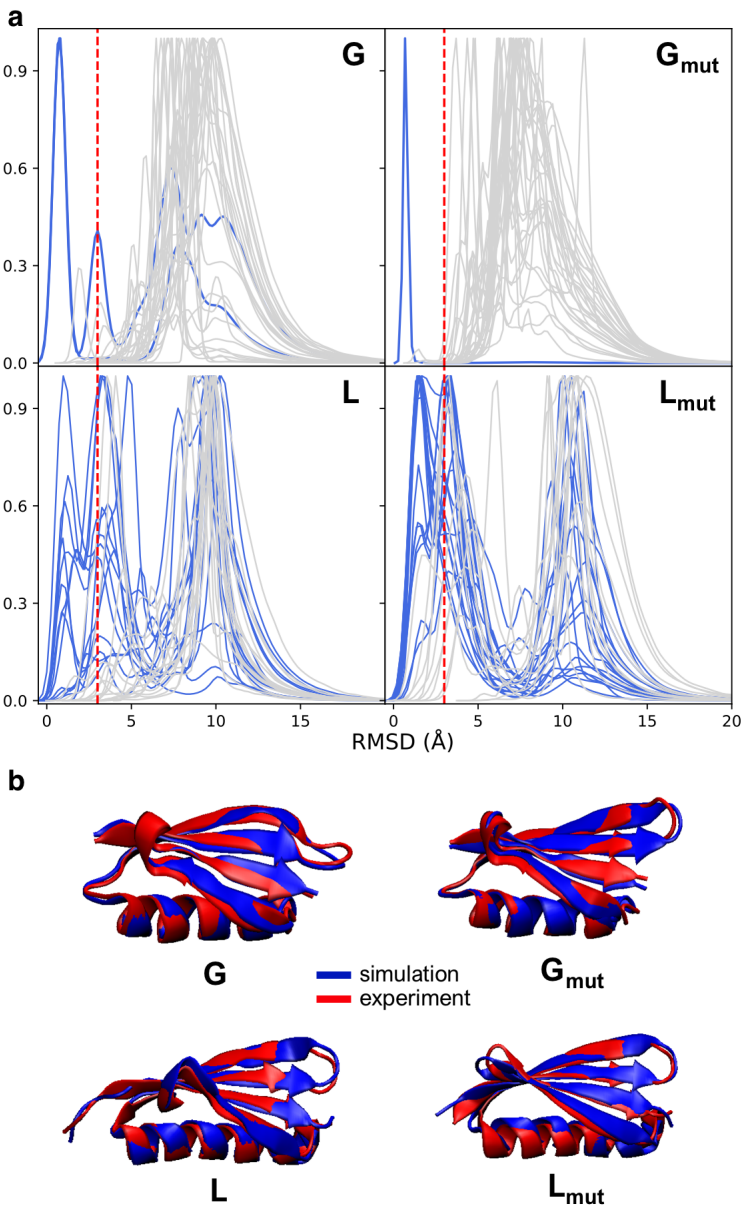
Supplementary Information for Deciphering the folding mechanism of protein G, L and their mutants

Liwei Chang[†,‡] and Alberto Perez[*,†,‡]

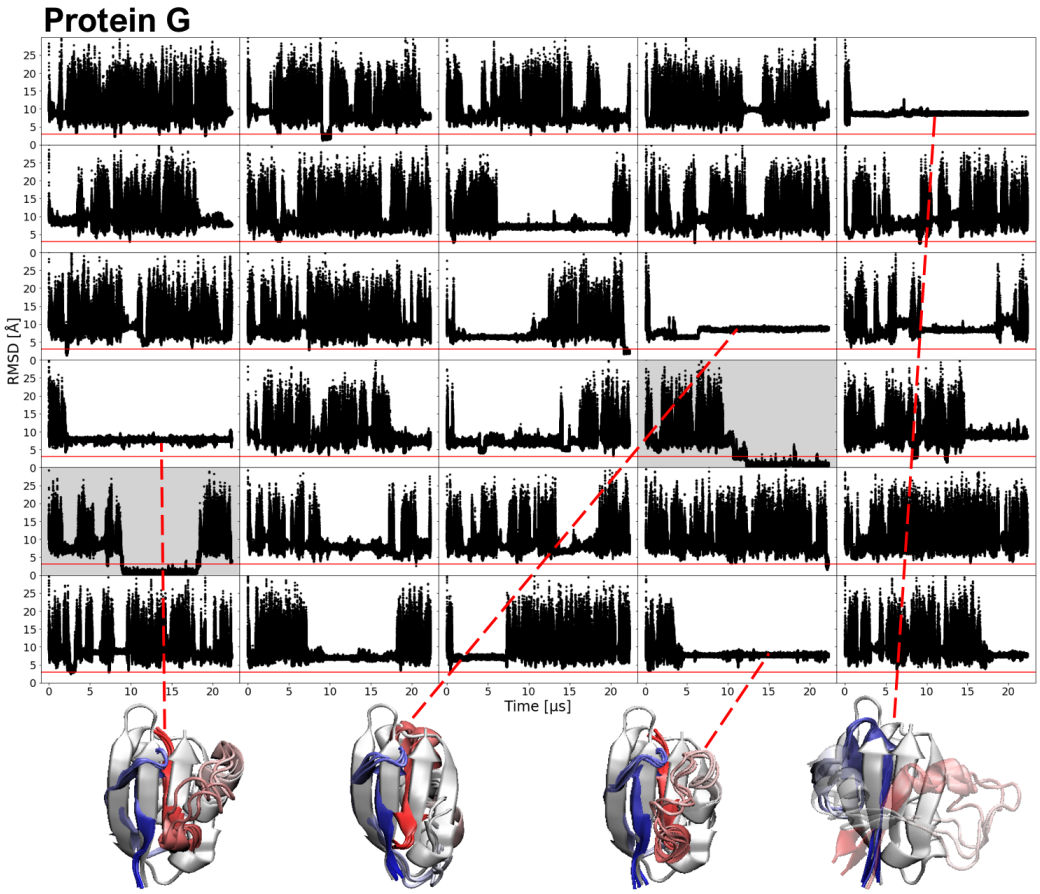
†*Department of Chemistry, University of Florida, Gainesville, FL 32611, USA.*

‡*Quantum Theory Project, University of Florida, Gainesville, FL 32611, USA.*

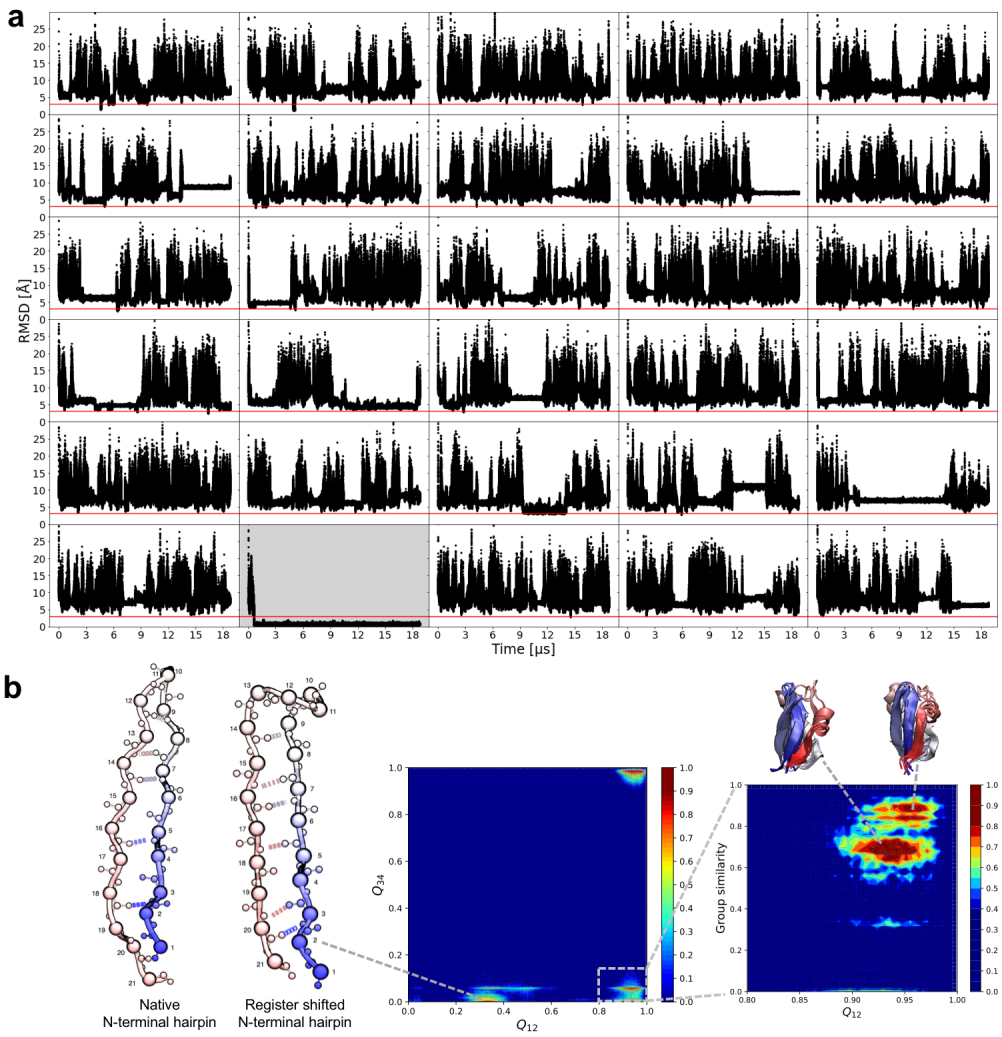
E-mail: perez@chem.ufl.edu



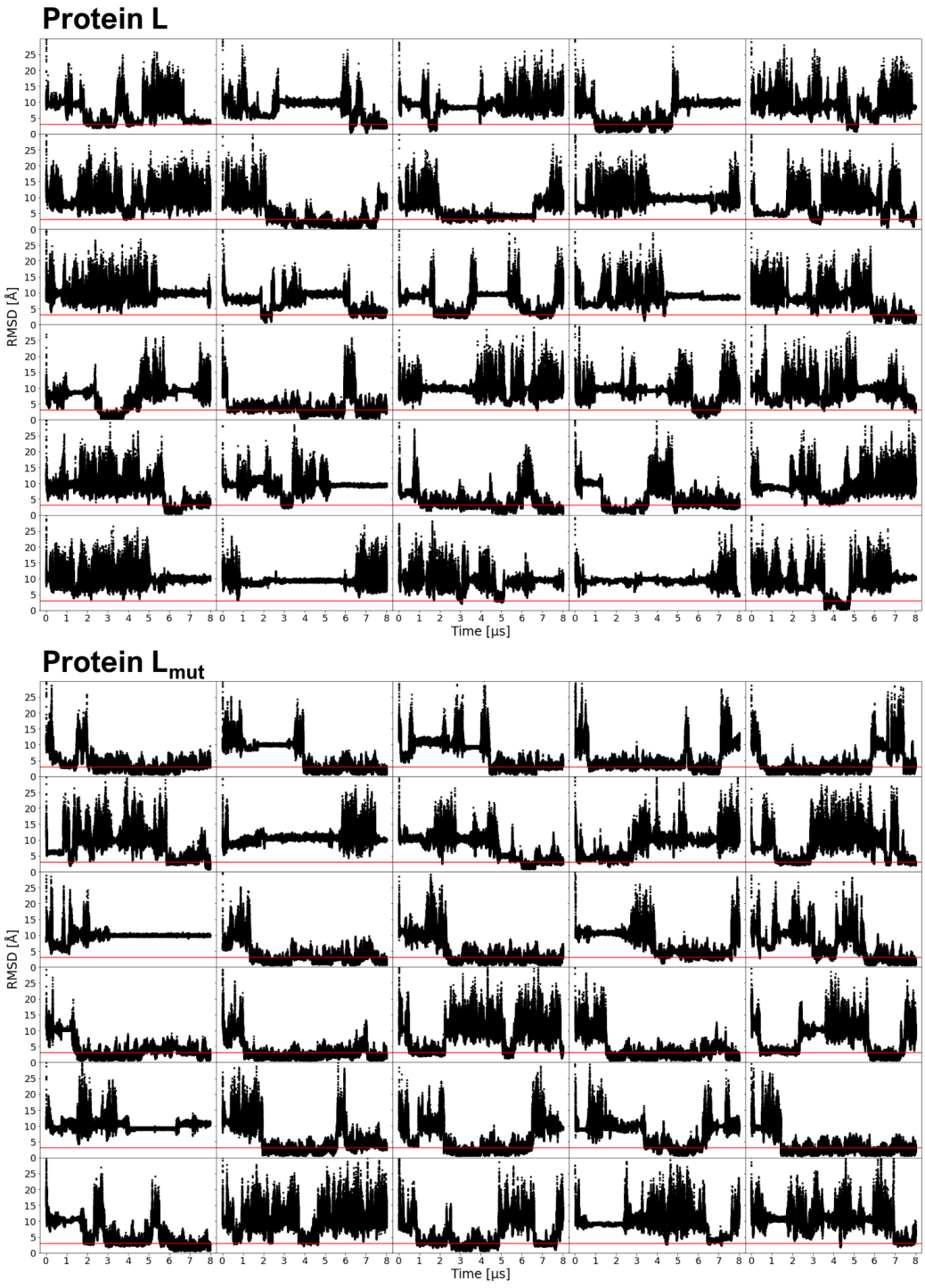
Supplementary Figure 1. Summary of all "walkers" and representative structures in MELD simulations. (a) Normalized histogram of RMSD against native structure for the 30 "walkers" in MELD simulations. Blue line indicates the walkers with high population in the native state (3 Å red dotted line). **(b)** Representative structures from MELD simulations by using hierarchical clustering at the lowest replica with Cpptraj[1] and clustering protocol can be found here.[2]



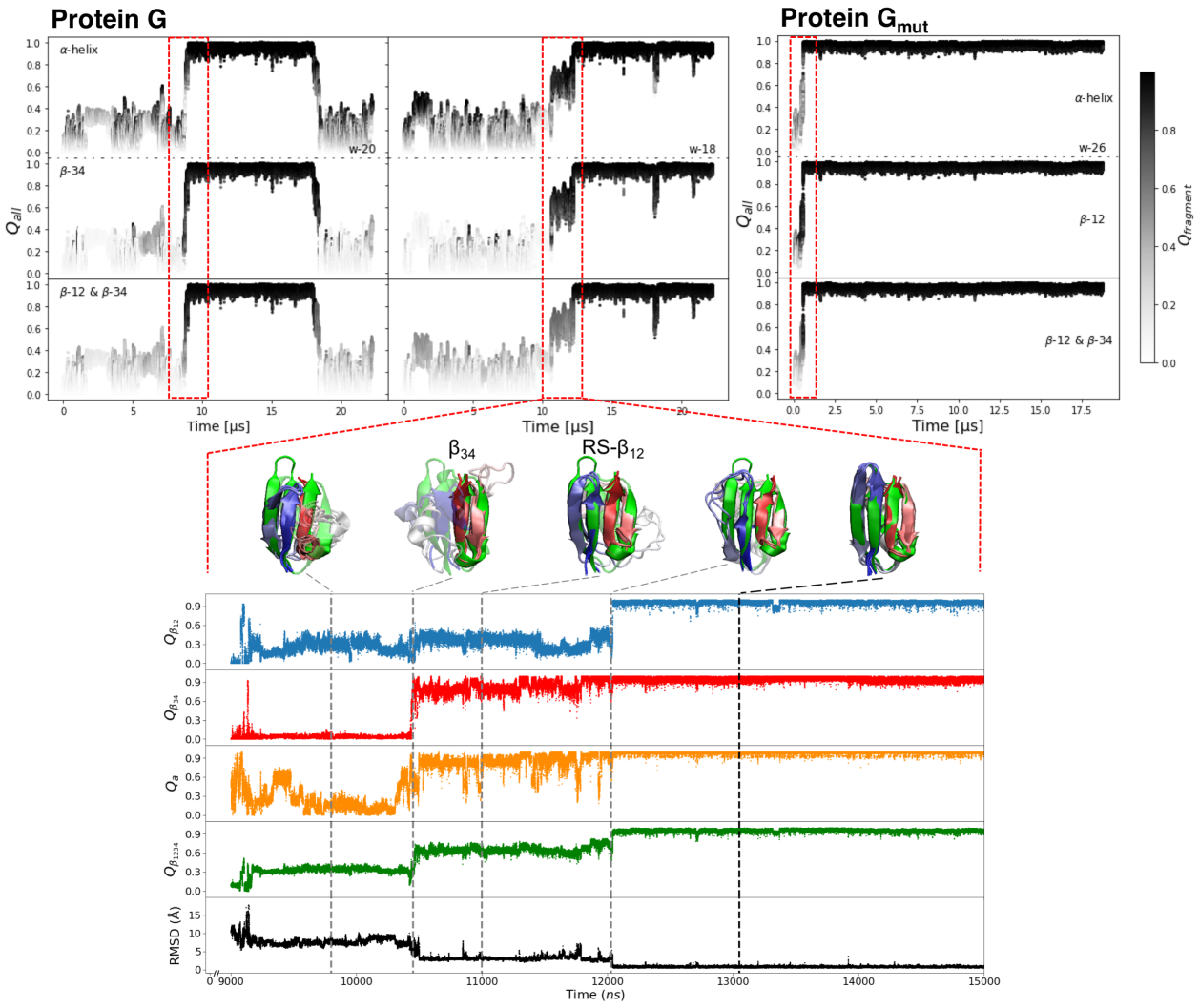
Supplementary Figure 2. MELD folding simulation summary for protein G. The time evolution for the RMSD of all walkers in the replica exchange simulation of protein G (top) and representatives of misfolded conformations (bottom). A horizontal red line indicates a 3 Å RMSD. Dashed lines connect long lived states and the corresponding non-native conformation.



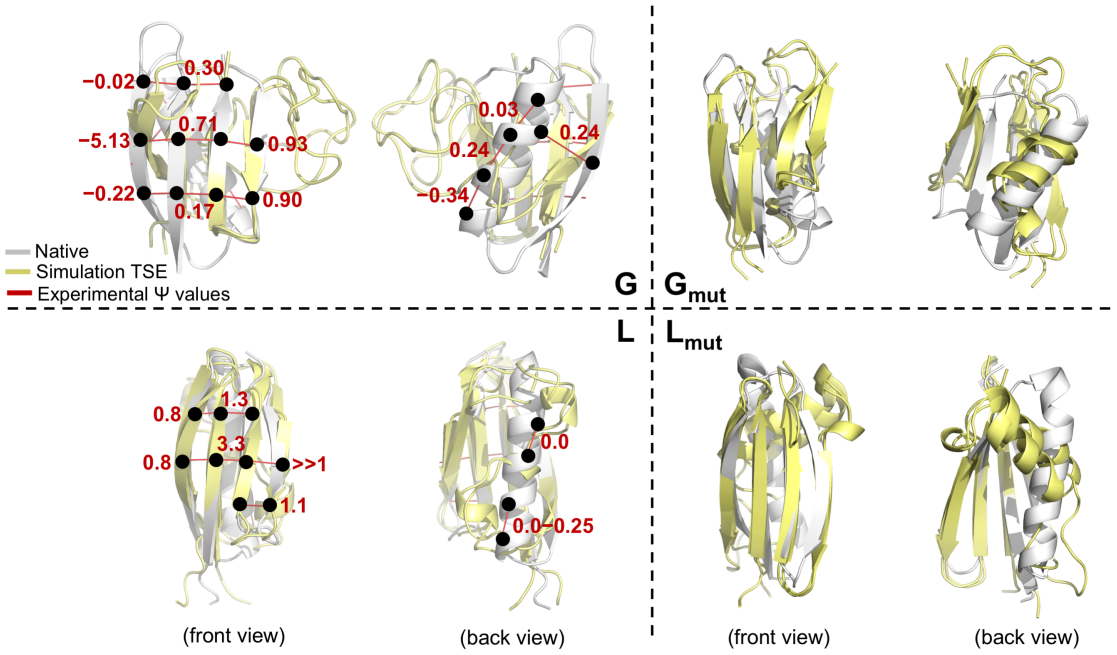
Supplementary Figure 3. MELD folding simulation summary for protein G mutant. (a) The time evolution for the RMSD of all walkers in the replica exchange simulation of protein G_{mut}. **(b)** Left: Representative conformations of the N-terminal β-hairpin in native and register-shifted conformations. Middle: sampling at the lowest replica of simulation projected on the native contact fraction of N-terminal and C-terminal hairpin. Right: Dominant conformations represented by group similarity[3] based on contacts between β_{12} and β_{34} for conformations sampling the N-terminal native hairpin conformation.



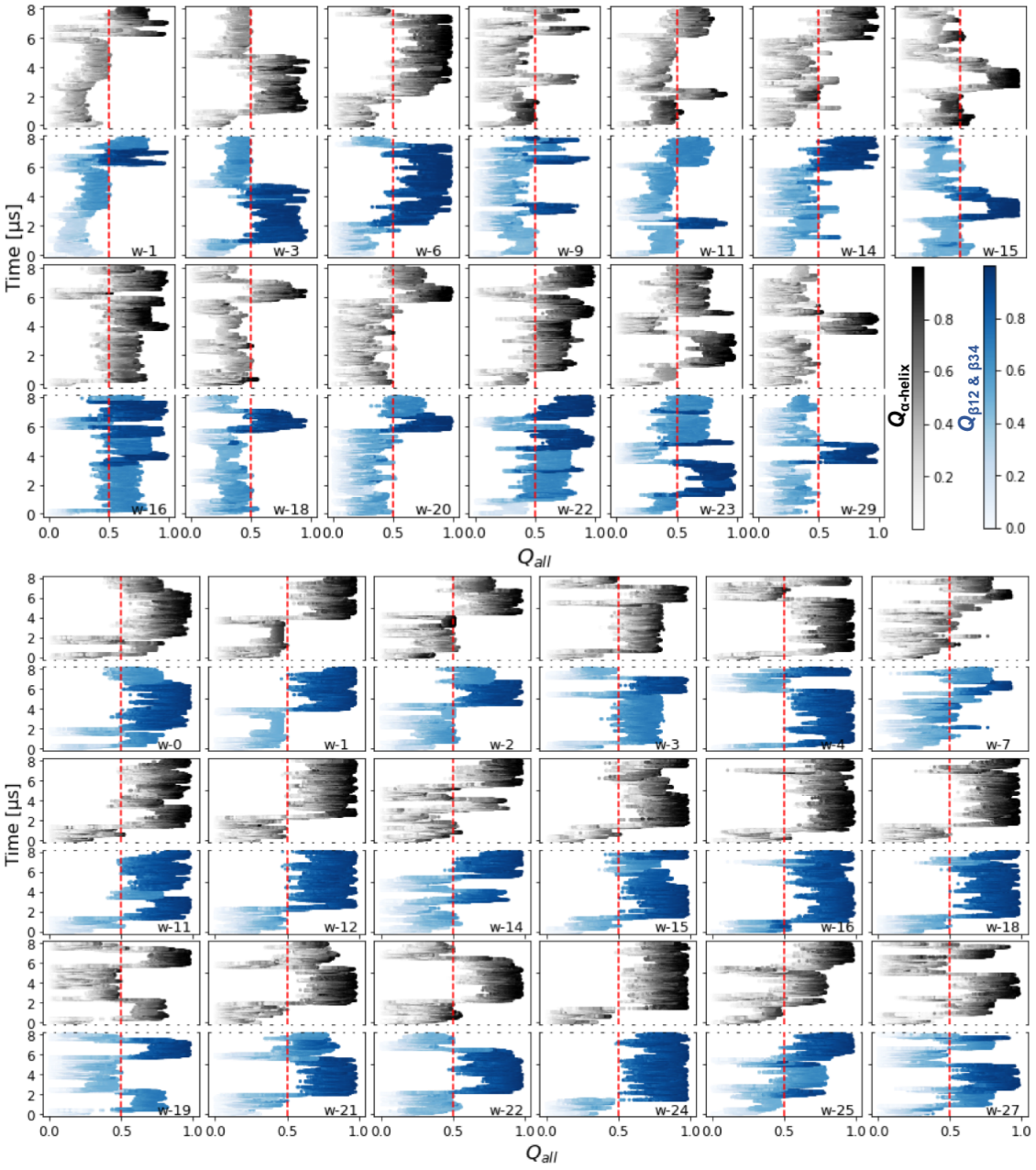
Supplementary Figure 4. MELD folding simulation summary for protein L/L$_{mut}$. The time evolution for the RMSD of all walkers in the replica exchange simulation of protein L (top) and protein L mutant (bottom). A horizontal red line denotes a 3 Å RMSD.



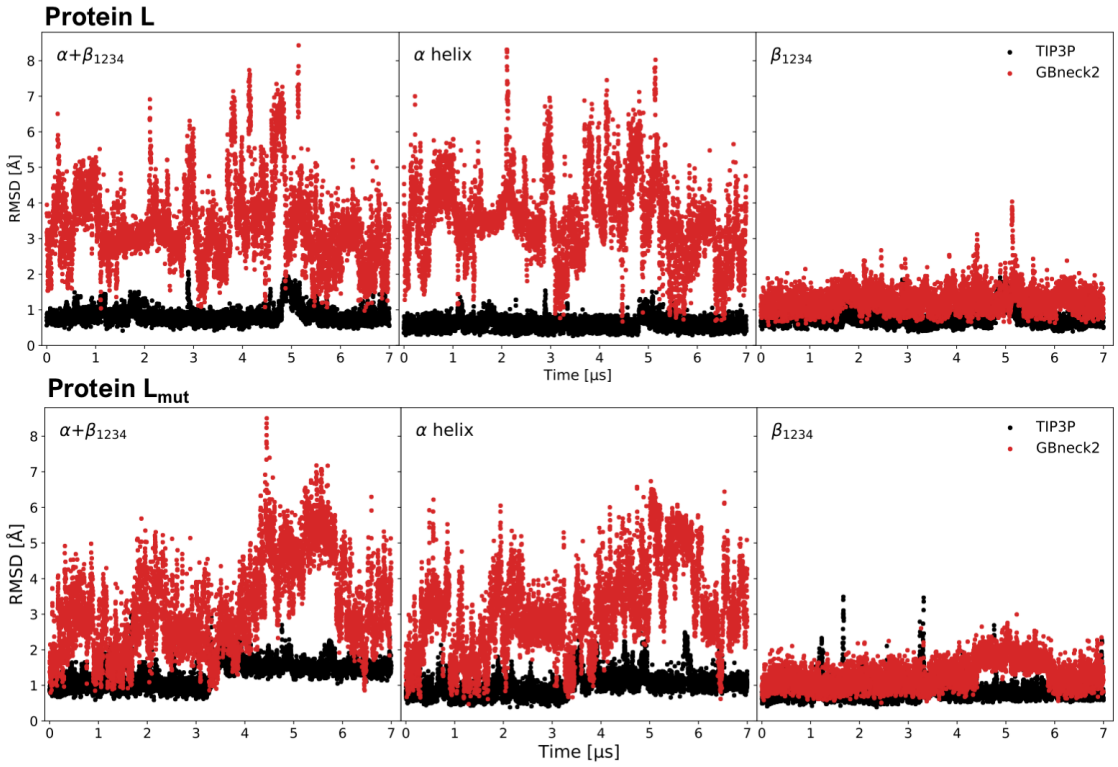
Supplementary Figure 5. The formation of native β-strands induces folding of protein G and its mutant. Top: the time evolution for native contact fractions of protein G and G_{mut} colored by native contact fractions of different fragments (helix, C-terminal hairpin or both hairpins) at walkers exhibiting folding transitions. Bottom: Native contact fractions of β_{12} (blue), β_{34} (red), α helix (orange), β_{1234} (green) and total RMSD (black) for protein G simulation at walker 18. Representative conformations are shown at different stages along the folding transition. See Figure 2 for walker 20 of protein G and walker 26 of protein G_{mut}.



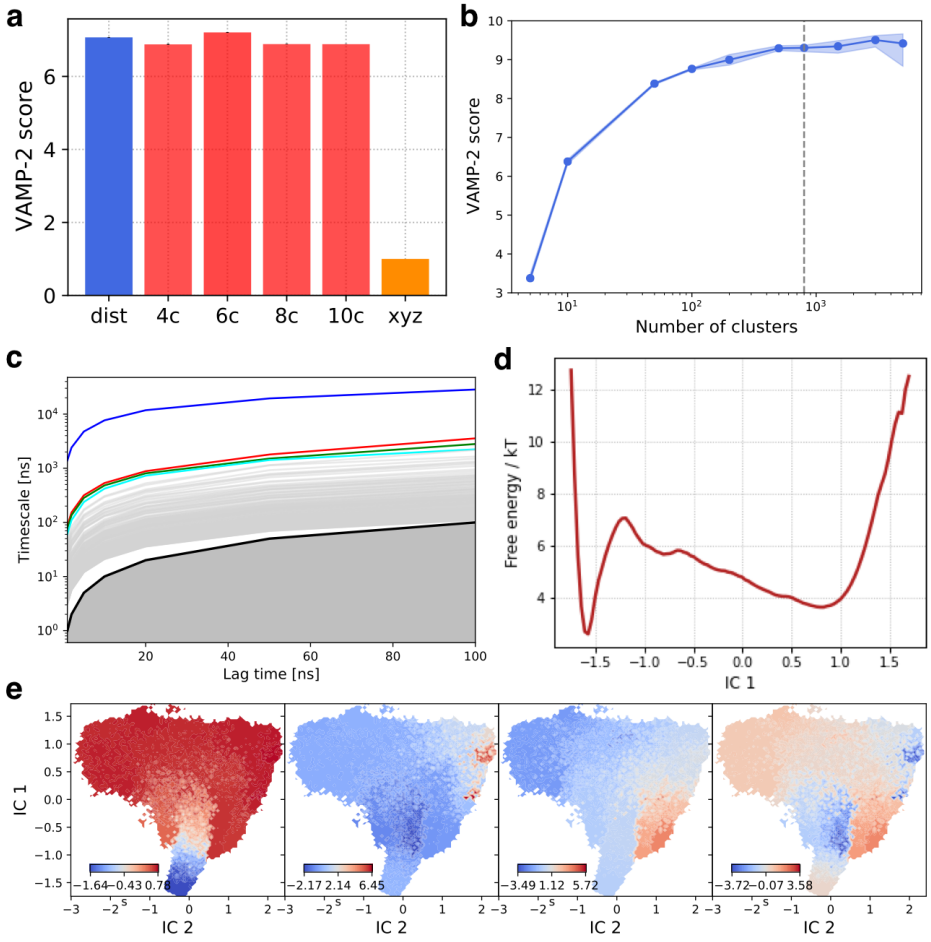
Supplementary Figure 6. MELD folding simulations provide structural models for experimental ψ analysis of transition state ensembles. The representative structures (yellow) are top cluster centers of the simulation during folding transition using the same clustering protocol in Figure S1, superposed with the native structure in white and experimental ψ values.[4,5]



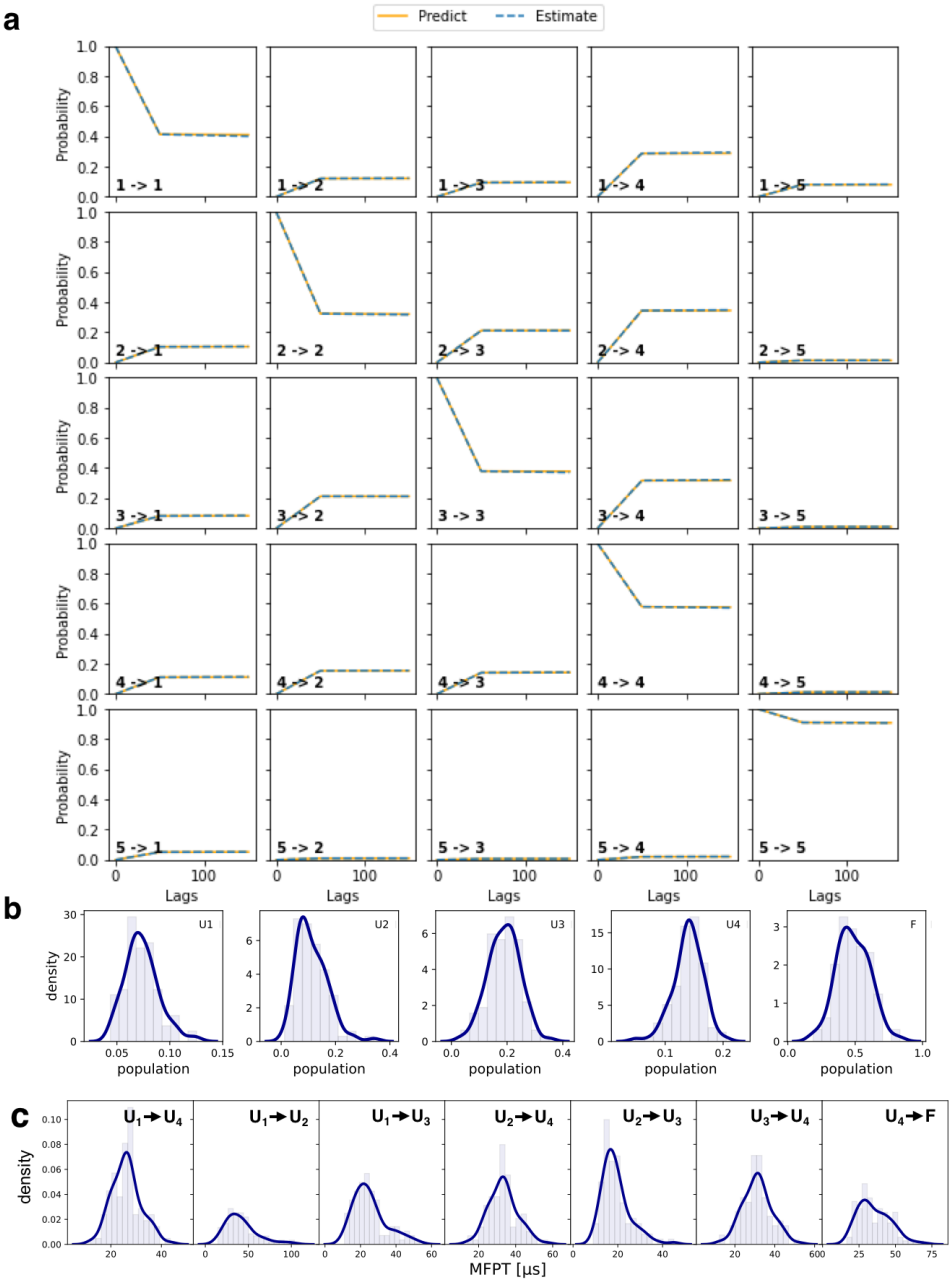
Supplementary Figure 7. The formation of native β-strands induces folding of protein L and its mutant. The time evolution for global native contact fractions of protein L (top) and L_{mut} (bottom) colored by native contact fractions of fragments at walkers with folding transitions. The native contact fraction of α-helix is shown in a white-to-black gradient and the native contact fraction average of β-12 and β-34 is shown in a white-to-blue gradient according to the fraction of native-like contacts in the helix or hairpins respectively.



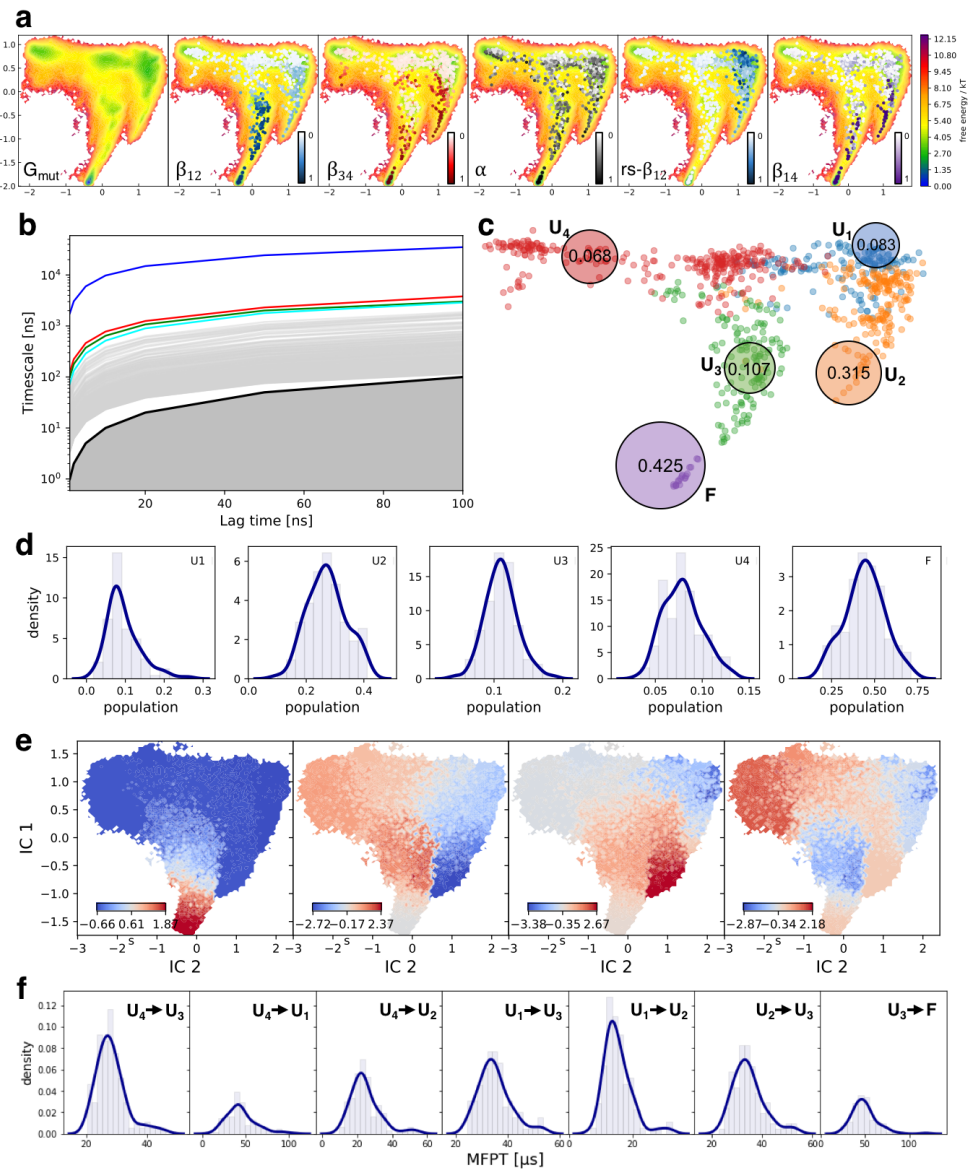
Supplementary Figure 8. Instability of α helix in protein L/L_{mut} in implicit solvent. The time evolution of RMSD for full protein (left), α helix (middle) and β_{1234} helix (bottom) simulated in explicit (black) and implicit solvent (red) at 300 K.



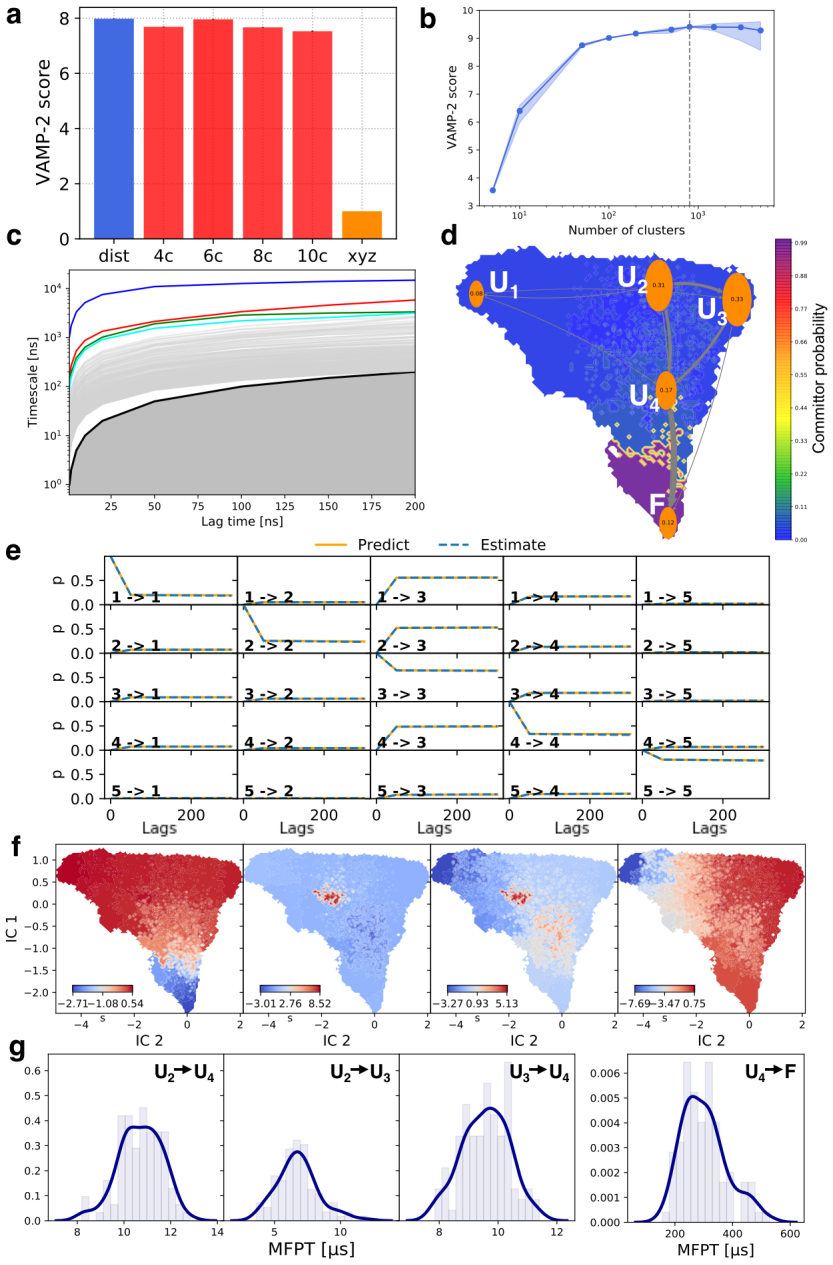
Supplementary Figure 9. Summary of MSM analysis for protein G_{mut}. **(a)** VAMP-2 scores for distance based features of pairwise residues (blue), contact based features (0 or 1) with cutoff 4-10 Å of pairwise residues (red), and aligned backbone cartesian coordinates (orange). **(b)** VAMP-2 scores for estimated Markov models with different number of clusters using k-means clustering method. **(c)** Implied timescale plot of estimated Markov models. **(d)** Reweighted one dimensional free energy profile along the first IC. **(e)** Right eigenvector values for the four slowest processes projected on the first two ICs.



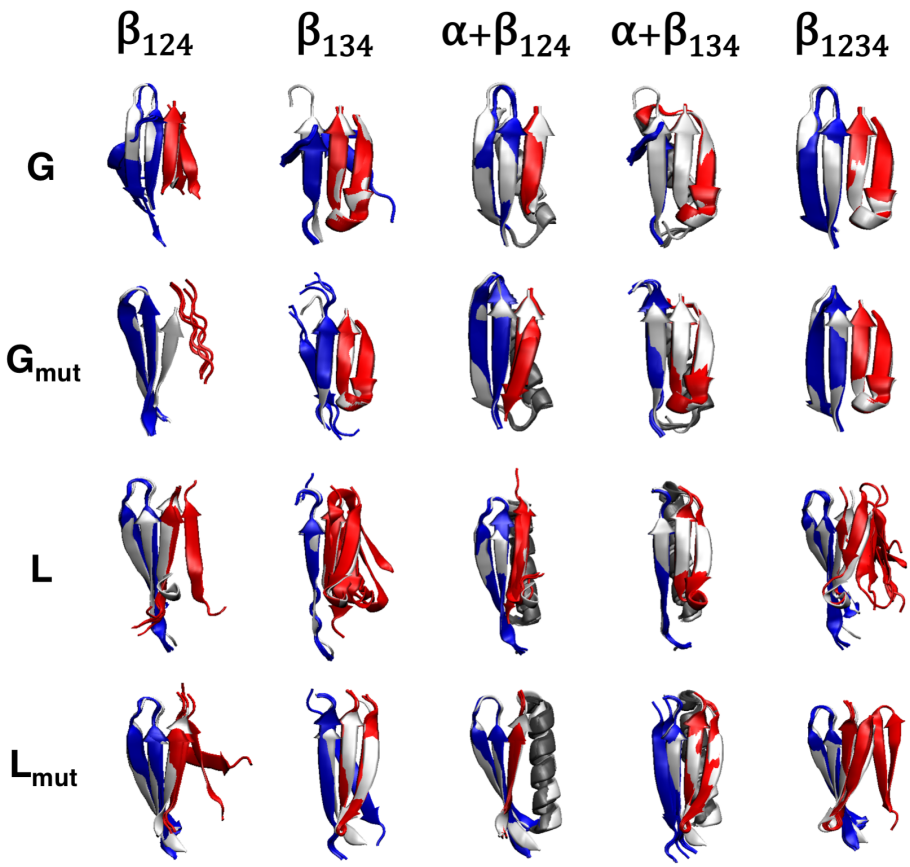
Supplementary Figure 10. Validation and uncertainty of MSM analysis for protein G_{mut}. **(a)** Comparison of the transition probability for every pair of macrostates between the estimate (using MSM computed at various lag times) and the prediction from final model. **(b)** Sampled population estimates of macrostates shown in Figure 4b (U_1 for green, U_2 for blue, U_3 for orange, U_4 for red and U_5 for purple). **(c)** Sampled mean first passage time for transitions between macrostates in Figure 4b .



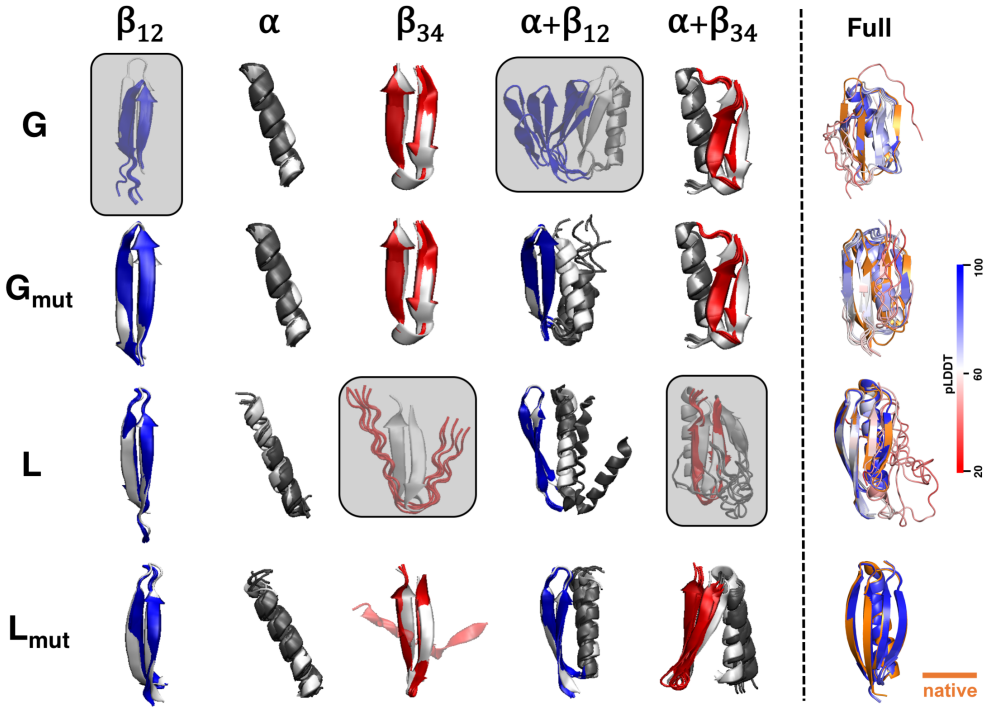
Supplementary Figure 11. MSM analysis for protein G_{mut} using contact based features. The analysis is performed using contact based features with a 6 Å cutoff. **(a)** Two dimensional representation of the free energy landscape projected on the first two ICs. The microstates are colored according to the native contact fraction for different fragments on a white (none)-to-color (all contacts) gradient. **(b)** Implied timescale plot of estimated Markov models at different lag times. **(c)** Population and microstate assignments based on a hidden Markov model with 5 states. **(d)** Sampled population estimates of macrostates shown in **(c)**. **(e)** Right eigenvector values for the four slowest processes projected on the first two ICs. **(f)** Sampled mean first passage time for transitions between macrostates in **(c)**.



Supplementary Figure 12. Summary of MSM analysis for protein G. (a) VAMP-2 scores for distance based features of pairwise residues (blue), contact based features (0 or 1) with a cutoff 4-10 Å cutoff for pairwise residues (red), and aligned backbone cartesian coordinates (orange). **(b)** VAMP-2 scores for estimated Markov models with different number of clusters using a k-means clustering method. **(c)** Implied timescale plot of estimated Markov models at different lag times. **(d)** Five-state hidden Markov state modeling with committor probability projected on the first two ICs. **(e)** Comparison of the transition probability for every pair of macrostates between the estimate (using MSM computed at various lag times) and the prediction from final model. **(f)** Right eigenvector values for the four slowest processes projected on the first two ICs. **(g)** Sampled mean first passage time for transitions with high flux values from transition path theory analysis.



Supplementary Figure 13. Disconnected protein fragment structure predictions from AlphaFold. Five structure models from AF predictions with MSA for fragment β_{124}, β_{134}, β_{124} with α helix, β_{134} with α helix, and β_{1234} for protein G, L and their mutants aligned with native structure (white).



Supplementary Figure 14. Protein fragment and full structure predictions from AlphaFold without MSA. Left: Five structure models from AF predictions for fragment β_{12} (blue), α helix (dark grey), β_{34} (red), β_{12} with α helix, and α helix with β_{34} for protein G, L and their mutants aligned with native structure (white). **Right:** Five structure predictions colored by pLDDT score aligned with native structure (orange) without MSA for input.

References

(1) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of Chemical Theory and Computation* **2013**, *9*, 3084–3095.

(2) Perez, A.; Morrone, J. A.; Brini, E.; MacCallum, J. L.; Dill, K. A. Blind protein structure prediction using accelerated free-energy simulations. *Science Advances* **2016**, *2*, e1601274.

(3) Chang, L.; Perez, A.; Miranda-Quintana, R. A. Improving the analysis of biological ensembles through extended similarity measures. *Physical Chemistry Chemical Physics* **2021**, *24*, 444–451.

(4) Yoo, T. Y.; Adhikari, A.; Xia, Z.; Huynh, T.; Freed, K. F.; Zhou, R.; Sosnick, T. R. The Folding Transition State of Protein L Is Extensive with Nonnative Interactions (and Not Small and Polarized). *Journal of Molecular Biology* **2012**, *420*, 220–234.

(5) Baxa, M. C.; Yu, W.; Adhikari, A. N.; Ge, L.; Xia, Z.; Zhou, R.; Freed, K. F.; Sosnick, T. R. Even with nonnative interactions, the updated folding transition states of the homologs Proteins G & L are extensive and similar. *Proceedings of the National Academy of Sciences* **2015**, *112*, 8302–8307.