BIOPHYSICS

Protein misfolding involving entanglements provides a structural explanation for the origin of stretched-exponential refolding kinetics

Yang Jiang[1]†, Yingzi Xia[2]†, Ian Sitarik[1], Piyoosh Sharma[2], Hyebin Song[3,4], Stephen D. Fried[2,5]*, Edward P. O'Brien[1,3,6]*

Stretched-exponential protein refolding kinetics, first observed decades ago, were attributed to a nonnative ensemble of structures with parallel, non-interconverting folding pathways. However, the structural origin of the large energy barriers preventing interconversion between these folding pathways is unknown. Here, we combine simulations with limited proteolysis (LiP) and cross-linking (XL) mass spectrometry (MS) to study the protein phosphoglycerate kinase (PGK). Simulations recapitulate its stretched-exponential folding kinetics and reveal that misfolded states involving changes of entanglement underlie this behavior: either formation of a nonnative, noncovalent lasso entanglement or failure to form a native entanglement. These misfolded states act as kinetic traps, requiring extensive unfolding to escape, which results in a distribution of free energy barriers and pathway partitioning. Using LiP-MS and XL-MS, we propose heterogeneous structural ensembles consistent with these data that represent the potential long-lived misfolded states PGK populates. This structural and energetic heterogeneity creates a hierarchy of refolding timescales, explaining stretched-exponential kinetics.

INTRODUCTION

A complete understanding of protein folding requires explaining the interrelationships between thermodynamics, kinetics, structural characteristics, and material properties at all relevant spatial, temporal, and energetic scales. For example, the folding of some small two-state proteins (Fig. 1A) (*1*, *2*) is qualitatively well-understood because (i) their thermodynamics and kinetics have been mathematically described (*1*, *2*), (ii) their transition states, free energies, and impact on kinetics have been detailed (*3*–*8*), (iii) atomic-scale models of folding have been presented with all-atom simulations (*9*–*13*), and (iv) their native structures can be accurately designed using machine learning (*14*, *15*).

Our understanding is incomplete, however, for proteins that do not fold in a two-state manner. An example is proteins that exhibit "stretched-exponential" folding kinetics (Fig. 1B) (*16*–*20*). Unlike two-state folding, where the survival probability of the nonnative state decreases exponentially over time as $\exp(-kt)$, these proteins show a decay following a stretched-exponential function, $\exp\left[-(kt)^{\beta}\right]$. While the thermodynamic and kinetic origins of stretched exponential folding kinetics are well understood, what remains unknown is the structural origin of this class of kinetics.

For two-state folding kinetics to arise, the rate of interconversion among nonnative conformations must be much faster than the kinetics of reaching the native state, providing a separation of timescales and allowing a "preequilibration" of the unfolded ensemble to be achieved (*2*). Structurally, the final transition (known as the transition path) to the native state must be fast, cooperative, and only populate intermediate states transiently (*2*). Under these conditions, although there are multiple pathways on the high-dimensional energy landscape, each with a different rate of folding, those rates sum to give an overall observed rate of the process across many copies of the protein that is accurately described by $\exp\left(-k_{\mathrm{F}}t\right)$ (Fig. 1A).

For stretched-exponential folding kinetics to arise, there must be large free energy barriers separating different regions of the nonnative state energy landscape such that the kinetics of interconversion between different nonnative states is slower than the interconversion of a given nonnative state to the native state (*20*). That is, the energy landscape partitions molecules into different folding channels that do not effectively interconvert with each other relative to the overall folding rate (Fig. 1B). This scenario prevents preequilibration of the nonnative ensemble of states, fundamentally changing the nature of the underlying kinetic mechanism such that the overall rate of folding is no longer a sum of microscopic rates, but instead is a sum (a convolution), of single exponential terms $\left[\sum_i A_i \exp\left(-k_i t\right)\right]$ (*16*, *21*). This sum can be approximated by a power law $\left(\exp\left[-\left(k_{\mathrm{F}}t\right)^{\beta}\right]\right)$ (*22*) characterized by just two terms, k_{F}, and the exponent β. When $\beta = 1$, the expression collapses back to two-state kinetics, while scenarios wherein $\beta < 1$ are described as stretched exponential kinetics.

The relationship is understood between stretched exponential kinetics and the thermodynamic scenarios on the free energy landscape that gives rise to them. However, the structural origins of these large energy barriers are unknown. It could be the case that for each protein, there is a unique structural origin. This is undoubtedly true at spatial resolutions on the scale of residues and individual secondary structural elements. However, as the history of two-state folding has taught us, at long enough scales, the finer details average out, and the entire classes of protein behavior can be described by widely applicable structural mechanisms. For example, the transition state ensemble characterized at the level of individual residues by Φ-value analysis can depend sensitively on the protein's primary

[1]Department of Chemistry, Pennsylvania State University, University Park, PA 16802, USA. [2]Department of Chemistry, Johns Hopkins University, Baltimore, MD 21218, USA. [3]Bioinformatics and Genomics Graduate Program, The Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA 16802, USA. [4]Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA. [5]Thomas C. Jenkins Department of Biophysics, Johns Hopkins University, Baltimore, MD 21218, USA. [6]Institute for Computational and Data Sciences, Pennsylvania State University, University Park, PA 16802, USA.
*Corresponding author. Email: sdfried@jhu.edu (S.D.F.); epo2@psu.edu (E.P.O.)
†These authors contributed equally to this work.

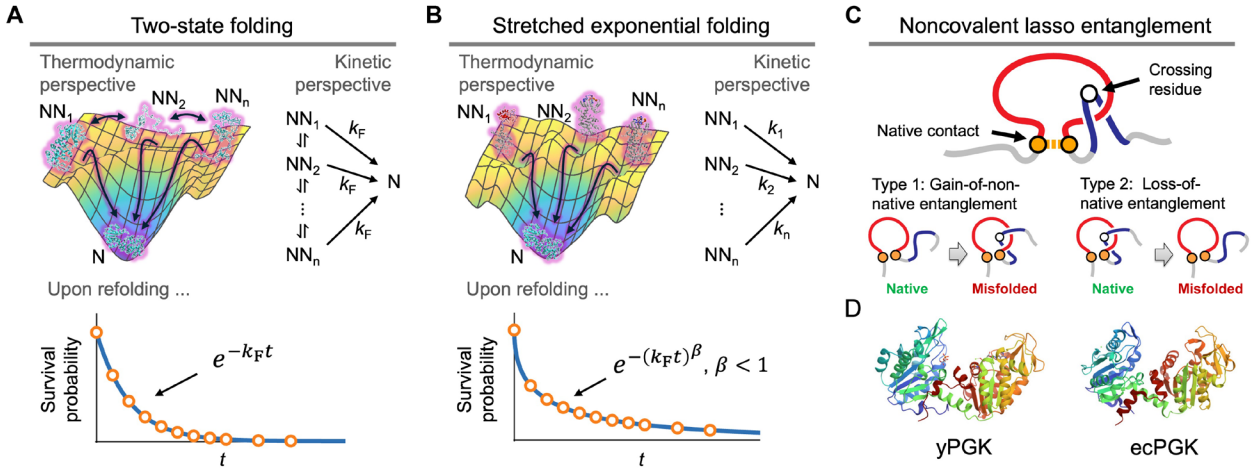

Fig. 1. Two protein folding scenarios and misfolding via noncovalent lasso entanglements. (**A**) Illustration of the two-state folding scenario. In the thermodynamic perspective, illustrated using a free energy surface, multiple nonnative states (NN_i) reside in a high-energy basin with no effective energy barriers between them, allowing rapid transitions among NNs. NNs fold into the native state (N), which is the global minimum, over a single energy barrier. In the kinetic perspective, fast transitions occur between NNs, folding into N with a single, overall rate constant k_F. Upon refolding, the time-dependent survival probability of NNs follows a single exponential function $\exp(-k_F t)$. (**B**) Illustration of the stretched exponential folding scenario. In the thermodynamic perspective, a free energy surface is colored from yellow to blue as the free energy decreases. Multiple NNs exist in distinct basins with large energy barriers between them, hindering fast interconversions. NNs fold into N via their own energy barriers. In the kinetic perspective, NNs fold into N with different rate constants k_i. Upon refolding, the survival probability of NNs follows a stretched exponential function $\exp\left[-(k_F t)^{\beta}\right]$ with $\beta < 1$. (**C**) Illustration of the components of a noncovalent lasso entanglement. The protein chain (gray) forms a native contact (orange) establishing a closed loop (red). A threading segment (blue) pierces this loop at the crossing residue (white). Two major types of misfolding include gain-of-nonnative entanglement and loss-of-native entanglement (*25*, *67*). (**D**) Crystal structures of homologous proteins PGK from *S. cerevisiae* [yPGK, Protein Data Bank (PDB) 1QPG] and *E. coli* (ecPGK, PDB 1ZMR).

structure, yet at larger scales, the thermodynamic and kinetic properties exhibit a simple two-state description (*4*, *6*, *7*). Therefore, we posit that for many proteins, there may be a general structural origin to stretched-exponential kinetics at sufficiently long length scales.

A further connection between stretched-exponential kinetics, intermolecular forces, and thermal energy was made when the Onuchic lab demonstrated that, within a lattice model description of a 27-residue protein, stretched exponential folding kinetics arose when the interaction energy between pairs of residues was much greater than the thermal energy (*23*). In this case, breaking the noncovalent interactions between pairs of residues caused a distribution of large free energy barriers and slow interconversion between nonnative states. This is likely to be an important component of a general explanation since enthalpy and entropy give rise to free energy barriers and are temperature dependent quantities. In this study, however, we aim to examine a potential structural basis beyond pairwise residue interaction strengths.

We hypothesize that a recently predicted class of protein misfolding (*24*, *25*) could contribute to the structural origin of stretched exponential folding kinetics because it has many relevant characteristics. This class of misfolding involves changes in a particular type of entanglement—noncovalent lassos (*26*), where the protein chain forms a loop closed by noncovalent interactions between residues and this loop is pierced one or more times by another segment on the protein chain. There are two types of misfolding involving these entanglements. Type 1 involves the formation, or gain, of a noncovalent lasso entanglement that is not present in the native structure. (From the nomenclature of Knot theory, this corresponds to a change of topology from $n = 0$, no threaded element, to $n = 1$, threaded. We emphasize, however, that noncovalent lasso entanglements are not mathematically nor structurally knots.) Type 2 involves the loss or failure to form of a noncovalent lasso that should be present in the native structure (Fig. 1C) (*25*), i.e., an $n = 1$ to $n = 0$ change in topology.

Several observations motivate this hypothesis. Previous simulations and experiments (*24*, *25*, *27*–*29*) indicate that these misfolded states can be long-lived kinetic traps (spanning short to very long timescales), exhibit large free energy barriers to unfolding, have populations that are sensitive to initial conditions—indicative of a lack of preequilibration—and importantly can be structurally heterogeneous for a given protein and therefore could give rise to a distribution of free energy barriers and a hierarchy of timescales that are necessary to exhibit stretched-exponential folding kinetics.

To test this hypothesis, we carried out simulations and experiments on *Saccharomyces cerevisiae* and *Escherichia coli* phosphoglycerate kinase (PGK; denoted as yPGK and ecPGK, respectively; see Fig. 1D), in which yPGK was previously shown to exhibit stretched exponential folding kinetics using two different experimental probes (*16*, *18*). Both PGK homologs exhibit native entanglements in their crystal structures (Table 1), suggesting the possibility that they could exhibit type 2 misfolding. We demonstrate that within our simulation model of PGK, stretched exponential folding kinetics arise from misfolding involving the formation of nonnative noncovalent lasso entanglements and the loss of native noncovalent lasso entanglements. We then conduct limited proteolysis and cross-linking mass spectrometry (XL-MS) and identify misfolded simulated structures that are consistent with these experimental data. Because such misfolding has been predicted to be widespread across the proteome of organisms (*24*, *26*), this structural mechanism may be relevant to a broad set of proteins.

RESULTS

Stretched exponential behavior occurs in simulations of *S. cerevisiae* PGK refolding

To test whether our simulation model recapitulates stretched exponential behavior observed in fluorescence resonance energy transfer (FRET) studies of refolded yPGK, we calculated the FRET efficiency time course from simulations of its refolding process upon temperature quench (see Eq. 3 and Materials and Methods). Fitting a stretched exponential function to this time course (see Eq. 6) yields a β value of 0.64 [95% confidence interval (CI) = [0.57, 0.73], $R^2 = 0.98$, P value < 0.0001, permutation test for β = 1; see Fig. 2A], indicating that the simulation model exhibits stretched exponential folding and yields a value similar to the in vitro experimental value of 0.59 ± 0.02 (*18*). (Note well that this quantitative agreement is almost certainly a coincidence, and what is most relevant for this study is that the β value is statistically less than 1, meaning that stretched-exponential behavior is occurring in the model.) In contrast, the fit quality gets worse (R^2 of 0.96) for a single exponential function that would describe a protein that folds in a two-state manner (see Eq. 5 and Fig. 2A). Thus, the simulation model of yPGK exhibits stretched exponential behavior.

Table 1. Representative native entanglements in the crystal structures of yPGK (PDB ID: 1QPG) and ecPGK (PDB ID: 1ZMR).

Native entanglements*		Closed loop†	Crossing residue‡	Linking number§
yPGK	1	50–160	20 (+)	+1
	2	57–107	118 (+)	+1
	3	221–335	208 (+)	+1
	4	254–305	239 (+)	+1
	5	298–316	284 (−)	−1
ecPGK	1	43–139	17 (+)	+1
	2	55–99	108 (+)	+1
	3	204–370	188 (+)	+1
	4	228–274	221 (+)	+1
	5	271–287	257 (−)	−1

*Representative native entanglements were obtained from our previous study (*26*). Numerous native entanglements were identified within the crystal structure and clustered on the basis of their positions along the protein sequence. Here, we provide the representative locations of these native entanglements reporting the minimal closed loop. The loss-of-native entanglements identified in the current study correspond to these representative native entanglements, with positions potentially shifting by several residues. †Native contact residues that form the minimally closed loop. The residue numbers are presented, using the numbering of the UniProt sequences (yPGK: P00560, ecPGK: P0A799). ‡Crossing residue number followed by its piercing chirality shown as +/− in parentheses. §Linking number defined as the sum of piercing chirality of crossing residues identified by Topoly (*65*).

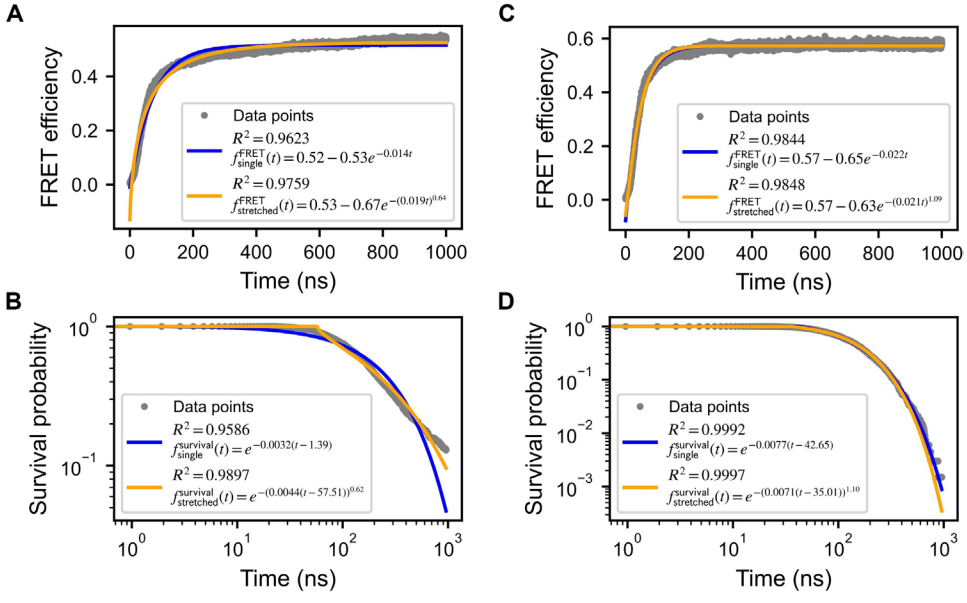

Fig. 2. Refolding kinetics in yPGK and ecPGK. (**A**) FRET efficiency fitting for yPGK during refolding. (**B**) Survival probability fitting for ecPGK during refolding. (**C**) FRET efficiency fitting for yPGK during refolding, excluding misfolded entangled states. (**D**) Survival probability fitting for ecPGK during refolding, excluding misfolded entangled states. Gray dots represent simulated FRET efficiencies in (A) and (C) and survival probability in (B) and (D). The fitted single exponential function (blue) and stretched exponential function (orange) are shown. The legend includes R^2 values and fitting functions. In (B) and (D), time and survival probability are on a logarithmic scale.

Stretched exponential behavior also occurs in simulations of *E. coli* PGK refolding

We next applied our simulation model to the homologous *E. coli* PGK protein, which we denote ecPGK. The time-dependent survival probability of the unfolded, nonnative population after a temperature quench was calculated (see Materials and Methods), and again a stretched exponential function was fit to the data (see Eq. 8). A β value of 0.62 (95% CI = [0.55, 0.64], $R^2 = 0.99$, P value < 0.0001, permutation test for $\beta = 1$; see Fig. 2B) was found. We again find a worse fit to a single exponential function ($R^2 = 0.96$; see Eq. 7 and Fig. 2B). We conclude that like yPGK, the refolding kinetics of ecPGK also displays a stretched exponential behavior in our model.

Misfolded entangled states give rise to stretched exponential behavior

As explained in Introduction, we hypothesize that a recently predicted form of misfolding might constitute the structural basis for this class of kinetics (Fig. 1). To test this, we first asked whether this type of misfolding occurred in our simulations. Using the Q (Eq. 1) and G (Eq. 2) order parameters, which are, respectively, the fraction of native contacts formed and a measure of the fraction of native contacts exhibiting a change in entanglement (see Materials and Methods), we find in our simulations a diverse set of misfolded states that exhibit changes in entanglements—indicated by the distinct subpopulations of nonzero G values in fig. S1.

Next, to test whether these misfolded entangled states are giving rise to stretched exponential behavior in yPGK and ecPGK, we removed those simulation trajectories where there are consecutively 1.5 ns of frames exhibiting consistent changes in entanglements. The G cutoff identifying these changes was chosen as the average upper boundary of the native basin in the Q-G space and had values of ≥0.03 for yPGK and ≥0.012 for ecPGK (fig. S1). We then fit the stretched exponential function (Eq. 6 for yPGK and Eq. 8 for ecPGK) on the refolding kinetic curve built on the remaining trajectories. For yPGK, we removed 546 trajectories of 1000 and found a β value of 1.09 (95% CI = [0.93, 1.27], $R^2 = 0.98$, P value = 0.21, permutation test for $\beta = 1$; see Fig. 2C), which is statistically indistinguishable from two-state single-exponential refolding kinetics (*16*). Similarly, for ecPGK, we removed 300 trajectories of 971 and got a value of 1.10 (95% CI = [1.00, 1.22], $R^2 = 1.00$, P value = 0.20, permutation test for $\beta = 1$; see Fig. 2D), again indistinguishable from two-state folding kinetics. These results demonstrate that the presence of these misfolded states results in stretched-exponential refolding kinetics, and their absence results in two-state behavior in our simulation model.

Two mass spectrometry techniques indicate that long-lived soluble, misfolded subpopulations exist in PGK

Our simulation results predict PGK populates misfolded subpopulations during refolding trajectories. To test this experimentally, we purified ecPGK, denatured it in 6 M guanidinium chloride (GdmCl), initiated refolding by 100-fold dilution to 0.06 M GdmCl, and interrogated the products of the refolding reactions with structural mass spectrometry techniques. These experiments used chemical denaturation (rather than thermal) to avoid aggregation. Previous work has suggested that ecPGK does not efficiently refold in the context of whole *E. coli* extracts (*30*, *31*); to gain more detailed information about the misfolded conformations populated during refolding, we used purified ecPGK in these assays and assessed the refolded ensembles with limited proteolysis mass spectrometry (LiP-MS) and XL-MS. ecPGK was allowed to refold for 1 hour after diluting GdmCl to native levels, and protein conformations were then assessed by pulse proteolysis for 1 min with proteinase K (PK; for LiP-MS) or chemical cross-linking with disuccinimidyl dibutyric urea (DSBU) for XL-MS.

These data allow us to identify structural changes in various portions of PGK relative to the native conformational ensemble, as measured by changes in protease susceptibility as well as changes in cross-linking propensity (XP) between pairs of residues (see cut-sites and cross-linked pairs data in dataset S1). LiP-MS identified significant changes at 42 PK cut-sites (see dark blue data points in Fig. 3A). XL-MS identified 16 unique residue pairs that exhibit a significant change in cross-linking (see dark blue data points in Fig. 3B, pairs that have identical residues are excluded as this can only occur in oligomers). These structural changes in the refolded ecPGK sample indicate that PGK populates long-lived, soluble misfolded states that persist for an hour.

Near-native misfolded states containing entanglement changes are long-lived states

Using a Markov state modeling approach (see Materials and Methods), we identified the native state [S10 in Fig. 4 (A and B)] and nine other metastable states [refer to Fig. 4 (A and B)] in the ecPGK refolding trajectories. Notably, states S2, S3, S4, S7, S8, and S9 exhibited G values ≥ 0.02, categorizing them as misfolded states involving a change in entanglement. In addition, states S6, S7, S8, and S9 displayed high Q values (≥0.60), indicating that structurally these states have a majority of their native state structure formed, which has the potential to lead to long-lived kinetically trapped states (*24*, *25*, *27*, *28*). Motivated by the long-lived changes in structure seen in the LiP- and XL-MS data on chemically refolded ecPGK, we first asked which of these metastable states are relatively long lived. To answer this question, we calculated the time it takes for structures in each of these metastable states to navigate back to the native state, as measured by its mean first passage time (MFPT; see Materials and Methods). Specifically, we simulated 100 trajectories initiated from randomly selected structures from each metastable state and measured its survival probability for 1 ms. The MFPTs were

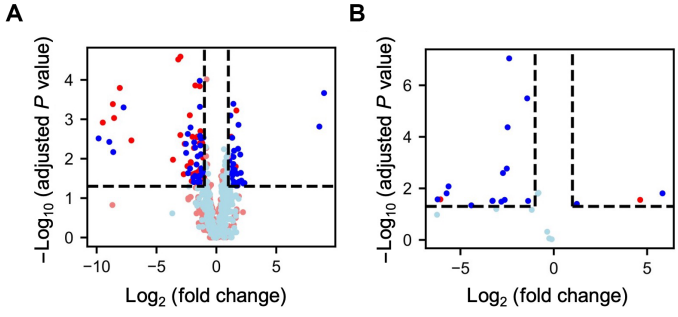

Fig. 3. Cut-sites and cross-linked residue pairs with significant changes between refolded (1 hour) and native ecPGK. (**A**) Cut-sites identified and quantified by LiP-MS, shown as abundance ratios between refolded ecPGK to native PGK. Full tryptic peptide cut-sites are in red, and half tryptic cut-site peptides are in blue. (**B**) Cross-linked residue pairs identified by XL-MS. Pairs with identical residues are in red, other pairs are in blue. Nonsignificant data points are light-colored; significant data points are dark-colored. Significance thresholds are marked by dashed lines (>2-fold change in abundance and adjusted P value < 0.05). Adjusted P values were computed using the Benjamini-Hochberg method (*64*).

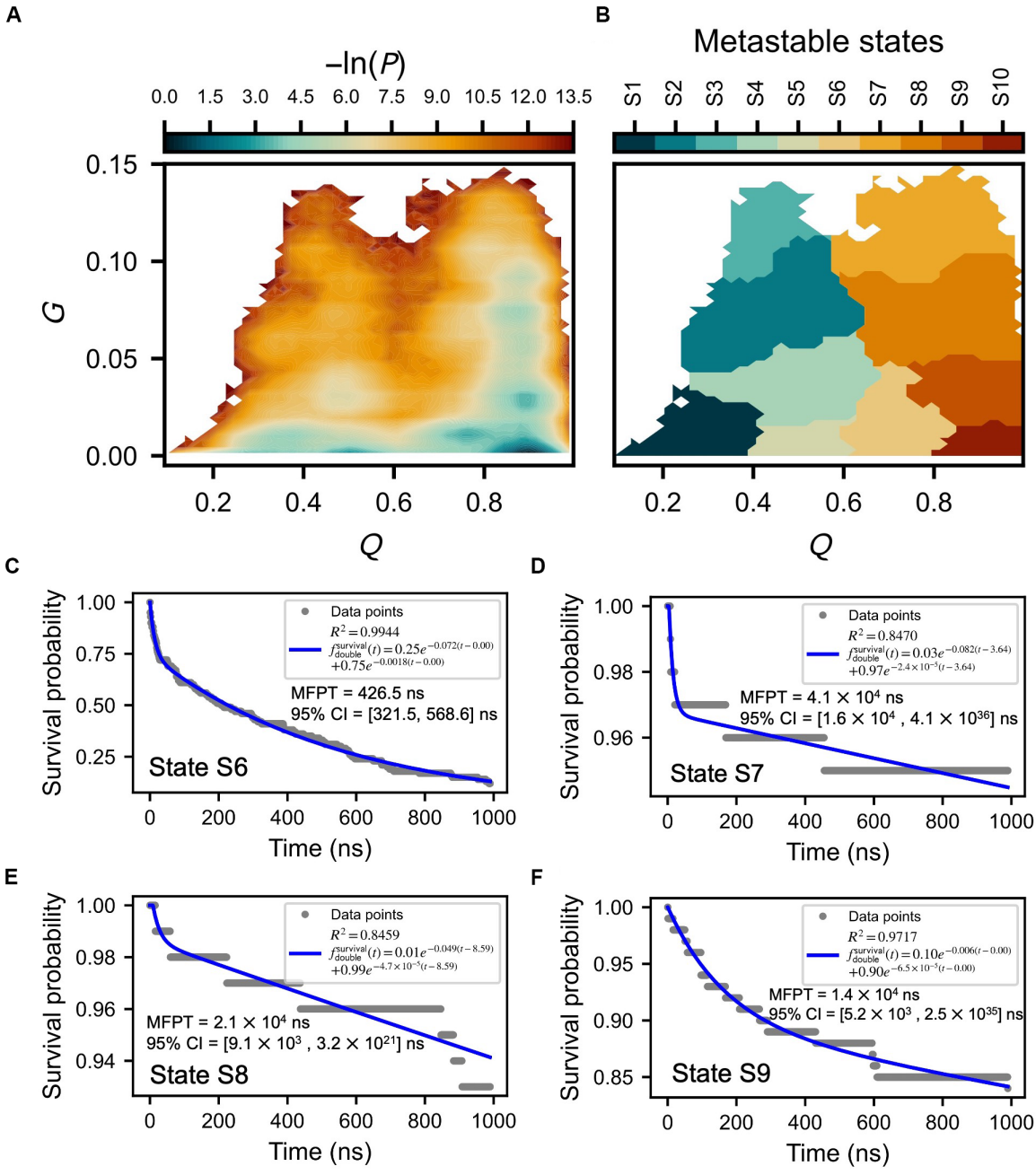

Fig. 4. Metastable states during the refolding process of ecPGK and the MFPTs for near-native states S6, S7, S8, and S9 toward the native state S10. (**A**) Log probability surface [−ln(*P*), where *P* represents the probability of sampling specific *Q* and *G* values] illustrating the refolding dynamics of ecPGK across order parameters *Q* and *G*. The color gradient reflects the probability distribution, transitioning from high (red) to low (green) free energy states. (**B**) Identified regions on the log probability surface corresponding to different metastable states, with the native state S10 located in the bottom right corner. (**C** to **F**) MFPTs for states S6, S7, S8, and S9, respectively, toward the native state S10, determined by fitting the survival probabilities (gray dots) using a double exponential function (blue curve). Each panel legend includes R^2 values and the fitting functions. Estimated MFPTs and 95% CIs (bootstrapping 10,000 times) are provided within the panels.

then determined by fitting this survival probability time course to a double exponential function [see Fig. 4 (C to F)]. We find that the near-native nonentangled state S6 exhibits a MFPT of approximately 400 ns, while the MFPTs of the near-native entangled states S7, S8, and S9 ranged from 1×10^4 to 4×10^4 ns, approximately 25- to 100-fold longer. We conclude that within our model, near-native states involving a change of entanglement are more likely to persist for relatively long timescales, and thus it is states S7, S8, and S9 that are most likely to contribute to the observed LiP-MS and XL-MS signals.

Metastable states that are consistent with the mass spectrometry data

Next, we asked whether long-lived metastable states S7, S8, and S9 exhibited structural changes that are consistent with the LiP- and

XL-MS data. To address this, we examine whether solvent-accessible surface area (SASA) differences between these states and the native ensemble (state S10) are statistically different in the regions where the PK cut-sites are located, and whether the XP (Eq. 11) differences between these states and the native state (see Materials and Methods) are statistically different for the pairs of residues in which cross-linking changes were experimentally observed. Of the 42 observed PK cut-sites in the LiP-MS data, 28 of them are consistent with the structural changes seen in one or more of these metastable states (corresponding to blue blocks in Fig. 5A and bolded values in dataset S2). For example, the residue D34 exhibits a change in protease susceptibility upon refolding, and we find a significant change in SASA around D34 in the simulation structures composing states S7 and S8 compared to the native state. Likewise, N219, which also exhibits a change in protease susceptibility, only exhibits a significant change in SASA in state S8 relative to the native state. Of the 16 observed cross-linking pairs in the XL-MS, 12 of them are consistent with the structural changes observed in one or more of these metastable states (see blue blocks in Fig. 5B and bolded values in dataset S2). Thus, of the 58 experimentally measured changes across LiP- and XL-MS, 40 (69%) can be explained by a combination of these three metastable states. We therefore conclude that the experimental signals arise from a heterogeneous ensemble of misfolded states. With the caveat that no single metastable state explains all the experimental signals, we asked which metastable state exhibits the greatest consistency with the experimental signals. We find that state S8 is the most consistent, with 33 consistent signals, compared to S7 and S9, each with 29 consistent signals (Fig. 5).

Next, we investigated whether the inconsistent, unexplained observations cluster along the primary structure. We found that our simulation structures are highly consistent with the LiP-MS PK cut-sites in the N-terminal domain of the protein (residues 1 to 164 and 374 to 387), with all 22 N-terminal cut-sites consistent with the structural changes in S7, S8, and S9. In contrast, only 6 of 20 PK cut-sites located toward the C terminus (residues 165 to 373) are consistent (see Fig. 5A). Three hypotheses could explain this lack of consistency: (i) The population of C-terminal domain misfolding is smaller than N-terminal domain misfolding in the simulations, resulting in a smaller sample size and decrease in statistical power to detect significant SASA changes; (ii) the misfolding that occurs toward the C terminus is highly native-like with subtle SASA changes resulting in a smaller effect size and decrease in statistical power; or (iii) any misfolding that occurs in the C-terminal domain in our simulations is located at positions different from where PK cut-sites are observed.

To test the first hypothesis, we examined the last 10 ns of simulation structures in states S7, S8, and S9 and found that 11,761 exhibit entanglement changes in the N-terminal domain, whereas only 3179 show changes in the C-terminal domain. This nearly fourfold difference in population is consistent with our first hypothesis. To test the second hypothesis, we calculated the effect size of SASA changes at the N-terminal PK cut-sites in the N-terminal misfolded structures and found it to be greater than that of the C-terminal cut-sites in the C-terminal misfolded structures (see fig. S2A). This indicates that misfolding in the C-terminal portion of PGK is more native-like, making SASA changes harder to detect in our simulations, consistent with our second hypothesis. To test the third hypothesis, we calculated the probability of entanglement changes seen in the simulations co-occurring at the observed PK cut-sites. We find that, depending on the metastable state, misfolding in the C-terminal domain does co-occur with most PK cut-sites (see fig. S2B), particularly in the regions 286 to 312. This is inconsistent with our third hypothesis, and we reject it. We conclude that the likely origin of the lack of consistency at the C-terminal PK cut-sites is due to limited statistical power arising from a smaller sample size and smaller effect size in the simulation model.

Representative structures that are maximally consistent with the experimental data

To identify representative simulation structures that are consistent with the structural mass spectrometry, we carried out a multistep process. We grouped the misfolded structures based on their entanglement

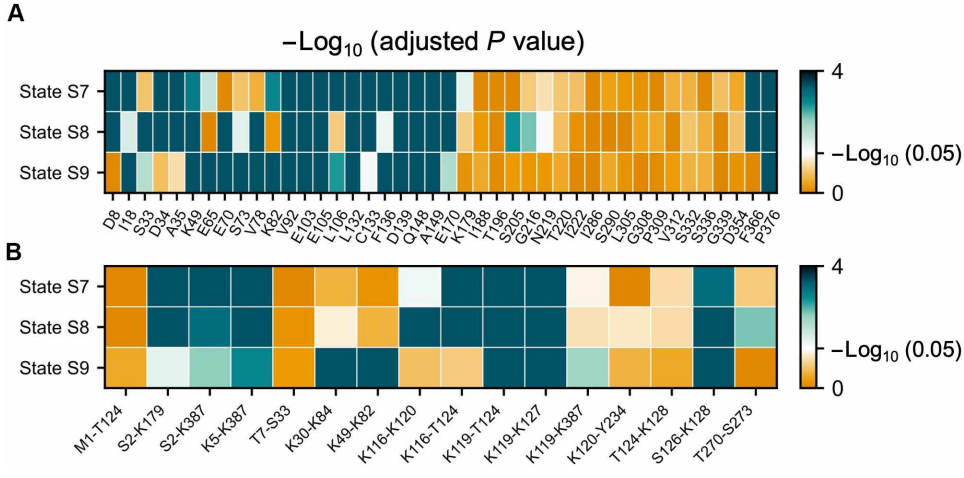

Fig. 5. Simulation structures exhibit structural changes consistent with structure mapping experiments. (**A**) $-\text{Log}_{10}$ (adjusted *P* values) for the statistical test of the null hypothesis that the mean protease susceptibility (measured as SASA) of the PK cut-site region (±5 residues) in the long-lived misfolded states S7, S8, and S9 is equal to that in the native state S10. (**B**) $-\text{Log}_{10}$ (adjusted *P* values) for the statistical test of the null hypothesis that the mean XP (using Eq. 11) of residue pairs in the long-lived misfolded states S7, S8, and S9 is equal to that in the native state S10. Adjusted *P* values are computed using the Benjamini-Hochberg method (*64*) for pooled *P* values of both tests. Values are colored from orange (adjusted *P* value = 1) to dark blue (adjusted *P* value = 0.0001), with white at 0.05. Values less than 0.05 are considered statistically significant.

changes and their consistency with the experimental data (see Materials and Methods, 91 clusters of entanglement changes were obtained and reported in dataset S3), resulting in 545, 2860, and 4799 unique groups, respectively, for states S7, S8, and S9. Each group has structures involving a unique set of entanglement changes and experimental signals. Next, for each group, we identified the structures that had the greatest population in the simulations. This resulted in a representative misfolded structural ensemble for each near-native misfolded state. Last, we selected a single representative structure from each representative ensemble that maximizes the consistency of the change in structure with the greatest number of experimental signals. The representative structures and representative misfolded structural ensembles can be interactively visualized and downloaded from https://obrien-lab.github.io/Misfolded-ecPGK-structural-ensemble-consistent-with-LiP-MS-and-XL-MS-experiments/ (summarized in dataset S4).

As a case study, we focus on the representative structure from misfolded state S7 (Fig. 6A) that is consistent with the highest number of experimental signals, totaling 22. Despite having extensive native structure, with 75% of native contacts formed ($Q = 0.75$), this misfolded structure is a kinetic trap marked by changes in noncovalent lasso entanglements ($G = 0.11$). We identified five clusters of entanglement changes within this structure: Three involve a gain of nonnative entanglements (Fig. 6, B to D), one involves a loss of native entanglement (Fig. 6E), and one involves only a switch in piercing chirality without altering the linking number, which we call a chirality-switch entanglement. In instances of gain-of-nonnative entanglements, segments of the protein chain nonnatively wrap around others, impeding proper folding. For example, as depicted in Fig. 6B (gain-of-nonnative entanglement, cluster ID 14), the closed loop spanning residues 95 to 109, previously unpierced in the crystal structure, wraps around residue 158 during misfolding, forming an entanglement that largely displaces it from its native position and leads to the loss of some native contacts. This consequently destabilizes the surrounding region while simultaneously creating a large transition state barrier to correct this misfolding requiring the unfolding of surrounding portions of the protein. Similar scenarios are observed in two other gain-of-nonnative entanglements where nearby residues get wrapped around by two different loops (Fig. 6, C and D). The sole loss-of-native entanglement identified in this structure (Fig. 6E) occurs concomitant with the gain of a nonnative entanglement in cluster ID 14—as the loop 95 to 109 wraps around residue 158, introducing a second but opposite piercing event to the native entanglement #2 in ecPGK (see Table 1), which cancels out the entanglement. The chirality-switch entanglement, ID 2 (Fig. 6F), is another consequence of the gain-of-nonnative entanglement ID 14. The native entanglement at the loops 48 to 104 is shallow due to the nearly coplanar β sheets at the crossing region. A minor structural change led to the loss of piercing at residue 108. The new piercing at residue 158 in the opposite direction resulted in a chirality-switch entanglement. This entanglement change is relatively trivial compared to the others mentioned above.

In this representative structure, a number of conformational changes are consistent with the measured changes from mass spectrometry. These changes include alterations in entanglement, resulting in shifts in SASA around PK cut-site residues D8, I18, D34, A35, V92, L106, L132, C133, F136, D139, Q148, A149, E170, F366, and P376 (highlighted as exposed cyan surfaces in Fig. 6G). In addition, changes in solvent-accessible surface distance (SASD) at residue pairs S2-K179, S2-K387, K5-K387, K116-K120, K116-T124, K119-T124, and K119-K127 (represented by magenta curves in Fig. 6H) are consistent with changes in XP. Entanglement changes with cluster IDs 14, 16, 24, and 44 sequestered the linker region and altered the conformation near the entangled regions, reducing the exposure of residues V92 and L106 and changing the SASDs of K116-K120, K116-T124, K119-T124, and K119-K127. This sequestered linker resulted in a flip in the orientation of the N-terminal domain and spatial separation of the domains, increasing the exposure of D8, I18, D34, A35, L132, C133, F136, D139, Q148, A149, E170, F366, and P376 and markedly increasing the SASDs of S2-K179, S2-K387, and K5-K387 to greater than 50 Å, making cross-linking impossible (XP = 0). These shifts in SASA and SASD can explain the experimentally observed changes in protease susceptibility and/or XP in these regions. These changes are statistically distinguishable from the native state ensemble S10 (see Fig. 6I). As an internal control, we note, as expected, that the crystal structure falls within the 95% CI of the native ensemble's structural properties generated from our simulation model, supporting the model's realism.

DISCUSSION

Our results provide a structural explanation for the origin of stretched exponential folding kinetics. PGK's power law distribution of folding kinetics, according to our model, arises from a diverse array of misfolded states that involve changes of entanglement status. These states either gain a nonnative entanglement or fail to form a native entanglement. The simulation model successfully reproduced the stretched exponential behavior observed in FRET dye–labeled yPGK (Fig. 2A) and predicted similar behavior in the refolding of ecPGK (Fig. 2B). Specifically, the β values (~0.6) less than 1.0 in the time-dependent FRET efficiency and survival probability (Eqs. 6 and 8) for both proteins are the hallmark of stretched exponential folding kinetics. Upon excluding simulation trajectories that populated these misfolded entangled states, the refolding kinetics exhibited two-state folding kinetics (Fig. 3), with a β value statistically no different than 1.0 (P values > 0.05). This clearly demonstrates that within our simulation model, these misfolded states give rise to stretched kinetics. We pinpointed, using LiP- and XL-MS, multiple positions along the primary structure of ecPGK that exhibit significant structural changes compared to the native state 1 hour after folding conditions were reestablished, providing experimental evidence that the population of soluble, misfolded subpopulations persist for long timescales. Applying rigorous statistical tests to our simulation structures, we identified misfolded structural ensembles that are consistent with the LiP- and XL-MS data (representative structures are exemplified by Fig. 6, and the structural ensemble can be interactively visualized at https://obrien-lab.github.io/Misfolded-ecPGK-structural-ensemble-consistent-with-LiP-MS-and-XL-MS-experiments/). The resulting ensembles indicate that these misfolded states are heterogeneous with changes of entanglement observed in multiple regions of PGK. Tellingly, no single state is consistent with all the experimental signals, indicating that PGK has a heterogeneous ensemble of misfolded structures.

The molecular origin of stretched exponential kinetics lies in the heterogeneity within the nonnative protein ensemble. The coexistence of multiple nonnative states that populate parallel folding pathways toward the native state, but do not interconvert with each other on fast timescales, introduces a spectrum of characteristic folding rates, giving rise to a power-law distribution in the ensemble folding kinetics. Hypothesized factors contributing to this

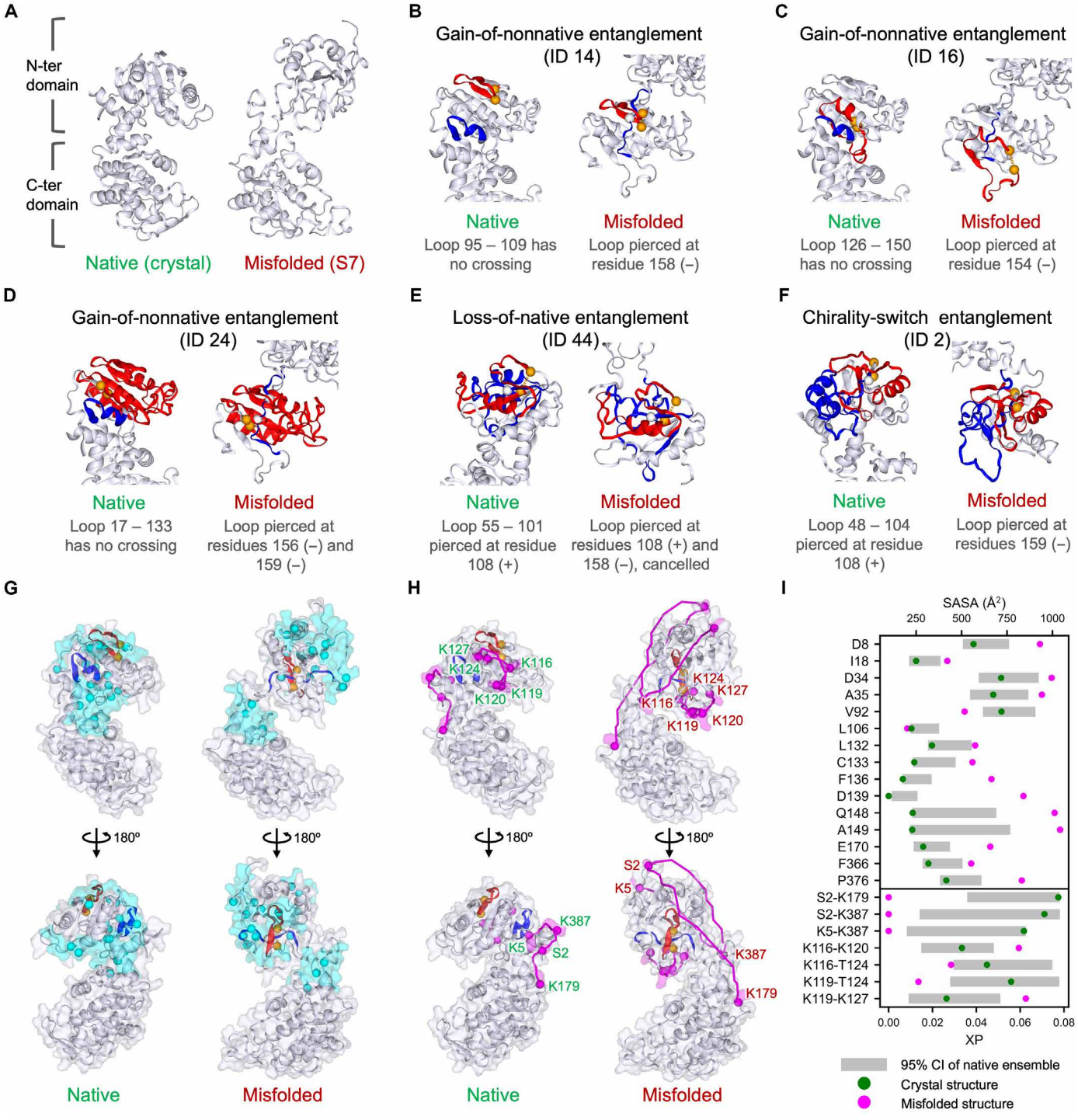

Fig. 6. Representative structure of misfolded state S7 generating consistent experimental signals. (**A**) Comparison between the native state (crystal structure, PDB 1ZMR) and misfolded state S7; (**B**) gain-of-nonnative entanglement ID 14; (**C**) gain-of-nonnative entanglement ID 16; (**D**) gain-of-nonnative entanglement ID 24; (**E**) loss-of-native entanglement ID 44; (**F**) chirality-switch entanglement ID 2; (**G**) consistent LiP-MS signals highlighted on the native structure (left) and misfolded structure (right). Entanglement changes ID 14 is highlighted. Protein backbone and surface are depicted in white, while PK cut-sites are represented as cyan balls. Residues (±5) around the cut-sites are transparent cyan surfaces; (**H**) consistent XL-MS signals highlighted on the native structure (left) and misfolded structure (right). Entanglement change ID 14 is highlighted. Protein backbone and surface are shown in white, residue pairs as magenta balls with transparent magenta surface, and SASDs identified by Jwalk as magenta curves; (**I**) SASA and XP values of crystal structure (green dots) and misfolded structure (magenta dots), alongside 95% CI of the distribution in native state ensemble S10 (gray bars), for all consistent experimental signals. SASAs were computed using the ±5 residues around the cut-sites. Throughout (B) to (H), closed loops are depicted in red, threading segments in blue, and native contacts as orange dashed lines with residues represented by orange balls. The native and misfolded structures are aligned on the N-terminal domain in (B) to (F) and on the C-terminal domain in (G) and (H).

heterogeneity (*20*), such as misligated ligands, proline isomerization, and intermolecular oligomeric interactions, were ruled out in earlier experimental studies of PGK (*16*, *17*), leaving intramolecular interactions and conformations as the origin of this phenomenon.

Our simulations indicate that the structural origin of the large free energy barriers separating these diverse misfolded states arises from the formation of native structure around the misfolded entangled regions. During refolding from a denatured state, changes in noncovalent lasso entanglements lead some PGK molecules to be caught in long-lived kinetic traps with significantly slower folding rates compared to near-native, non-entangled states. These kinetic traps persist because large portions of the primary structure are properly folded and stable (gray, folded portions in Fig. 6), requiring them to unfold before the native state can be reached. This thermally activated process gives rise to the large free energy barriers between different misfolded states and the distinct channels on the energy landscape they fold through. Such "backtracking" (*32*) from more-ordered to less-ordered structure to allow proper folding has been seen in a variety of contexts in simulations and experiments (*33*–*39*). This study provides a new perspective on the likely nature of these nonnative states and why they exhibit metastability on a wide range of timescales.

We note that chaperones can facilitate folding by mediating such backtracking (termed "unfoldase activity") (*40*, *41*), although they must first identify a misfolded epitope, such as exposed hydrophobic surfaces. Since misfolded states with changes in entanglement do not always have this property because of how structurally native-like they can be, some of them have the potential to bypass cellular proteostasis pathways (*24*, *27*, *28*). This is consistent with the previous finding that ecPGK does not efficiently refold over hour timescales, even in the presence of the chaperones DnaK and GroEL (*31*).

The entangled kinetic traps observed in PGK are not unique to this protein. Our previous simulation studies indicate that approximately one-third of the *E. coli* cytosolic proteome can populate similar entangled kinetic traps during refolding and/or cotranslational folding (*24*). The likelihood of encountering these traps appears to correlate with protein size, as larger proteins with more complex native structural topologies are more prone to becoming kinetically trapped (*25*). An interesting area for future research would be to understand whether there are any sequence biases that might predispose subdomains to populate these entangled kinetic traps.

The long-lived kinetic traps identified in this study align with the kinetic partitioning mechanism (KPM) identified decades ago (*42*). KPM describes the slow folding of biomolecules through indirect off-pathway processes, exhibiting three-stage multipathway kinetics on a rugged free energy surface caused by topological frustration. The three stages include nonspecific collapse into a relatively compact phase, kinetic ordering into near-native low-energy misfolded structures, and final transition to the native state. This study adds details to the KPM model and demonstrates that stretched exponential folding kinetics can also arise from later stages of folding. First, our study has identified a role for entanglements as a structural source of topological frustration that contributes to the second and third stages of the KPM. Second, while stretched exponential kinetics can arise in the first stage of the KPM (*42*), we have found that such kinetics can also arise at later stages of the folding process. This can result in slow folding kinetics on timescales of seconds, hours, or longer (*28*).

In this study, we used a rigorous statistical analysis to assess the consistency between mass spectrometry experimental signals and our predicted misfolded states. Unlike our previous studies (*24*, *27*–*29*), we adopted a two-tailed statistical test rather than a one-tailed test, which at a practical level means we only considered whether a structural change was significant, not the direction of the change (for example, a residue becoming less exposed or more exposed). This decision was made because of our growing appreciation of the complex nature of molecular scenarios that can give rise to the experimental signals, rendering the interpretation of directional changes in protease susceptibility or XP ambiguous. As an illustration, consider the LiP-MS results indicating significantly reduced abundances in the refolded sample relative to the untreated sample [that is, the negative $\log_2$ (fold change) values in Fig. 3 and dataset S1]. One interpretation is that the reduced peptide abundance in the refolded sample could be that it arises during refolding from a decrease in protease susceptibility due to reduced solvent exposure in this region (resulting in a lower SASA) preventing the protease for accessing and cleaving these segments. However, an alternative scenario that cannot be ruled out is that, upon refolding, the relevant protein segment was highly unstructured (resulting in a higher SASA) making it easy for PK to extensively cleave and digest this segment leaving very little to be detected by the mass spectrometry instrument, leading to lower peptide abundances. Thus, the same differential experimental signal can be interpreted in contradictory ways. Similarly, in XL-MS, structurally interpreting increases or decreases in XP is hampered by the potential for intermolecular cross-linking and the influence of one cross-link on another when a residue is shared between them. Thus, we chose to only interpret significant changes in the LiP- and XL-MS data, not the direction of the change in peptide abundances between untreated to treated samples.

There are limitations to our study. The structure-based coarse-grained force field we used is biased toward stabilizing native contacts. While more heterogeneous ensembles might be achieved by allowing stabilizing nonnative interactions, such heterogeneity could introduce even more timescales that are likely to reinforce stretched-exponential kinetic behavior. In a recent study, it was shown that entanglement-based misfolded states are populated in aqueous, all-atom, physics-based simulations of protein folding (*25*), and they can persist for hours to days in all-atom simulations (*25*, *27*). Another limitation is the accelerated timescales inherent to the coarse-grained models and low-friction Langevin dynamics we used in this study, which makes a comparison of simulation and experimental timescales challenging (*27*). This acceleration can explain why PGK's simulated folding timescale is on the order of microseconds, while LiP- and XL-MS indicate subpopulations that remain nonnative for at least 1 hour.

More broadly, there is a small but growing body of evidence for the existence of this class of misfolding. Alteration of protein biochemical properties, susceptibility to limited proteolysis, and differential client protein interactions with chaperones can be explained or, in some cases, predicted due to such misfolding (*24*, *25*, *30*, *31*). Ultimately, however, obtaining high-resolution structures will be the most convincing evidence. A major challenge to this is the aforementioned structural heterogeneity of these misfolded states. Therefore, single-particle cryo–electron microscopy, with its ability to cluster subpopulations, is perhaps the most promising technique to achieve this.

To conclude, we have identified a general structural mechanism that can give rise to stretched-exponential kinetics. We hypothesize that this recently identified class of protein misfolding could contribute to our

understanding of other aspects of protein folding, such as the possibility that they are relevant to the molten globule intermediate state observed during the folding of most globular proteins and studied extensively in the 1990s.

MATERIALS AND METHODS

Temperature quenching simulations

To simulate the protein unfolding and refolding process in the temperature jump experiment, we performed temperature quenching simulations using a Gō-based coarse-grained model (*24*, *27*, *43*–*48*). This coarse-grained model represents each amino acid residue as a single interaction site centered on the Cα position and uses a structure-based potential energy function (*27*). The force field parameters were tuned to reproduce the structural stability for a given protein (*27*). Specifically, for yPGK and ecPGK, we iteratively evaluate the structural stability by running 10 parallel 1-μs molecular dynamic simulations starting from the crystal structures [Protein Data Bank (PDB) 1QPG for yPGK and PDB 1ZMR for ecPGK] at 303 and 310 K, respectively, with different sets of parameters obtained from the previous work (*27*). The minimum values of the parameters that can maintain more than 68% of the native contacts over 98% of the simulation time were chosen to parameterize the proteins (*27*). The force field parameters can be found in table S1.

In the temperature quenching simulations, the crystal structure was initially unfolded at 800 K for 60 ns, followed by a refolding simulation at 295 K for yPGK and 310 K for ecPGK, each lasting 1 μs. To improve statistical significance, we conducted 1000 independent simulations for each protein simultaneously, using different random seeds. All simulations were performed using Langevin dynamics with a collision frequency of 0.05 ps^{-1} and a time step of 15 fs, implemented with OpenMM (*49*).

Metastable states clustering

To determine the entangled states, we used previously proposed order parameters, Q and G, to characterize the folding status of the protein (*27*). The calculation of Q, which represents the fraction of native contacts, is defined as follows

$$Q = \frac{\sum_{i\in I}\sum_{j\in J}\Theta\left(i,j\,|\,\text{Current}\right)}{\sum_{i\in I}\sum_{j\in J}\Theta\left(i,j\,|\,\text{Native}\right)} \tag{1}$$

where i and j are the residue indices and satisfy $j > i + 3$; I and J are both the set of residues within secondary structure elements (α-helical or β strands); $\Theta\left(i,j\,|\,\text{Current}\right)$ and $\Theta\left(i,j\,|\,\text{Native}\right)$ are step functions that equal 1 when residue i and j have native contact and 0 when i and j do not have native contact in the current structure and native structure, respectively. Native contacts are considered formed when the distance between the Cα atoms of residues i and j does not exceed 1.2 times their native distance and the native distance does not exceed 8 Å.

G serves as an order parameter that quantifies the change in entanglement compared to the native structure and is calculated as

$$G = \frac{1}{N}\sum_{(i,j)}\Theta\left[\left(\mathrm{i},\mathrm{j}\right)\in \mathrm{nc}\cap \mathrm{g}\left(\mathrm{i},\mathrm{j}\right)\neq \mathrm{g}^{\mathrm{native}}\left(\mathrm{i},\mathrm{j}\right)\right] \tag{2}$$

where (i, j) is one of the native contacts in the native crystal structure; nc is the set of native contacts formed in the current structure; $g\left(i,j\right)$ and $g^{\text{native}}\left(i,j\right)$ are, respectively, the total linking number of the native contact (i, j) in the current and native structures estimated using a discrete version of Gauss double integration (*27*); N is the total number of native contacts within the native structure; and the selection function Θ equals 1 when the condition is true and 0 when it is false.

Q and G were calculated for each frame in the refolding trajectories. Subsequently, the trajectories for each protein were projected onto the Q versus G space, and we applied the k-means algorithm (*50*) to group them into 400 clusters (microstates). A Markov state model (MSM) was constructed, and the clusters were further coarsened into a reduced number of metastable states using the PCCA+ algorithm (*51*). All clustering and MSM construction processes were performed using the PyEmma package (*52*).

We excluded 29 trajectories from subsequent analyses of ecPGK due to the presence of a misfolded structure characterized by reversed chirality in the secondary structure packing within the N-terminal domain [also known as "mirror image" conformation (*53*, *54*), depicted in fig. S3 and details in Supplementary Methods]. Given that the N-terminal domain of ecPGK is asymmetric, unlike the mirror image phenomenon observed in small symmetric proteins (*53*, *54*), this occurrence is likely an artifact resulting from the simplified coarse-grained force field (*55*).

FRET efficiency fitting

The FRET efficiency (E), which was evaluated in the reference experiment (*18*), was computed using the following function

$$E(t) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+\left[\frac{R_i(t)}{R_0}\right]^6} \tag{3}$$

where t is the refolding time; N is the number of trajectories; $R_i(t)$ is the donor-acceptor distance in the ith trajectory at time t, and R_0 is the Förster distance. According to the reference experiment (*18*), the PGK was labeled with AcGFP1 at N terminus and mCherry at C terminus. Therefore, the Förster distance R_0 was set to 50 Å, which is on the average magnitude of the Förster distance between widely used FRET pairs that use mCherry as acceptor (*56*). The donor-acceptor distance was estimated as the sum of protein's end-to-end distance $d_i(t)$ and the radius-of-gyration of AcGFP1 and mCherry, i.e.

$$R_i(t) = d_i(t) + R_{\text{AcGFP1}} + R_{\text{mCherry}} \tag{4}$$

where $R_{\text{AcGFP1}} = 15.5$ Å (*57*) and $R_{\text{mCherry}} = 15.6$ Å (*58*).

To examine the nonexponential behavior, the time course of the FRET efficiency was fitted using both a single exponential function and a stretched exponential function. The single exponential function is given by

$$f_{\text{single}}^{\text{FRET}}(t) = y_0 + Ae^{-kt} \tag{5}$$

and the stretched exponential function is expressed as

$$f_{\text{stretched}}^{\text{FRET}}(t) = y_0 + Ae^{-(kt)^{\beta}} \tag{6}$$

Here, y_0, A, k, and β are fitting parameters. We constrained $y_0 \geq 0$, $A \leq 0$, $k \geq 0$ and $\beta \geq 0$ during the least-square curve fitting process. The initial values $y_0 = 1$, $A = -1$, $k = 0.1$, and $\beta = 0.1$ were used in the fitting.

The 95% CIs for the estimated β values were calculated through bootstrapping the FRET efficiency time course 10,000 times. In each iteration, the time series of the end-to-end distance was resampled with replacement, followed by computing the FRET efficiency using Eq. 3 and fitting the stretched exponential function in Eq. 6 to the resampled data. The *P* value was computed using a permutation test to assess the null hypothesis that the β value equals 1. To construct a simulated scenario under this null hypothesis, we first generated a synthetic time course of FRET efficiency by fitting a single exponential curve (Eq. 5) to the observed data. Subsequently, we derived a simulated trajectory of end-to-end distance based on this synthetic FRET efficiency and replicated this trajectory *N* times.

To assess the null hypothesis that the β value equals 1, the *P* value was computed using the permutation test described as follows: (i) We pooled the simulated trajectories and the actual simulation data, creating a combined dataset of 2 *N* observations. (ii) Through 10,000 iterations, we randomly resampled this combined dataset without replacement, effectively permuting the indices. (iii) Within each permutation, we evenly partitioned the data into two sets. (iv) For each partition, we calculated the absolute difference in β values obtained from fitting the two subsets using Eq. 6. (v) The *P* value was then estimated as the frequency of observing a difference in the resampled data equal to or greater than the one obtained in the actual data.

Survival probability fitting

The survival probability of nonnative ecPGK versus the refolding simulation time was obtained by computing the fraction of trajectories in which the structure at a given time frame has not yet reached the native state (i.e., the first-passage time of this trajectory is longer than the given time frame). Here, reaching the native state was identified if the protein is continuously assigned as the native state by the metastable state analysis for 1.5 ns.

The time course of the survival probability was fitted using both a single exponential function and a stretched exponential function with a lag time t_0. The single exponential function is given by

$$f_{\text{single}}^{\text{survival}}(t) = \begin{cases} 1, & 0 \le t < t_0 \\ e^{-k(t-t_0)}, & t \ge t_0 \end{cases} \tag{7}$$

and the stretched exponential function is expressed as

$$f_{\text{stretched}}^{\text{survival}}(t) = \begin{cases} 1, & 0 \le t < t_0 \\ e^{-[k(t-t_0)]^{\beta}}, & t \ge t_0 \end{cases} \tag{8}$$

Here, t_0, k, and β are fitting parameters. We constrained $t_0 \ge 0$, $k \ge 0$, and $\beta \ge 0$ during the least-square curve fitting process. The initial values $t_0 = 0$, $k = 1$, and $\beta = 1$ were used in the fitting.

The 95% CIs for the estimated β values were calculated through bootstrapping the temperature quenching trajectories 10,000 times. In each iteration, the time series of the survival probability was resampled with replacement, followed by fitting the stretched exponential function in Eq. 8 to the resampled data. To assess the null hypothesis that the β value equals 1, the *P* value was computed using the same permutation test as described in the FRET efficiency fitting. Here, we first generated a synthetic time course of survival probability by fitting a single exponential curve (Eq. 7) to the observed data. Subsequently, we derived *N* simulated first-passage times based on this synthetic survival probability.

MFPT estimation

To estimate the MFPT for the near-native states S6, S7, S8, and S9, we randomly sampled 100 structures from the original temperature quenching simulation trajectories for each state, followed by performing an independent simulation for each starting structure at 310 K for 1 μs. To better estimate the MFPTs, we fitted the survival probability of these states using the double exponential function (*24*, *27*, *28*), which is given by

$$f_{\text{double}}^{\text{survival}}(t) = \begin{cases} 1 & 0 \le t < t_0 \\ f_1 e^{-k_1(t-t_0)} + (1-f_1) \cdot e^{-k_2(t-t_0)}, & t \ge t_0 \end{cases} \tag{9}$$

Here, t_0, k_1, k_2, and f_1 are fitting parameters. We constrained $t_0 \ge 0$, $k_1 \ge 0$, $k_2 \ge 0$, and $0 \le f_1 \le 1$ during the least-square curve fitting process. The initial values $t_0 = 0$, $k_1 = 1$, $k_2 = 0$, and $f_1 = 1$ were used in the fitting. MFPTs were then estimated as

$$\text{MFPT} = t_0 + \frac{f_1}{k_1} + \frac{1-f_1}{k_2} \tag{10}$$

The 95% CIs for the estimated MFPTs were calculated through bootstrapping the simulation trajectories 10,000 times. In each iteration, the time series of the survival probability was resampled with replacement, followed by fitting the double exponential function in Eq. 9 to the resampled data and then computing MFPT using Eq. 10.

LiP-MS and XL-MS experiments on ecPGK

The wild-type *E. coli pgk* gene was cloned onto a pET21(+) vector with a C-terminal His6 tag, and the protein was expressed from *E. coli* BL21(DE3) and purified by Ni-NTA chromatography. Native samples were prepared by diluting 2 μl of a frozen stock [consisting of 20 mM Hepes-NaOH (pH 7.4), 100 mM NaCl, 2 mM $MgCl_2$, 10% (v/v) glycerol, and protein (10 mg/ml)] 100-fold with 198 μl of native dilution buffer [20 mM Hepes-NaOH (pH 7.4), 100 mM NaCl, 2 mM $MgCl_2$, 1.01 mM tris(2-carboxyethyl)phosphine (TCEP), and 0.06 M GdmCl]. Refolded samples were prepared by taking a frozen stock prepared without glycerol, reducing the protein to dryness in a centrifugal concentrator, resuspending in an equal volume of 6 M GdmCl and 10 mM TCEP, incubating for 24 hours, and then diluting 2 μl of unfolded protein 100-fold with 198 μl of refolding dilution buffer [20 mM Hepes-NaOH (pH 7.4), 100 mM NaCl, 2 mM $MgCl_2$, 0.91 mM TCEP, and 0.1% (v/v) glycerol]. Refolding reactions proceeded for 1 hour before, and the native samples (used as a reference) were interrogated by either LiP-MS or XL-MS. Preparation of native and refolded ecPGK for LiP-MS and XL-MS was identical except dithiothreitol was used in place of TCEP (in identical concentrations) for LiP-MS studies.

For limited proteolysis, ecPGK samples were combined with PK in a 1:100 (w/w) enzyme:substrate ratio, incubated for 1 min, and quenched by boiling at 105°C for 5 min. For cross-linking, ecPGK samples were cross-linked by adding a 100 mM stock of DSBU to a final concentration of 1 mM, incubated for 1 hour, and then quenched by addition of tris-HCl (pH 7.5) to a final concentration of 20 mM. To perform quantitative cross-linking experiments, native ecPGK was cross-linked with natural abundance DSBU, refolded ecPGK was cross-linked with an isotopically heavy form of DSBU (synthesis reported in Supplementary Methods), and heavy/light pairs were pooled together. PK-proteolyzed and chemically cross-linked samples were prepared for liquid chromatography–mass spectrometry using standard procedures for in-solution

digest and solid-phase extraction, and data-dependent acquisition on a Q-Exactive HF-X was carried out using standard proteomic procedures (see Supplementary Methods). LiP-MS data were analyzed using a recently developed workflow in which ions are assigned and quantified with FragPipe (*59*) and processed with FLiPPR (*60*), and XL-MS data were analyzed using a procedure described in Supplementary Methods. Because of the large number of PGK peptides identified, we found that increasing the number of technical replicates from three to five was important to retain statistical significance following correction for multiple hypothesis testing.

Consistency test

We assessed the statistical consistency between simulated conformations in near-native entangled states (S7, S8, and S9) and experimental data from LiP-MS and XL-MS. For LiP-MS, we examined whether the identified PK cut-sites exhibit statistically significant changes in protease susceptibility in one of the misfolded states compared with the native state. For a given cut-site K in the misfolded state i, the null hypothesis is that the mean protease susceptibility, which is measured as the SASA of the region $K-5$ to $K+5$, denoted as $\langle \mathrm{SASA} \rangle_K^i$, is equal to that in the native state, i.e., $\langle \mathrm{SASA} \rangle_K^i = \langle \mathrm{SASA} \rangle_K^{\mathrm{Native}}$. The SASA values were calculated using the FreeSASA library (*61*) for the all-atom structures. These all-atom structures were obtained through back-mapping from the coarse-grained representation using our in-house back-mapping tool (*27*, *48*). A two-tailed permutation test was performed to estimate the P value. The permutation test was performed as follows: (i) combine the two sets of SASA values into one set and resample the combined set 10,000 times without replacement, i.e., permutate the indices, resulting in 10,000 resampled combined sets; (ii) for each resampled combined set, partition the first half as the resampled SASA for state i and the second half as the resampled SASA for the native state; (iii) for each resampled dataset, compute the statistic $|\mu_1 - \mu_2| / \sqrt{s_1^2/n_1 + s_2^2/n_2}$, where $\mu_1 = \langle \mathrm{SASA} \rangle_K^i$, $\mu_2 = \langle \mathrm{SASA} \rangle_K^{\mathrm{Native}}$, s_1 and s_2 are the corresponding SDs, and n_1 and n_2 are the sample sizes; (iv) Estimate the P value as the frequency of observing the resampled statistic that is greater than or equals to the observed statistic.

In the XL-MS analysis, we tested whether the identified residue pairs exhibit statistically significant changes in XP in one of the misfolded states compared with the native state. For a given residue pair (j, k) in the misfolded state i, the null hypothesis is that the mean XP, denoted as $\langle \mathrm{XP} \rangle_{j,k}^i$, is equal to that in the native state, i.e., $\langle \mathrm{XP} \rangle_{j,k}^i = \langle \mathrm{XP} \rangle_{j,k}^{\mathrm{Native}}$. A similar permutation test was performed to estimate the P value for this hypothesis test.

The XP is estimated using a modified scoring function from matched and nonaccessible crosslink (MNXL) score (*62*). The original MNXL score, designed for assessing XP in Lys-Lys pairs using the SASD, had limitations when non-Lys residues were involved in cross-linking. In addition, our XL-MS experiment used a cross-linker molecule that is 1.1 Å longer than the one used in the literature. To address these scenarios, we introduced modifications to the MNXL score, proposing the following equation of XP for a given residue pair (j, k) in a structure s

$$XP(j,k,s) = \begin{cases} N\left[\mu(A_j,A_k), \sigma^2(A_j,A_k)\right], & \text{if } J(j,k,s) \le C(A_j,A_k) \\ 0, & \text{else} \end{cases} \quad (11)$$

Here, N is the probability density at $J(j,k)$ in a normal distribution with mean $\mu(A_j, A_k)$ and SD $\sigma(A_j, A_k)$. A_j and A_k denote the amino acid types of the two residues. $J(j,k)$ is the SASD between residues j and k and computed using the Jwalk algorithm (*62*). $C(A_j, A_k)$ is the cutoff distance between amino acid types A_j and A_k. We shifted the mean and cutoff distance in the MNXL score to reflect the length differences in the cross-linker molecule and the amino acid side chains. The SD was then rescaled accordingly to ensure a fixed value at $\mu - 3\sigma$ in the normal distributions. The amino acid–dependent values in this equation are summarized in table S2.

The analysis used structures from states S7, S8, and S9, extracted from the last 50-ns temperature quenching trajectories with a lag time of 20 frames. To address autocorrelation in the data, we applied a lag time determined by the plateaued variance observed in block averaging analysis (*63*) of SASA and XP trajectories (see fig. S4). Among the last 50-ns trajectories, 25 frames were excluded because of failure in SASA computation caused by poor side-chain conformations reconstructed from CG structures. Raw P values were adjusted for multiple tests using the Benjamini-Hochberg method (*64*). The overall consistency for a near-native entangled state was then assessed on the basis of the number of adjusted P values below 0.05.

Clustering entanglement changes

To gain more structural insights to the near-native entangled states that exhibit significant consistency with the LiP-MS and XL-MS data, we developed a clustering and grouping algorithm to differentiate structures manifesting diverse statuses of noncovalent lasso entanglements. Initially, for the crystal structure, we identified native entanglements formed across all native contacts via Topoly (*65*). We recorded the loop position (i, j) and the crossing residues $(k_1, k_2, \ldots, k_N)$ along with their chirality (+ or −) for each of the two tails. We excluded trivial crossings where the same residue or adjacent residues (within two residues) pierce the loop multiple times. Using the same protocol, we identified entanglements in predicted structures only for native contacts with a Gaussian linking number ≥ 0.5 or that established a native entanglement. The changes in entanglement status for a loop in the predicted structure were determined by comparing linking numbers and chirality of each tail with those in the crystal structure. The status is denoted in the format "L*C*," where "L" indicates the absolute linking number and "C" indicates chirality. In the notation, there are two "*" symbols, which refer to the changes in the absolute linking number and chirality, respectively. Changes in absolute linking number can be no change (#), gain (+), or loss (−), while changes in chirality can be no change (#) or switching (~). For example, if a closed loop had 0 linking number for both tails in the crystal structure but gained linking number to +1 and −1 for N-terminal and C-terminal tails, respectively, the status of this loop will be denoted as ($L + C$~, $L + C$~).

The changes of entanglements were identified for all loops in all the predicted structures analyzed. Subsequently, we conducted clustering analyses for those entanglement changes sharing the same status sequentially, considering N-terminal crossing residues, C-terminal crossing residues, loop location, and crossing contamination. An agglomerative clustering approach was used for crossing residues and loop location clustering.

For crossing residues on the N- or C-terminal tail, the pairwise distance for loops m and n is defined as

$$d_{mn}^{\text{Crossing}} = \max\left\{d_{mn}^{\text{ref}}, d_{mn}\right\} \tag{12}$$

where d_{mn}^{ref} and d_{mn} are the absolute difference between the median crossing residue indices of loops m and n in the crystal structure and the predicted structure, respectively. If a loop lacks crossing residues, the median index is assigned −1. If one loop lacks crossing residues while the other has them, d_{mn}^{ref} or d_{mn} will be automatically assigned to 10 residues. Clustering used the "average" linkage method, and clusters were finalized at a distance cutoff of 20 residues.

For loop location, the pairwise distance for loops m (L_m) and n (L_n) is defined as

$$d_{mn}^{\text{Loop}} = \begin{cases} 0.5, & L_m \subseteq L_n \text{ or } L_n \subseteq L_m \\ \dfrac{\max\{L_m \cup L_n\} - \min\{L_m \cup L_n\} + 1}{|L_m| + |L_n|}, & \text{else} \end{cases} \tag{13}$$

where $|L_m|$ is the length of the loop m, and $L_m \cup L_n$ is the combined set of the loop residue indices. The distance equals 1.0 when two loops are adjacent, becomes less than 1.0 when they overlap and greater than 1.0 when they are distant. Clustering used the average linkage method, and clusters were finalized at a distance cutoff of 1.0.

The last clustering analysis for "crossing contamination" aimed to separate two loops where one involves crossing residues of the other [e.g., loop (10, 20) is contaminated by the crossing residue 15 of another loop]. The pairwise distance between loops m and n is defined as

$$\begin{aligned} d_{mn}^{\text{Contamination}} = \max\Big\{ & d\left(L_m, k_1^n\right), d\left(L_m, k_2^n\right), \ldots, \\ & d\left(L_m, k_{N_n}^n\right), d\left(L_n, k_1^m\right), \ldots, d\left(L_n, k_{N_m}^m\right) \Big\} \end{aligned} \tag{14}$$

where $d\left(L_m, k_s^n\right)$ is the distance reflecting how deep the loop L_m is contaminated by the sth crossing residue k_s^n of the loop L_n. It is defined as

$$d\left(L_m, k_s^n\right) = \begin{cases} \dfrac{\min\left\{j_m - k_s^n, k_s^n - i_m\right\}}{j_m - i_m}, & i_m \le k_s^n \le j_m \\ 0, & \text{else} \end{cases} \tag{15}$$

On the basis of this pairwise distance, we performed a divisive clustering, followed by an agglomerative clustering for the loops. In divisive clustering, loops were divided into subsets where the distance between two loops in the same subset was less than a cutoff of 0.1 (i.e., the maximum contamination depth is less than 10% of the loop length). Agglomerative clustering using the "complete" linkage was then applied on these subsets, and the final clusters were chosen at the cutoff of 0.1. In the agglomerative clustering approach performed in all the steps, the permuCLUSTER algorithm (*66*) was used to mitigate the impact of input order, with 100 permutations.

Representative structure selection

Each cluster acquired in the previous clustering step represents a unique entanglement change with different types and locations on the protein. The last 10-ns simulation structures were labeled using the metastable state ID, clusters of changes in entanglements they have, and the experimental signals they produce, followed by grouping based on the labels. Here, we identify a simulation structure capable of generating an experimental signal only if it demonstrates an SASA greater than the 97.5th percentile or less than the 2.5th percentile of all native state structures for a LiP-MS signal or exhibits an XP score outside the corresponding percentiles for an XL-MS signal. These groups represent the misfolded structural ensemble that can generate the observed LiP-MS and/or XL-MS signals.

For each group, we select the structure with the highest microstate probability as the representative. In case of a tie, priority is given to the one with the highest Q value, followed by the highest G value. These structures form the representative misfolded structural ensemble. In addition, for each of the metastable states S7, S8, and S9, we further choose a single representative structure that can generate the greatest number of experimental signals. If there is a tie, then preference is given to the structure with the highest microstate probability, followed by the highest Q and G values.

Supplementary Materials

The PDF file includes:

Supplementary Methods
Figs. S1 to S7
Tables S1 and S2
Legends for datasets S1 to S4
References

Other Supplementary Material for this manuscript includes the following:

Datasets S1 to S4

REFERENCES AND NOTES

1. R. Zwanzig, Two-state models of protein folding kinetics. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 148–150 (1997).
2. G. C. Rollins, K. A. Dill, General mechanism of two-state protein folding kinetics. *J. Am. Chem. Soc.* **136**, 11420–11427 (2014).
3. A. R. Fersht, Transition-state structure as a unifying basis in protein-folding mechanisms: Contact order, chain topology, stability, and the extended nucleus mechanism. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 1525–1529 (2000).
4. S. B. Ozkan, I. Bahar, K. A. Dill, Transition states and the meaning of Φ-values in protein folding kinetics. *Nat. Struct. Biol.* **8**, 765–769 (2001).
5. C. D. Snow, Y. M. Rhee, V. S. Pande, Kinetic definition of protein folding transition state ensembles and reaction coordinates. *Biophys. J.* **91**, 14–24 (2006).
6. T. R. Weikl, K. A. Dill, Transition-states in protein folding kinetics: The structural interpretation of Φ values. *J. Mol. Biol.* **365**, 1578–1586 (2007).
7. T. R. Weikl, Transition states in protein folding kinetics: Modeling Φ-values of small β-sheet proteins. *Biophys. J.* **94**, 929–937 (2008).
8. R. D. M. Travasso, P. F. N. Faísca, A. Rey, The protein folding transition state: Insights from kinetics and thermodynamics. *J. Chem. Phys.* **133**, 125102 (2010).
9. F. Ding, N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, E. I. Shakhnovich, Direct molecular dynamics observation of protein folding transition state ensemble. *Biophys. J.* **83**, 3525–3532 (2002).
10. J. Gsponer, A. Caflisch, Molecular dynamics simulations of protein folding from the transition state. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 6719–6724 (2002).
11. J. Zhang, W. Li, J. Wang, M. Qin, L. Wu, Z. Yan, W. Xu, G. Zuo, W. Wang, Protein folding simulations: From coarse-grained model to all-atom model. *IUBMB Life* **61**, 627–643 (2009).
12. S. Piana, K. Lindorff-Larsen, D. E. Shaw, Protein folding kinetics and thermodynamics from atomistic simulation. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 17845–17850 (2012).
13. V. Pande, "Protein folding studied with molecular dynamics simulation" in *Encyclopedia of Biophysics* (Springer Berlin Heidelberg, 2013); http://link.springer.com/10.1007/978-3-642-16712-6_730, pp. 2016–2020.
14. S. Ovchinnikov, P.-S. Huang, Structure-based protein design with deep learning. *Curr. Opin. Chem. Biol.* **65**, 136–144 (2021).
15. J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, D. Baker, Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
16. J. Sabelko, J. Ervin, M. Gruebele, Observation of strange kinetics in protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 6031–6036 (1999).

17. S. Osváth, J. J. Sabelko, M. Gruebele, Tuning the heterogeneous early folding dynamics of phosphoglycerate kinase. *J. Mol. Biol.* **333**, 187–199 (2003).
18. A. Dhar, K. Girdhar, D. Singh, H. Gelman, S. Ebbinghaus, M. Gruebele, Protein stability and folding kinetics in the nucleus and endoplasmic reticulum of eucaryotic cells. *Biophys. J.* **101**, 421–430 (2011).
19. T. W. Kim, S. J. Lee, J. Jo, J. G. Kim, H. Ki, C. W. Kim, K. H. Cho, J. Choi, J. H. Lee, M. Wulff, Y. M. Rhee, H. Ihee, Protein folding from heterogeneous unfolded state revealed by time-resolved X-ray solution scattering. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 14996–15005 (2020).
20. O. Bilsel, Heterogeneous folding and stretched kinetics. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 18915–18917 (2020).
21. S. Osváth, L. Herényi, P. Závodszky, J. Fidy, G. Köhler, Hierarchic finite level energy landscape model: To describe the refolding kinetics of phosphoglycerate kinase. *J. Biol. Chem.* **281**, 24375–24380 (2006).
22. D. C. Johnston, Stretched exponential relaxation arising from a continuous sum of exponential decays. *Phys. Rev. B* **74**, 184430 (2006).
23. N. D. Socci, J. N. Onuchic, P. G. Wolynes, Protein folding mechanisms and the multidimensional folding funnel. *Proteins* **32**, 136–158 (1998).
24. D. A. Nissley, Y. Jiang, F. Trovato, I. Sitarik, K. B. Narayan, P. To, Y. Xia, S. D. Fried, E. P. O'Brien, Universal protein misfolding intermediates can bypass the proteostasis network and remain soluble and less functional. *Nat. Commun.* **13**, 3081 (2022).
25. Q. V Vu, I. Sitarik, Y. Jiang, D. Yadav, P. Sharma, S. D. Fried, M. S. Li, E. P. O'Brien, A newly identified class of protein misfolding in all-atom folding simulations consistent with limited proteolysis mass spectrometry. bioRxiv 500586 [Preprint] (2022). https://doi.org/10.1101/2022.07.19.500586.
26. V. Rana, I. Sitarik, J. Petucci, Y. Jiang, H. Song, E. P. O'Brien, Non-covalent lasso entanglements in folded proteins: Prevalence, functional implications, and evolutionary significance. *J. Mol. Biol.* **436**, 168459 (2024).
27. Y. Jiang, S. S. Neti, I. Sitarik, P. Pradhan, P. To, Y. Xia, S. D. Fried, S. J. Booker, E. P. O'Brien, How synonymous mutations alter enzyme structure and function over long timescales. *Nat. Chem.* **15**, 308–318 (2023).
28. R. Halder, D. A. Nissley, I. Sitarik, Y. Jiang, Y. Rao, Q. V. Vu, M. S. Li, J. Pritchard, E. P. O'Brien, How soluble misfolded proteins bypass chaperones at the molecular level. *Nat. Commun.* **14**, 3689 (2023).
29. P. D. Lan, D. A. Nissley, I. Sitarik, Q. V. Vu, Y. Jiang, P. To, Y. Xia, S. D. Fried, M. S. Li, E. P. O'Brien, Synonymous mutations can alter protein dimerization through localized interface misfolding involving self-entanglements. *J. Mol. Biol.* **436**, 168487 (2024).
30. P. To, B. Whitehead, H. E. Tarbox, S. D. Fried, Nonrefoldability is Pervasive Across the E. *coli* Proteome. *J. Am. Chem. Soc.* **143**, 11435–11448 (2021).
31. P. To, Y. Xia, S. O. Lee, T. Devlin, K. G. Fleming, S. D. Fried, A proteome-wide map of chaperone-assisted protein refolding in a cytosol-like milieu. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2210536119 (2022).
32. S. Gosavi, L. L. Chavez, P. A. Jennings, J. N. Onuchic, Topological frustration and the folding of interleukin-1β. *J. Mol. Biol.* **357**, 986–996 (2006).
33. L. L. Chavez, S. Gosavi, P. A. Jennings, J. N. Onuchic, Multiple routes lead to the native state in the energy landscape of the β-trefoil family. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 10254–10258 (2006).
34. S. Gosavi, P. C. Whitford, P. A. Jennings, J. N. Onuchic, Extracting function from a β-trefoil folding motif. *Proc. Natl. Acad. Sci.U.S.A.* **105**, 10384–10389 (2008).
35. D. T. Capraro, M. Roy, J. N. Onuchic, P. A. Jennings, Backtracking on the folding landscape of the β-trefoil protein interleukin-1β? *Proc. Natl. Acad. Sci. U.S.A.* **105**, 14844–14848 (2008).
36. J. I. Sułkowska, P. Sułkowski, J. Onuchic, Dodging the crisis of folding proteins with knots. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 3119–3124 (2009).
37. S. Gosavi, Understanding the folding-function tradeoff in proteins. *PLOS ONE* **8**, e61222 (2013).
38. J. D. Leuchter, A. T. Green, J. Gilyard, C. G. Rambarat, S. S. Cho, Coarse-grained and atomistic MD simulations of RNA and DNA folding. *Isr. J. Chem.* **54**, 1152–1164 (2014).
39. K. T. Halloran, Y. Wang, K. Arora, S. Chakravarthy, T. C. Irving, O. Bilsel, C. L. Brooks, C. R. Matthews, Frustration and folding of a TIM barrel protein. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 16378–16383 (2019).
40. J. Macošek, G. Mas, S. Hiller, Redefining molecular chaperones as chaotropes. *Front. Mol. Biosci.* **8**, 683132 (2021).
41. S. K. Sharma, P. De Los Rios, P. Christen, A. Lustig, P. Goloubinoff, The kinetic parameters and energy cost of the Hsp70 chaperone as a polypeptide unfoldase. *Nat. Chem. Biol.* **6**, 914–920 (2010).
42. D. Thirumalai, D. K. Klimov, S. A. Woodson, Kinetic partitioning mechanism as a unifying theme in the folding of biomolecules. arXiv:cond-mat/9704067 [cond-mat.soft] (1997).
43. E. P. O'Brien, J. Christodoulou, M. Vendruscolo, C. M. Dobson, Trigger factor slows co-translational folding through kinetic trapping while sterically protecting the nascent chain from aberrant cytosolic interactions. *J. Am. Chem. Soc.* **134**, 10920–10932 (2012).
44. A. K. Sharma, B. Bukau, E. P. O'Brien, Physical origins of codon positions that strongly influence cotranslational folding: A framework for controlling nascent-protein folding. *J. Am. Chem. Soc.* **138**, 1180–1195 (2016).
45. B. Fritch, A. Kosolapov, P. Hudson, D. A. Nissley, H. L. Woodcock, C. Deutsch, E. P. O'Brien, Origins of the mechanochemical coupling of peptide bond formation to protein synthesis. *J. Am. Chem. Soc.* **140**, 5077–5087 (2018).
46. D. A. Nissley, E. P. O'Brien, Structural origins of FRET-observed nascent chain compaction on the ribosome. *J. Phys. Chem. B* **122**, 9927–9937 (2018).
47. S. E. Leininger, F. Trovato, D. A. Nissley, E. P. O'Brien, Domain topology, stability, and translation speed determine mechanical force generation on the ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 5523–5532 (2019).
48. D. A. Nissley, Q. V. Vu, F. Trovato, N. Ahmed, Y. Jiang, M. S. Li, E. P. O'Brien, Electrostatic interactions govern extreme nascent protein ejection times from ribosomes and can delay ribosome recycling. *J. Am. Chem. Soc.* **142**, 6103–6110 (2020).
49. P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, V. S. Pande, OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **13**, e1005659 (2017).
50. J. MacQueen, "Some methods for classification and analysis of multivariate observations" in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Oakland, CA, USA, 1967), vol. 1, pp. 281–297.
51. S. Röblitz, M. Weber, Fuzzy spectral clustering by PCCA+: Application to Markov state models and data classification. *Adv. Data Anal. Classif.* **7**, 147–179 (2013).
52. M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, F. Noé, PyEMMA 2: A software package for estimation, validation, and analysis of Markov models. *J. Chem. Theory Comput.* **11**, 5525–5542 (2015).
53. J. K. Noel, A. Schug, A. Verma, W. Wenzel, A. E. Garcia, J. N. Onuchic, Mirror images as naturally competing conformations in protein folding. *J. Phys. Chem. B* **116**, 6880–6888 (2012).
54. K. Kachlishvili, G. G. Maisuradze, O. A. Martin, A. Liwo, J. A. Vila, H. A. Scheraga, Accounting for a mirror-image conformation as a subtle effect in protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8458–8463 (2014).
55. J. I. Kwiecińska, M. Cieplak, Chirality and protein folding. *J. Phys. Condens. Matter* **17**, S1565–S1580 (2005).
56. N. Akrap, T. Seidel, B. G. Barisas, Förster distances for fluorescence resonant energy transfer between mCherry and other visible fluorescent proteins. *Anal. Biochem.* **402**, 105–106 (2010).
57. C. E. J. Dieteren, S. C. A. M. Gielen, L. G. J. Nijtmans, J. A. M. Smeitink, H. G. Swarts, R. Brock, P. H. G. M. Willems, W. J. H. Koopman, Solute diffusion is hindered in the mitochondrial matrix. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 8657–8662 (2011).
58. A. Huang, H. Yao, B. D. Olsen, SANS partial structure factor analysis for determining protein–polymer interactions in semidilute solution. *Soft Matter* **15**, 7350–7359 (2019).
59. F. Yu, S. E. Haynes, G. C. Teo, D. M. Avtonomov, D. A. Polasky, A. I. Nesvizhskii, Fast quantitative analysis of timsTOF PASEF data with MSFragger and IonQuant. *Mol. Cell. Proteomics* **19**, 1575–1585 (2020).
60. E. Manriquez-Sandoval, J. Brewer, G. Lule, S. Lopez, S. D. Fried, FLiPPR: A processor for limited proteolysis (LiP) mass spectrometry data sets built on FragPipe. *J. Proteome Res.* **23**, 2332–2342 (2024).
61. S. Mitternacht, FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Res* **5**, 189 (2016).
62. J. M. A. Bullock, J. Schwab, K. Thalassinos, M. Topf, The importance of non-accessible crosslinks and solvent accessible surface distance in modeling proteins with restraints from crosslinking mass spectrometry. *Mol. Cell. Proteomics* **15**, 2491–2500 (2016).
63. A. Grossfield, P. N. Patrone, D. R. Roe, A. J. Schultz, D. Siderius, D. M. Zuckerman, Best practices for quantification of uncertainty and sampling quality in molecular simulations [Article v1.0]. *Living J. Comput. Mol. Sci.* **1**, 5067 (2019).
64. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B. Methodol.* **57**, 289–300 (1995).
65. P. Dabrowski-Tumanski, P. Rubach, W. Niemyska, B. A. Gren, J. I. Sulkowska, Topoly: Python package to analyze topology of polymers. *Brief. Bioinform.* **22**, bbaa196 (2021).
66. W. A. van der Kloot, A. M. J. Spaans, W. J. Heiser, Instability of hierarchical cluster analysis due to input order of the data: The PermuCLUSTER solution. *Psychol. Methods* **10**, 468–476 (2005).
67. Y. Jiang, C. M. Deane, G. M. Morris, E. P. O'Brien, It is theoretically possible to avoid misfolding into non-covalent lasso entanglements using small molecule drugs. *PLoS Comput. Biol.* **20**, e1011901 (2024).
68. D. Frishman, P. Argos, Knowledge-based protein secondary structure assignment. *Proteins* **23**, 566–579 (1995).
69. M. Q. Müller, F. Dreiocker, C. H. Ihling, M. Schäfer, A. Sinz, Cleavable cross-linker for protein structure analysis: Reliable identification of cross-linking products by tandem MS. *Anal. Chem.* **82**, 6958–6968 (2010).

70. M. L. Mendes, L. Fischer, Z. A. Chen, M. Barbon, F. J. O'Reilly, S. H. Giese, M. Bohlke-Schneider, A. Belsom, T. Dau, C. W. Combe, M. Graham, M. R. Eisele, W. Baumeister, C. Speck, J. Rappsilber, An integrated workflow for crosslinking mass spectrometry. *Mol. Syst. Biol.* **15**, e8994 (2019).
71. S. Lenz, L. R. Sinn, F. J. O'Reilly, L. Fischer, F. Wegner, J. Rappsilber, Reliable identification of protein-protein interactions by crosslinking mass spectrometry. *Nat. Commun.* **12**, 3564 (2021).

Acknowledgments: All computer simulations and data analysis in this work have been carried out on high-performance computing collectively known as Roar Collab, which is operated by the Institute for Computational and Data Sciences at The Pennsylvania State University. S.D.F. acknowledges support from a Camille Dreyfus Teacher-Scholar Award and a Sloan Fellowship. **Funding:** This work was supported by National Science Foundation MCB-2045844 (S.D.F.), National Institutes of Health DP2-GM140926 (S.D.F.), National Science Foundation MCB-2031584 (E.P.O.), and National Institutes of Health R35-GM124818 (E.P.O.). **Author contributions:** Conceptualization: Y.J., I.S., and E.P.O. Methodology: Y.J., Y.X., I.S., P.S., H.S., S.D.F., and E.P.O. Investigation: Y.J., Y.X., and P.S. Resources: Y.X. Data curation: Y.J., Y.X., I.S., and P.S. Validation: Y.J., Y.X., I.S., and E.P.O. Formal analysis: Y.J., Y.X., H.S., and E.P.O. Software: Y.J., Y.X., and I.S. Visualization: Y.J. and I.S. Supervision: S.D.F. and E.P.O. Writing—original draft: Y.J., Y.X., S.D.F., and E.P.O. Writing—review and editing: Y.J., Y.X., I.S., P.S., H.S., S.D.F., and E.P.O. Funding acquisition: S.D.F. and E.P.O. Project administration: Y.J., S.D.F., and E.P.O. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. All codes and input files used for the simulations and data analyses are available in the Github repositories https://github.com/obrien-lab/Misfolded-ecPGK-structural-ensemble-consistent-with-LiP-MS-and-XL-MS-experiments (archived on 10.5281/zenodo.14579793) and https://github.com/obrien-lab/cg_simtk_protein_folding (archived on 10.5281/zenodo.14579799). The representative structural ensemble of ecPGK that consistent with the experimental signals can be interactively visualized and downloaded using the webapp https://obrien-lab.github.io/Misfolded-ecPGK-structural-ensemble-consistent-with-LiP-MS-and-XL-MS-experiments/. The mass spectrometry data generated in this study have been deposited to the PRIDE repository with the dataset identifiers PXD053579 for LiP-MS and PXD053582 for XL-MS. The expression plasmids for expression of PGK in *E. coli* can be provided by S.D.F. pending scientific review and a completed material transfer agreement.

Submitted 28 August 2024
Accepted 6 February 2025
Published 14 March 2025
10.1126/sciadv.ads7379