Title: Non-Refoldability is Pervasive Across the *E. coli* Proteome

Authors: Philip To,[1] Briana Whitehead,[2] Haley E. Tarbox,[1] Stephen D. Fried[1,2]*

Affiliations:

[1] Department of Chemistry, Johns Hopkins University, Baltimore, MD, USA

[2] Department of Biophysics, Johns Hopkins University, Baltimore, MD, USA

*Correspondence to: sdfried@jhu.edu

Abstract: Decades of research on protein folding have primarily focused on a privileged subset of small proteins that can reversibly refold out of a denaturant. However, these studies are not representative of the complexity of natural proteomes, which consist of many proteins with more sophisticated architectures. Here, we introduce an experimental approach to probe protein refolding for whole proteomes. We accomplish this by first unfolding and refolding *E. coli* lysates, and then interrogating the resulting protein structures using a permissive protease that preferentially cleaves at flexible regions. Using mass spectrometry, we globally analyze the digestion patterns to assess structural differences between native and "refolded" proteins. These studies show that roughly half of the *E. coli* proteome cannot reassemble their native states following chemical denaturation. Our results imply that thermodynamics alone cannot specify the native structures of all proteins, signaling a pervasive role for kinetic control in shaping protein biogenesis.

One Sentence Summary: Many proteins cannot put themselves back together again after being taken apart, suggesting that their sequences do not encode sufficient information to fold them.

Short Title: Non-refoldable Proteins

Main Text:

The foundational paradigm of protein folding is that the primary sequence of a protein contains all the information needed for it to adopt a specific native structure (*1*, *2*). This remarkable property is explained by Anfinsen's thermodynamic hypothesis (*3*, *4*), which asserts that because native states minimize the Gibbs free energy of a protein molecule under physiological conditions, proteins can reliably navigate to their native structures and remain in those states by following a free energy gradient (*5*).

Most experiments of protein folding are conducted on purified, small, single-domain soluble proteins, and employ reporters such as circular dichroism (CD), nuclear magnetic resonance (NMR), or Förster resonance energy transfer (FRET) to follow the proportion of protein molecules that are folded as a function of time, temperature, denaturant concentration, or sequence (*6–8*). Studies of this kind, conducted over the previous 50 years, have yielded immense insight into the molecular determinants that underpin stable globular folds (*9*, *10*). However, our reliance on the thermodynamic hypothesis and two-state models as ground truths to interpret these experiments has severely limited our ability to study the folding of complex proteins or to even consider alternative non-thermodynamic scenarios (*11, 12*).

In the following, we describe an approach that leverages the high-throughput sequencing capability of mass spectrometry to monitor the refolding kinetics of whole proteomes with a high level of structural resolution. Our experiments reveal that following denaturation, many proteins are incapable of navigating back to their native structures. Importantly, the distribution we observe of this behavior signals that co-translational folding plays an indispensable role in the assembly of several classes of proteins.

Approach.

In our experiments (Fig. 1A), *E. coli* cells are lysed under native conditions and the clarified extract divided so that a portion is retained in its native state; another portion is unfolded by addition of high concentrations of chemical denaturants (8 M urea or 6 M guanidinium chloride (GdmCl)) and refolded by lowering the denaturant concentration (by dilution or by dialysis). Following a period of time, the structures of the proteins in these complex mixtures are probed by subjecting them to pulse proteolysis with proteinase K (PK). A low protease loading, active for a brief period of time (1 min), ensures that PK cleaves at only

exposed or unstructured sites of target proteins, thereby encoding structural information of the protein's conformation into cleavage sites (*13*). Proteins are then diluted, fully trypsinized, and prepared for analysis by liquid chromatography/tandem mass spectrometry (LC-MS/MS). Half-tryptic peptides (HTPs), in which one cut-site is deemed to arise from trypsin (which only cuts after Arg and Lys) and the other cut-site from PK (which can cut between any two residues), reveal a location where PK cleaved. By quantifying the relative abundance of HTPs arising from refolded proteins as compared to native proteins, one can assess sites in a protein where the local conformation is different. This method, called limited proteolysis MS (LiP-MS) has been instrumental in exploring conformational change (*14*), thermostability (*15*), and allosteric binding (*16*) on the proteome-scale, and here we adapt it to probe protein folding.

A reversibly-refolding protein will have no 'memory' of being unfolded, and the refolded protein will equilibrate to the same ensemble of conformations. Therefore, the profile (and quantity) of HTPs will be identical between the native and refolded samples. On the other hand, aggregated proteins might be expected to be more resistant to PK cleavage, and soluble misfolded proteins (with less compacted hydrophobic cores) might be expected to be more susceptible to PK cleavage. To critically test these hypotheses, we first performed our LiP-MS method on two purified model proteins that are known refolders: Staphylococcal nuclease (SNase) and Ribonuclease H from *Thermus thermophilus* (*Tt*RNase H; fig. S1). SNase refolded by dilution out of urea has a CD spectrum that superimposes on that of the native protein recombinantly expressed from *E. coli* (Fig. 2A). Likewise, when refolded and native SNase are probed with LiP-MS, they generate a set of 147 distinct tryptic and half-tryptic peptides (Fig. 2C) that are all present in equal abundances within our cut-offs for significance (abundance ratio greater than 2-fold, P-value by Welch's t-test < 0.01). Importantly, when the same analysis is conducted on SNase spiked into *E. coli* lysate (Fig. 2D), the conformations of native and refolded protein are again indistinguishable. We repeated these studies on *Tt*RNase H, and found that its native and refolded forms generated overlapping CD spectra (Fig. 2B), and no significant differences in PK cleavage patterns – both in isolation (Fig. 2E; over 176 peptides) and when spiked into *E. coli* lysate (Fig. 2F; over 147 peptides). These studies show that LiP-MS provides a consistent picture with CD for refolding proteins, although does so with much greater structural resolution (providing independent quantifications at many distinct sites across the protein), and that thermodynamic refolding can be observed in a complex mixture (*17*). Moreover, when these

same experiments are conducted on SNase and *Tt*RNase refolded via slow dialysis from denaturant (rather than rapid dilution), we again found that the native and refolded forms were indistinguishable (fig. S2), consistent with the notion that if a protein's native state ensemble populates its free energy landscape's thermodynamic minimum, correct refolding can occur regardless of the details of how the denaturant is removed.

Refoldability of the Proteome.

We next planned to carry out a similar experiment in which total soluble cellular extracts are unfolded, refolded, and then compared to a native lysate. However, a limitation that we considered is that many proteins in the *E. coli* lysate may aggregate during refolding. Proteins that aggregate could hypothetically interfere with other proteins in the mixture that could in principle refold on their own. To address this possibility, we unfolded and refolded *E. coli* lysates under a range of conditions and performed pelleting assays seeking conditions that minimized aggregation (figs. S3-S4). We found that precipitation could be minimized when refolding is conducted at an elevated pH (8.0) and low concentration (0.23 mg mL^{-1}). We speculate this may be because the average isoelectric points (pI) among *E. coli* proteins is 6.6 (*18*), and so at pH 8.0 proteins will be preponderantly negatively charged, disfavoring nonspecific associations. Thermal denaturation followed by slow cooling generates substantially more precipitation (fig. S4C), suggesting this may be an unideal strategy for probing the refoldability of individual proteins in complex mixtures (*19*). Dynamic light scattering (DLS) experiments confirm that thermally-refolded lysates are highly enriched with large particle sizes (>0.2 μm; fig. S5), whilst chemically-refolded lysates had particle diameter distributions somewhat larger than (or similar to) native lysates. Collectively, these studies suggest refolding via removal of chemical denaturants at high dilution afford proteins in lysates an opportunity to refold (figs. S5-S6).

We performed proteome-wide studies (Fig. 1A), unfolding *E. coli* lysates in both urea and GdmCl, initiating refolding by dilution, and recording refolding kinetics over time by pulsing samples with PK after discrete time increments (Fig. 1B). In these studies, we analyzed samples both with and without PK treatment (controls). Comparison of the native and refolded controls enable the determination of overall changes in protein abundance (due to, for example, precipitated protein in the refolded sample) and are used as normalization factors (fig. S7A-B),

so that abundance differences in peptide fragments can be attributed solely to structural changes in refolded proteins. Moreover, because our conclusions rely heavily on missing data (a non-refolding protein will generate signature peptide fragments that are absent in its native form), we developed a new filtering algorithm for label free quantification that can reproducibly quantify peptide abundance over nine orders of magnitude (fig. S7C–F, Supplementary Text 1).

Inspection of the dataset in which lysates were diluted out of GdmCl reveals that the distribution of abundance ratios for the 21,386 peptides sequenced (Fig. 3A-B) follows a trimodal distribution, with half (50.6%) of the peptides present at similar abundance in the native and refolded samples, and a significant portion present only in native or refolded (and not detectable in the other). This distribution implies that many regions of *E. coli* proteins *failed* to refold, generating signature fragments that are completely absent in the native sample. HTPs (signifying exposed sites) are asymmetrically distributed and much more abundant in the refolded sample (Fig. 3B). This finding is consistent with the view that most proteins that fail to refold under our conditions form soluble disordered entities (rather than aggregate), which are more susceptible to PK cleavage.

We define a protein as 'non-refoldable' if we can map two or more peptides to it with significantly different abundances in the refolded sample (though our primary findings are not sensitive to this cut-off (Table S1)). Proteins with only one peptide mapped to it are considered to have too little data to make an assessment. In total, we quantified 21,386 peptides (mapping to 1,426 proteins) in our experiment, of which 20,274 peptides (from 1,389 proteins) passed our stringent filters and were quantified by our LFQ algorithm. Among the set of proteins with two or more mapped peptides, 477 (43%) were found not to be thermodynamic refolders (Fig. 3C). Between technical replicates, ~90% of proteins were assigned the same refoldability status (fig. S7G-J).

We next sought to identify features of proteins that correlate with non-refoldability (Table S2, Supplementary Text 2). Indeed, multimeric proteins showed a much lower proficiency to refold, whilst monomeric proteins had a high propensity to refold (P-value $< 10^{-16}$ by chi-square test; Fig. 3D). Since natural protein folding on a polysome provides a means to couple folding and assembly, operon structure could facilitate biogenesis of multimeric proteins in a manner not recapitulated by *in vitro* refolding (*20*, *21*). Dimers with simple or polar

interfaces such as superoxide dismutase refold more easily (Fig. 3E) than dimers with complex interfaces (Fig. 3F) for which synthesis, folding, and assembly may require more coordination (*22*, *23*). We found that trimers are much less refoldable than dimers and remarkably, also less refoldable than tetramers or hexamers. This may be because higher symmetry allows tetramers (or hexamers) to form in multiple steps, whereas refolding of trimers likely requires an inefficient termolecular process (*24*).

We also observed that proteins with more domains tend to refold significantly less efficiently (Fig. 3G), consistent with a prevailing view that the vectorial synthesis of translation decouples domains into independently folding units (*25–27*), thereby preventing non-native inter-domain interactions that could form during refolding from denaturant. Strikingly, proteins with split domain architecture such as inosine 5'-monophosphate dehydrogenase (Fig. 3H) refolded exceptionally poorly (83% non-refoldable), suggesting a scenario where co-translational coordination is most critical (fig. S8A). Moreover, we observe evidence for coupling of domain refolding, such that domains in a multidomain protein typically all refold or all fail to refold (fig. S8B-D).

Assigning the quantified peptides into domains (rather than proteins) revealed fold-types with strikingly different levels of intrinsic refoldability (Fig. 3I, Table S3). Satisfyingly, fold-types exemplified by model proteins are typically efficient refolders (e.g., ribonuclease H-like (*28*; 30% non-refoldable) and immunoglobulin-like domains (*29*; 0% non-refoldable)), whereas other less-characterized folds display markedly lower levels of refoldability (e.g., ferritin-like (75%) and class II aaRS-like (80%)).

Proteins of greater molecular weight and essential proteins were found to be less refoldable (fig. S10A,G). However, these observations are confounded by the fact that we generally quantify more peptides for massive/essential proteins, potentially making it "easier" to detect a significant structural difference (fig. S10C,I) within them. Importantly, this bias does not affect the other key trends noted (figs. S9B, S10F, S11C), though we do generally observe more peptides from proteins with more domains (fig. S9D, Supplementary Text 3). Alternatively, by lumping together all peptides associated with a given attribute (without regard to which specific protein they came from), this bias vanishes (Tables S4-S5). Most of the key

trends we observe are *more* statistically significant at the "peptide level" (figs. S9A,E, S10E, S11E).

Unexpectedly, proteins with cofactors were not generally worse refolders, suggesting that apo-proteins are generally adept at refolding into an intermediate that can select the correct cofactor from a complex mixture (fig. S11D-E). Assigning proteins into cellular compartments and cellular functions shows ribosomal proteins and those involved in translation are the poorest refolders, whilst soluble regions of inner membrane proteins are surprisingly efficient refolders (figs. S11, S12; see Supplementary Text 4). Most of these trends (with several intriguing exceptions) were recapitulated in separate experiments in which the proteome was refolded out of urea (figs. S13-S15, Tables S6-S7), and most of the proteins quantified in both the urea and GdmCl experiments had the same refoldability status (fig. S15).

Kinetics of Refolding Across the Proteome.

We found that most proteins that *can* refold do so within the first minute (Fig. 4A, fig. S16, Table S8). This is consistent with the corpus of classical protein folding studies, which typically record folding transitions on the ms–s timescales (*30*, *31*). Nevertheless, we identified 130 "slow" refolders whose structural dynamics can be probed on the timescale of this experiment (fig. S16C). For instance, triose phosphate isomerase (TIM) is refoldable, however the head of the protein barrel that abuts the active site refolds on the minute-timescale, as observed by the decay kinetics monitored at four independent sites (Fig. 4B). This finding is consistent with previous work that showed TIM requires a minute to regain catalytic activity (*32*). Overall, the refolding proteome was more susceptible to PK after 1 minute of refolding (fig. S16A,B) and as time progressed up to 2 h, it became more native-like (fig. S16F). We could identify kinetic trends toward native structures at 3569 distinct sites, of which the majority (64%) became native between 1–5 min (fig. S16G, Supplementary Text 5). Several proteins evinced refolding on multiple slow timescales (min–h) at distinct locations (fig. S17, S18), which could be attributed to formation of a split domain (fig. S17A,B), kinetic decoupling between distinct domains (fig. S18), or uptake of a divalent metal cation, which appears to frequently be rate-limiting (figs. S17C,D, S18A,B, S19).

Proteins with more subunits (except trimers) or heavier proteins were not substantially slower at refolding than their monomeric or lighter peers at the min-timescale (Fig. 4C,D, fig.

S20A,D). This suggests that the translational entropy loss and the greater conformational entropy deficit associated with refolding these kinds of proteins do not pose insurmountable barriers *per se*. We found intriguing relationship between refolding kinetics and isoelectric point (pI; Fig. 4E, fig. S10D-F, S16D, S17, S20C). Proteins with low pI (<5) are highly negatively charged and fold slowly, likely because of the internal charge repulsion within the chain. Proteins with pI close to cellular pH (7 < pI < 8) have a high propensity to fold very rapidly: this phenotype likely evolved as a defense mechanism against aggregation because charge-neutral proteins are the most prone to aggregate (*33*, *34*).

Different fold-types showed distinct kinetic behaviors. Ribonuclease H-like domains and OB-folds tend to refold slowly, whereas Rossmann folds, 3-helix bundles and thioredoxin folds are enriched amongst fast refolders (Fig. 4F, fig. S19A,B, fig. S20E-G). TIM barrels were generally poorer refolders than other fold-types, particularly after given 5 min of refolding time.

Quite unexpectedly, we found an *inverse* correlation between non-refoldability and which proteins are substrates for the *E. coli* chaperonin GroL/GroS *in vivo* (class III proteins (*35*); Table S9), such that class III proteins are much less likely to be non-refoldable (39%) than non-GroL/GroS substrates (77%). This observation drives home the point that chaperonins cannot fundamentally alter an individual protein's free energy landscape (rendering a non-refoldable protein refoldable), but rather mitigate aggregation risk (*36*). This function is not required in our experiments because we minimize aggregation through elevated pH and high dilution (fig. S3-6). Hence, chaperonins appear to select substrates that – if shielded from aggregation – are capable of refolding on their own via a thermodynamic gradient. The observation that TIM barrels are known to be common chaperonin substrates might be explained by the fact these abundant proteins have a high propensity to remain partially folded after 5 min, a timescale over which aggregation might occur in more crowded environments (*34*). Satisfyingly, we find that most (70%) of a small set of "kinetically stable" proteins are non-refoldable (P-value = 7×10^{-4} by Fisher's exact test)(*37*), suggesting that non-refoldable proteins may generally unfold so slowly that cells would simply never need them to refold.

Conclusion.

Mass spectrometry proteomics has evolved in recent years to identify and quantify thousands of proteins in complex mixtures, though this power has seldom been applied to

address biophysical questions. Here we apply these tools to survey protein refolding globally, sensitively, and with higher structural resolution than afforded by many biophysical techniques. Our data show that many *E. coli* proteins – especially those with many subunits or particular domains – cannot relocate their native states solely by following a thermodynamic gradient. We speculate that the free energy landscapes of such proteins may be laden with local (or even global) minima that correspond to non-native soluble structures, for which biological processes with kinetic control (e.g., translation) are needed to avoid. In such a scenario, native states of many proteins may not be thermodynamically stable but kinetically *persistent* (*12*, *37*). In conclusion, we have greatly expanded the number of proteins whose folding has been interrogated, and have shown that many proteins do not follow the same paradigm as extensively-studied model systems (*38*).

References and Notes:

1. M. Karplus, Behind the folding funnel diagram. *Nature Chem. Biol*. **7**, 401–404 (2011).
2. K. Dill, H. Chan, From Levinthal to Pathways to Funnels. *Nature Struct. Biol.* **4**, 10–19 (1997).
3. C. B. Anfinsen, Principles That Govern the Folding of Protein Chains. *Science* **181**, 223–230 (1973).
4. C. B. Anfinsen, E. Haber, M. Sela, F. H. White, The Kinetics of Formation of Native Ribonuclease During Oxidation of the Reduced Polypeptide Chain. *Proc. Natl. Acad. Sci. USA* **47**, 1309–1314 (1961).
5. K. A. Dill, S. B. Ozkan, M. S. Shell, T. R. Weikl, The Protein Folding Problem. *Annu. Rev. Biophys.* **37**, 289–316 (2008).
6. C. M. Dobson, P. A. Evans, The folding of hen lysozyme involves partially structured intermediates and multiple pathways. *Nature* **358**, 302–307 (1992).
7. F. M. Hughson, P. E. Wright, R. L. Baldwin, Structural Characterization of a Partly Folded Apomyoglobin Intermediate. *Science* **249**, 1544–1548 (1990).
8. E. A. Lipman, E. A, B. Schuler, O. Bakajin, W. A. Eaton, Single-Molecule Measurement of Protein Folding Kinetics. *Science* **301**, 1233–1235 (2003).
9. A. Matouschek, J. Kellis, L. Serrano, A. Fersht, Mapping the Transition State and Pathway of Protein Folding by Protein Engineering. *Nature* **340**, 122–126 (1989).
10. S. Padmanabhan, S. Marqusee, T. Ridgeway, T. M. Laue, R. L. Baldwin, Relative Helix-Forming Tendencies of Nonpolar Amino Acids. *Nature* **344**, 268–270 (1990).

11. Z. Lai, J. McCulloch, H. A. Lashuel, J. W. Kelly, Guanidine Hydrochloride-Induced Denaturation and Refolding of Transthyretin Exhibits a Marked Hysteresis. *Biochemistry* **36,** 10230–10239 (1997).

12. S. S. Jaswal, J. L. Sohl, J. H. Davis, D. A. Agard, Energetic landscape of α-lytic protease optimizes longevity through kinetic stability. *Nature* **415**, 343–346 (2002).

13. S. Schopper, A. Kahraman, P. Leuenberger, Y. Feng, I. Piazza, O. Müller, P. J. Boersema, P. Picotti, Measuring Protein Structural Changes on a Proteome-Wide Scale Using Limited Proteolysis-Coupled Mass Spectrometry. *Nature Protoc.* **12**, 2391–2410 (2017).

14. Y. Feng, G. De Franceschi, A. Kahraman, M. Soste, A. Melnik, P. J. Boersema, P. P. de Laureto, Y. Nikolaev, A. P. Oliveira, P. Picotti, Global Analysis of Protein Structural Changes in Complex Proteomes. *Nature Biotech.* **32**, 1036–1044 (2014).

15. P. Leuenberger, S. Ganscha, A. Kahraman, V. Cappelletti, P. J. Boersema, C. von Mering, M. Claassen, P. Picotti, Cell-Wide Analysis of Protein Thermal Unfolding Reveals Determinants of Thermostability. *Science* **355**, eaai7825, (2017).

16. I. Piazza, K. Kochanowski, V. Cappelletti, T. Fuhrer, E. Noor, U. Sauer, P. Picotti, A Map of Protein-Metabolite Interactions Reveals Principles of Chemical Communication." *Cell* **172**, 358–361 (2018).

17. V. Bianco, M. Alonso-Navarro, D. Di Silvio, S. Moya, A. L. Cortajarena , I. Coluzza, Proteins are Solitary! Pathways of Protein Folding and Aggregation in Protein Mixtures. *J. Phys. Chem. Lett.* **10**, 4800–4804 (2019).

18. L. P. Kozlowski, Proteome-pI: proteome isoelectric point database. *Nucleic Acids Res.* **45**, D1112–D1116 (2017).

19. A. E. Varela, J. F. Lang, Y. Wu, M. D. Dalphin, A. J. Stangl, Y. Okuno, S. Cavagnero, Kinetic Trapping of Folded Proteins Relative to Aggregates under Physiologically Relevant Conditions. *J. Phys. Chem. B*. **122**, 7682–7698 (2018).

20. Y.-W. Shieh, P. Minguez, P. Bork, J. J. Auburger, D. L. Guilbride, G. Kramer, B. Bukau, Operon structure and cotranslational subunit association direct protein assembly in bacteria. *Science* **350**, 678–680 (2015).

21. A. Shiber, K. Döring, U. Friedrich, K. Klann, D. Merker, M. Zedan, F. Tippmann, G. Kramer, B. Bukau, Cotranslational assembly of protein complexes in eukaryotes revealed by ribosome profiling. *Nature* **561**, 268–272 (2018).

22. E. Natan, J. N. Wells, S. A. Teichmann, J. A. Marsh, Regulation, evolution and consequences of cotranslational protein complex assembly. *Curr. Opin. Struct. Biol.* **42**, 90–97 (2017).

23. A. Schwarz, M. Beck, The benefits of cotranslational assembly: A structural perspective. *Trends Cell. Biol.* **29**, 791–803 (2019).

24. S. E. Ahnert, J. A. Marsh, H. Hernandez, C. V. Robinson, S. A. Teichmann, Principles of assembly reveal a periodic table of protein complexes. *Science* **350**, aaa2245–aaa2245 (2015).

25. J. L. Chaney, P. L. Clark, Roles for Synonymous Codon Usage in Protein Biogenesis. *Annu. Rev. Biophys.* **44**, 143–166 (2015).

26. M. B. Borgia, A. Borgia, R. B. Best, A. Steward, D. Nettels, B. Wunderlich, B. Schuler, J. Clarke, Single-molecule fluorescence reveals sequence-specific misfolding in multidomain proteins. *Nature* **474**, 662–665 (2011).

27. T. A. Thanaraj, P. Argos, Ribosome-mediated translational pause and protein domain organization. *Protein Sci.* **5**, 1594–1612 (1996).

28. C. Cecconi, E. A. Shank, C. Bustamante, S. Marqusee, Direct observation of the three-state folding of a single protein molecule. *Science* **309**, 2057–2060 (2005).

29. S. J. Hamill, A. Steward, J. Clarke, The folding of an immunoglobulin-like greek key protein is defined by a common-core nucleus and regions constrained by topology. *J. Mol. Biol.* **297**, 165–178 (2000).

30. K. L. Maxwell, *et al*. Protein Folding: Defining a "Standard" Set of Experimental Conditions and a Preliminary Kinetic Data Set of Two-State Proteins. *Protein Sci.* **14**, 602–616 (2005).

31. N. S. Bogatyreva, A. A. Osypov, D. N. Ivankov, KineticDB: a Database of Protein Folding Kinetics. *Nucleic Acids Res.* **37**, D342–D346 (2009).

32. S. G. Waley, Refolding of triosephosphate isomerase. *Biochem J.* **135**, 165–172 (1973).

33. M. Calamai, N. Taddei, M. Stefani, G. Ramponi, F. Chiti, Relative Influence of Hydrophobicity and Net Charge in the Aggregation of Two Homologous Proteins. *Biochemistry* **42**, 15078–15083 (2003).

34. P. R. Majhi, R. R. Ganta, R. P. Vanam, E. Seyrek, K. Giger, P. L. Dubin, Electrostatically driven protein aggregation: β-lactoglobulin at low ionic strength. *Langmuir* **22**, 9150–9159 (2006).

35. M. J. Kerner, D. J. Naylor, Y. Ishihama, T. Maier, H.-C. Chang, A. P. Stines, C. Georgopoulos, D. Frishman, M. Hayer-Hartl, M. Mann, F. U. Hartl, Proteome-Wide Analysis of Chaperonin-Dependent Protein Folding in Escherichia Coli. *Cell* **122**, 209–220 (2005).

36. Y. E. Kim, M. S. Hipp, A. Bracher, M. Hayer-Hartl, F. U. Hartl, Molecular Chaperone Functions in Protein Folding and Proteostasis. *Annu. Rev. Biochem.* **82**, 323–355 (2013).

37. K. Xia, M. Manning, H. Hesham, Q. Lin, C. Brystoff, W. Colón, Identifying the subproteome of kinetically stable proteins via diagonal 2D SDS/PAGE. *Proc. Natl. Acad. Sci. USA* **104**, 17329–17334 (2007).

38. E. Braselmann, J. L. Chaney, P. L. Clark, Folding the proteome. *Trends Biochem Sci*. **38**, 337–344 (2013).

39. I. M. Keseler, *et al*. The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. *Nucleic Acids Res.* **45**, D543–550 (2017).

40. G.-W. Li, D. Burkhardt, C. Gross, J. Weissman, Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* **24**, 624–635 (2014).

41. T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, H. Mori, Construction of *Escherichia coli* K-12 in-frame single-gene knockout mutants: the Keio collection. *Mol. Sys. Biol.* **2**, 2006.0008 (2006).

42. J. Gough, K. Karplus, R. Hughey, C. Chothia, Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* **313**, 903–919 (2001).

43. A. P. Pandurangan, J. Stahlacke, M. E. Oates, B. Smithers, J. Gough, The superfamily 2.0 database: A significant proteome update and a new webserver. *Nucleic Acids Res.* **47**, D490–D494 (2019).

Acknowledgments: We thank S. Marqusee and D. Barrick for thoughtful conversations, and P. Mortimer for support with and maintenance of mass spectrometer instrumentation. We are grateful to B. García-Moreno and co-workers for providing samples of SNase and *Tt*RNase H; **Funding:** We acknowledge HFSP (RGY0074/2019) for funding. B.W. thanks the Program in Molecular Biophysics training grant (NIH T32) and H.E.T. thanks the Chemistry-Biology Interface training grant for support (NIH T32). **Author contributions:** P.T. performed proteomics experiments with assistance from B.W. and H.E.T. P.T. and S.D.F. performed analyses. S.D.F wrote the manuscript with input from all authors; **Competing interests:** Authors declare no competing interests; **Data and materials availability:** Raw proteomic data are available via the ProteomeXchange under accession codes XXXXXXX. Data used to construct figures are found in Supplementary Tables S1-S9 and Supplementary Data S1-S24. Python programs are available on GitHub at XXXXX.

Supplementary Materials:

Materials and Methods

Texts S1-S5

Figures S1-S20

Tables S1-S9

Data S1-S24

Reference 39–43

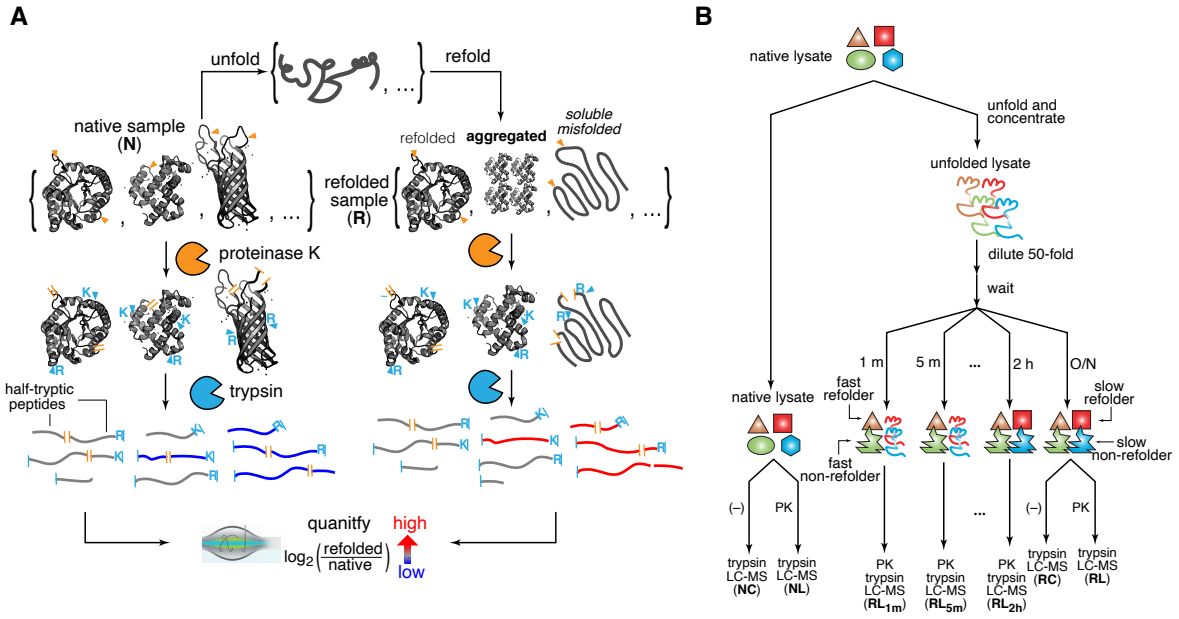

Fig. 1. Limited proteolysis mass spectrometry (LiP-MS) to interrogate refolding of the proteome. (**A**) Lysates from *E. coli* are prepared under native conditions, globally unfolded and refolded. The structures of the refolded proteins are probed by pulse proteolysis with proteinase K (PK) and compared to that of their native forms. Label free quantification (LFQ) of half-tryptic peptides reveals sites across the proteome where local conformation differ between a protein's native and refolded forms. (**B**) By varying the time between dilution and pulse proteolysis, kinetics of proteome refolding is monitored. (N, native; R, refolded; C, control; L, limited proteolysis)

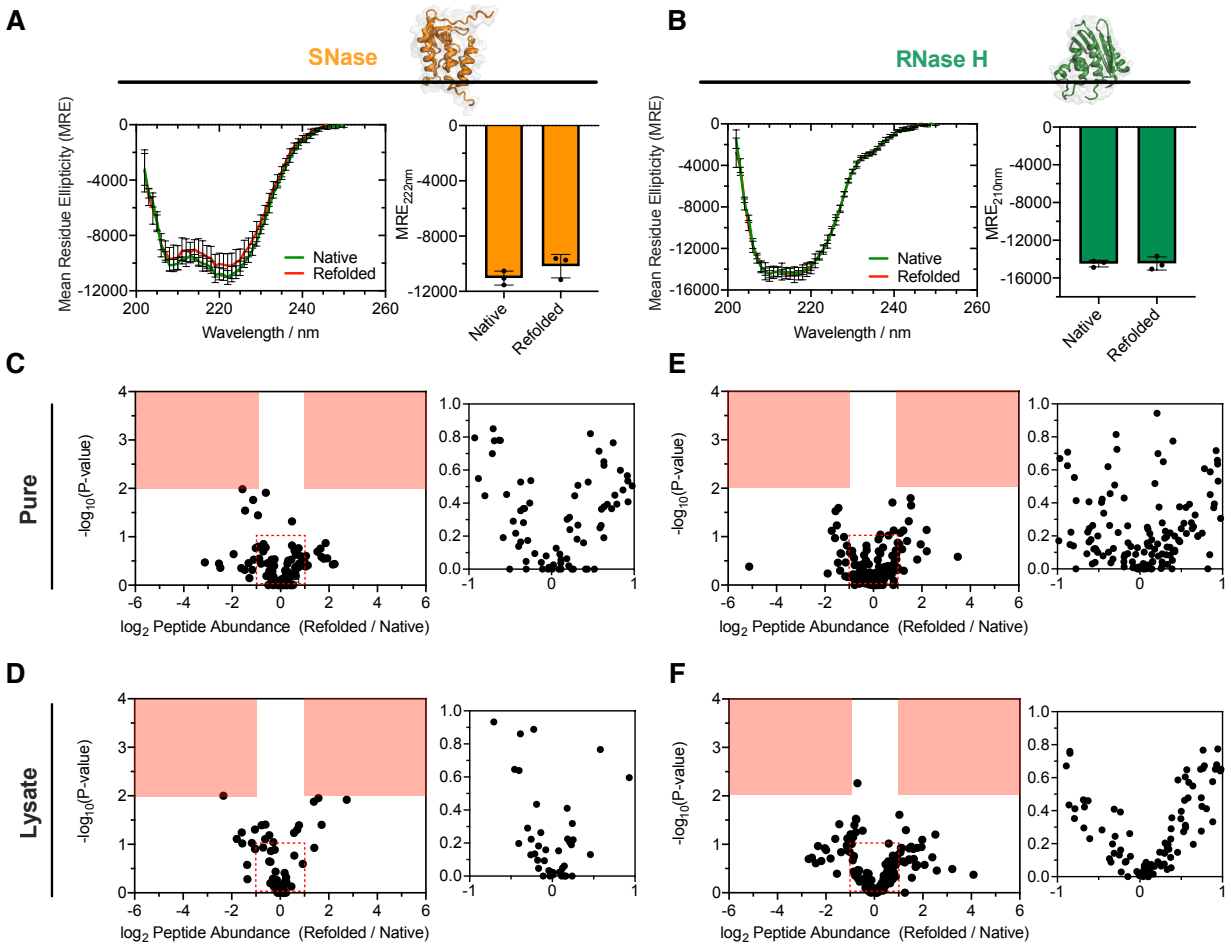

Fig. 2. Refolding of Small Model Proteins. (**A, B**) Circular Dichroism (CD) spectrum of (**A**) Staphylococcal nuclease (SNase) or (**B**) Ribonuclease H from *Thermus thermophilus* (*Tt*RNase H), natively expressed from *E. coli* (green), and following unfolding with 8 M urea and 50-fold dilution (red). Bar chart shows no significant difference in MRE at 222 nm (SNase) or 210 nm (RNase H) ($n = 3$). (**C**) Volcano plot comparing peptide abundances from native and refolded SNase ($n = 3$). Effect sizes reported as ratio of averages, and P-values are based on Welch's t-test. Red regions designate significance (effect-size > 2, P-value < 0.01). Inset shows large number of points clustered near the origin. The data suggest no significant difference in the structure of native SNase and the conformation produced when it is diluted out of urea. (**D**) As in **C**, except SNase was spiked into *E. coli* lysate, providing a complex background. (**E,F**) Volcano plot showing peptides quantified by LiP-MS derived from native and refolded *Tt*RNase H, both as a purified protein (**E**), and spiked into *E. coli* lysate (**F**).

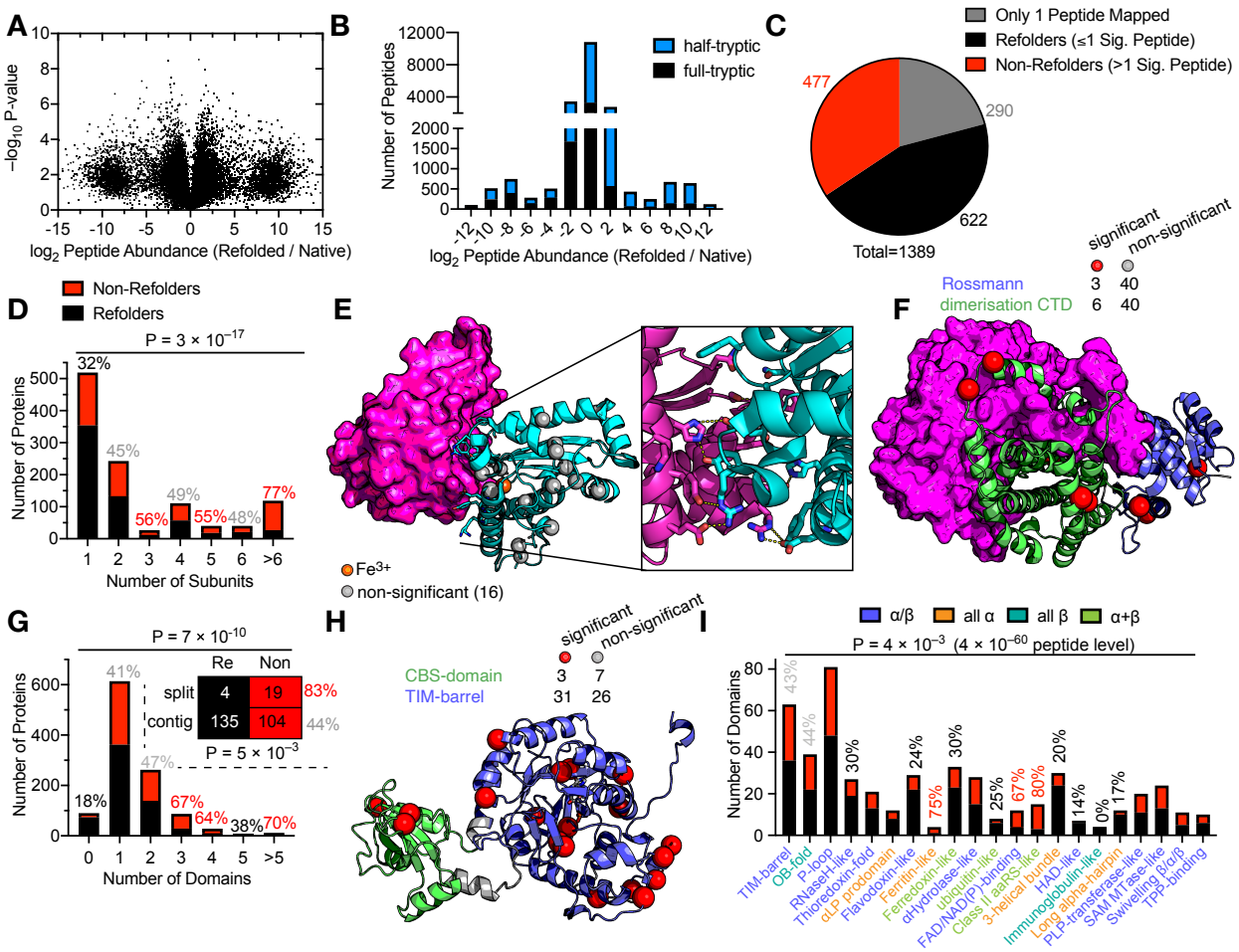

Fig. 3. Refoldability of the *E. coli* Proteome. (**A**) Volcano plot comparing peptide abundances from native and refolded *E. coli* lysates normalized for protein abundance ($n = 3$). Effect sizes reported as ratio of averages, and P-values are based on Welch's t-test. (**B**) Histogram of abundance ratios for half-tryptic and full-tryptic peptides. Half-tryptic peptides are enriched in the refolded proteome. (**C**) Overall number of refolding proteins out of 1099 *E. coli* proteins. 290 proteins only furnished one peptide and hence too little data to make an assessment. (**D**) Proteins that are part of complexes with multiple subunits – but especially trimers and pentamers – are significantly less refoldable than monomeric proteins (P-value = 3×10^{-17} by chi-square test). Percentages represent fraction of proteins that are non-refoldable of given classification. Red percentages represent groups that are more non-refoldable than average; black represents groups that are more refoldable than average; grey represents groups that are comparable to average. (**E**) X-ray structure of Fe-superoxide dismutase (SodB, PDB: 2NYB), a refoldable homodimeric protein, showing a simple interface mediated primarily by polar interactions. (**F**) X-ray structure of 6-phosphogluconate dehydrogenase (Gnd, PDB: 2ZYA), a non-refoldable homodimeric protein in which C-terminal domains (green) interdigitate. (**G**) Proteins with many domains, but particularly three or four, are significantly more non-refoldable than proteins with zero or one domain (P-value = 7×10^{-10} by chi-square test). *Inset:* contingency table for 2-domain proteins comparing the refoldability of proteins in which domains are contiguous or in which one is split by an intervening domain. Split domains are strikingly less refoldable. (**H**) X-ray structure of inosine 5'-monophosphate dehydrogenase (GuaB, PDB: 1ZFJ), a protein in which a split TIM-

barrel domain evinces high level of non-refoldability compared to the intervening domain. (**I**) Certain fold-types (e.g., ribonuclease H-like and immunoglobulin domains) are much more refoldable than others (e.g., ferritin-like and class II aaRS/biotin synthase-like). P-value = 4×10^{-3} by chi-square test though counts for many fold-types are low. Statistical significance is drastically enhanced by conducting analysis at the peptide level (see fig. S9).

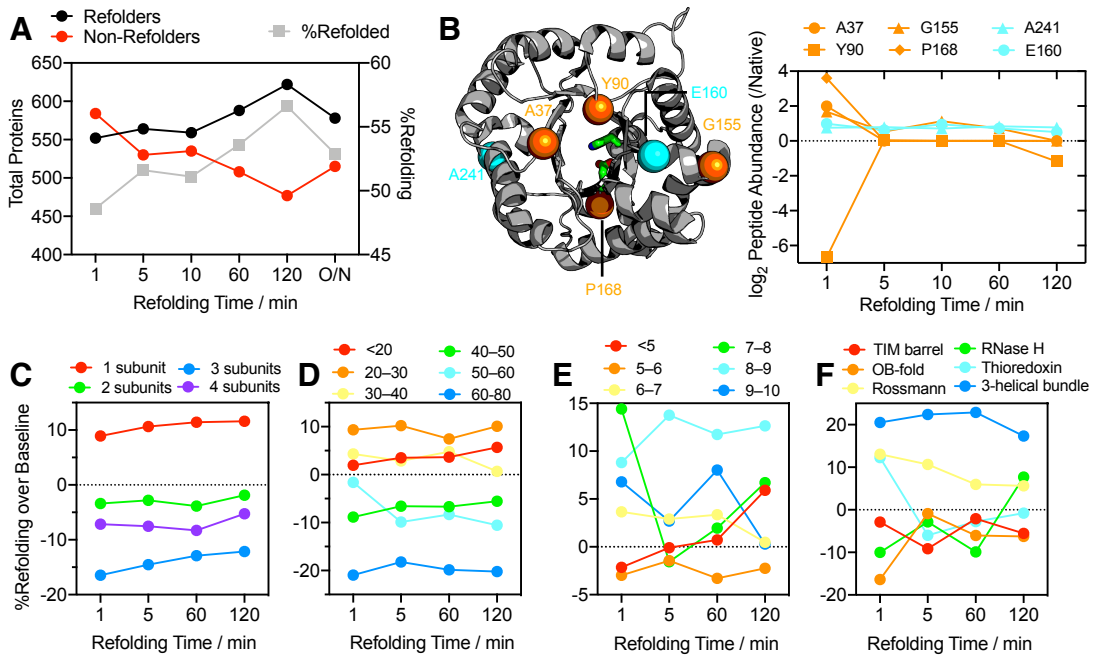

Figure 4. Proteome Refolding Kinetics. (**A**) From 1 min to 2 h, the number of refolding proteins increases. Overnight incubation may lead to increased degradation. (**B**) Structure of triose phosphate isomerase (*at left*) with the positions of six residues whose accessibility change significantly from 1 min to 2 h (*at right*). Orange residues change conformation in a minute-timescale process associated with forming one side of the TIM-barrel; green residues represent the catalytic residues; Ala241 (cyan) is located at a kink between two helices and becomes native-like slowly. (**C-F**) Variation in refoldability as a function of time for several classifications: protein subunit count (**C**); protein molecular weight in kDa (**D**); protein isoelectric point (**E**); domain fold-type (**F**). To highlight differences, the percent of refolding proteins (domains) is reported as a difference compared to the overall refoldability of the proteome at that time point. (**C**) Dimers and tetramers do not generally require more time to refold, though trimers do. (**D**) More massive proteins do not generally require more time to refold, though the smallest proteins (<20 kDa) do. (**E**) Polyanionic proteins (pI < 5) tend to refold slowly; proteins with pI between 7-8 tend to refold rapidly. (**F**) Rossman folds and 3-helical bundles fold rapidly; ribonuclease H-like domains and OB-folds fold slowly. TIM barrels are less refoldable than most domains, particularly at 5 minutes.