Supporting Information for

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

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

[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

Table of Contents

Supplementary Text

Text S1. Description of LFQ Filtering Algorithm

For proteomic studies, two separate LFQ analyses were conducted: one comprising of 3 replicates of a refolded control sample and 3 replicates of a native control sample (in which the samples were not digested with Proteinase K), and a second comprising of 3 replicates of a refolded LiP sample and 3 replicates of a native LiP sample. These 3-vs.-3 LFQs were each conducted in Proteome Discoverer (PD) version 2.4, using the standard parameters of the Minora feature mapper. From each .pdResult file, we generated an Excel output: a three-tiered output file of the results in the hierarchy of Protein>Peptide Group>Consensus Feature (referred to as the PD output file).

The script called Analyzer_v17_v10+pd24.py (available on GitHub) requires as inputs the PD output file for the control experiment and the LiP experiment.

We use PD's in-built algorithms to infer protein abundance differences based on the available peptide group data for the control experiment. If a protein abundance difference is greater than 2-fold and the P-value calculated by PD is less than 0.01, this was considered to be a significant protein abundance difference, in which case the value of (protein abundance-Refolded)/(protein abundance-Native) is used as a normalization constant for all peptides that map to that protein in the analysis of the LiP experiment. If either of those thresholds are not met (or if no quantification data is available for said protein, or if that protein was not identified in the control experiment), then no normalization is conducted for the peptides that map to that protein.

The main role of the analyzer is to convert the raw extracted ion chromatogram (XIC) peak intensities derived from PD's Minora feature mapper and convert them into an abundance difference to be used for our downstream analysis. Because our analysis relies explicitly on quantification at the peptide level (rather than the protein level) we opted to perform this with in-house scripts. Furthermore, our experiments are characterized by two additional complications not present for most proteomic applications. Firstly, our native and refolded samples – at the peptide level – are very different from each other because of the large number of proteinase K sites that are present only in the native or refolded samples. Hence, retention times can be very different for a given peptide between the two sample-types due to chromatographic matrix effects. Secondly, the fragments that are characteristic of a distinct structure will be absent (or at very low abundance) in either the refolded or native sample.

The hierarchy of the PD output file provides for each protein, a set of peptide groups that maps to that protein; and for each peptide group, a set of consensus features that map to that peptide group. A consensus feature refers to a set of MS1 XICs that were feature-mapped together into a single grouping, regardless of whether it was successfully assigned to a particular peptide sequence based on MS2. Hence, each consensus feature consists of a retention time window, a precursor m/z, a charge state, a number of peptide-spectrum matches (PSMs), and a set of intensities for each of the 6 (or more) runs being compared in the LFQ.

Firstly, we merged together the ion counts of all consensus features that matched to the same peptide, with the same modification state and charge state. This operation was necessary because we found that peptides would oftentimes have fairly distinct retention times in the native and refolded samples (owing to the fact that these samples contain very different peptides overall, and hence experience significant column matrix effects). Hence, this summation in effect overrides the standard requirement of separating consensus features based on pre-determined retention time windows. Next, we assessed for each consensus feature whether it should be considered based on the following criteria. If it had zero missing values, it was kept without conditions. If it had one missing value, it was kept though the missing value was discarded (not filled with zero). If it had three non-zero values corresponding to the refolded (or native) replicates and three missing values corresponding to the native (or refolded) replicates, it would be considered an 'all-or-nothing' concensus feature and would be kept, except the missing values would be filled with 1000 ion counts (an estimate of the limit of detection).

We next calculated the ratio associated with that consensus feature (average of the refolded extracted ion intensities divided by the average of the native extracted ion intensities), and a P-value according to the t-test with Welch's correction for non-equal population variance.

Some peptide groups were associated with more than one consensus feature. This occurs frequently if multiple charge states associated with that peptide are detected. It

can also happen if the peptide undergoes stochastic methionine oxidation. We considered all consensus features for a given peptide group together, including those arising from distinct methionine oxidation levels and different charge states.

If the ratios associated with the various consensus features did not agree in sign (e.g., in the 2+ charge state, the peptide was more abundant in native, but in the 3+ charge state with a methionine oxidation, it was more abundant in refolded), then we assigned the peptide group a ratio of unity (i.e., the data were inconsistent and therefore not used to test against the null hypothesis). If all the ratios agreed in sign, then we took the median of the available ratios as the overall ratio for that peptide group. Moreover, to determine the P-value associated with that ratio we used Fisher's method to combine the P-values associated with the different consensus features. We did not adjust this P-value for multiple hypothesis testing because each set of extracted ion intensities is only used to bear on one hypothesis: whether the peptide in question implies identical structure at a given location between the refolded and native forms of a protein, or distinct.

This filtering procedure discards a significant number of data points with very large (or small) abundance ratios (specifically, those that were observed in only *some* of the replicates of one sample-type), though it leaves behind a subset that are highly reproducible (fig. S3).

Analyzer compiles each sequenced peptide, along with its associated metadata, the identity of the peptide as tryptic or half-tryptic (and if so, the location of the

proteinase K cleavage site), abundance ratio, normalized abundance ratio, and P-value and outputs it into a _out.txt file.

Analyzer further performs a protein(domain)-level assessment, whereby it counts the total number of peptides associated with a protein (domain), the total number peptides that are deemed significant (more than 2-fold abundance difference between native and refolded but less than 64-fold, P-value less than 0.01), and the total number of peptides that are deemed 'all-or-nothing' (more than 64-fold abundance difference between native and refolded, P-value less than 0.0158). The threshold of 64 was chosen because manual inspection of the data showed that we did not observe more than 64-fold differences in peptides' ion abundances unless one of the consensus features was one in which missing values were filled with 1000. These data are compiled into a _summary17_10+Protein.txt file and a _summary17_10+Domain.txt file. We considered a protein (domain) to be non-refoldable if it had 2 or more significant peptides. We moreover only considered proteins (domains) for the analysis overall if there were 2 or more peptides in total mapped to it. However the primary claims of the analysis are not sensitive to any of these cut-offs (Data S1).

The summary files for the experiments (divided by protein and divided by domain) are provided in Data S1-S4. The _out.txt files can be obtained upon reasonable request. The .raw files are available through proteomeXchange.

Text S2. The "N_{tot} bias"

The primary caveat one must bear in mind in interpreting our results is the N_{tot} bias. We define N_{tot} as the total number of peptides identified and quantified for a given protein (or domain). Stated simply, because we define a protein (domain) to be non-refoldable if it has two or more peptides that possess significant abundance differences between native and refolded samples ($N_{sig} > 1$), a protein (domain) will be more likely to be judged non-refoldable if we quantify more peptides for it. We note that our control studies on SNase and *Tt*RNase H (Fig. 2, fig. S1) indicate that it is, in principle, possible to identify 100-200 unique peptides for a protein, all of which are non-significant. Moreover, the protein aceE, identified as a refolder in our primary data set (Data S1), admitted 0 significant peptides out of 210 unique peptides quantified; showing that it is possible for very large/complex with high coverage to still be counted as refoldable using our definition. Nevertheless, we found it prudent to be mindful of this potential source of bias.

We devised two ways to address this bias (see figs. S7-S8). On one hand, we considered grouping together all significant and non-significant peptides that correspond to a given protein classification (e.g., monomers) without regard to which protein they came from. We refer to this as a peptide-level analysis. The advantage of this approach is that all bias is removed associated with N_{tot} ; moreover, it obviates the need to define a 'minimum' number of significant peptides a protein requires to be labeled non-refoldable. A second analysis we performed consisted of comparing the distributions of the proteins' N_{tot} associated with various classifications. As long as

proteins in different classifications that are being compared to each other do not have N_{tot}'s that are significantly differently distributed, then we expect this bias to not affect that particular comparison.

If the classifications are confounded by large differences in N_{tot}, then the peptide-level analysis will not correlate with the protein(domain)-level analysis; these comparisons are "probably biased." For classifications that do not have large differences in N_{tot}, the peptide-level analysis is significant and mirrors the protein-level trends; these comparisons are "robust."

For instance, the differences in refoldability of multimers compared to monomers is robust, because the trends at the protein level (Fig. 4A) are recapitulated at the peptide level (fig. S8A): monomers are the most refoldable and also are associated with a low fraction of all the peptides mapped to them being significant. Ipso facto, trimers are the least refoldable and also are associated with a high fraction of all the peptides mapped to them being significant. This theme is recapitulated across the subunit categories, with the only exception of pentamers. Moreover, on the whole, different subunit classifications are not associated with different number of peptides quantified per protein (fig. S8D), with the exception of one pairwise comparison between tetramers and monomers. Hence, the findings concerning the relationship between refoldability and subunit composition is robust.

We find similar robust trends concerning the relationship between refoldability and location (fig. S7A, D, G) and isoelectric point (fig. S7B, E, H). Locations associated with higher levels of refoldability (ribosomes and membranes) also tend to have a lower

fraction of their peptides being significant (fig. S7D). Moreover, proteins in different locations are not associated with different numbers of peptides mapped per protein (fig. S7G), with the exception of one pairwise comparison between ribosomal and inner membrane proteins. Similarly, pI ranges associated with higher levels of refoldability (7–8, 8–9, 9–10) are also associated with a lower fraction of peptides being significant (fig. S7E), and proteins in different pI ranges are not associated with different number of peptides mapped per protein (fig. S7H), with the exception of one pairwise comparison between pI 5–6 and pI 7–8.

For fold-types and cofactors, our analyses are somewhat more ambiguous because many of these categories are associated with smaller counts. On one hand, different fold-types are *not* associated with significantly different numbers of peptides quantified per domain (fig. S8H). However, it is more challenging to ascertain a clear pattern between domain refoldability and fraction of peptides associated with a given domain that are significant, owing to the small counts associated with many of the fold-types (fig. S8G). Satisfyingly though, many of the fold-types that are the most refoldable do indeed have a rather small fraction of their peptides being significant (right-hand side of fig. S8G).

As for cofactors, inspection of fig. S8C shows that for the most part, cofactors associated with high levels of refoldability (e.g., iron-sulfur clusters and heme) are also associated with lower fractions of their peptides being significant. Likewise, the more non-refolding metalloproteins also are associated with more of their peptides, on the whole, being significant. The primary caveat is TPP-proteins, which were generally

amongst the least refoldable but were not associated with a higher frequency of significant peptides (fig. S8C). This can be explained clearly by inspecting the distribution for the total number of peptides quantified (fig. S8F), where it is apparent that TPP-proteins received unusually high coverage compared to other cofactor classifications, which could explain their higher apparent non-refoldability. On the other hand, other co-factor categories have similar distributions for the total number of peptides quantified per protein.

The most biased trend is that between refoldability and molecular weight. Unsurprisingly, more massive proteins tend to be more non-refoldable (fig. S7C), though this trend is not mirrored at the peptide-level, and indeed proteins with molecular weight >100 kDa actually generate significant peptides at a low frequency (fig. S7F). This can be attributed to N_{tot}-bias, because the more massive a protein is, the more total peptides we tend to quantify it, as shown clearly in fig. S7I.

The trend between refoldability and number of domains is somewhat more complicated. Proteins with more domains also tend to be more non-refoldable (Fig. 4C), and this is quite well recapitulated at the peptide-level as well (fig. S8B), with 0-domain proteins generating significant peptides at a low rate, and 5-domain proteins generally significant peptides at a high rate. However, this being said, proteins with more domains also tend to result in more peptides being quantified (fig. S8E), in a manner that is similar to proteins with greater molecular weight resulting in higher numbers of quantified peptides (fig. S7I). Therefore, the finding that proteins with more domains are less refoldable is potentially biased. A conservative approach to view this

finding therefore might be to say: The more domains the protein has, the more likely that one of them is a non-refolder, thereby rendering the protein a non-refolder.

Text S3. A Note on Membrane Proteins

Membrane proteins are not well represented in our study because we lyse our cells under native conditions (without detergents or denaturants) and clarify lysates prior to further experiments (see fig. S5). That being said, our study includes 35 proteins localized to the inner membrane and 6 localized to the outer membrane (according to EcoCyc annotations) and these proteins were generally associated with high levels of refoldability.

Inspection of the data shows that the inner membrane proteins detected in our study are primarily those which are soluble proteins that are localized to the inner membrane through association with membrane proteins. For instance, we observe many peptides for AtpA, AtpC, AtpD, AtpG, and AtpH, the components of the cytoplasmic F_1 module – and none for AtpE and AtpB, the components of the membrane-bound F_O module. We also tend to observe the ATP-binding subunit of several ATP-binding cassette (ABC) transporters which are also soluble components (such as DppD and DppF from the dipeptide transporter, PotA from the spermidine transporter, and OppD and OppF from the tripeptide transporter), but not their transmembrane counterparts. For complex I, we only observe several subunits in the cytosolic peripheral arm (NuoB, NuoE, NuoF, and NuoI), but none of the subunits in the transmembrane region. Hence, it is understandable that during lysis, some of these soluble portions became detached from their membrane-embedded partners. All but four of the inner membrane proteins were refoldable, and three of the four non-refolders

belonged to ATP synthase. This observation may reflect an authentic trend, though it is hard to make a strong case with so few examples.

The 6 outer membrane proteins that were quantified were BamA, OmpC, OmpX, Pal, SlyB, and YnfB. We presume that these proteins were detached from the membrane during lysis, and were sparingly soluble at the very low concentrations that they would have been present at in the lysates. All of these proteins were found to be 'refoldable.' This finding could be explained by the possibility that in the native samples they were already unfolded (as one would expect for an outer-membrane beta-barrel protein in solution), and that following 'refolding' they returned to the same unfolded ensemble.

Text S4. Description of Analysis of Refolding Kinetics

To probe the overall refolding kinetics of the *E. coli* proteome, we merged the _out.txt files associated with each time point. Each time point was referenced against a common native-LiP (NL) sample and normalized against a common set of normalization factors obtained from a control study (without proteinase K treatment). For time points in which a given peptide was not identified or quantified – resulting in its absence in the corresponding _out.txt file – the peptide was assigned default values of ratio = 1, P-value = 1 (resulting in (0,0) on log scales). As an aside, we adapted analyzer_v17_v10+pd24.py to be able to process LFQ experiments with arbitrarily large number of channels; however, we obtained the most consistent results by performing separate 3-vs-3 LFQs and merging the data together. To create the plots in fig. S10A-B, we counted the number of significant (half-)tryptic peptides at each timepoint.

To create the plots in fig. S10D-H, we calculated the percent of proteins that are refolding in each classification at a given time point. This did not require merging data-sets (Data S1 and S2).

To define slow and very slow refolders (Fig. 5), we merged the summary data for proteins (domains) that were simultaneously identified in the 1 (or 5) min and the 120 min timepoints (Data S3). Proteins (domains) were only considered for this analysis if two or more peptides were quantified in both time points.

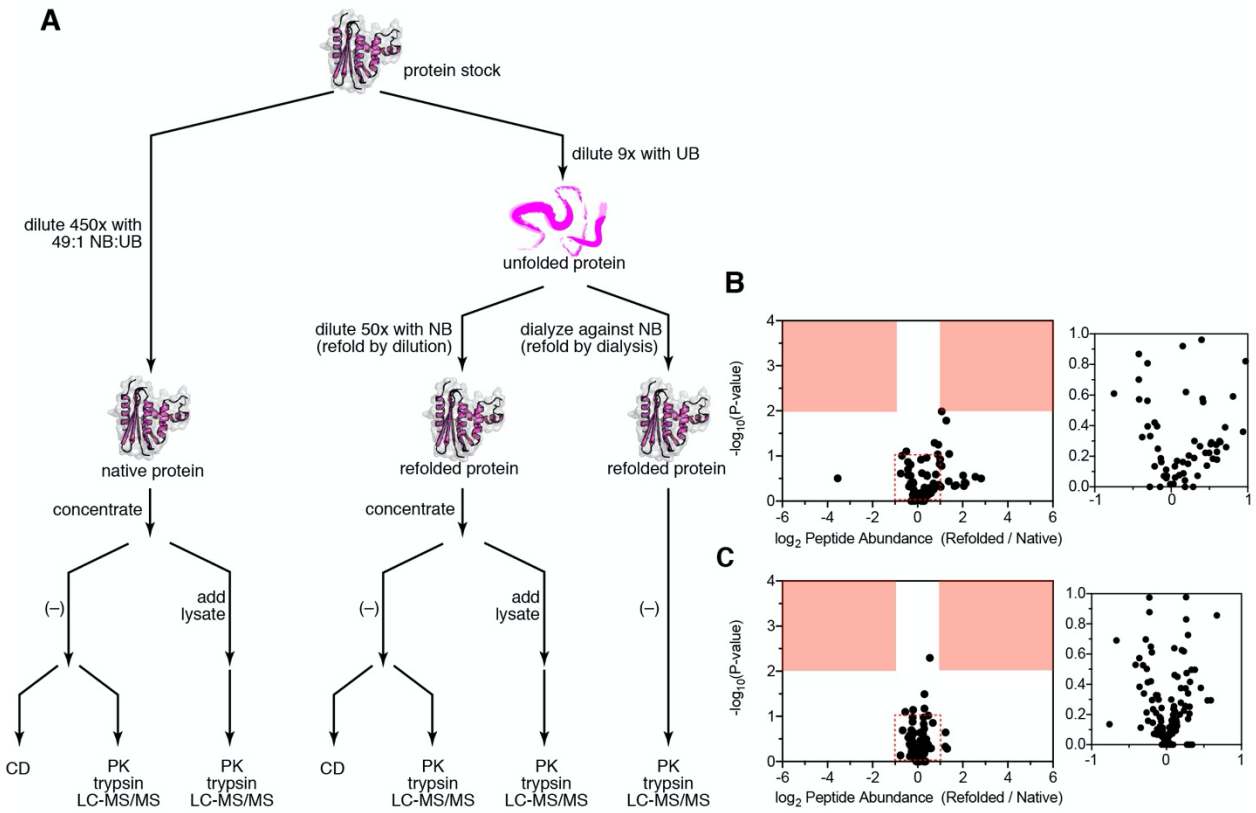

Fig. S1. Refolding of Model Proteins. (**A**) Workflow for experiments on purified proteins. A concentrated protein stock was either diluted 450-fold with a 49:1 mixture of native buffer (NB) and unfolding buffer (UB) to generate a native sample, or diluted 9-fold with unfolding buffer, incubated, and diluted 50-fold further to generate a refolded sample; see Methods. (**B, C**) Volcano plot comparing native proteins to their refolded forms after unfolding and refolded by dialysis, in technical triplicate. Peptide abundances from native and refolded Staphylococcal nuclease (SNase, **B**), and ribonuclease H from *Thermus thermophilus* (*Tt*RNase H, **C**) ($n = 3$). Effect sizes reported as ratio of averages, and P-values are based on Welch's 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 *Tt*RNase H and the conformation produced when it is refolded by dialysis out of urea.

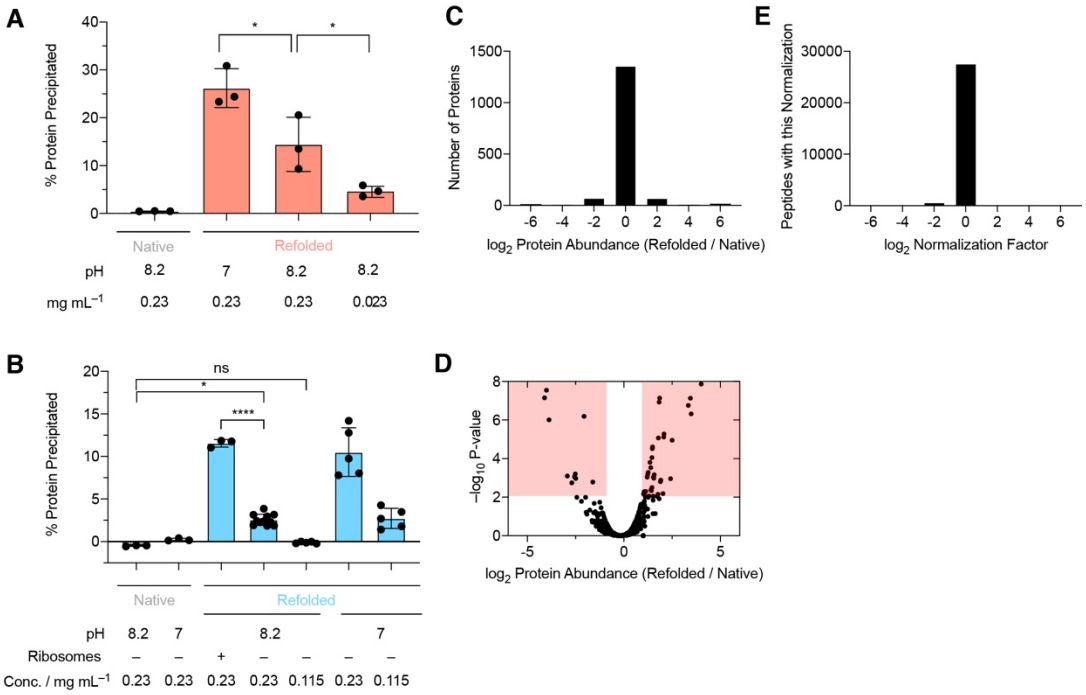

Fig. S2. Pelleting Assays to Monitor Aggregation Formed upon Refolding *E. coli* Lysates. (**A-B**) Assays to measure levels of precipitation formed upon refolding clarified *E. coli* extracts with the BCA assay; see Materials and Methods. Statistical analyses conducted by ANOVA with Tukey's multiple comparison follow-up test. * $P < 0.05$; **** $P < 0.0001$. (**A**) Thermal denaturation followed by slow cooling results in high levels of precipitation. This effect could be partially mitigated by performing thermal unfolding/refolding at pH 8.0 and at very low concentration. (**B**) Chemical denaturation followed by dilution results in lower (but still significant) levels of precipitation. This effect can be mitigated substantially if ribosomes are removed from the clarified extract prior to unfolding and by lowering concentration. The resulting amount of precipitation is not detectable different from samples that were never unfolded. (**C-E**) Few proteins precipitate under the optimized refolding conditions used for LC-MS/MS studies. (**C**) Control samples in which native and GdmCl-refolded *E. coli* lysates were not subjected to limited proteolysis were generated in order to determine overall protein abundance differences between these two samples. Histogram showing the majority of the proteins were present in equal abundance between native and refolded samples. (**D**) Volcano plot showing that 9 proteins were significantly less abundant in the refolded sample (abundance ratio > 2, P-value < 0.01 by Welch's t-test), which could be attributed to precipitation. (**E**) Each peptide derived from a protein that was present with significantly different abundance between the native and refolded samples (abundance ratio > 2, P-value < 0.01) was adjusted with a normalization factor. A histogram of all peptide normalization factors shows the vast majority (~99%) were unity.

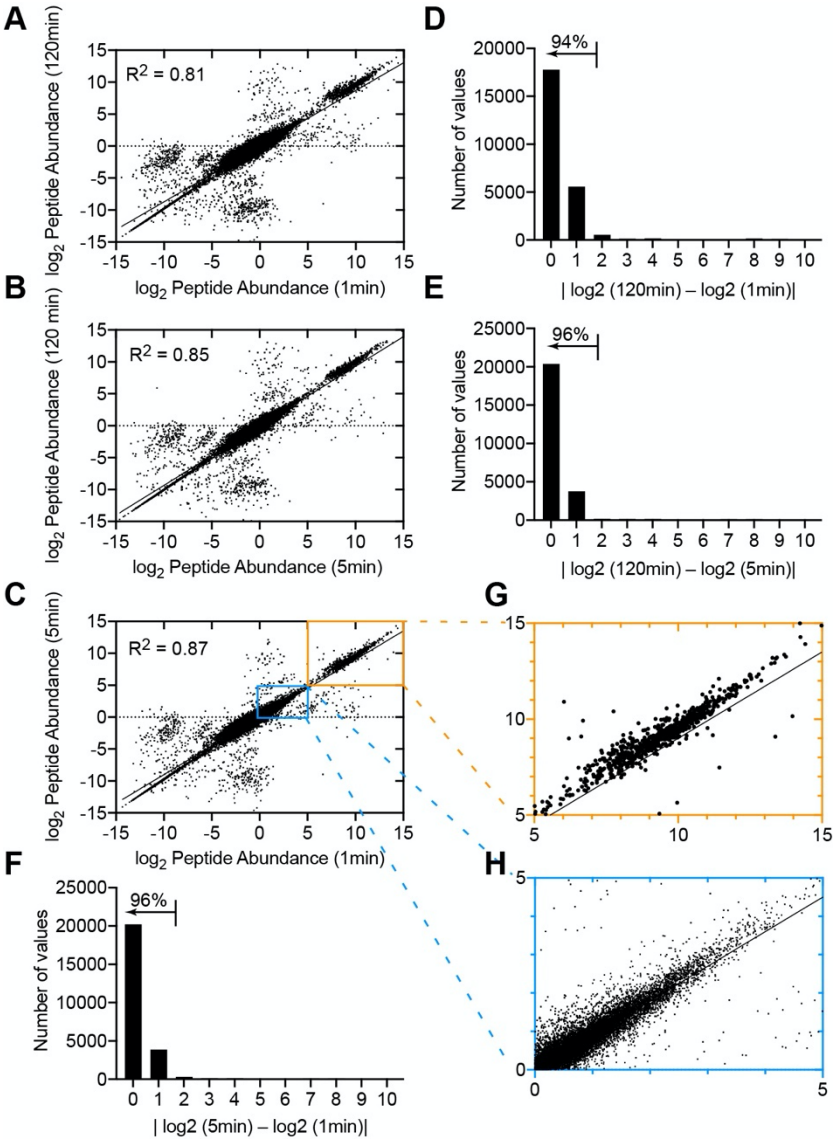

Fig. S3. Reproducibility Across Timepoints of Proteome-wide Refolding from GdmCl. (**A, B, C**) The reproducibility in peptide abundance ratios (refolded/native) for peptides identified and quantified simultaneously in two experiments in which samples were refolding out of GdmCl for different periods of time. Each of these experiments in turn consisted of three biological replicates of native and three biological replicates of refolded. Pearson correlation coefficients are provided; (**A**) reproducibility between samples that were refolding for 120 min and samples that were refolding for 1 min; (**B**) reproducibility between samples that were refolding for 120 min and samples that were refolding for 5 min; (**C**) reproducibility between samples that were refolding for 5 min and samples that were refolding for 1 min. (**D, E, F**) Histograms showing the differences in the log abundance ratio between peptides that were simultaneously identified and quantified in two experiments in which samples were refolding out of GdmCl for different periods of time; (**D**) between samples that were refolding for 120 min and samples that were

refolding for 1 min (ratios for 94% of peptides were within 2.8-fold); (**E**) between samples that were refolding for 120 min and samples that were refolding for 5 min (ratios for 96% of peptides were within 2.8-fold); (**F**) reproducibility between samples that were refolding for 5 min and samples that were refolding for 1 min (ratios for 96% of peptides were within 2.8-fold). (**G**, **H**) Insets showing details of scatter plot **C**, with one highlighting a region with smaller peptide abundance ratios (**G**), and another highlighting a region with large peptide abundance ratios (**H**).

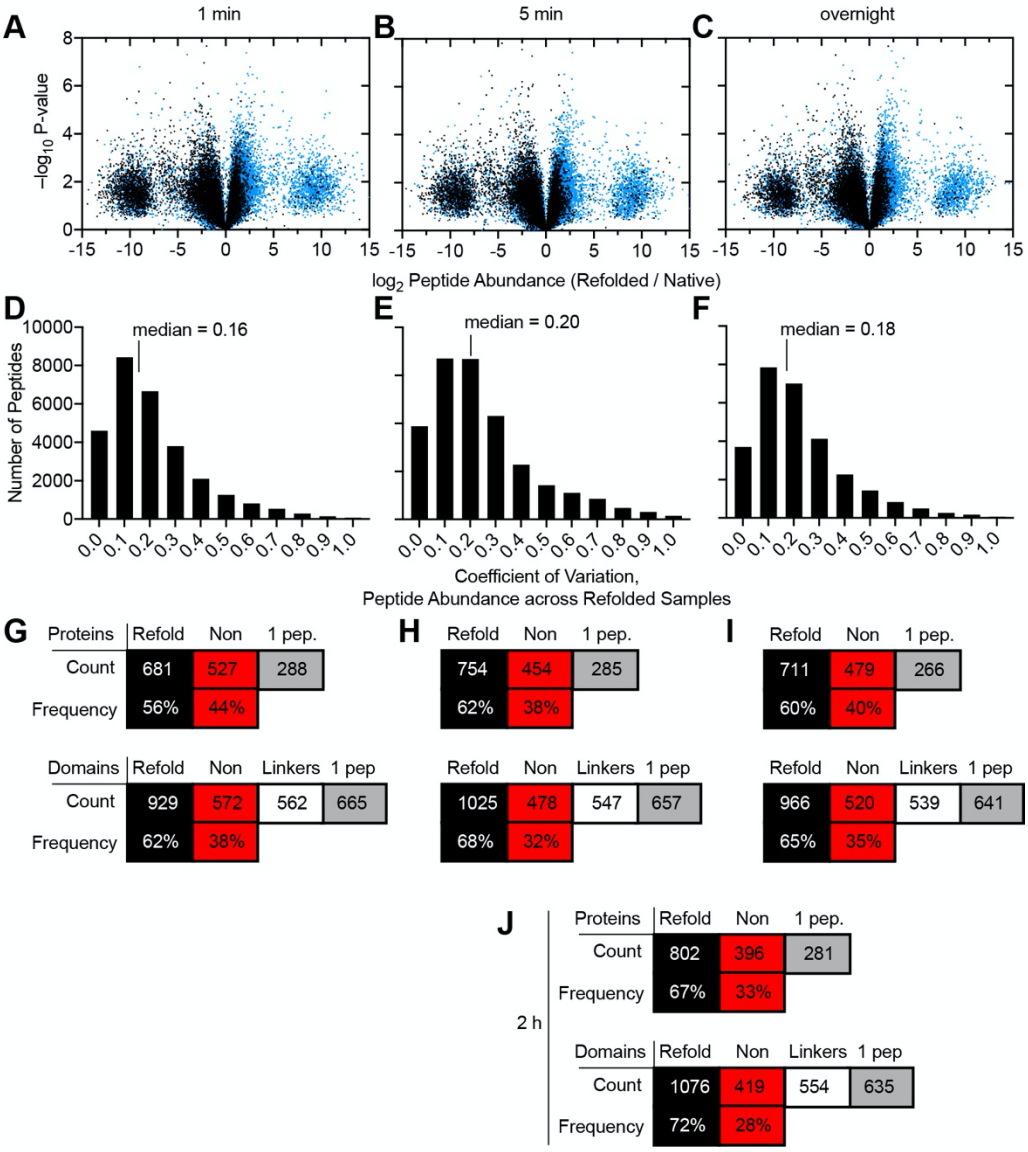

Fig. S4. Summary Statistics for 1 min, 5 min, and overnight refolding timepoints. (**A-C**) Volcano plot comparing peptide abundances from 3 native and 3 refolded *E. coli* lysates normalized for protein abundance, in which the refolding reaction was incubated for distinct periods of time between refolding and probing with proteinase K. Effect sizes reported as ratio of averages, and P-values are based on Welch's t-test. Replicates are from separate bacterial cultures. Plots correspond to the (**A**) 1 min; (**B**) 5 min; and (**C**) overnight refolding times. (**D-F**) Histograms of the coefficients of variation for peptide abundances detected in 3 independent proteome-wide refolding reactions for different refolding times. Histograms correspond to the (**D**) 1 min; (**E**) 5 min; and (**F**) overnight refolding times. (**G-I**) Summary tables showing the number of refoldable proteins, non-refoldable proteins, and proteins which had only 1 peptide detected (and hence were excluded from the analysis), as well as their frequencies for different refolding times. Below are shown summary tables at the domain level. Regions that fell outside of annotated domains (linkers) are tabulated as well but were not used for

further analysis. Summary tables correspond to the (**G**) 1 min; (**H**) 5 min; (**I**) overnight refolding times. (**J**) The analogous summary table for the 120 min timepoint (the reference dataset used for most of the main text) included for comparison.

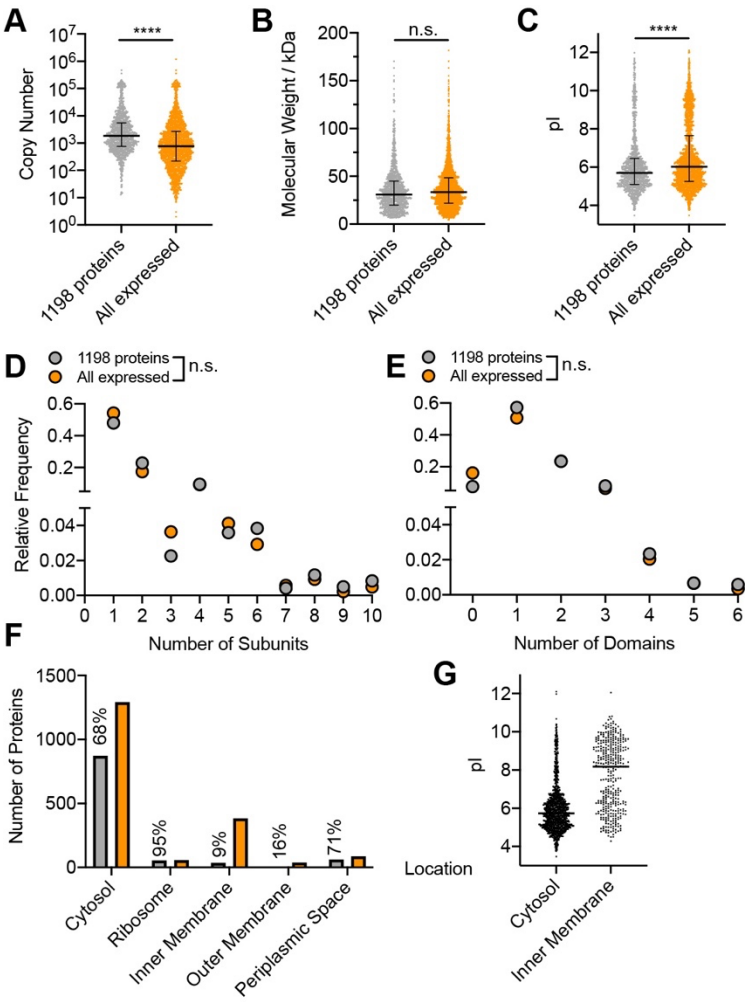

Fig. S5. Assessment of Bias in Study's Dataset. Comparison of properties between the 1,198 proteins that comprise the primary dataset in this study to the full set of 2,270 proteins expressed in *E. coli* during log phase in rich media (ref. 22). (**A**) Distributions of protein copy number. The set of 1,198 proteins is enriched for higher-abundance proteins (median copy number is 1846, compared to 769; $P < 0.0001$ by Kolmogorov-Smirnov test). (**B**) Distributions of molecular weight. The set of 1,198 proteins is similarly distributed with respect to molecular weight (median molecular weight is 31 kDa compared to 33 kDa). (**C**) Distributions of pI. The set of 1,198 proteins is de-enriched for proteins with higher pI (median pI is 5.7 compared to 6.0; $P < 0.0001$ by Kolmogorov-Smirnov test), see panel **G**. (**D**) Histograms of number of subunits. The set of 1,198 proteins is similarly distributed with respect to number of subunits (average subunit count is 4.0 compared to 4.0). (**E**) Histograms of number of domains. The set of 1,198 proteins is similarly distributed with respect to number of domains (average number of domains is 1.45 compared to 1.32). (**F**) Number of proteins, divided by location. The set of 1,198 proteins over-represents ribosomes, proportionately represents cytosolic and periplasmic proteins, and highly under-represents membrane proteins. (**G**) Distribution of pI, divided by location. The absence of membrane proteins explain the difference in the pI distrubtions in panel **C**.

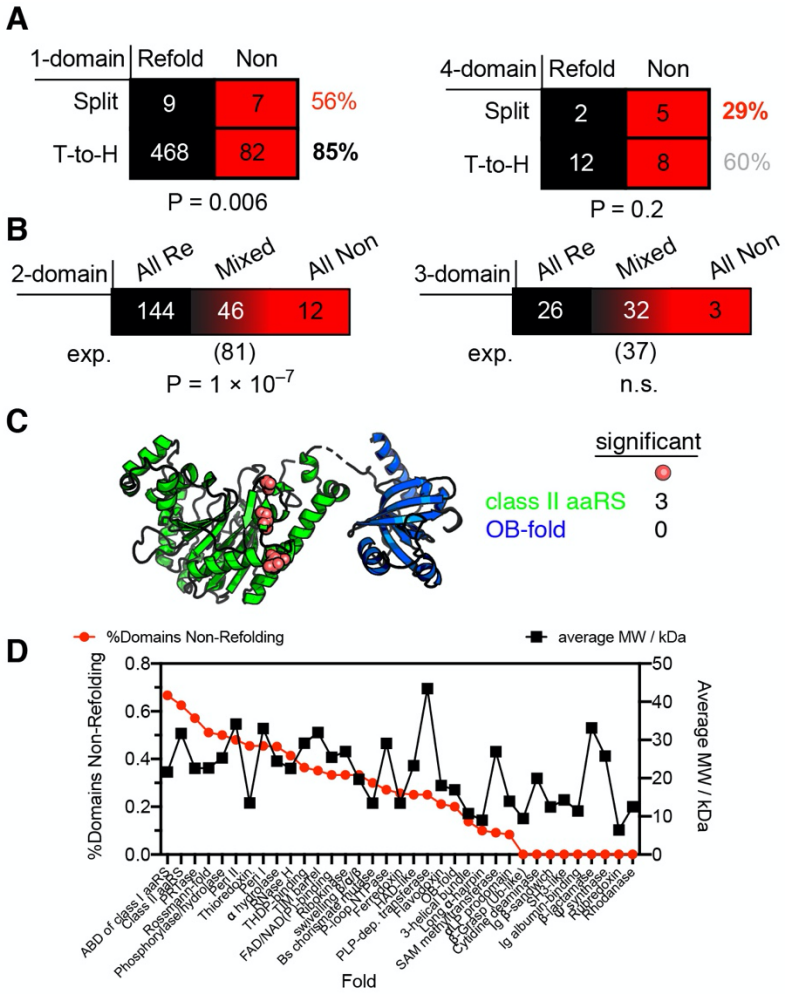

Fig. S6. Split Domains and Domain Coupling. (**A**) Splits in domain organization affect refoldability. Contingency tables showing the number of proteins with 1 or 4 domain(s) that are refoldable (or non-refoldable) divided into sub-classifications of whether the protein has domains that are split into multiple ranges with intervening sequences (split) or are arranged in a tail-to-head manner (T-to-H), as based on domain ranges from SUPERFAMILY2. Split 1-domain and 4-domain proteins tend to be *less* refoldable. The effect is statistically significant for 1-domain proteins (P = 0.006 by Fisher's exact test). Though the effect size is large for 4-domain proteins (~2-fold), the smaller number of observations makes the result not statistically significant. (**B**) Contingency tables showing the number of 2- or 3- domain proteins in which all domains are refoldable, all domains are non-refoldable, or a mixture. Particularly for 2-domain proteins, there is strong evidence for 'domain coupling' in that there are significantly fewer 2-domains proteins in which 1 domains folds and the other does not than expected (1.8-fold, P = 1 × 10^{-7} by Fisher's exact test). (**C**) X-ray structure of lysyl-tRNA synthetase (LysS; PDB: 1BBW), a 2-domain protein in which one domain (an OB-fold) is refoldable and the other (a class II aaRS/biotin synthase-like domain) is not. Although cases like LysS are less common (where one domain apparently refolds whereas the other does not), these two fold-types have high intrinsic differences in refoldability at the domain-level. (**D**) List of fold-types in

order to increasing level of refoldability (percent of domaisn red). Overlaid in black is the average molecular weight for each of the fold-types (amongst the examples included for which there is data). Refoldable fold-types are not larger or smaller than non-refoldable fold-types.

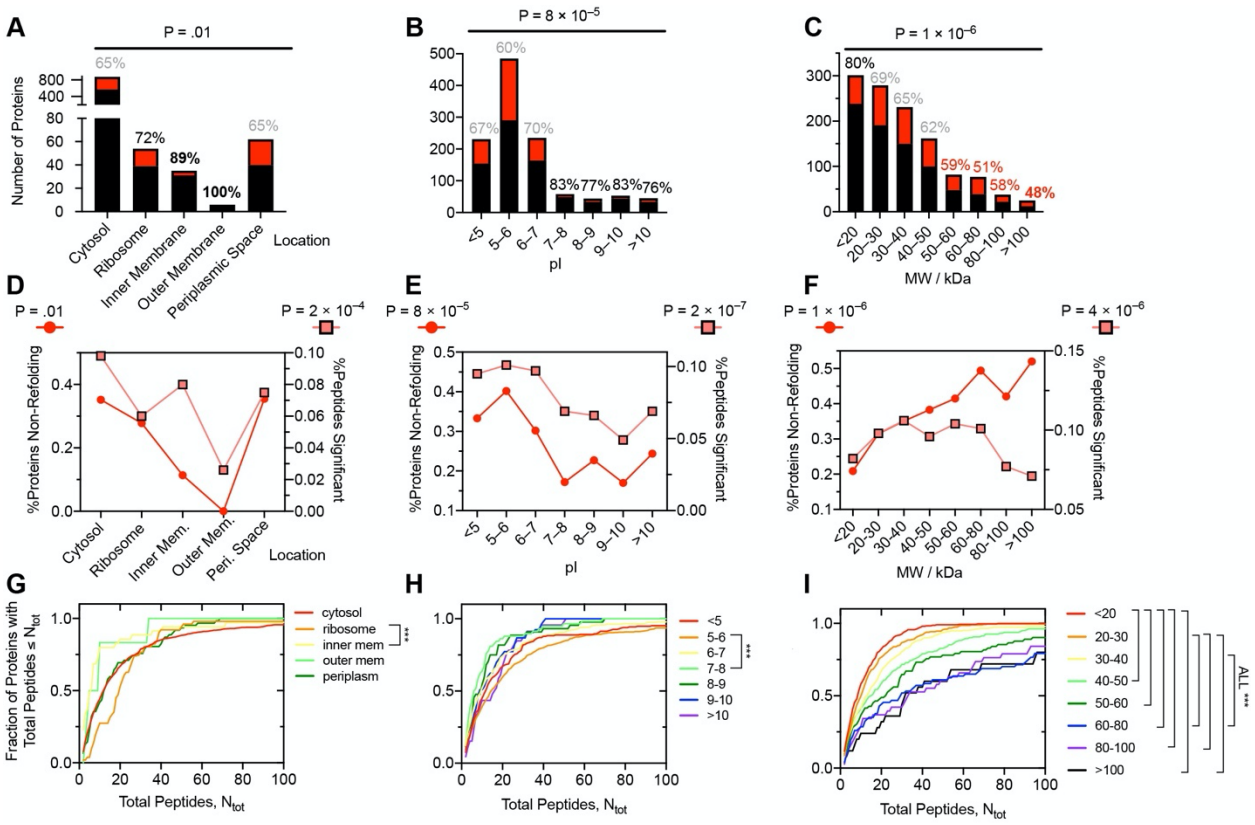

Fig. S7. Correlations between Refoldability and Location, pI, and Molecular Weight, and Checking for Sequence Coverage Bias (the "N_{tot} Bias"). A protein is more likely to be called non-refoldable if we identify and quantify more peptides associated with that protein (see Text S2). We call this the N_{tot}-bias (N_{tot} is the total number of peptides quantified for a particular protein). If a class of proteins is called more non-refoldable simply because it has greater coverage (e.g., more peptides were quantified per protein), then this bias will be apparent in two ways: the overall frequency of significant peptides for that class will not correlate with the frequency of non-refoldable proteins (e.g., **D-F**), and proteins of particular classes will have their N_{tot}'s distributed significantly differently than the other classes in that comparison group (e.g., **G-I**). Overall, the correlation between refoldability and molecular weight is likely biased (**F, I**), but the correlation with location and pI are not (**D, E, G, H**).

(**A-C**) Bar charts showing the number of (Non-)Refoldable proteins as a function of location (**A**), pI (**B**), and molecular weight (**C**). (**A**) Soluble protein regions that are localized to membranes are more refoldable than cytosolic or periplasmic proteins (see Text S3). (**B**) Proteins with pI between 7–8 and 9–10 are more refoldable than other groups. (**C**) Proteins of increasing molecular weight are more non-refoldable.

(**D-F**) Plots showing the frequency of non-refoldable proteins of a given classification (red circles) and the frequency of peptides that have significantly different abundance in refolded samples over native samples (pink squares). P-values based on chi-square

tests at the protein and peptide level for each analysis are given. (**D**) Locations with more non-refoldable proteins are also associated with higher frequencies of significant peptides (except for the inner membrane). (**E**) pI groups with more non-refoldable proteins are also associated with higher frequencies of significant peptides. (**F**) Proteins of high molecular weight tend to be non-refoldable, but do not generate significant peptides at a higher frequency, suggesting bias.

(**G-I**) Cumulative frequency distributions showing the fraction of proteins with N_{tot} total peptides quantified or fewer for each classification. (**G**) By the Kruskal-Wallis test, proteins with different locations are generally not associated with significant differences in their N_{tot}, with the exception of one pairwise comparison between ribosomal and inner membrane proteins ($P < 0.0001$). (**H**) By the Kruskal-Wallis test, proteins with different pI are generally not associated with significant differences in their N_{tot}, with the exception of one pairwise-comparison between pI 5–6 and pI 7–8 ($P = 0.0001$). (**I**) In contrast, proteins of greater molecular weight generate more peptides in a clear monotonic trend. By the Kruskall-Wallis test, many of the pairwise-comparisons are significant ($P < 0.0001$).

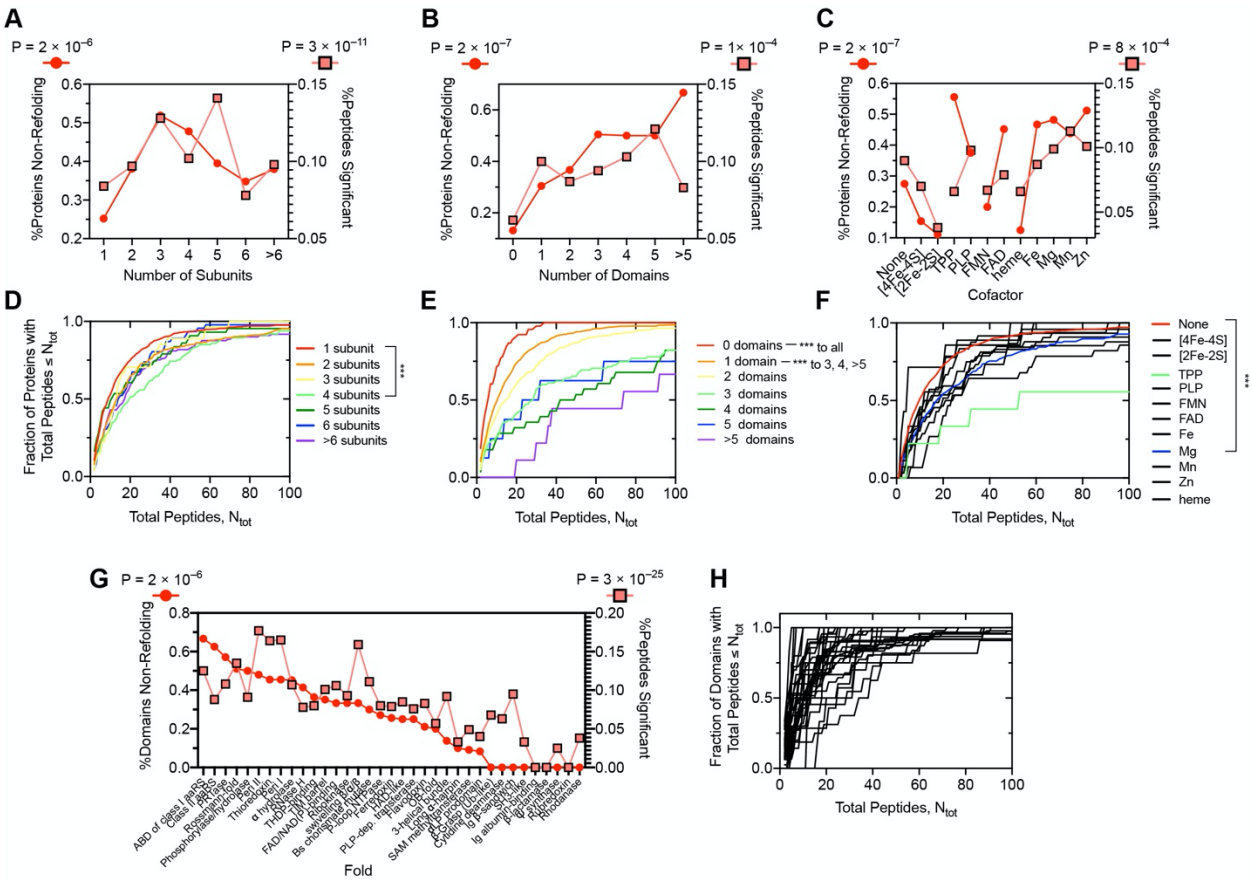

Fig. S8. Checking for Sequence Coverage Bias (the "N_{tot} Bias") in Correlations between Refoldability and Composition, Domains, Cofactors, and Fold-Type. (**A-C, G**) Plots showing the frequency of non-refoldable proteins (or domains) of a given classification (red circles) and the frequency of peptides that have significantly different abundance in refolded samples over native samples (pink squares). P-values based on chi-square tests at the protein and peptide level for each analysis are given. (**A**) Proteins that are monomers or part of hexamers are the most refoldable, and produce significant peptides at the lowest frequency. This correlation is robust across the series with the exception of pentamers. (**B**) Proteins with more domains are more non-refoldable, and also generate significant peptides at a higher frequency. The exception is proteins with > 5 domains. (**C**) Cofactors that are associated with higher levels of non-refoldability also produce significant peptides at higher frequency. The exception is TPP-proteins, which can be attributed to the fact that TPP proteins tend to have high peptide coverage (see **F**). (**G**) A weak correlation is apparent between the frequency of fold-types that are non-refolding and the frequency of significant peptides associated with that fold-type, most likely owing to small counts in these categories.

(**D-F, H**) Cumulative frequency distributions showing the fraction of proteins with N_{tot} total peptides quantified or fewer for each classification. (**D**) By the Kruskal-Wallis test, proteins with different subunit counts are generally not associated with significant

differences in their N_{tot}, with the exception of one pairwise-comparison between monomeric and tetrameric proteins ($P < 0.0001$). (**E**) Proteins with more domains generate more peptides in a clear monotonic trend that is similar to the molecular weight trend. By the Kruskall-Wallis test, pairwise-comparisons involving 0-domain proteins ($P < 0.0003$) and 1-domain proteins are significant ($P < 0.005$). (**F**) By the Kruskal-Wallis test, proteins with different cofactors are generally not associated with significant differences in their N_{tot}, with the exception of one pairwise-comparison between apo proteins and magnesium proteins ($P < 0.0001$). Although it is not statistically significant by the Kruskall-Wallis test, TPP proteins by inspection are associated with higher coverage levels, potentially explaining their higher apparent non-refoldability. (**H**) By the Kruskal-Wallis test, domains of different fold-types are not associated with significant differences in their N_{tot}.

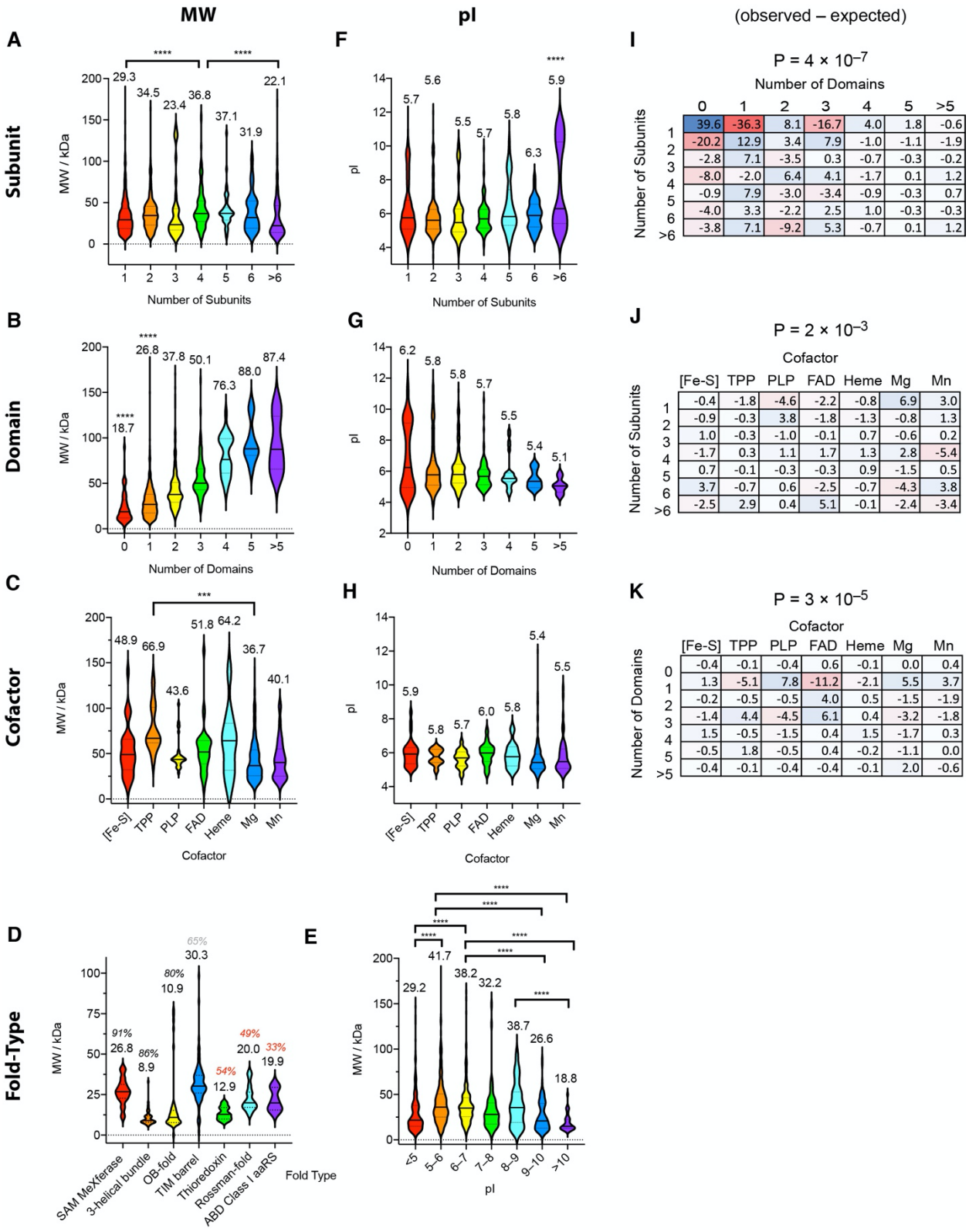

Fig. S9. Cross Correlation Analyses. (**A-E**) Plots showing the distribution of molecular weights of detected proteins for various classifications: (**A**) number of subunits; (**B**) number of domain; (**C**) cofactors; (**D**) fold-tye; (**E**) isoelectric point (pI).

Median molecular weight for each category is provided. (**A**) By the Kruskal-Wallis test, proteins with different numbers of subunits do not have significantly different molecular weights, with the exception of two pairwise-comparisons between monomeric and tetrameric proteins, and between tetrameric proteins and proteins part of complex with six or more subunits ($P < 0.0001$). (**B**) Proteins with more domains have larger molecular weights in a clear monotonic trend. However, by the Kruskall-Wallis test, only pairwise-comparisons involving 0-domain proteins and 1-domain proteins are significant. (**C**) By the Kruskal-Wallis test, proteins with different cofactors do not have significantly different molecular weights, with the exception of TPP-containing proteins, which are generally heavier (and also tend to be less refoldable). Note that heme-containing proteins also tend to have higher molecular weights, but are amongst the most refoldable proteins, indicating that molecular weight does not explain the relationship between cofactor and refoldability. (**D**) Molecular weight distributions of several fold-types in decreasing order of refoldability. Median molecular weights and percent refolding are given. By the Kruskal-Wallis test, fold-types are associated with different molecular weight distributions, but these do not explain the relationship between fold-type and refoldability. (**E**) By Kruskall-Wallis test, many pI groups are associated with different molecular weight distributions. Particularly, proteins with pI between 5–6 and 6–7 have on average higher molecular weights. These two pI groups also correspond to higher levels of non-refoldability, suggesting a confounding bias with molecular weight.

(**F-H**) Plots showing the distribution of pI's of detected proteins for various classifications: (**F**) number of subunits; (**G**) number of domains; (**H**) cofactors. Median pI for each category is provided. (**F**) By the Kruskal-Wallis test, proteins with different numbers of subunits do not have significantly different pI distributions, with the exception of proteins part of complexes with six or more subunits (due to ribosomal proteins). (**G**) By the Kruskal-Wallis test, proteins with different numbers of domains do not have significantly different pI distributions. (**H**) By the Kruskal-Wallis test, proteins with different cofactors do not have significantly different pI distributions.

(**I–K**) Contingency tables showing observed minus expected counts for the number of detected proteins with given pair of properties: (**I**) table showing the number of proteins with given number of subunits and given number of domains relative to expected; (**J**) table showing the number of proteins with given number of subunits and given cofactor relative to expected; (**K**) table showing the number of proteins with given number of domains and given cofactor relative to expected. Overall P-values based on chi-square test are given. (**I**) 0-domain proteins are highly enriched to be monomeric; 1-domain and 3-domain proteins are enriched to be multimeric. (**J**) There is very little enrichment for specific cofactors to co-occur with specific subunit counts. (**K**) FAD-containing proteins are enriched amongst multi-domain proteins.

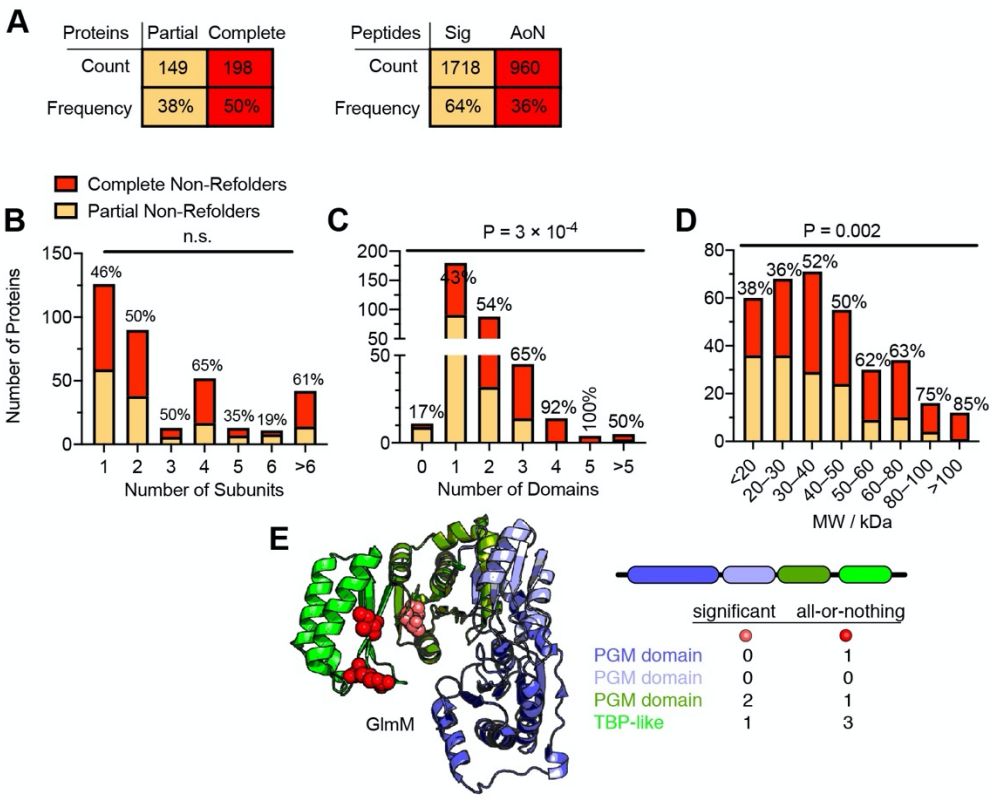

Fig. S10. Complete Non-Refolders and All-or-Nothing Peptides. (**A**) Count and frequency of partial and complete non-refolders, and of significant (sig) and all-or-nothing (AoN) peptides, according to the following definitions. All-or-Nothing peptides were not detected in all three biological replicates of either the native/refolded samples, and were detected in all three biological replicates on the other sample type. Proteins that have two or more all-or-nothing peptides are defined to be 'complete non-refolders' whereas those that have two or more significant peptides that are not all-or-nothing (64-fold > abundance ratio > 2-fold, P-value < 0.01 by Welch's t-test) and only 1 or 0 all-or-nothing peptides are defined to be 'partial non-refolders.' (**B**) Amongst non-refolders, there was no significant correlation between subunit composition and propensity to be a complete non-refolder, which is also the case for the majority of the classifications (see Data S1). (**C**) Non-refoldable proteins with more domains are more likely to be complete non-refolders ($P = 3 \times 10^{-4}$ by chi-square test). (**D**) Non-refoldable proteins with greater molecular weight are more likely to be complete non-refolders (P = 0.002 by chi-square test). (**E**) X-ray structure of GlmM (PDB: 6GYZ), a four-domain protein in which the final TBP-like domain is a complete non-refolder, a third PGM domain is a partial non-refolder, and the other two domains are refolding. Proteins with more domains are more likely to be complete non-refolders because they're more likely to have a domain that is a complete non-refolder.

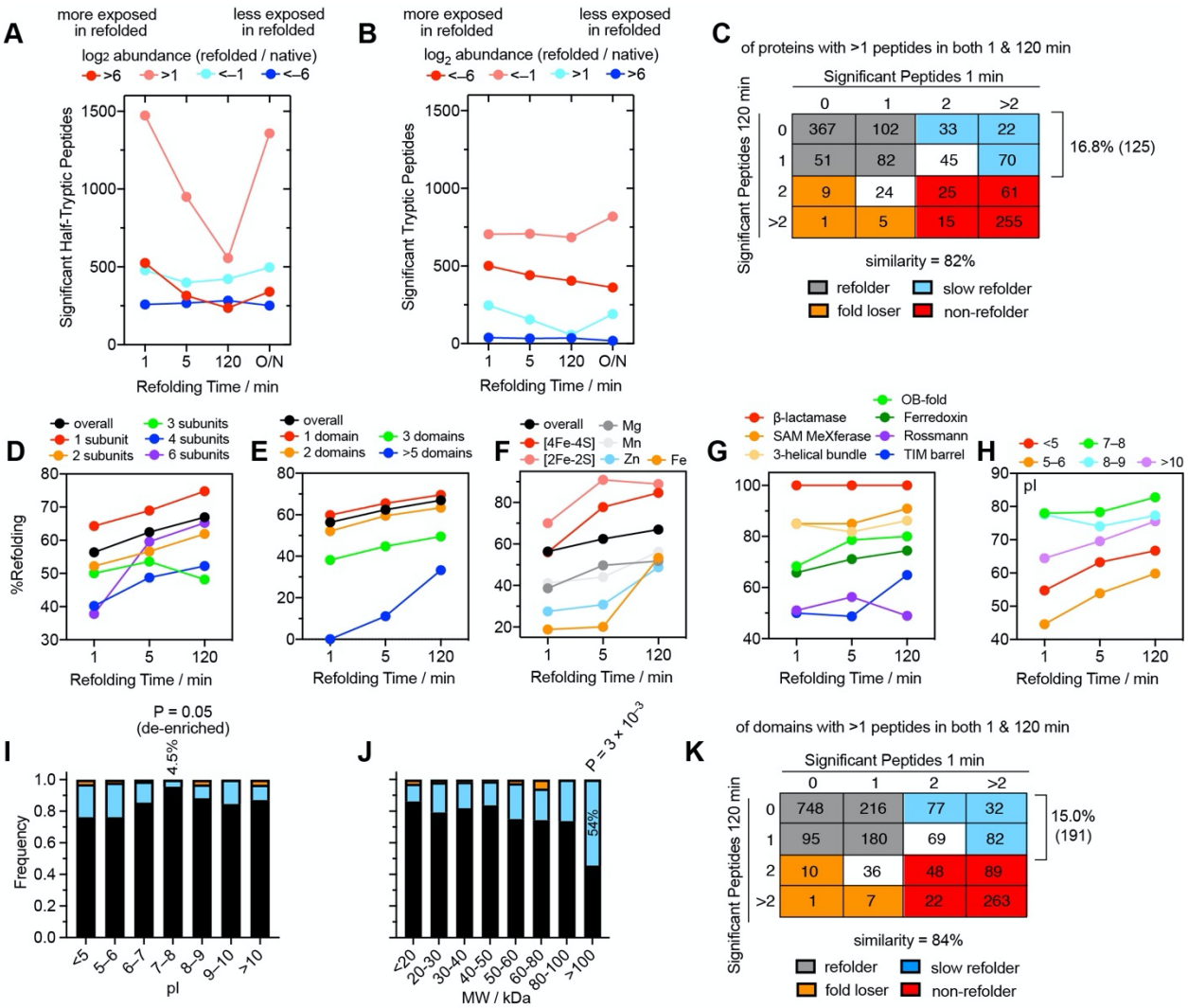

Fig. S11. Properties of Slow Refolders. (**A**) Plot showing the number of significant half-tryptic peptides identified at each time point following refolding experiments from GdmCl. Traces are colored based on the abundance ratio, with dark red points representing sites that are greatly more exposed in the refolded form and dark blue points representing sites that are greatly more exposed in the native form. The number of sites that are more exposed in the refolded form decrease up to 2 h (as the sample becomes more native-like), and then increase after overnight incubation. This is most likely due to reassembly of proteases that slowly degrade other proteins. In contrast, sites that are more exposed in the native form are fewer and do not vary as much over time. (**B**) Analogous to panel **A**, except showing the number of significant full-tryptic peptides. The color scheme is inverted to reflect that full-tryptic peptides that are less abundant in the refolded sample correspond to sites that are *more* exposed in the refolded form (the opposite of half-tryptic peptides). (**C**) A total of 1401 proteins were simultaneously identified in the 1-min and 120-min refolding experiments, of which 1167 had two or more peptides identified for that protein at both time points, permitting an

independent assessment of its refoldability. Table shows the number of proteins that had a given number of significant peptides mapped to it in the 1 min and 120 min time points. 602 were identified as refolders at both time points and 356 were identified as non-refolders at both time points. Overall, 82% of proteins had the same refolding status at both time points. Cells highlighted in blue correspond to proteins that required more than 1 min to refold ('slow refolders'). Cells highlighted in orange correspond to proteins that could refold rapidly but then misfolded afterwards ('fold losers'). For this analysis, proteins that were 'borderline' (had 1 significant peptide at one time point and 2 significant peptides at the other) were discounted (in white). Overall, 16.8% of these proteins are slow refolders (125 total). (**D-H**) Percent of proteins (or domains, for **G**) that are refolding as a function of time for several classifications: protein subunit count, (**D**); number of domains, (**E**); cofactor (**F**); fold-type, (**G**); and protein isoelectric point, pI (**H**). (**D**) Multimers generally do not require more time to refold than monomers, with the exception of hexamers. (**E**) Multidomain proteins generally do not require more time to refold than 1-domain proteins, with the exception of 6- and 7- domain proteins. (**F**) Iron-sulfur cluster-containing proteins are enriched for slow refolders. In a refolding buffer with additional supplemented Mg^{2+} (but with no other metal), Mg-metalloproteins evince a kinetic profile similar to the overall trend, whilst Mn-, Zn-, and Fe-metalloproteins refold very slowly, oftentimes requiring more than 5 min to become native-like. (**G**) SAM methyltransferases, 3-helical bundles, and β-lactamases are fast and efficient refolders. OB-folds and ferredoxin-folds mostly refold within 1 min but several members require more time to refold. TIM barrels are uniquely slow refolders, with many members requiring more than 5 min to refold. (**H**) Polyanionic and polycationic proteins (pI < 6 or pI > 10) tend to refold slowly compared to proteins with pI between 7–9. (**I, J**) Frequency bars showing the fraction of the proteins of a given classification that were refolders at both time time-points (black), slow refolders (refolded within 2 h but not 1 min, blue), and fold losers (refolded within 1 min but not at 2 h, orange). Note that these analyses only cover proteins that were identified at both time points with two more peptides each. (**I**) Proteins with pI between 7–8 are significantly de-enriched in the population of slow-refolders (3.7-fold, P = 0.05 by chi-sqaure test) suggesting they tend to refold quickly. The same is true for proteins with pI between 8–9 (2-fold) but the effect is not statistically significant. (**J**) Proteins with molecular weights >100 kDa are significantly enriched in the population of slow refolders (3.2-fold, P = 0.003 by chi-square test), but not proteins in other molecular weight ranges. (**K**) Analogous to **C** but conducted on domains rather than proteins. A total of 2561 domains were simultaneously identified in the 1-min and 120-min refolding experiments, of which 1975 had two or more peptides identified for that protein at both time points, permitting an independent assessment of its refoldability. Table shows the number of domains that had a given number of significant peptides mapped to it in the 1 min and 120 min time points. 1239 were identified as refolders at both time points and 422 were identified as non-refolders at both time points. Overall, 84% of domains had the same refolding status at both time points, and 15% of these domains are slow refolders (191 total).

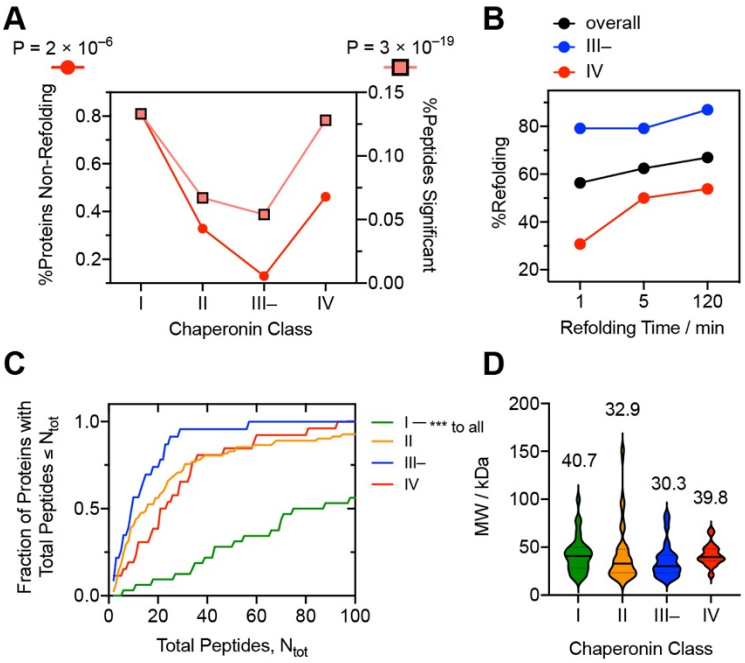

Fig. S12. Peptide-level and Kinetic Analysis of Chaperonin Classes. (**A**) Plot showing the frequency of non-refoldable proteins of a given chaperonin class (red circles) and the frequency of peptides that have significantly different abundance in refolded samples over native samples (pink squares). P-values based on chi-square tests at the protein and peptide level for each analysis are given. Class III^{-} proteins are the most refoldable, class II and class IV proteins are intermediate, and class I proteins are the most non-refoldable; these trends are recapitulated well at the peptide level. (**B**) Percent of proteins that are refolding as a function of time, grouped by chaperonin class. Class IV proteins tend to be slower refolders. (**C**) Cumulative frequency distributions showing the fraction of proteins with N_{tot} total peptides quantified or fewer for each chaperonin class. By the Kruskal-Wallis test, class II, III^{--}, and IV proteins are not associated with significant differences in their N_{tot}, though class I proteins do tend to have larger N_{tot} ($P < 0.0005$) owing to the fact that they tend to be more abundant. (**D**) Molecular weight distributions associated with each chaperonin class; there are no statistically significant differences.

Data S1. DataS1_GdmClRefolding_Protein.xlsx

Protein-level summary files for refolding for 1 min, 5 min, 120 min, and overnight incubations following dilution from 6 M GdmCl. Observed and expected counts and chi-square analyses. Legend tab provides definitions of each column.

Data S2. DataS2_GdmClRefolding_Domain.xlsx

Domain-level summary files for refolding for 1 min, 5 min, 120 min, and overnight incubations following dilution from 6 M GdmCl. Observed and expected counts and chi-square analyses. Legend tab provides definitions of each column.

Data S3. DataS3_TimeCourses.xlsx

Merged summary files showing the proteins (domains) that were simultaneously identified in both the 1 min (5 min) and 120 min timepoints, and were therefore used to assess status as (very) slow refolder. Legend tab provides definitions of each column.

Data S4. DataS4_GdmClRefolding_v1_Protein.xlsx

Analogous to Data S1 for an earlier (v1) of the experiment in which refolding was carried out without ribosome depletion and at slightly higher concentration. The presence of ribosomes in lysates result in much greater levels of aggregation.