A New Class of Protein Misfolding is Observed in All-atom Folding Simulations

Quyen V. Vu[1], Yang Jiang[2], Ian Sitarik[2], Mai Suan Li[1, 3], Edward P. O'Brien[2, 4, 5*]

[1]Institute of Physics, Polish Academy of Sciences; Al. Lotnikow 32/46, 02-668 Warsaw, Poland.

[2]Department of Chemistry, Penn State University; University Park, Pennsylvania, United States.

[3]Institute for Computational Sciences and Technology; Quang Trung Software City, Tan Chanh Hiep Ward, District 12, Ho Chi Minh City, Vietnam.

[4]Bioinformatics and Genomics Graduate Program, The Huck Institutes of the Life Sciences, Penn State University; University Park, Pennsylvania, United States.

[5]Institute for Computational and Data Sciences, Penn State University; University Park, Pennsylvania, United States.

*Corresponding author. Email: epo2@psu.edu

Abstract: A small number of mechanisms intrinsic to a protein's primary structure are known to cause monomeric protein misfolding. Coarse-grained simulations predict a new mechanism of misfolding exists involving off-pathway, non-covalent lasso entanglements, which are separate and distinct from protein knots. These misfolded states can be long-lived kinetic traps, and in some cases are structurally similar to the native state according to the coarse-grained simulations. Here, we examine whether such misfolded states occur in long-time-scale, physics-based all-atom simulations of protein folding. We find they do indeed form, estimate they can persist for weeks, and some have characteristics similar to the native state. These results indicate monomeric proteins can exhibit subpopulations of misfolded, self-entangled states that can explain long-time-scale changes in protein structure and function in vivo.

One-Sentence Summary: Entangled misfolded states form in physics-based all-atom simulations of protein folding and have characteristics similar to the native state.

Only a small number of mechanisms intrinsic to a protein's primary structure are known to cause monomeric protein misfolding (Table 1). Recently, high-throughput, coarse-grained simulations of protein synthesis and folding of the E. coli proteome have suggested there exists an additional widespread mechanism of misfolding[1–3]. Proteins can populate off-pathway misfolded states that involve a change in entanglement[1]. An entanglement is defined by the presence of two structural components: a loop formed by a protein backbone segment closed by a non-covalent native contact, and another protein segment that is threaded through this loop[4–6] (Figs. 1a and 1b). One-third of protein domains have such entanglements present in their native state[6].

The misfolded states observed in the aforementioned simulations involved either the gain of a non-native entanglement (*i.e.*, the formation of an entanglement that is not present in the native ensemble, Table S1) or the loss of a native entanglement (*i.e.*, an entanglement present in the native state fails to form, Table S1)[1–3]. This newly predicted class of misfolding offers an explanation for the decades old observations that non-functional protein molecules can persist for long-time scales in the presence of chaperones and not rapidly aggregate nor be degraded[7–9].

The coarse-grained simulations suggest these misfolded states are often long-lived because to correctly fold they need to change their entanglement state: *e.g.*, an entangled state would need to disentangle, which is energetically costly as already folded portions of the protein would need to unfold (see gray segments in Figs. 1b and 1c). And that these states are likely to bypass cellular quality control mechanisms and remain long-lived in vivo[7–9] because in some cases they are similar in size, shape, and structure to the native state (compare Figs. 1b and 1d).

Two criticisms of these findings are that they are based on models with limited spatial resolution and use force field approximations that have the potential to affect the results. It could be the case, for example, that these states are never populated in a physics-based all-atom model. Further, the coarse-grained model's force field was structure-based, meaning the native state structure was encoded in the van der Waals energy term to energetically favor the native state over other states[10]. Both of these approximations have the potential to affect these earlier findings.

Here, we examine whether a more fundamental, physics-based, all-atom model of the protein folding process exhibits these self-entangled states; and if they do, we test whether those states are off-pathway long-lived traps with properties similar to the native state. To do this, we analyze previously reported all-atom protein folding trajectories of ubiquitin[11] and N-terminal domain (NTD) of λ-repressor[12] starting from their native and unfolded states (Fig. 2a and b). Ubiquitin is a small protein (76 residues) found in eukaryotic organisms and regulates a range of processes including protein degradation and the cell cycle. Ubiquitin folds on the millisecond (ms) timescale under typical conditions in vitro[13] and in about 3 ms in molecular dynamics simulations[11]. The NTD of λ-repressor in the published trajectories[12] is 80 residues in length, binds DNA, and folds on the microsecond timescale[12,14]. Entanglements are not present in the native structure of either of these proteins.

We analyzed whether non-native entanglements are populated during the folding trajectories of Ubiquitin and λ-repressor (Figs. 2c and d) totaling 8 milliseconds and 643 microseconds of simulation time. We observed several non-native entangled states (denoted by high, non-zero G values (eq. S4) in Figs. 2c and d) during folding transitions (indicated by shifts from low to high Q values in Figs. 2c and d). Thus, protein self-entangled states are observed in this class of models. Next we asked whether these entangled states are off-pathway, which means the entangled state

must first unfold to reach the native state. We never observe in these trajectories a direct transition from a (high G, low Q) pair of values to a (low G, high Q) pair of values from one time point to the next. This means these misfolded states are off pathway. These entangled states are clearly short-lived (lasting just a few nanoseconds according to Figs. 2c and d), seemingly at odds with the prediction that these can be kinetically trapped states[2]. However, these simulations were performed at high temperatures (390 K for Ubiquitin and 350 K for λ-repressor), near their melting temperatures in the all-atom force field. Therefore, these states may be able to rapidly disentangle due to these high temperatures.

To test if these states are kinetic traps at a physiological temperature, we performed unrestrained molecular dynamics simulations at 310 K (37 °C). We selected entangled structures that had at least 60% of their native contacts (Q) formed as starting structures, resulting in 21 and 12 structures, respectively, for Ubiquitin and λ-repressor. Structurally, these entangled states are qualitatively similar (Table S2), 20 out of the 21 ubiquitin structures have closed-loops located towards the C-terminus and the threading segment composed of the N-terminus. For λ-repressor, the loop forms towards the C-terminal end, and is very similar in all conformations, and the N-terminal threads through it. Representative structures for these entangled states are shown in Figs. 2e and f. Three trajectories were started from each of these conformations and the simulations were run for 700 ns, or until the entanglement was lost (*i.e.,* $G = 0$ was reached).

For the 63 trajectories (=21 x 3) of ubiquitin, 71% (45 out of 63) contain a non-native entanglement that persists for 700 ns (fig. S2a and table S2). On average then, the time to disentangle is approximately 2.1 μs (estimated using eq. S6, setting $t = 700$ ns). For λ-repressor, 17% of trajectories (6 out of 36) persist in an entangled state for 700 ns (fig. S2b and table S2). The average time to disentangle for λ-repressor is approximately 390 ns (estimated using eq. S6 setting $t = 700$ ns), 50 times faster than for ubiquitin entanglements.

Relative to their time-scale of protein folding, these entangled states are not long-lived kinetic traps even at 310 K. Based on the previous coarse-grained simulation results, this is to be expected[1], as ubiquitin and NTD are atypically small (the median protein length in yeast is 379 residues[15]). In such very small proteins, entanglements compose a large proportion of the total protein structure present – and hence, it is easier to disentangle as most of the protein structure is misfolded and less stable. For example, for the entanglements reported in Table S2, 79% of ubiquitin's primary structure is involved in the entanglement (*i.e.*, either part of the loop or threading segment, Fig. 2g), and 83% of λ-repressor's primary structure is involved in the entanglement (Fig. 2h). In contrast, a typical length protein with a misfolded entanglement present will need to unfold large amounts of already folded structure to allow disentanglement (Figs. 3a-d). Unfolding for these proteins is energetically more costly, and hence near-native entanglements in large proteins are more likely to be long-lived kinetic traps[2].

Unrestrained all-atom simulations are incapable of folding such large proteins on tractable simulation time scales. Therefore, to test this "size-effect" hypothesis we calculated the lifetime of a near-native entangled state for a larger protein, E. coli's 4-Diphosphocytidyl-2-C-Methyl-D-Erythritol Kinase (Figs. 3a-d; 283 residues; Gene: ispE). This protein was chosen from our dataset of proteins that exhibit entanglements in coarse-grained simulations[10]. We carried out in silico temperature jump experiments on an all-atom near-native entangled state of this enzyme that was created by back-mapping coarse-grained conformations to an all-atom representation.

Arrhenius analysis was then carried out (Fig. 3e), corrected for the observed acceleration in unfolding rates in all-atom simulations (see Methods), to estimate the lifetime of this entangled state at 298 K. (This temperature was used to allow comparison to experimental unfolding data measured at this temperature; see SI Method for full details). These simulations indicate this enzyme's entangled state takes 1.008×10^{6} s (95% Confidence Interval: [682.11s, 0.985×10^{9}s]) to disentangle, corresponding to 280 hours or 11.6 days (Fig. 3e). This indicates that for a typical sized protein, these self-entanglements are long-lived kinetic traps.

Finally, we analyzed how native-like are these three protein's entangled states. The radius of gyration of these states are comparable to the native state (within 5.5%, 5.2%, and 2.4%, on average, respectively, of the native state Ubiquitin, λ-repressor, and 4-Diphosphocytidyl-2-C-Methyl-D-Erythritol Kinase values), indicating the entangled and native states are of similar sizes. These entangled states are estimated to have a solubility similar to the native state (Table S2) according to a model that accounts for surface area chemical properties[1]. This indicates these misfolded states are likely to remain as soluble in solution as the native fold. Next, we analyzed the secondary structure of proteins using Stride software[16] and find that the non-native entangled states contain up to 80% of the secondary structure found in the native state (Table S2). Thus, the size, solubility, and secondary structure analyses indicate these misfolded states are similar to the native state.

In conclusion, non-native self-entanglements occur during protein folding in all-atom models, some of these states are long-lived, and have properties similar to the native state. This new class of protein misfolding opens up novel avenues of research including new targets for drug design, expanding our understanding of protein homeostasis in cells, and the impact of these states on protein function.

References

1. Jiang, Y. *et al.* How synonymous mutations alter enzyme structure and function over long time scales. *bioRxiv* 2021.08.18.456802 (2022) doi:10.1101/2021.08.18.456802.
2. Nissley, D. A. *et al.* Universal protein misfolding intermediates can bypass the proteostasis network and remain soluble and less functional. *Nat. Commun.* **13**, 3081 (2022).
3. Halder, R., Nissley, D. A., Sitarik, I. & O'Brien, E. P. Subpopulations of soluble, misfolded proteins commonly bypass chaperones: How it happens at the molecular level. *bioRxiv* (2021) doi:10.1101/2021.08.18.456736.
4. Baiesi, M., Orlandini, E., Trovato, A. & Seno, F. Linking in domain-swapped protein dimers. *Sci. Rep.* **6**, 1–11 (2016).
5. Baiesi, M., Orlandini, E., Seno, F. & Trovato, A. Exploring the correlation between the folding rates of proteins and the entanglement of their native states. *J. Phys. A Math. Theor.* **50**, (2017).
6. Baiesi, M., Orlandini, E., Seno, F. & Trovato, A. Sequence and structural patterns detected in entangled proteins reveal the importance of co-translational folding. *Sci. Rep.* **9**, 1–12 (2019).
7. Komar, A. A., Lesnik, T. & Reiss, C. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Lett.* **462**, 387–391 (1999).

8. Zhou, M. *et al.* Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature* **494**, 111–115 (2013).

9. Zhou, M., Wang, T., Fu, J., Xiao, G. & Liu, Y. Nonoptimal codon usage influences protein structure in intrinsically disordered regions. *Mol. Microbiol.* **97**, 974–987 (2015).

10. Nissley, D. A. *et al.* Electrostatic Interactions Govern Extreme Nascent Protein Ejection Times from Ribosomes and Can Delay Ribosome Recycling. *J. Am. Chem. Soc.* **142**, 6103–6110 (2020).

11. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. Atomic-level description of ubiquitin folding. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 5915–5920 (2013).

12. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science (80-.).* **334**, 517–520 (2011).

13. Sivaraman, T., Arrington, C. B. & Robertson, A. D. Kinetics of unfolding and folding from amide hydrogen exchange in native ubiquitin. *Nat. Struct. Biol.* **8**, 331–333 (2001).

14. Yang, W. Y. & Gruebele, M. Folding at the speed limit. *Nature* **423**, 193–197 (2003).

15. Brocchieri, L. & Karlin, S. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res.* **33**, 3390–3400 (2005).

16. Frishman, D. & Argos, P. Knowledge-based protein secondary structure assignment. *Proteins Struct. Funct. Bioinforma.* **23**, 566–579 (1995).

17. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. THEORY OF PROTEIN FOLDING: The Energy Landscape Perspective. *Annu. Rev. Phys. Chem.* **48**, 545–600 (1997).

18. Sarkar, D., Kang, P., Nielsen, S. O. & Qin, Z. Non-Arrhenius Reaction-Diffusion Kinetics for Protein Inactivation over a Large Temperature Range #. *ACS Nano* **13**, 8669–8679 (2019).

19. Gosavi, S., Chavez, L. L., Jennings, P. A. & Onuchic, J. N. Topological frustration and the folding of interleukin-1β. *J. Mol. Biol.* **357**, 986–996 (2006).

20. Sherman, M. Y. & Qian, S. B. Less is more: Improving proteostasis by translation slow down. *Trends Biochem. Sci.* **38**, 585–591 (2013).

21. Kiefhaber, T., Kohler, H. H. & Schmid, F. X. Kinetic coupling between protein folding and prolyl isomerization. I. Theoretical models. *J. Mol. Biol.* **224**, 217–229 (1992).

22. Kikis, E. A. The intrinsic and extrinsic factors that contribute to proteostasis decline and pathological protein misfolding. *Advances in Protein Chemistry and Structural Biology* vol. 118 (Elsevier Ltd, 2019).

23. Salahuddin, P., Siddiqi, M. K., Khan, S., Abdelhameed, A. S. & Khan, R. H. Mechanisms of protein misfolding: Novel therapeutic approaches to protein-misfolding diseases. *J. Mol. Struct.* **1123**, 311–326 (2016).

24. Lafita, A., Tian, P., Best, R. B. & Bateman, A. Tandem domain swapping: determinants of multidomain protein misfolding. *Curr. Opin. Struct. Biol.* **58**, 97–104 (2019).

Acknowledgments: We thank D.E Shaw's group for providing access to their MD trajectories.

Funding:

National Science Centre, Poland grant 2019/35/B/ST4/02086 (MSL)

Department of Science and Technology, Ho Chi Minh City, Vietnam grant 13/2020/HĐ-QPTKHCN (MSL)

National Science Foundation grant MCB-1553291 (EPO)

National Institutes of Health grant R35-GM124818 (EPO)

This research was supported in part by PLGrid and TASK Infrastructure in Poland

Author contributions:

Conceptualization: EPO

Methodology: YJ, IS, EPO

Investigation: QVV, YJ, IS, MSL, EPO

Visualization: QVV, YJ, IS

Funding acquisition: MSL, EPO

Project administration: MSL, EPO

Supervision: MSL, EPO

Writing – original draft: QVV, YJ, IS, EPO

Writing – review & editing: QVV, YJ, IS, MSL, EPO

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: Full details of all-atom molecular dynamics simulations and entanglement calculations are included in the supporting information (PDF).

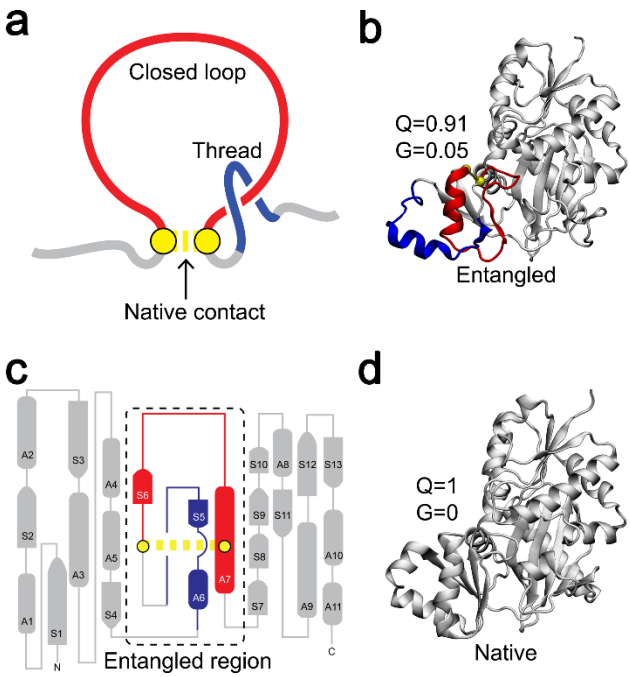

Fig. 1. Visualizing entanglements through various representations. (a) An illustration of the two geometric elements that compose an entanglement: the closed loop is colored in red and the threading segment is in blue. The loop is closed by a non-covalent contact between two residues (yellow). (b) 3-D structure of an entangled state from the protein D-alanine—D-alanine Ligase B (DDLB; the closed loop and crossing section of the threading segment of their entangled regions are colored in red and blue, respectively) taken from previously reported coarse-grained simulations[1]. Q and G correspond, respectively, to the fraction of native contacts formed in the structure, and the fraction of native contacts that exhibit a change in entanglement. (c) A flattened 2-d structure representation of the entangled state of DDLB shown in panel (b). (d) A 3-D structure of the crystal structure of DDLB. All native contacts are formed in the crystal structure (Q=1) and there is no entanglement change (hence, $G = 0$). In panels (a), (b) and (c) the coloring scheme of the relevant elements is the same.

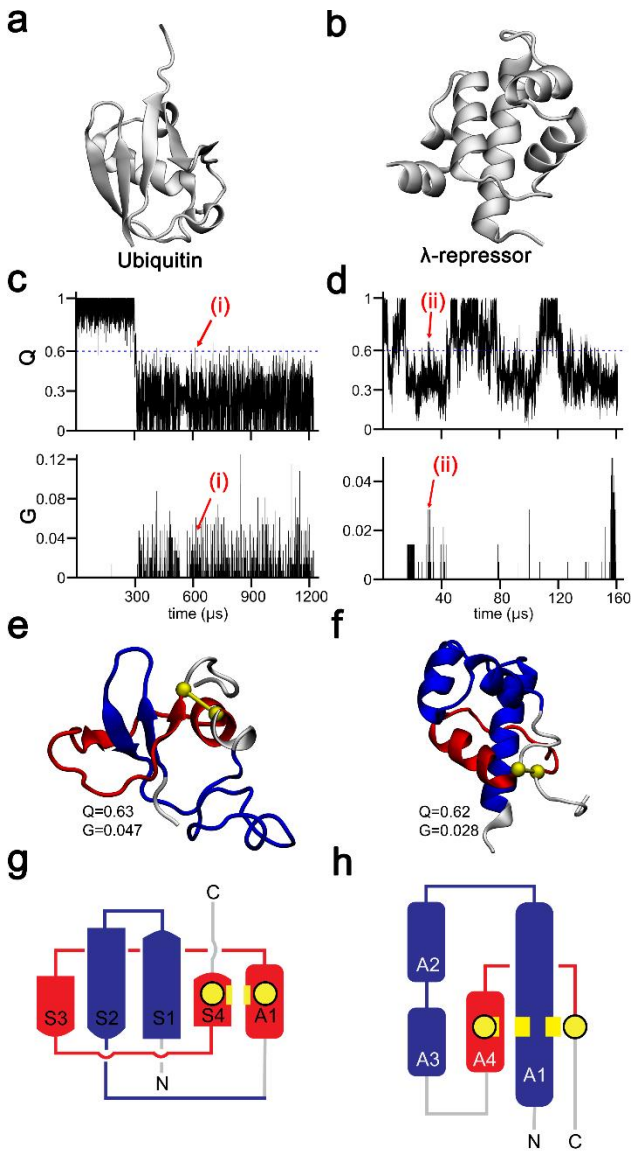

Fig. 2. Misfolded entanglements are observed in all-atom protein folding simulations. (a) Ubiquitin's native structure. (b) λ-repressor's native structure. (c) Upper panel: Ubiquitin's fraction of native contacts Q versus time from all-atom simulations, the blue-dashed line represents $Q = 0.6$. Lower panel: Ubiquitin's degree of entanglement G versus time for the same trajectory. (d) Same as Fig. 2c, but for λ-repressor. (e) 3-D structure of a Ubiquitin entangled state observed at the time point labeled (i) in Fig. 2c, with entanglement elements colored as in Fig. 1a. (f) Same as Fig. 2e, but for λ-repressor, and from point labeled (ii) in Fig. 2d. (g) A flattened secondary structure representation for the entangled state in Fig. 2e. (h) Same as Fig. 2g, but for λ-repressor.

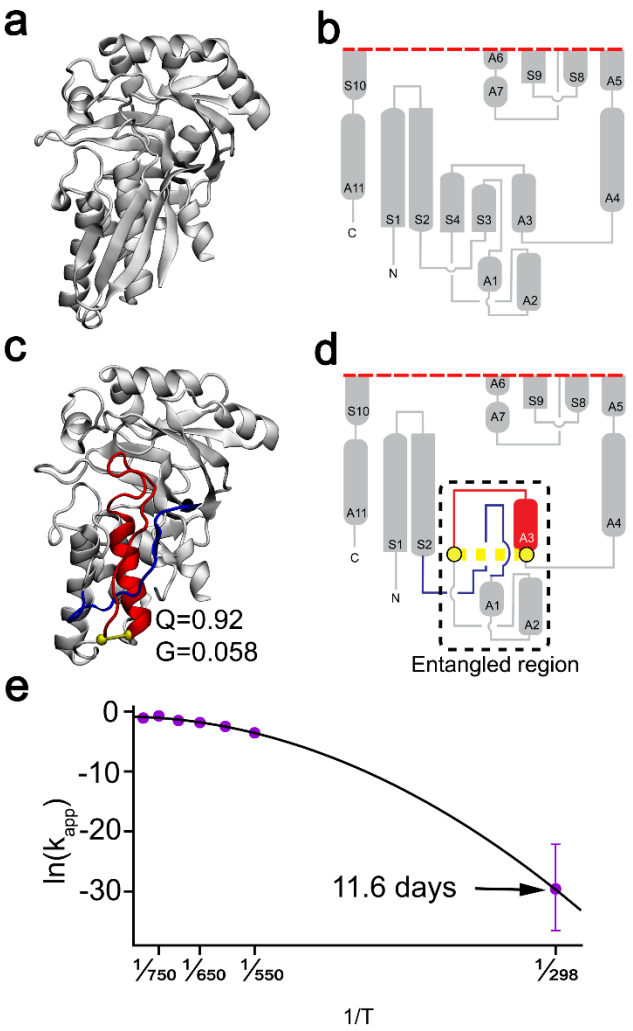

Fig. 3. A misfolded entanglement in a larger protein is estimated to persist for many days. (a) Native structure of 4-Diphosphocytidyl-2-C-Methyl-D-Erythritol Kinase (PDB ID: 2WW4). (b) Flattened secondary structure of 4-Diphosphocytidyl-2-C-Methyl-D-Erythritol Kinase in the native state. Structure above the dashed red line is not shown as it does not involve the entanglement that forms. (c) Entangled structure of 4-Diphosphocytidyl-2-C-Methyl-D-Erythritol Kinase. 92% of native contacts are present in this structure. (d) Flattened secondary structure of 4-Diphosphocytidyl-2-C-Methyl-D-Erythritol Kinase in the entangled state shown in panel (c). (e) Arrhenius plot, with a super-Arrhenius fit[17,18], error bars of the natural logarithm of disentanglement rate at high temperatures are smaller than symbols, the extrapolated lifetime of the gain in entanglement at 298 K is about 11.6 days (with corrected for the acceleration in unfolding rates observed in simulations). Error bars represent the 95% confidence interval about the mean.

Table 1. Mechanisms of protein misfolding[19–24]

Intrinsic to the primary structure	Extrinsic to the primary structure
• Proline isomerization • Intra-chain domain swapping • Out-of-register β-strands • Incorrect helix packing • Mispacking of sidechains • Alternative 2nd or 3rd structure • Backtracking (frustration) • Incorrect disulfide bond formation	• Transcription & Translation errors • mRNA splicing errors • Perturbation of temperature, oxidative stress, pH, osmotic pressure, protein concentration, chaperones, processing enzymes • Soluble & insoluble aggregation • Genetic mutations • Erroneous complex formation • Improper/incomplete degradation • Cofactor and ligand concentrations • Post-translational modifications • Inter-chain domain swapping