq \frac{\sum_X \sum_{\{y\}_X} \mathbf{1}_{u}(g_y)\, e^{-F(i,X)/k_{\rm B}T}}{\sum_X \sum_{\{y\}_X} e^{-F(i,X)/k_{\rm B}T}}, \qquad \text{(S6)}$$

where $\mathbf{1}_{uv}(g)$ and $\mathbf{1}_u(g)$ indicate the presence of edge (u,v) and vertex u, respectively, in microstate g, and $\{y\}_X$ is the set of all visited microstates with order parameter X. The use of a biasing potential allows the Markov chain to explore the entire free-energy landscape rapidly without getting stuck for long intervals in local free-energy minima. The fact that the underlying Markov chain obeys detailed balance ensures that the ensemble average within each coarse-grained state (i,X) converges to its equilibrium value given a sufficient number of Monte Carlo steps, $n_{\rm MC}$. Typically, we choose $n_{\rm MC} \simeq 1000$ per coarse-grained state (i,X).

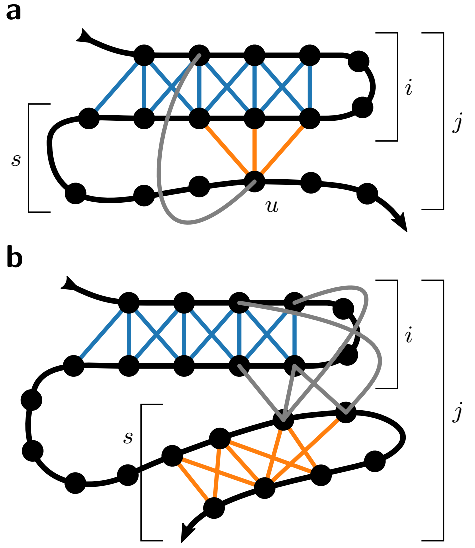

FIG. S1. **Schematic of mean-field barrier calculations.** (a) In the first mechanism, a single vertex u is added to the existing structured region i to form a new loop in the polymer backbone. (b) In the second mechanism, a pre-assembled substructure s makes contact with the existing structured region i; in this case, substructure s has no residues in common with configuration i. In both cases, after the formation of this initial contact, the polymer is in topological configuration j. See text for details.

S3. MEAN-FIELD BARRIER CALCULATIONS

To compute the free-energy barrier between a pair of topological configurations i and j, we assume that the initial configuration i is in local equilibrium. Making a contact between the existing structure in configuration i and the new substructure s, which is part of configuration j, necessarily requires the formation of a new loop in the polymer backbone; after this initial contact, folding can proceed in topological configuration j by making further native contacts at a much smaller entropic cost per contact. The barrier calculation should therefore account for all the ways in which this initial contact between the structured region of configuration i and the new substructure s can be made. This calculation is carried out in a mean-field approximation, where the effective strength of an interaction between a residue from the new substructure s and a residue v in the existing structured region depends on the local equilibrium in configuration i, $\langle \mathbf{1}_v \rangle_i$; this approximation is described below. Fluctuations within configuration i are taken into account by Boltzmann-averaging this barrier calculation over all values of the order parameter X in this configuration.

The addition of a new substructure to the existing structured region in configuration i can occur by one of two mechanisms, depending on the way the substructures interact in topological configuration j. The first mechanism applies in cases where the contacts associated with the new substructure s directly involve residues that are already present in topological configuration i. As a result, the first step in the assembly of the new substructure involves the addition of a residue u that participates in substructure s but is not part of the existing structured region i (see Figure S1a). Assuming that the value of the order parameter for the existing structure is X, the mean-field free energy of this configuration depends on the loss of conformation entropy due to bringing a residue u into contact with structured region in configuration i, $\langle \Delta S_u \rangle_{i,X}$, as well as the mean-field energies of all native contacts between u and residues in region i,

$$\frac{\Delta F^{\dagger}_{i,X\to j}}{k_{\mathrm{B}}T} = -\ln \sum_{u} \exp\left\langle \frac{\Delta S_u}{k_{\mathrm{B}}} \right\rangle_{i,X} \times \left\{ \exp\left[-\sum_{v\in\mathcal{V}(i)} \left(\frac{\epsilon_{uv}}{k_{\mathrm{B}}T} \right) \langle \mathbf{1}_v \rangle_{i,X} \right] - 1 \right\}, \tag{S7}$$

where $\mathcal{V}(i)$ indicates the set of residues that contribute to configuration i. The first sum in Eq. (S7) runs over all residues $\{u\}$ that participate in one of the contacts comprising substructure s and are not in the set $\mathcal{V}(i)$.

The second mechanism applies in cases where the new and existing substructures do not have any residues in common (see Figure S1b). Instead, these substructures interact in the native state via edges that are not part of any substructure (i.e., gray edges in Figure S1b). To calculate the barrier in this case, we assume that both the initial configuration i and the new substructure s are in local equilibrium. In the mean-field approximation, the free energy of all microstates in which the substructure s makes contact with the locally equilibrated structured region in configuration i, assuming that the value of the order parameter for the existing structure is X, is

$$\frac{\Delta F^{\dagger}_{i,X\to j}}{k_{\mathrm{B}}T} = F_s - \left\langle \frac{\Delta S_s}{k_{\mathrm{B}}} \right\rangle_{i,X} - \ln\left\{ \exp\left[-\sum_{\substack{u\in\mathcal{V}(s)\\ v\in\mathcal{V}(i)}} \langle \mathbf{1}_u \rangle_s \left(\frac{\epsilon_{uv}}{k_{\mathrm{B}}T} \right) \langle \mathbf{1}_v \rangle_{i,X} \right] - 1 \right\}, \tag{S8}$$

where F_s is the free energy of the isolated substructure s and $\langle \Delta S_s \rangle_{i,X}$ is the entropic penalty due to bringing s into contact with the structured region in configuration i.

We compute the apparent barrier between configurations i and j by summing over all values of the order parameter X,

$$\frac{\Delta F^{\dagger}_{i\to j}}{k_{\mathrm{B}}T} = -\ln \sum_{X} \exp\left[\frac{-\Delta F^{\dagger}_{i,X\to j} - (F_{i,X} - F_i)}{k_{\mathrm{B}}T} \right]. \tag{S9}$$

The term $(F_{i,X} - F_i)$ accounts for the free-energy difference between microstates at a specific value of the order parameter and the total free energy of topological configuration i, F_i. To obey detailed balance, the barrier for the reverse transition is $\Delta F^{\dagger}_{j\to i} = \Delta F^{\dagger}_{i\to j} - (F_j - F_i)$.

S4. TRANSITION-PATH THEORY

Given the continuous-time Markov chain specified by the rate matrix in Eq. (3), we can use transition-path theory (4) to calculate properties of the ensemble of folding trajectories. The stationary distribution of the Markov chain, $\pi(i, X)$, is equivalent to the Boltzmann distribution, $\pi_i = \exp(-F_i/k_{\mathrm{B}}T)/\sum_j \exp(-F_i/k_{\mathrm{B}}T)$. All folding transition paths originate in the unfolded configuration, $A = \varnothing$, and terminate in the configuration with the maximum number of substructures, B. Here we reproduce a number of equations from Ref. 4 for completeness.

First, we calculate $p_{\mathrm{fold}}(i)$, the equilibrium probability that a dynamical trajectory will reach configuration B, starting from configuration i, before returning to configuration A. By definition, p_{fold} is equal to zero and one in configurations A and B, respectively. Using the rate matrix k_{ij}, p_{fold} is computed for all intermediate configurations by solving the linear system

$$\sum_j k_{ij} p_{\mathrm{fold}}(j) = 0 \quad \forall i \in (A \cup B)^c, \tag{S10}$$

where $(A \cup B)^c$ indicates all configurations that are neither A nor B. The reactive flux through every transition $i \to j$ is

$$f(i \to j) = \begin{cases} \pi_i \big[1 - p_{\mathrm{fold}}(i)\big] k_{ij} p_{\mathrm{fold}}(j) & \text{if } i \neq j, \\ 0 & \text{if } i = j. \end{cases} \tag{S11}$$

The net reactive flux through the transition $i \to j$ is $f^{+}_{ij} \equiv \max(f_{ij} - f_{ji}, 0)$. From this calculation, we can determine the overall folding rate,

$$k_{\mathrm{fold}} = \frac{\sum_{j\neq A} f^{+}_{Aj}}{\pi_A}. \tag{S12}$$

In the two-state approximation, the apparent free-energy barrier between configurations A and B is $\Delta F^{\dagger}_{AB} = -\ln(2k_{\mathrm{fold}})$. Lastly, the fraction of time spent in configuration i in the transition-path ensemble is

$$p_{AB}(i) = \pi_i p_{\mathrm{fold}}(i)[1 - p_{\mathrm{fold}}(i)]. \tag{S13}$$

S5. THEORETICAL ϕ AND ψ-VALUE CALCULATIONS

Theoretical ϕ and ψ-values are calculated as described in Eqs. (5) and (6). For the rate calculation, k_{fold}, the unfolded, A, and folded, B, configurations are chosen as described in Sec. S4. The inverse temperature $(k_{\mathrm{B}}T)^{-1}$ is chosen to equate the free-energies of the folded ensemble, which includes contributions from all native contacts, and the unfolded ensemble; we take the unfolded ensemble to include both the random coil configuration, $\varnothing$, and all individual substructures in isolation, $F_{\mathrm{unfolded}} = -k_{\mathrm{B}}T \ln\left[1 + \sum_s \exp(-F_s/k_{\mathrm{B}}T)\right]$, where the index s runs over all substructures. (Exceptions are made for proteins 1rnb and 2vil, where, due to the stability of partially folded intermediate configurations, the free-energy differences between the native and unfolded ensembles are set to -2.5 and -1 $k_{\mathrm{B}}T$, respectively. These choices ensure that the native states are globally stable.)

The mutations considered in our comparisons with experimental measurements are listed in Tables S1–S3 and shown in Figures S7–S9. Unless otherwise noted, we assume that the experimental errors on ϕ and ψ-values are ± 0.1. We leave ϕ-values that are less than -0.1 or greater than 1.1 out of comparisons with the theoretical predictions. (ϕ-values in the range $[-0.1 : 0]$ or $[1 : 1.1]$ are set to 0 or 1, respectively.) For ψ-value comparisons, we set values greater than 1 to unity.

S6. ANALYSIS OF ATOMISTIC MOLECULAR DYNAMICS SIMULATIONS

For the analysis of atomistic simulation data, we adopt a history-dependent native-contact definition (2): a contact is formed when heavy atoms from two residues pass within 3.5 Å of one another and broken when all heavy atoms of the same residues move farther than 5.5 Å apart. To reduce the contribution of transient fluctuations further, we disregarded contacts lasting less than 5 ns; changing this threshold by ±5 ns does not meaningfully affect the results of the subsequent calculations. Native contacts were defined on the basis of the crystal structure as described in the Materials and Methods for direct comparison with the theoretical results. We determined the largest structured region at every 1-ns time step by decomposing the graph of native contacts into connected components. We then calculated a one-dimensional free-energy landscape as a function of the number of native contacts using all time steps from the available trajectories. Folding transition paths are defined as the portions of the trajectories that transit from the free-energy minimum of the unfolded ensemble on this landscape to the free-energy minimum of the folded ensemble without returning to the free-energy minimum of the unfolded ensemble. Unfolding transition paths are defined analogously, starting from the free-energy minimum of the folded ensemble.

For the configuration lifetime calculations shown in Figure 5, we identified all substructures with at least 6 contacts present in the largest structured region. We verified that every such substructure is completely contained within the largest structured region, i.e., no contacts from a substructure that forms part of the largest structured region are found outside of this region, in more than 99.8% of all time steps. For the commitment calculations shown in Figure 6, we used the stricter criterion for substructure formation described in the Results.

We calculated ϕ-values from the simulated transition paths using the method described in Ref. 7,

$$\psi_{uv}^{\mathrm{simulation}} \simeq p(\mathbf{1}_{uv}|\mathrm{TP}), \tag{S14}$$

$$\phi_{u}^{\mathrm{simulation}} = \sum_{v} \psi_{uv}^{\mathrm{simulation}}/d_u, \tag{S15}$$

where $p(\mathbf{1}_{uv}|\mathrm{TP})$ is the probability of observing a native contact between residues u and v at any time step in the transition-path ensemble and d_u is the number of native contacts formed by residue u. We estimated the variability in the predicted ϕ and ψ-values across the observed transition paths by performing bootstrapping simulations in which the 10 observed transition paths were sampled with replacement; the standard deviation of $\phi_u^{\mathrm{simulation}}$ estimated in this way is shown in Figure 7b. The calculations shown in Figure 7b are slightly different from the results presented in Ref. 7 because our definitions of native contacts are not identical. Because $\phi_u^{\mathrm{simulation}}$ is calculated directly from $\psi_{uv}^{\mathrm{simulation}}$, we obtain the same correlation coefficient with the theoretical predictions for both sets of values.

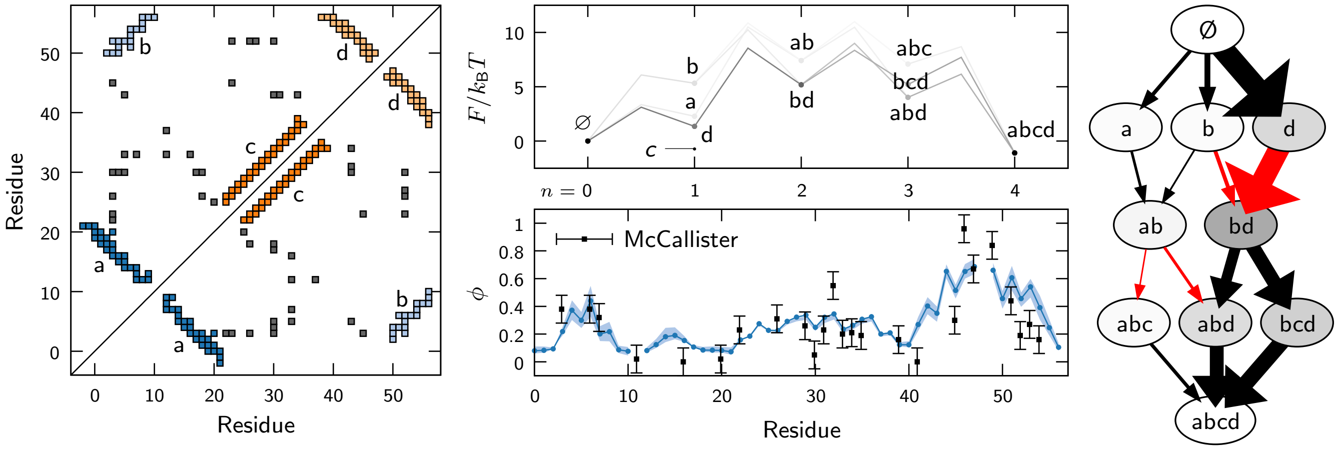

FIG. S2. **Predicted folding landscape for protein G (1igd) and comparison with experimental ϕ-values (11).** The configuration abcd is the native ensemble in this case, because all residues contribute the one of the four substructures. The free-energy landscape and folding network are drawn as in Figures 3c and 4a, respectively.

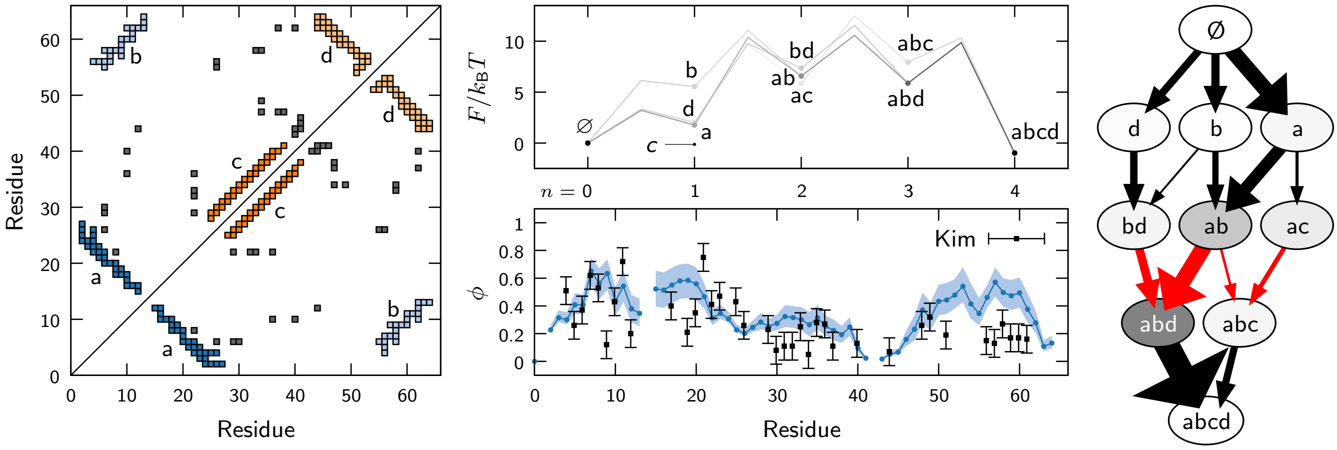

FIG. S3. **Predicted folding landscape for protein L (1k53) and comparison with experimental ϕ-values (12).** The configuration abcd is the native ensemble in this case, because all residues contribute the one of the four substructures. The free-energy landscape and folding network are drawn as in Figures 3c and 4a, respectively.

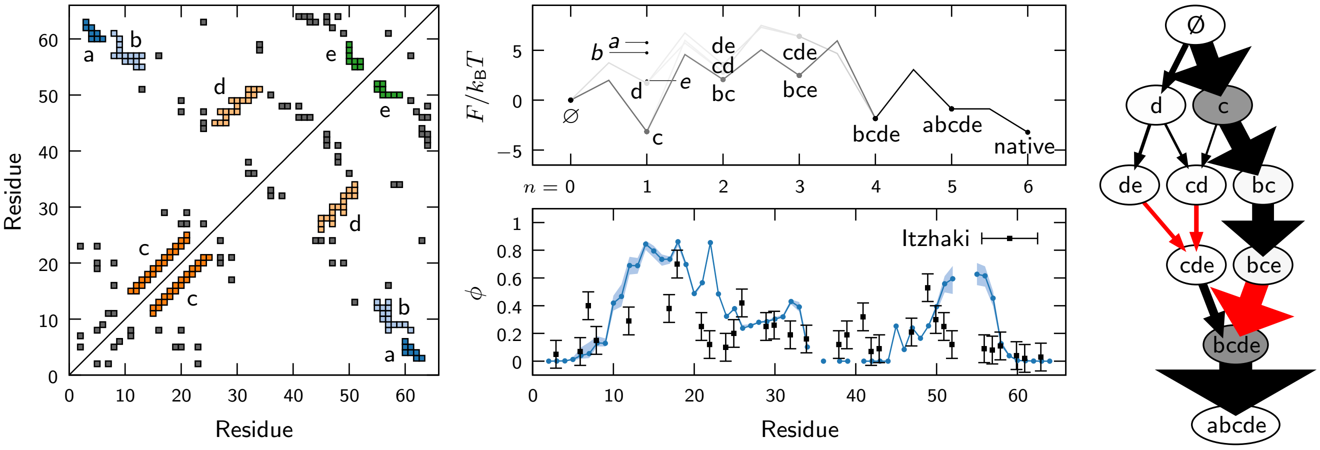

FIG. S4. **Predicted folding landscape for chymotrypsin inhibitor 2 (2ci2) and comparison with experimental ϕ-values (13).** The free-energy landscape and folding network are drawn as in Figures 3c and 4a, respectively.

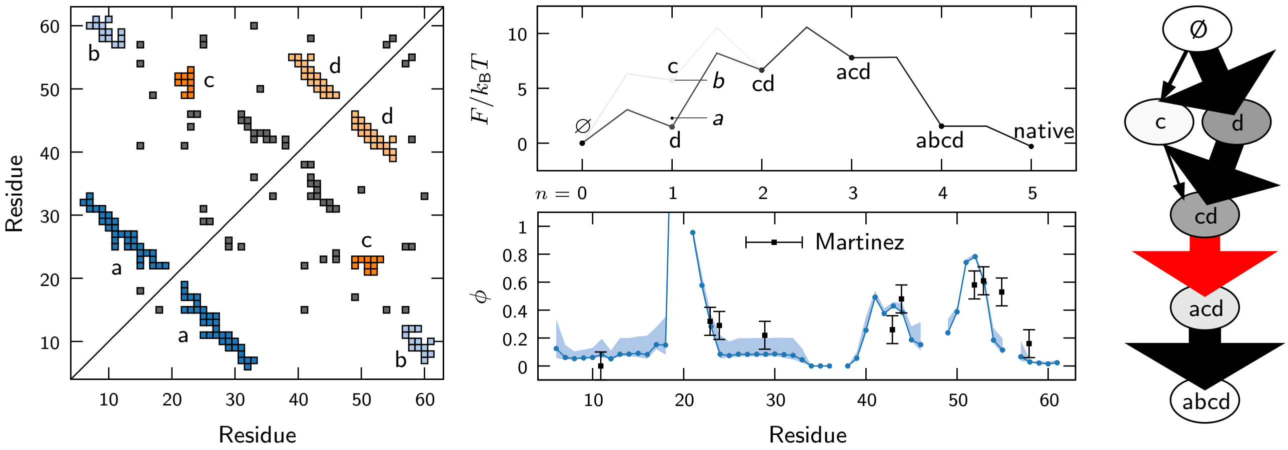

FIG. S5. **Predicted folding landscape for the Src SH3 domain (1shg) and comparison with experimental ϕ-values (14).** The free-energy landscape and folding network are drawn as in Figures 3c and 4a, respectively.

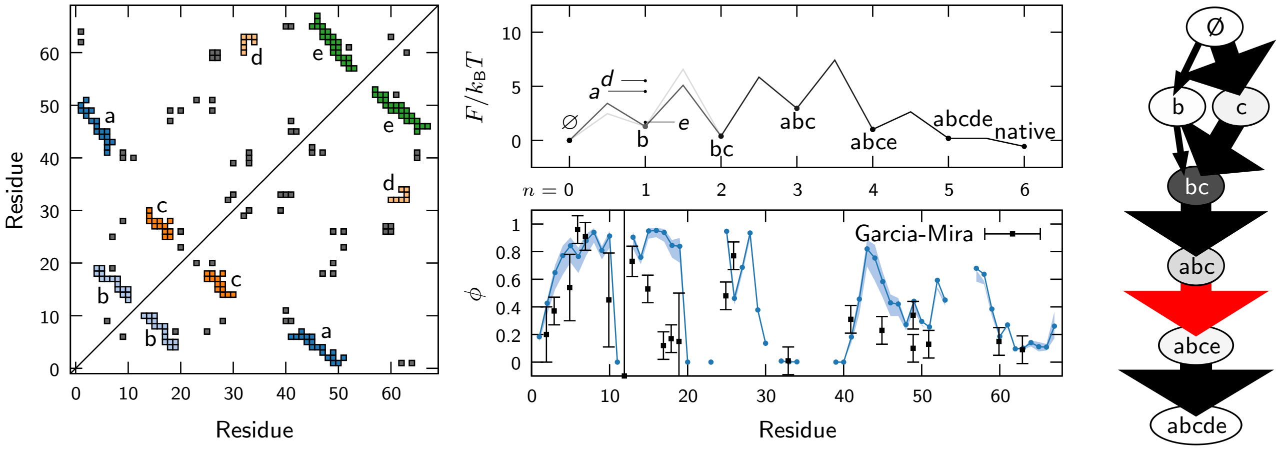

FIG. S6. **Predicted folding landscape for the cold-shock protein (1csp) and comparison with experimental ϕ-values (15).** The free-energy landscape and folding network are drawn as in Figures 3c and 4a, respectively.

1enh	ϕ_{expt}	ϕ_{pred}
F8A	0.42±0.1	0.15
L13A	0.51±0.1	0.31
A14G	0.79±0.1	0.39
F20A	0.36±0.1	0.24
Y25G	0.28±0.1	0.22
L26A	0.46±0.1	0.51
L38A	0.48±0.1	0.37
G39A	0.92±0.1	0.34
L40A	0.95±0.1	0.13
A43G	1.00±0.1	0.47
A54G	0.62±0.1	0.80

1igd	ϕ_{expt}	ϕ_{pred}
I6A	0.38±0.1	0.44
L7A	0.32±0.1	0.20
T16A	0.00±0.1	0.15
A20G	0.02±0.1	0.08
D22A	0.23±0.1	0.16
A26G	0.31±0.1	0.23
V29A	0.26±0.1	0.34
K31G	0.23±0.1	0.32
Q32G	0.55±0.1	0.35
Y33A	0.20±0.1	0.24
A34G	0.21±0.1	0.26
N35G	0.19±0.1	0.31
V39A	0.16±0.1	0.12
G41A	0.00±0.1	0.27
D46A	0.96±0.1	0.65
D47A	0.67±0.1	0.69
T49A	0.84±0.1	0.66
T51A	0.44±0.1	0.61
T53A	0.27±0.1	0.54
V54A	0.16±0.1	0.39

1shg	ϕ_{expt}	ϕ_{pred}
A11G	0.00±0.1	0.08
V23A	0.32±0.1	0.28
T24A	0.29±0.1	0.08
D29A	0.22±0.1	0.08
K43A	0.26±0.1	0.43
V44A	0.48±0.1	0.38
F52A	0.58±0.1	0.78
V53A	0.61±0.1	0.59
A55G	0.53±0.1	0.11
V58A	0.16±0.1	0.03

1k53	ϕ_{expt}	ϕ_{pred}
V4A	0.51±0.1	0.30
T5A	0.26±0.1	0.41
I6A	0.37±0.1	0.41
K7A	0.62±0.1	0.65
A8G	0.53±0.1	0.56
N9A	0.12±0.1	0.63
L10A	0.43±0.1	0.43
I11A	0.72±0.1	0.54
F12A	0.20±0.1	0.38
T17A	0.40±0.1	0.55
T19A	0.21±0.1	0.58
A20G	0.35±0.1	0.56
E21A	0.75±0.1	0.47
F22A	0.41±0.1	0.31
K23A	0.47±0.1	0.35
T25A	0.43±0.1	0.23
F26G	0.26±0.1	0.20
A29G	0.23±0.1	0.25
T30A	0.08±0.1	0.27
S31G	0.11±0.1	0.32
E32G	0.11±0.1	0.32
A33G	0.25±0.1	0.30
Y34A	0.05±0.1	0.26
A35G	0.28±0.1	0.32
Y36A	0.27±0.1	0.25
A37G	0.11±0.1	0.22
L40A	0.13±0.1	0.10
N44A	0.07±0.1	0.05
T48A	0.26±0.1	0.37
V49A	0.32±0.1	0.32
V51A	0.19±0.1	0.44
Y56A	0.15±0.1	0.46
T57A	0.13±0.1	0.57
L58A	0.27±0.1	0.50
N59A	0.17±0.1	0.47
I60A	0.17±0.1	0.49
K61A	0.16±0.1	0.38

2ci2	ϕ_{expt}	ϕ_{pred}
T3G	0.05±0.1	0.00
P6A	0.07±0.1	0.04
E7A	0.40±0.1	0.05
L8A	0.15±0.1	0.13
S12G	0.29±0.1	0.69
K17G	0.38±0.1	0.74
K18G	0.70±0.1	0.86
L21A	0.25±0.1	0.57
Q22G	0.12±0.1	0.85
K24G	0.10±0.1	0.32
P25A	0.20±0.1	0.38
E26A	0.42±0.1	0.24
I29A	0.25±0.1	0.29
I30G	0.26±0.1	0.30
L32A	0.19±0.1	0.43
V34G	0.16±0.1	0.10
V38A	0.12±0.1	0.00
T39A	0.19±0.1	0.00
E41A	0.32±0.1	0.00
Y42G	0.07±0.1	0.00
R43A	0.09±0.1	0.00
V47A	0.21±0.1	0.24
L49A	0.53±0.1	0.26
F50A	0.30±0.1	0.39
V51A	0.25±0.1	0.56
D52A	0.12±0.1	0.59
N56A	0.09±0.1	0.62
I57A	0.08±0.1	0.45
A58G	0.11±0.1	0.13
V60G	0.04±0.1	0.00
P61A	0.02±0.1	0.00
V63G	0.03±0.1	0.00

1csp	ϕ_{expt}	ϕ_{pred}
L2A	0.20±0.2	0.43
K5A	0.54±0.24	0.84
K7A	0.91±0.1	0.88
N10A	0.45±0.34	0.91
K13A	0.73±0.11	0.90
F15A	0.53±0.1	0.95
F17A	0.12±0.1	0.94
E19A	0.15±0.35	0.84
D25A	0.48±0.1	0.95
I33A	0.01±0.1	0.00
L41A	0.31±0.1	0.18
Q45A	0.23±0.1	0.58
F49A	0.34±0.1	0.44
I51A	0.13±0.1	0.26
A60G	0.15±0.1	0.19
V63A	0.09±0.1	0.10

1ubq	ϕ_{expt}	ϕ_{pred}
I3A	0.30±0.1	0.78
V5A	0.50±0.1	0.86
T7A	0.80±0.1	0.85
I13A	0.50±0.1	0.71
L15A	0.50±0.1	0.69
V17A	0.50±0.1	0.40
T22A	0.50±0.1	0.27
I23A	0.40±0.1	0.24
V26A	0.30±0.1	0.51
L27A	0.10±0.1	0.54
A28G	1.00±0.1	0.66
I30A	0.50±0.1	0.60
Q41A	0.00±0.1	0.57
L43A	0.30±0.1	0.50
L50A	0.00±0.1	0.06
L56A	0.10±0.1	0.05
I61A	0.00±0.1	0.08
L67A	0.00±0.1	0.79
L69A	0.30±0.1	0.73

1imp	ϕ_{expt}	ϕ_{pred}
A13G	0.98±0.1	0.83
F15A	0.57±0.1	0.58
L16A	0.52±0.1	0.57
L18A	0.40±0.1	0.55
V19A	0.32±0.1	0.39
L33A	0.27±0.1	0.33
L36A	0.25±0.1	0.37
V37A	0.15±0.1	0.24
L52A	0.03±0.1	0.00
V68A	0.23±0.1	0.10
V71A	0.36±0.1	0.07
A76G	0.37±0.2	0.05
A77G	0.37±0.1	0.05
F83A	0.31±0.1	0.52

1tiu	ϕ_{expt}	ϕ_{pred}
I2A	0.45±0.1	0.00
V4A	0.29±0.1	0.00
L8A	0.28±0.1	0.03
V13A	0.00±0.1	0.01
V15A	0.01±0.1	0.01
A19G	0.38±0.1	0.25
I23A	0.82±0.1	0.15
L25A	0.42±0.1	0.00

TABLE S1. **List of ϕ-value mutations.** Data points are from the following references, modified as described in Sec. S5: 1enh (16), 1igd (11), 1shg (14), 1k53 (12), 2ci2 (13), 1csp (15), 1ubq (6) refolding, 1imp (17) and 1tiu (18).

V30A	0.45±0.1	0.00
G32A	0.51±0.1	0.74
L36A	0.50±0.1	0.53
L41A	0.40±0.1	0.00
C47A	0.42±0.1	0.48
H56A	0.52±0.1	0.51
L58A	0.79±0.1	0.65
L60A	0.67±0.1	0.16
C63A	0.23±0.1	0.04
M67A	0.13±0.1	0.11
V71A	0.63±0.1	0.68
A82G	0.16±0.1	0.30
L84A	0.05±0.1	0.31
V86A	0.01±0.1	0.02

1btb	ϕ_{expt}	ϕ_{pred}
Q9G	0.72±0.2	0.04
I13A	0.45±0.2	0.06
Q18G	0.69±0.2	0.09
A25G	0.68±0.2	0.04
A36G	0.70±0.2	0.42
L37A	0.59±0.2	0.60
L41A	0.45±0.2	0.58
V45A	0.47±0.2	0.21
L49A	0.47±0.2	0.58
V50G	0.77±0.2	0.39
F56A	0.35±0.2	0.13
Q58G	0.11±0.2	0.08
Q61G	0.09±0.2	0.08
T63A	0.38±0.2	0.05
A67G	0.30±0.2	0.29
E68A	0.52±0.2	0.52
V70A	0.41±0.2	0.47
Q72G	0.81±0.2	0.85
A77G	0.90±0.2	0.73
A79G	0.63±0.2	0.81
T85A	0.51±0.2	0.64

1fkb	ϕ_{expt}	ϕ_{pred}
V2A	0.39±0.1	0.39
V4A	0.27±0.1	0.40
T21A	0.40±0.1	0.45
V23A	0.52±0.1	0.47
V24A	0.38±0.1	0.45
T27A	0.28±0.1	0.41
F36A	0.15±0.1	0.29
L50A	0.39±0.1	0.33
V55A	0.08±0.1	0.32
I56A	0.19±0.1	0.34
R57G	0.14±0.1	0.37
E60G	0.13±0.1	0.26
E61G	0.20±0.1	0.36
V63A	0.51±0.1	0.36
T75A	0.24±0.1	0.41
I76A	0.34±0.1	0.40
I91A	0.04±0.1	0.00
L97A	0.23±0.1	0.40
V98A	0.27±0.1	0.44
V101A	0.57±0.1	0.44
L106A	0.35±0.1	0.42

1rnb	ϕ_{expt}	ϕ_{pred}
N5A	0.09±0.1	0.13
T6G	0.21±0.1	0.38
V10A	0.33±0.1	0.35
L14A	0.59±0.1	0.37
T26G	0.00±0.1	0.07
V36A	0.00±0.1	0.00
N58A	0.94±0.1	0.00
N77A	0.00±0.1	0.11
N84A	0.16±0.1	0.00
S91A	0.93±0.1	0.53
S92A	0.95±0.1	0.52

3chy	ϕ_{expt}	ϕ_{pred}
A36G	0.75±0.1	0.66
D38G	0.60±0.1	0.33
A42G	0.68±0.1	0.62
D64A	0.11±0.1	0.28
A97G	0.00±0.1	0.09
A98G	0.03±0.1	0.07
T112G	0.12±0.1	0.04

2vil	ϕ_{expt}	ϕ_{pred}
L3A	0.35±0.1	0.00
V7A	0.45±0.1	0.00
I18A	0.49±0.1	0.44
I23A	0.65±0.1	0.62
M28A	0.58±0.1	0.65
C44A	0.85±0.1	0.64
V46A	0.69±0.1	0.66
L47A	0.43±0.1	0.58
L48A	0.62±0.1	0.67
I61A	0.05±0.1	0.71
L65A	0.24±0.1	0.61
E73A	0.69±0.1	0.56
A77G	0.52±0.1	0.57
A78G	0.56±0.1	0.58
T81A	0.75±0.1	0.57
M84A	0.68±0.1	0.57
L114A	0.03±0.1	0.01

TABLE S2. **List of ϕ-value mutations (continued).** Data points are from the following references, modified as described in Sec. S5: 1btb (19), 1fkb (20), 1rnb (21), 3chy (22) and 2vil (23).

1igd	ψ_{expt}	ψ_{pred}
K4–T51	0.17±0.1	0.55
I6–T53	0.71±0.1	0.51
N8–T55	0.30±0.1	0.26
T16–Y33	0.24±0.1	0.08
K28–Q32	0.24±0.1	0.35
Q32–D36	0.03±0.1	0.33
T44–T53	0.93±0.1	0.61
D46–T51	0.90±0.1	0.66

1k53	ψ_{expt}	ψ_{pred}
N9–T19	0.75±0.2	0.57
N9–N59	1.00±0.4	0.68
I11–K61	1.00±0.1	0.57
K28–E32	0.26±0.1	0.31
A35–T39	0.00±0.1	0.25
D50–N59	1.00±0.1	0.43
A52–T57	1.00±0.1	0.50

1ubq	ψ_{expt}	ψ_{pred}
Q2–E16	0.53±0.1	0.66
Q2–E64	0.03±0.1	0.59
F4–T12	1.00±0.1	0.90
F4–T66	0.75±0.1	0.90
K6–T12	1.00±0.1	0.89
K6–T66	1.00±0.1	0.88
K6–H68	0.52±0.1	0.89
E24–A28	0.48±0.1	0.64
A28–D32	0.90±0.1	0.66
R42–Q49	0.07±0.1	0.40
R42–H68	0.26±0.1	0.72
R42–V70	0.57±0.1	0.67
F44–Q49	0.02±0.1	0.40
I44–V70	1.00±0.1	0.81

2acy	ψ_{expt}	ψ_{pred}
D10–N81	0.70±0.1	0.22
E12–N79	1.00±0.1	0.23
K24–A28	0.01±0.1	0.08
A28–K32	0.00±0.1	0.08
W38–Q50	1.00±0.1	0.80
Q40–V97	0.13±0.1	0.07
S56–H60	0.34±0.1	0.35
R59–E63	1.00±0.1	0.42

TABLE S3. **List of ψ-value mutations.** For ubiqutin (1ubq), two experimental ψ-values (residue pairs 2–16 and 44–70) involve residues that do not form native contacts in the crystal structure. We calculated theoretical ψ-values for the nearest native contacts in our model, replacing these pairs with contacts 1–16 and 44–68, respectively. Data points are from the following references, modified as described in Sec. S5: 1igd (24), 1k53 (25), 1ubq (5) and 2acy (26).

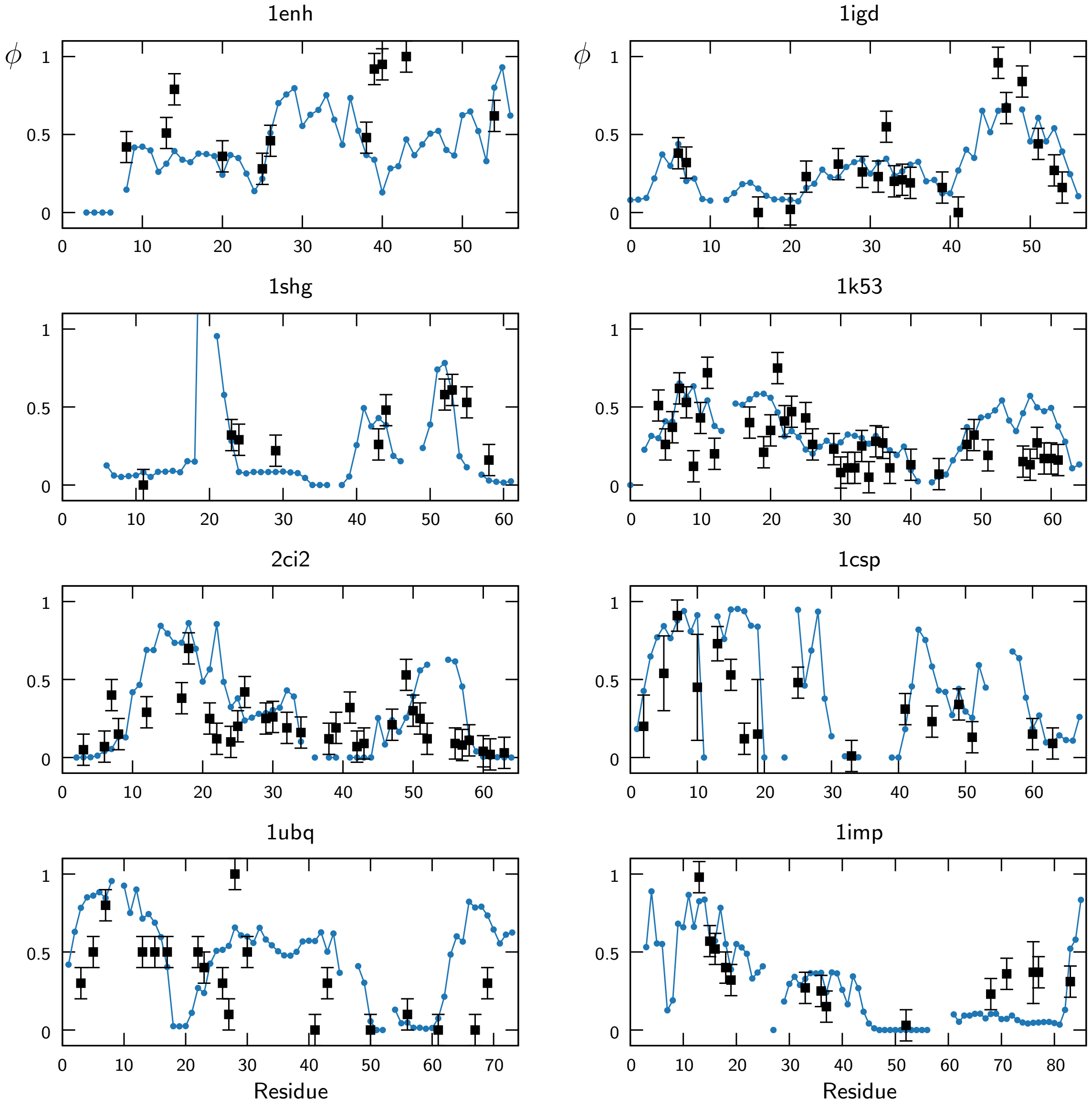

FIG. S7. **Comparison of predicted and experimental ϕ-values.** Predictions are indicated by blue circles, and experimental points are shown as black squares. Experimental errors are assumed to be 0.1 unless otherwise indicated; see Tables S1 and S2 for a list of data points.

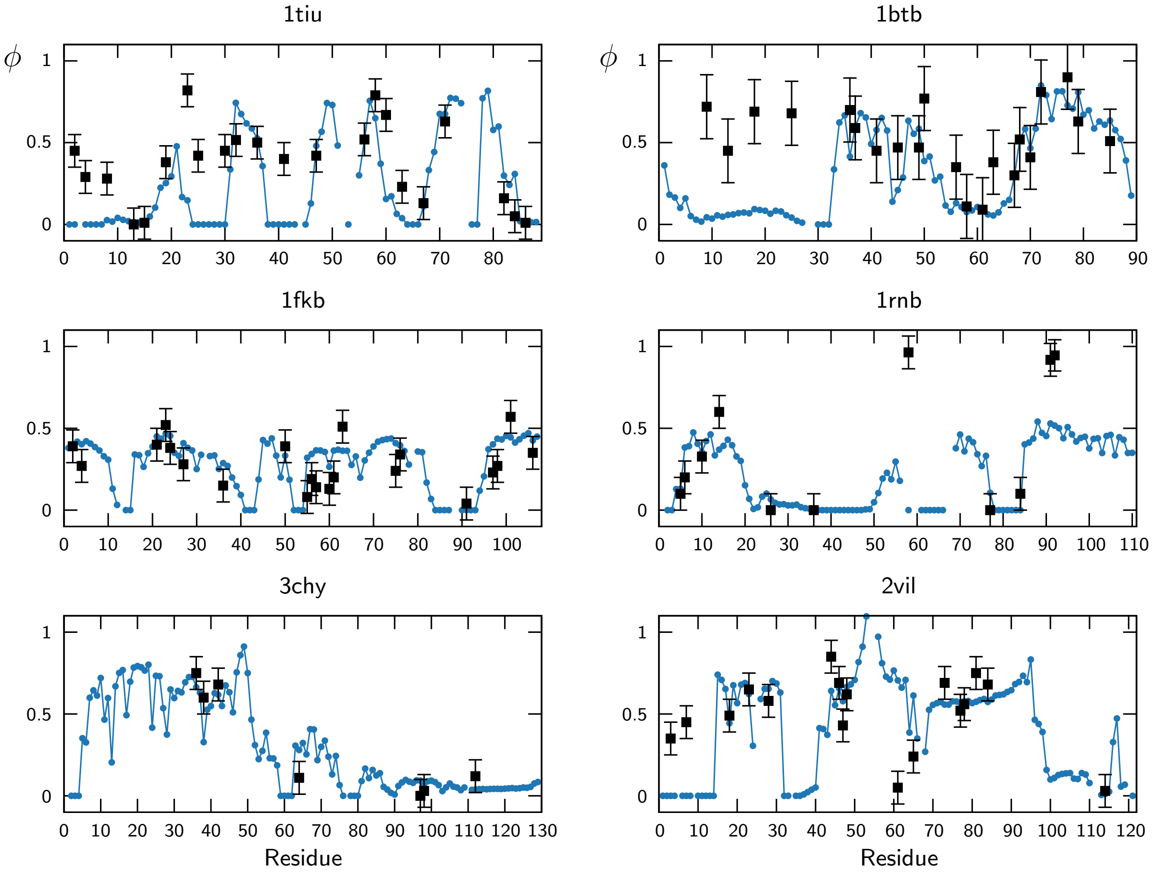

FIG. S8. **Comparison of predicted and experimental ϕ-values (continued).** Predictions are indicated by blue circles, and experimental points are shown as black squares. Experimental errors are assumed to be 0.1 unless otherwise indicated; see Tables S1 and S2 for a list of data points.

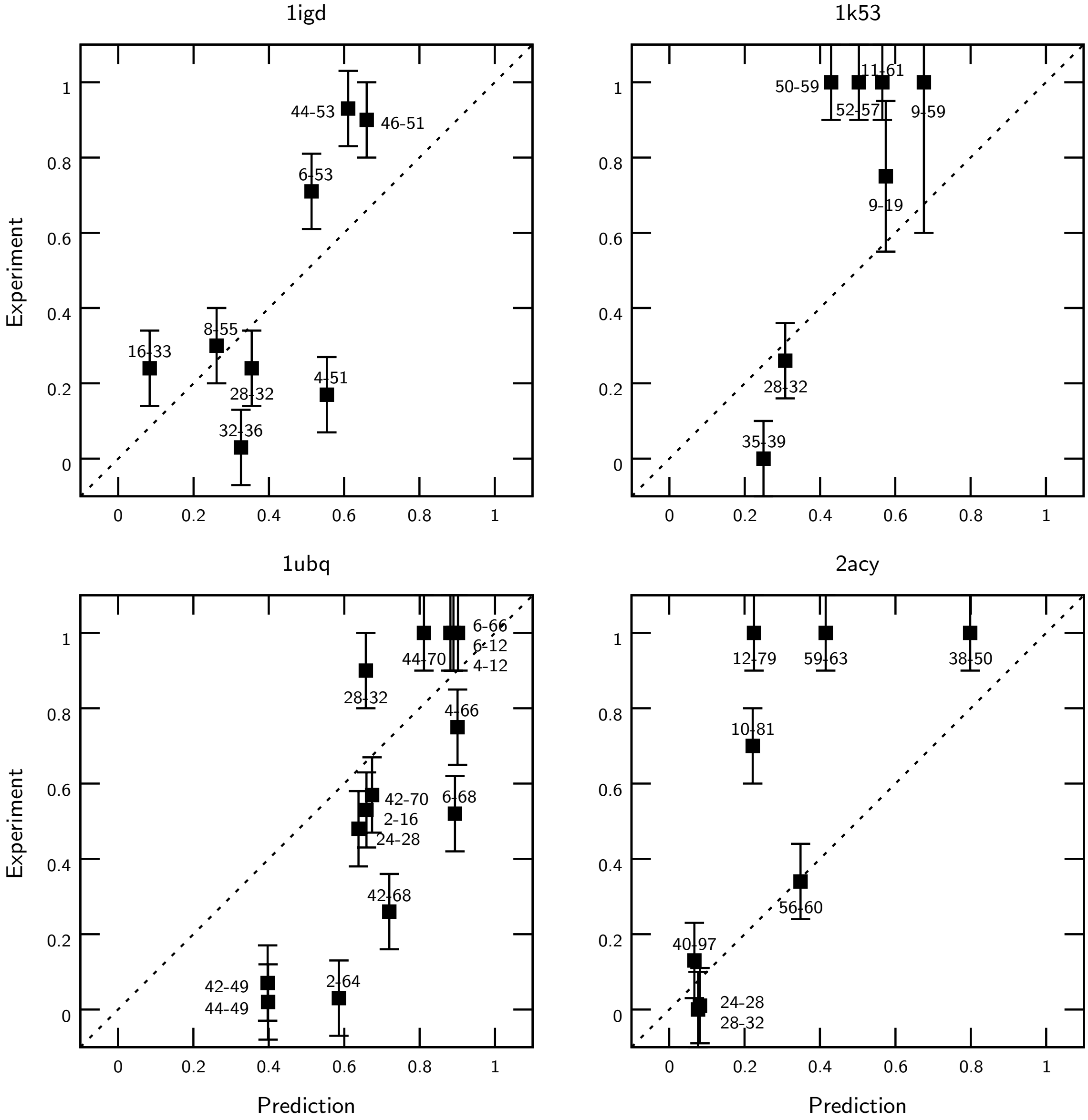

FIG. S9. **Scatter plots for ψ-value comparisons.** For each point, the pair of mutated residues is indicated. Experimental errors are assumed to be 0.1 unless otherwise indicated; see Table S3 for a list of data points.

SUPPORTING REFERENCES

[1] Onuchic, J. N., and P. G. Wolynes, 2004. Theory of protein folding. *Curr. Opin. Struct. Biol.* 14:70–75.

[2] Best, R. B., G. Hummer, and W. A. Eaton, 2013. Native contacts determine protein folding mechanisms in atomistic simulations. *Proc. Natl. Acad. Sci. U.S.A.* 110:17874–17879.

[3] Jacobs, W. M., A. Reinhardt, and D. Frenkel, 2015. Communication: Theoretical prediction of free-energy landscapes for complex self-assembly. *J. Chem. Phys.* 142:021101.

[4] Metzner, P., C. Schütte, and E. Vanden-Eijnden, 2009. Transition path theory for Markov jump processes. *Multiscale Model. Sim.* 7:1192–1219.

[5] Sosnick, T. R., R. S. Dothager, and B. A. Krantz, 2004. Differences in the folding transition state of ubiquitin indicated by φ and ψ analyses. *Proc. Natl. Acad. Sci. U.S.A.* 101:17377–17382.

[6] Went, H. M., and S. E. Jackson, 2005. Ubiquitin folds through a highly polarized transition state. *Protein Eng. Des. Sel.* 18:229–237.

[7] Best, R. B., and G. Hummer, 2016. Microscopic interpretation of folding ϕ-values using the transition path ensemble. *Proc. Natl. Acad. Sci. U.S.A.* 113:3263–3268.

[8] Wang, F., and D. P. Landau, 2001. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.* 86:2050.

[9] Frenkel, D., and B. Smit, 2001. Understanding molecular simulation: From algorithms to applications. Academic Press.

[10] Belardinelli, R. E., and V. D. Pereyra, 2007. Wang-Landau algorithm: A theoretical analysis of the saturation of the error. *J. Chem. Phys.* 127:184105.

[11] McCallister, E. L., E. Alm, and D. Baker, 2000. Critical role of β-hairpin formation in protein G folding. *Nat. Struct. Mol. Biol.* 7:669–673.

[12] Kim, D. E., C. Fisher, and D. Baker, 2000. A breakdown of symmetry in the folding transition state of protein L. *J. Mol. Biol.* 298:971–984.

[13] Itzhaki, L. S., D. E. Otzen, and A. R. Fersht, 1995. The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: Evidence for a nucleation–condensation mechanism for protein folding. *J. Mol. Biol.* 254:260–288.

[14] Martínez, J. C., and L. Serrano, 1999. The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nat. Struct. Mol. Biol.* 6:1010–1016.

[15] Garcia-Mira, M. M., D. Boehringer, and F. X. Schmid, 2004. The folding transition state of the cold shock protein is strongly polarized. *J. Mol. Biol.* 339:555–569.

[16] Gianni, S., N. R. Guydosh, F. Khan, T. D. Caldas, U. Mayor, G. W. White, M. L. DeMarco, V. Daggett, and A. R. Fersht, 2003. Unifying features in protein-folding mechanisms. *Proc. Natl. Acad. Sci. U.S.A.* 100:13286–13291.

[17] Friel, C. T., A. P. Capaldi, and S. E. Radford, 2003. Structural analysis of the rate-limiting transition states in the folding of Im7 and Im9: Similarities and differences in the folding of homologous proteins. *J. Mol. Biol.* 326:293–305.

[18] Fowler, S. B., and J. Clarke, 2001. Mapping the folding pathway of an immunoglobulin domain: Structural detail from phi value analysis and movement of the transition state. *Structure* 9:355–366.

[19] Nölting, B., R. Golbik, J. L. Neira, A. S. Soler-Gonzalez, G. Schreiber, and A. R. Fersht, 1997. The folding pathway of a protein at high resolution from microseconds to seconds. *Proc. Natl. Acad. Sci. U.S.A.* 94:826–830.

[20] Fulton, K. F., E. R. Main, V. Daggett, and S. E. Jackson, 1999. Mapping the interactions present in the transition state for unfolding/folding of FKBP12. *J. Mol. Biol.* 291:445–461.

[21] Serrano, L., A. Matouschek, and A. R. Fersht, 1992. The folding of an enzyme: III. Structure of the transition state for unfolding of barnase analysed by a protein engineering procedure. *J. Mol. Biol.* 224:805–818.

[22] López-Hernéndez, E., and L. Serrano, 1996. Structure of the transition state for folding of the 129 aa protein CheY resembles that of a smaller protein, CI-2. *Fold. Des.* 1:43–55.

[23] Choe, S. E., L. Li, P. T. Matsudaira, G. Wagner, and E. I. Shakhnovich, 2000. Differential stabilization of two hydrophobic cores in the transition state of the villin 14T folding reaction. *J. Mol. Biol.* 304:99–115.

[24] Baxa, M. C., W. Yu, A. N. Adhikari, L. Ge, Z. Xia, R. Zhou, K. F. Freed, and T. R. Sosnick, 2015. Even with nonnative interactions, the updated folding transition states of the homologs Proteins G & L are extensive and similar. *Proc. Natl. Acad. Sci. U.S.A.* 112:8302–8307.

[25] Yoo, T. Y., A. Adhikari, Z. Xia, T. Huynh, K. F. Freed, R. Zhou, and T. R. Sosnick, 2012. The folding transition state of protein L is extensive with nonnative interactions (and not small and polarized). *J. Mol. Biol.* 420:220–234.

[26] Pandit, A. D., A. Jha, K. F. Freed, and T. R. Sosnick, 2006. Small proteins fold through transition states with native-like topologies. *J. Mol. Biol.* 361:755–770.