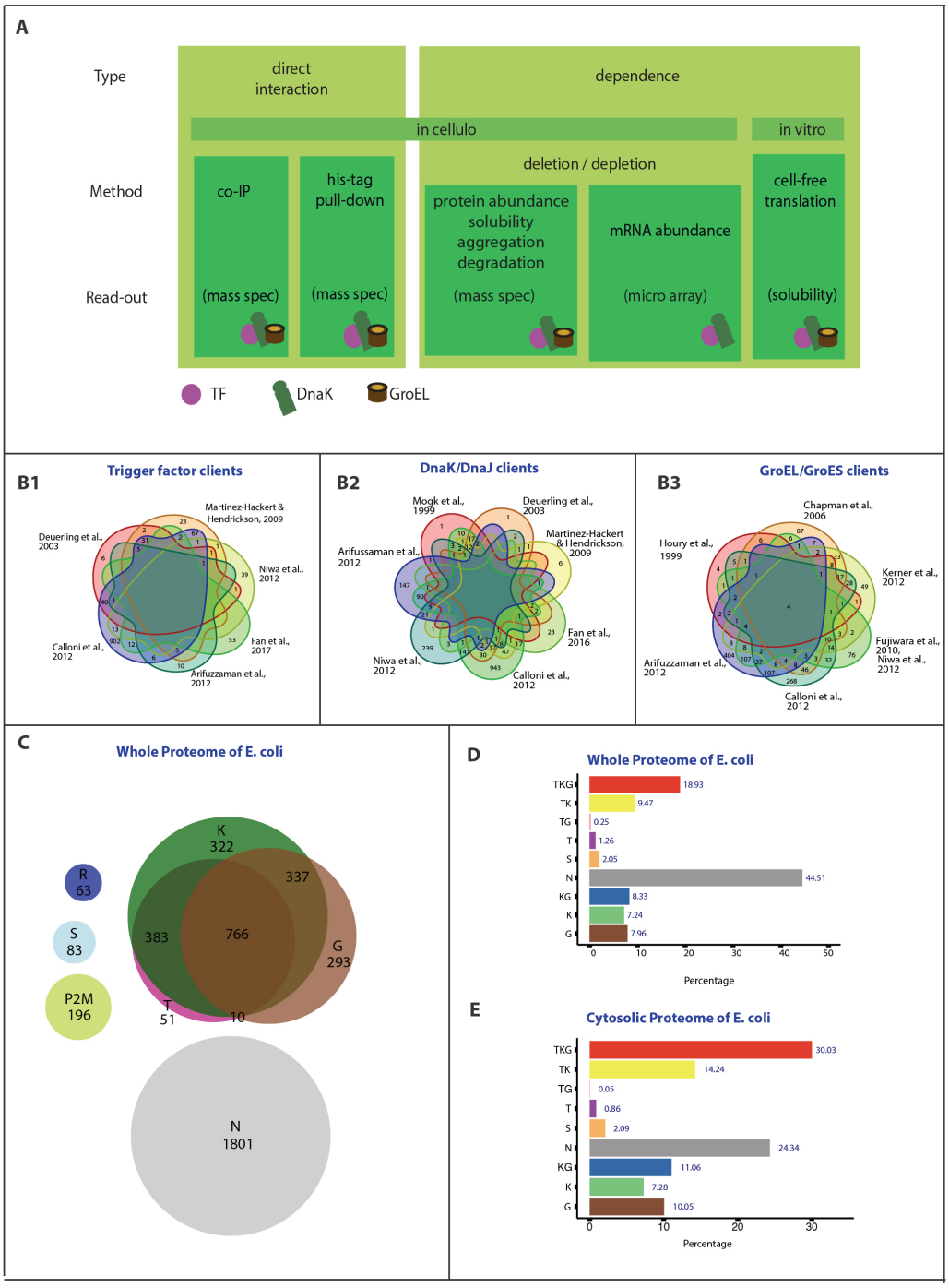
SUPPLEMENTARY MATERIALS

# Differential proteostatic regulation of insoluble and abundant proteins

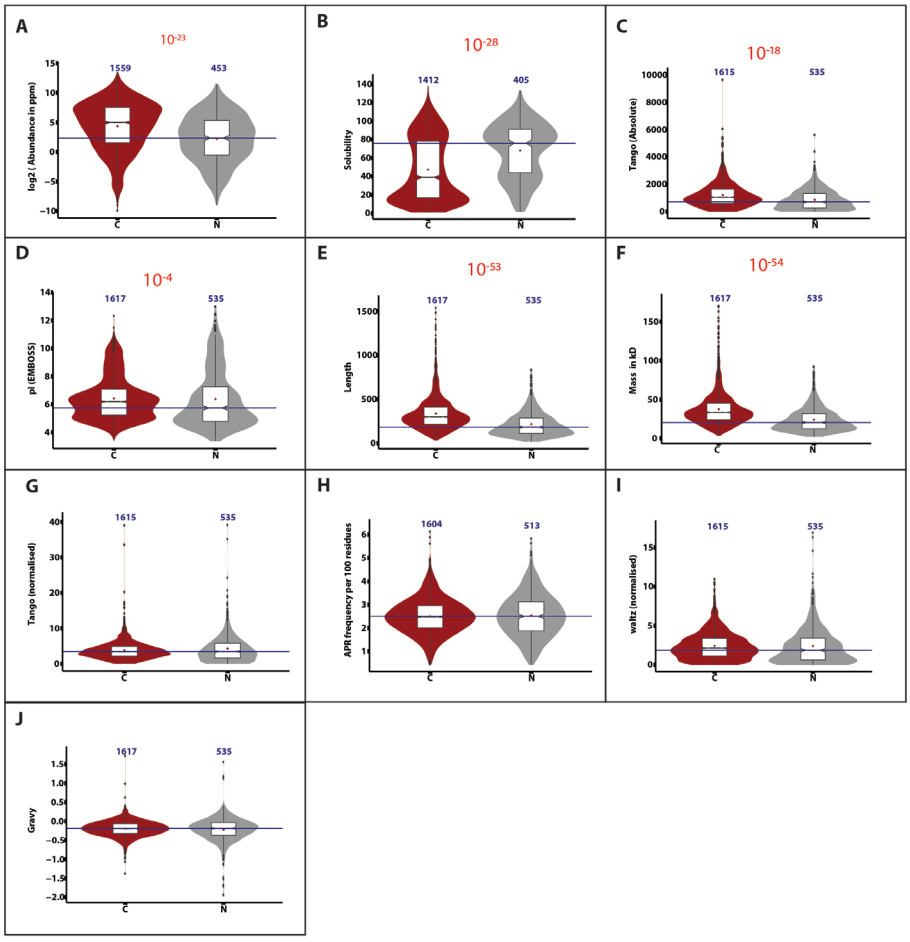
Reshmi Ramakrishnan, Bert Houben, Frederic Rousseau* & Joost Schymkowitz*
Switch Laboratory, VIB Center for Brain and Disease Research, and Department of Cellular and Molecular Medicine, KULeuven, Herestraat 49, 3000 Leuven, Belgium.

# Supplementary Figures



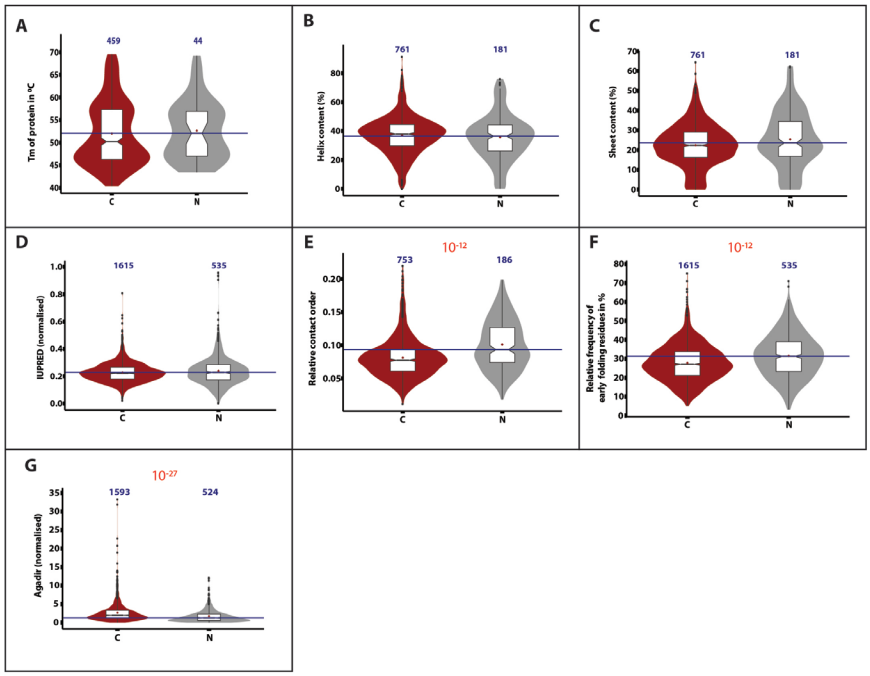

**Supplementary Figure 1: Overview of the chaperone interactome data collected for the study. (A) Overview of the types of experimental data collected for our database**: Studies are divided into two groups based on the type of evidence they provide for chaperone association: "direct interaction" or "dependence" (indirect). "Direct interaction" refers to those studies that directly show protein-protein interaction with chaperones and mainly entails pull-down assays (either his-tag pull-down or co-IP) followed by mass spectrometric analysis. All the direct interaction studies in our dataset were performed *in cellulo*. "Dependence" on the one hand refers to techniques that use *in cellulo* deletion or depletion of a specific chaperone and infer chaperone association through the detection of changes in protein solubility, aggregation status, proteome or transcriptome abundance, increased degradation or altered dependence on other chaperones. On the other hand, "dependence" is also inferred from *in vitro* translation studies that detect changes in solubility upon addition of a particular chaperone. Symbolic depictions of the three major cytosolic *E. coli* chaperone systems – Trigger Factor (TF), DnaK and GroEL – indicate for each type of analysis whether experimental data on that particular chaperone system was available. The amalgamation of data from different studies/methods helps to encompass chaperone dependency at a proteome-wide range.

**(B) Correspondence of data from different sources**: Venn diagrams show overlap between chaperone client sets determined by each study. Studies are grouped per chaperone system under consideration B1) Overlap between studies identifying Trigger Factor clients. B2) Overlap between studies identifying DnaK clients B3) Overlap between studies identifying GroEL/ES clients. **(C) Chaperone dependency overview for the *E. coli* proteome**: Euler diagram indicates sizes of protein sets dependent on a specific chaperone, as well as the overlap between these sets. "K", "T" and "G" represent the sets of proteins classified by at least one of the studies as clients of DnaK, Trigger Factor or GroEL respectively. "N" indicates the group of non-binders, proteins that don't show chaperone dependency in any of the studies considered. The proteins contained in the "S" group showed contradictory results i.e. increased mRNA or protein abundance upon chaperone deletion, and no other apparent chaperone dependencies in any of the studies. "R" indicates the group of redundant proteins identified using the CD-HIT algorithm at 90% sequence identity, which were removed from further analyses. **(D) Percentage of *E. coli* proteome classified to each of the chaperone flux groups. (E) Percentage of *E. coli* cytosolic proteins classified to each of the chaperone flux groups.**

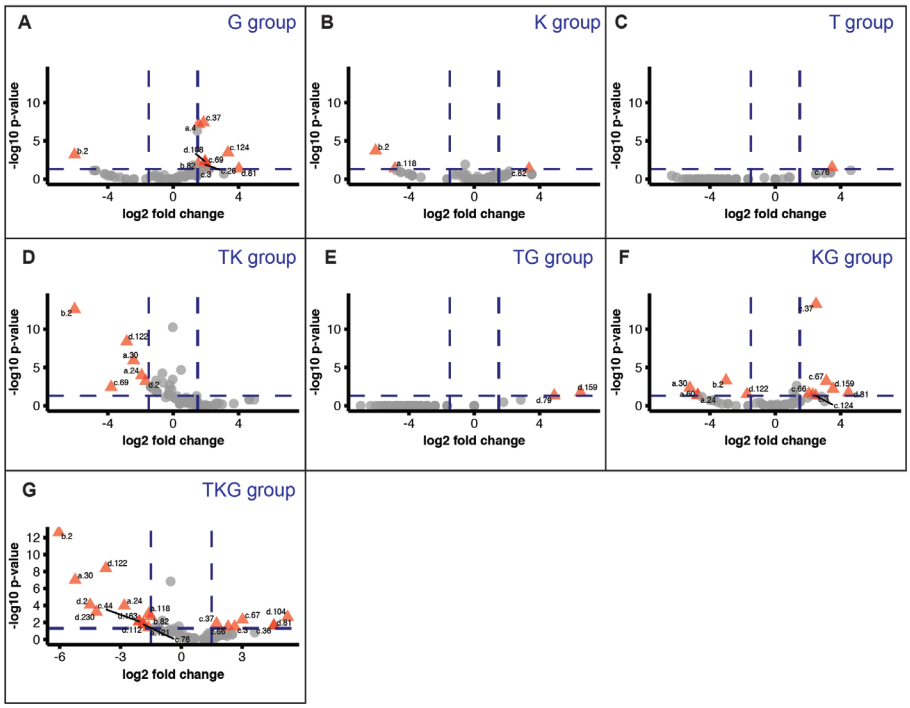



**Supplementary Figure 2: Distribution of protein-specific factors in chaperone-dependent folders (C) versus spontaneous folders (N).** Distributions are shown through a combination of boxplots (in white) and violin plots (colored). The lower and upper hinges of the boxplots indicate first and third quantiles, the bands across boxes show the median. The upper whiskers indicate the largest value no further than 1.5 times the Inter-Quartile Range (IQR) from the upper hinge, the lower whiskers show the smallest value at least 1.5 * IQR from the lower hinge. Notches extend 1.58 * IQR / sqrt(n) from either side of the median and represent a 95% confidence interval for the median value. To ascertain multimodality of distributions, violin plots were added. Violin plots correspond to rotated kernel density plots extending from each side of the boxplot and show a probability density of the data at different y-values. Red dots correspond to arithmetic means per group. For easy comparison, a horizontal blue line was added at the height of the median value of the non-binders group in each plot. The number of proteins in each group is shown above the respective violin plots. Statistical significance was assessed through a non-parametric Kruskal-Wallis test followed by *post-hoc* pairwise Wilcoxon testing with Bonferroni correction for multiple comparisons. The resulting p-values for comparison with the non-binders group are shown above each plot. **(A)** Abundance

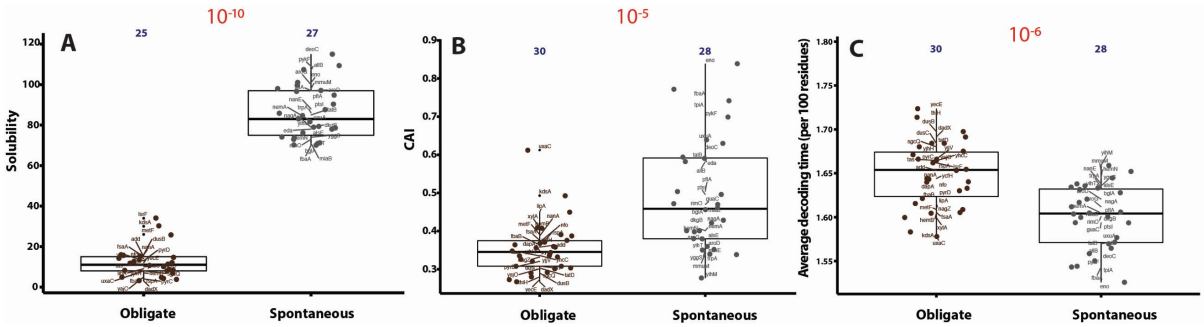
in p.p.m. determined by mass spectrometry (Wang, et al., 2012). **(B)** Solubility during cell-free translation, expressed as percent soluble protein (Niwa, et al., 2009). **(C)** Total aggregation propensity as predicted by TANGO (Fernandez-Escamilla, et al., 2004). **(D)** Isoelectric point (pI). **(E)** Protein length in amino acids. **(F)** Molecular weight (kDa). **(G)** Length-normalized aggregation propensity (TANGO). **(H)** APR frequency (TANGO). **(I)** Distribution of length-normalized Waltz score (Maurer-Stroh, et al., 2010). **(J)** Hydrophathy score (GRAVY).



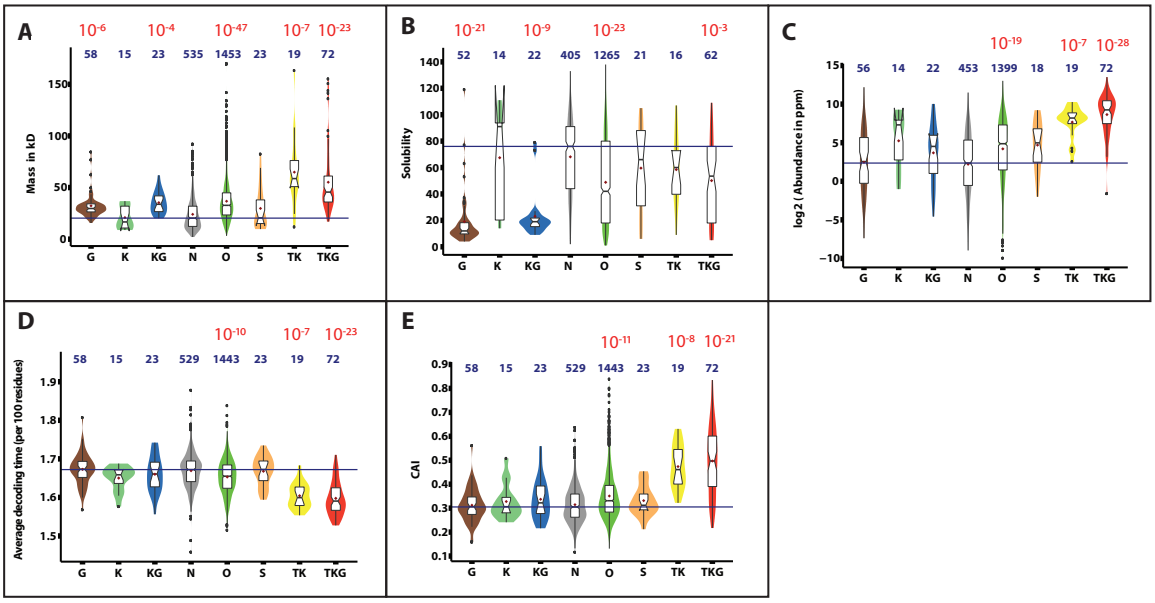

**Supplementary Figure 3: Distribution of protein-specific factors in chaperone dependent folders (C) versus spontaneous folders (N).** Plots are constructed as those in Figure 3. **(A)** Melting temperature (Tm, ºC) (Leuenberger, et al., 2017). **(B)** Percentage of residues in helices in the native structure. **(C)** Percentage of residues in sheets in the native structure. **(D)** Prediction of the intrinsically disordered propensity of the sequence from the summed IUPRED score (Dosztanyi, et al., 2005). **(E)** Relative contact order (Plaxco, et al., 1998), i.e. the average distance in sequence between residues that interact in the native state, normalized to the chain length. **(F)** Frequency of sequence segments capable of nucleating protein folding (foldons), predicted by EFoldMine (Raimondi, et al., 2017). **(G)** Alpha-helix propensity of the sequence as predicted by Agadir (%) (Munoz and Serrano, 1997).




**Supplementary Figure 4: Enriched SCOP superfamilies (Lo Conte, et al., 2000; Pandurangan, et al., 2018) in the different chaperone fluxes.** The data are displayed as so-called volcano plots, which show the fold enrichment in the x-axis and the p-value of the significance of the enrichment (Fisher Exact) as a power of ten in reverse order on the y-axis. The dotted lines indicate the preset cutoffs employed to identify significant results. Each point on the plot corresponds to a superfamily, the significantly enriched superfamilies are shown in red and have their SCOP code indicated. For a full list, including human-readable names, please refer to supplementary table 1.




**Supplementary Figure 5: Reanalysis of the TIM barrel superfamily.** The data are shown as box plots constructed in the same way as those in Figure 3. **(A)** Solubility during cell-free translation (Niwa, et al., 2009). **(B)** Codon Adaptation Index (CAI) (Sun, et al., 2013). **(C)** Decoding time (Dana and Tuller, 2014). **(D)** Translation efficiency (Li, et al., 2014). **(E)** Relative fraction of lysine residues. **(F)** Relative fraction of arginine residues.




**Supplementary Figure 6: Validation through a reanalysis of cross-sectional data.** After removing the proteins in the outer layers, which in total constitute 1453 proteins (O), the chaperone substrates dropped from 1617 to only 187, divided over the categories as follows; 72 (TKG), 19 (TK), KG (23), K (15), G (58), T (0), TG (0). We verified that for the main findings we found the same conclusions than for the more inclusive approach: (A) mass, (B), (C) abundance, (D) decoding time and (E) codon adaptation index.

# Supplementary Methods

## Data acquisition and database construction

The *E. coli* K12 reference proteome (4305 proteins) and the amino acid sequences of the proteins therein were obtained from UniProt (UniProt, 2008) (proteome ID: UP000000625). Chaperone dependency classifications were determined as outlined in detail in the "Results" section and mapped to the reference proteome. The dataset was then expanded with a set of experimentally determined, transcriptome- and proteome-wide features: protein solubility and cell-free expression yield were obtained from the *in vitro* translation analyses of Niwa et al. (Niwa, et al., 2009); Intracellular protein abundance was acquired from the mass-spectrometry-based integrated *E. coli* dataset from the PaxDb database (Wang, et al., 2012) and from ribosome-profiling data obtained by Li et al., (Li, et al., 2014); The latter study also provided data on mRNA abundance and translation efficiency; Genome-wide transcriptomic microarray analyses by Esquerre et al. (Esquerre, et al., 2016) provided a second mRNA abundance scale, as well as mRNA half-life measurements. Protein melting temperatures (Tm) were obtained from limited proteolysis and mass spectrometry analyses performed by Leuenberger et al. (Leuenberger, et al., 2017).

In order to assess nucleotide sequence characteristics, sequences were obtained from the European Nucleotide Archive (ENA) (Harrison, et al., 2018). Protein structures for the calculation of Contact Order (CO) (Plaxco, et al., 1998) were retrieved from the Protein Data Bank (PDB) (Westbrook, et al., 2003). When available, the optimal resolution structure with a coverage of at least 40% was retrieved for each protein. Based on nucleotide sequence, primary amino acid sequence and protein structures, the feature-space was expanded using a combination of database cross-references, simple calculations and advanced bio-informatics tools: Aggregation propensity and aggregation-prone regions were identified using the TANGO algorithm (Fernandez-Escamilla, et al., 2004); The WALTZ algorithm was employed to predict amylogenic regions (Maurer-Stroh, et al., 2010); Alpha-helix propensity in the unfolded state was calculated using the thermodynamics algorithm AGADIR (Munoz and Serrano, 1997). Intrinsic disorder calculations were performed using IUPRred (Dosztanyi, et al., 2005); The EFoldMine method (Raimondi, et al., 2017) was used to determine for each protein the percentage of reidues predicted to be capable of initiating protein folding. To this end, residues were classified as being part of a foldon if their EFoldMine early folding score exceeds 1.63; GRAVY (grand average of hydropathy) scores were determined by calculating the average hydropathy per protein using the method developed by Kyte & Doolittle (Kyte and Doolittle, 1982); Average decoding times were calculated based on decoding time scales devised by Dana & Tuller (Dana and Tuller, 2014); Isoelectric point values were obtained as the average of different scales using the standalone version of the Isoelectric Point Calculator (IPC) (Kozlowski, 2016); Codon Adaptation Index (CAI) was calculated using the CodonW software (http://codonw.sourceforge.net//culong.html). Structural classification and protein topology were mapped from the SCOPe (Chandonia, et al., 2017) and SUPERFAMILY

(Pandurangan, et al., 2018) databases. Secondary structure content was calculated from UniProt secondary structure annotations based on a consensus between PDB structures; relative CO (Plaxco, et al., 1998) was calculated from the PDB structures by determining the average sequence distance between amino acids that form native contacts, divided by protein length.

## Data clean-up

Using the CD-HIT algorithm (Fu, et al., 2012) all redundant protein sequences were removed at 90% sequence identity. UniProt annotation was used to remove ribosomal proteins were to reduce the known bias in charge and size distributions. Chaperones, Heat Shock Proteins (HSPs) and proteases were also removed considering the possibility of functional association between them. Chaperones and HSPs were removed based on UniProt annotation, proteases were filtered out using the *E. coli* protease database compiled by the Ehrmann lab (https://www.uni-due.de/zmb/members/ehrmann/e-coli-proteases/all-proteases.shtml) (Clausen, et al., 2002). Finally, non-cytosolic proteins were removed, using EcoCyc (Keseler, et al., 2017) subcellular location classifications, specifically the SmartTable entitled "All cytosolic proteins of *E. coli* K-12 substr. MG1655". Combined, all these clean-up steps left us with a database of 2198 proteins.

## Data analysis and visualization

All parsing, mapping, data preprocessing and calculations were performed by scripts written in Python 2.7.10. Statistical analysis was performed using R free statistical software version 3.3.1 and most of the graphical outputs were created using ggplot2 package version 2.1.0.

### Translational efficiency variant design

GFP variants were designed based on the codon decoding time scale devised by Tuller and colleagues (Dana and Tuller, 2014). The original GFP cDNA sequence from *Aequorea Victoria*, was obtained from GenBank (M62653.1) and GFP cycle 3 mutations F99S, M153T and V163A were introduced, as well as the A206K mutation in order to minimize GFP dimerization (Crameri, et al., 1996; Zacharias, et al., 2002). An N-terminal V5 tag and C-terminal His tag were added. The "fast" and "slow" variants were designed by exchanging each codon in the GFP sequence (excluding tags) for its counterpart with the lowest or highest decoding time, respectively. All three constructs were produced by Genscript and subcloned into the Invitrogen pBAD/Myc-His vector, controlled by an arabinose-inducible promotor. For visualization of unfolded GFP in inclusion bodies, a Tetracystein tag was added to the inclusion bodies through site-directed mutagenesis.

### *In vitro* translation and solubility analysis

*In vitro* translation assays were performed using the New England Biolabs Inc. PURExpress® *in vitro* protein synthesis kit. Where mentioned, reactions were supplemented with DnaK mix or GroE mix, obtained from Cosmo Bio as part of their PUREfrex® system. Both the PURExpress® and PUREfrex® systems are based on the PUREsystem™ devised by Shimizu and colleagues (Shimizu, et al., 2001). Linear template DNA with a T7 promotor for expression in the PURExpress® system was produced through PCR according to the manufacturer's instructions. Protein solubility upon cell-free translation was determined as previously described by Taguchi and colleagues (Niwa, et al., 2009). Briefly, cell-free translation was performed for 1 hour at 37°C, with or without the addition of DnaK or GroE mix following the manufacturers instructions, after which soluble and insoluble fractions were separated through centrifugation at 21000 g for 30 mins at 4°C. Total and soluble fractions were diluted 1:10 in 8M urea to completely unfold and dissolve any protein produced, and protein levels in each fraction were determined through SDS-PAGE followed by Western blotting. Blots were developed using chemiluminescence after incubation with primary anti-GFP antibody (Cell Signaling Technologies antibody 2555S) and secondary HRP-conjugated antibody. Blots were quantified using Bio-Rad's Image Lab™ Software. Soluble expression was determined by calculating the ratio of soluble over total protein.

***In cellulo* expression and solubility analysis**

The vectors we designed were transformed into *E. coli* K12 MG1655. Where indicated, co-transformation with vector pKJE7 obtained from the Takara Bio was performed. This vector overexpresses DnaK, DnaJ and GrpE upon arabinose induction. For protein expression and solubility analysis, bacterial strains were grown overnight in Lysogeny Broth (LB) supplemented with ampicillin for GFP expression and both ampicillin and chloramphenicol for co-expression of the GFP constructs with pKJE7. The overnight cultures were diluted 1:100 in fresh LB supplemented with the appropriate antibiotics and grown to an OD of about 0.6, after which expression was induced with 0.2 % arabinose. Expression was allowed to proceed for 3 hours after which cells were lysed in in B-PER™ reagent supplemented with 0.1 mg/ml lysozyme (Sigma-Aldrich), cOmplete™ Protease Inhibitor Cocktail (Sigma-Aldrich) and Pierce™ universal nuclease for cell lysis (ThermoFisher). Cells were lysed on ice for 30 mins, after which soluble and insoluble fractions were separated through centrifugation at 17100 g for 30 mins at 4°C. Supernatant was removed and the insoluble fraction dissolved in an equal volume of 8M urea. GFP in soluble and insoluble fractions was then quantified through SDS-PAGE followed by Western blotting. Blots were developed using chemiluminescence after incubation with primary anti-GFP antibody (Cell Signaling Technologies antibody 2555S) or anti-DnaK antibody (USBio D8076) and secondary HRP-conjugated antibody. Blots were quantified using Bio-Rad's Image Lab™ Software. Soluble GFP fractions were determined by calculating the ratio of soluble over total (soluble + insoluble) protein.

**Structured Illumination Microscopy (SIM)**

After three hours of overexpression, cells were fixed by adding 2.5 % Paraformaldehyde and 0.04 % glutaraldehyde (final concentrations) to culture media, followed by incubation at room temperature for 15 mins and 30 mins on ice. Cells were then washed in PBS and resuspended in GTE buffer (50 mM Glucose, 25 mM Tris, and 10 mM EDTA, pH 8.0). Directly preceding microscopic analysis, cells were transferred to a glass slide and covered with a coverslip. For staining of tetracystein-tagged GFP with the ReAsH reagent (ThermoFisher), ReAsH-EDT2 reagent was added 1 hour after induction. Cells were fixed two hours later as described above, with the inclusion of an additional wash step with 1X BAL wash buffer in PBS, before being transferred to a glass slide for SIM imaging. Imaging was performed using a Zeiss Elyra S.1 system.

**Supplementary References**


Chandonia, J.M., Fox, N.K. and Brenner, S.E. SCOPe: Manual Curation and Artifact Removal in the Structural Classification of Proteins - extended Database. *J Mol Biol* 2017;429(3):348-355.

Clausen, T., Southan, C. and Ehrmann, M. The HtrA family of proteases: Implications for protein composition and cell fate. *Molecular Cell* 2002;10(3):443-455.

Crameri, A.*, et al.* Improved green fluorescent protein by molecular evolution using DNA shuffling. *Nature Biotechnology* 1996;14(3):315-319.

Dana, A. and Tuller, T. The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Research* 2014;42(14):9171-9181.

Dosztanyi, Z.*, et al.* IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 2005;21(16):3433-3434.

Esquerre, T.*, et al.* The Csr system regulates genome-wide mRNA stability and transcription and thus gene expression in Escherichia coli. *Sci Rep* 2016;6:25057.

Fernandez-Escamilla, A.M.*, et al.* Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature Biotechnology* 2004;22(10):1302-1306.

Fu, L.*, et al.* CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;28(23):3150-3152.

Harrison, P.W.*, et al.* The European Nucleotide Archive in 2018. *Nucleic Acids Res* 2018.

Keseler, I.M.*, et al.* The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. *Nucleic Acids Res* 2017;45(D1):D543-D550.

Kozlowski, L.P. IPC - Isoelectric Point Calculator. *Biol Direct* 2016;11(1):55.

Kyte, J. and Doolittle, R.F. A Simple Method for Displaying the Hydropathic Character of a Protein. *Journal of Molecular Biology* 1982;157(1):105-132.

Leuenberger, P.*, et al.* Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. *Science* 2017;355(6327).

Li, G.W.*, et al.* Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. *Cell* 2014;157(3):624-635.

Lo Conte, L.*, et al.* SCOP: a structural classification of proteins database. *Nucleic Acids Res* 2000;28(1):257-259.

Maurer-Stroh, S.*, et al.* Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nature Methods* 2010;7(3):237-U109.

Munoz, V. and Serrano, L. Development of the multiple sequence approximation within the AGADIR model of alpha-helix formation: comparison with Zimm-Bragg and Lifson- Roig formalisms. *Biopolymers* 1997;41(5):495-509.

Niwa, T.*, et al.* Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proceedings of the National Academy of Sciences of the United States of America* 2009;106(11):4201-4206.

Pandurangan, A.P.*, et al.* The SUPERFAMILY 2.0 database: a significant proteome update and a new webserver. *Nucleic Acids Res* 2018.

Plaxco, K.W., Simons, K.T. and Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol* 1998;277(4):985-994.

Raimondi, D.*, et al.* Exploring the Sequence-based Prediction of Folding Initiation Sites in Proteins. *Scientific Reports* 2017;7.

Shimizu, Y.*, et al.* Cell-free translation reconstituted with purified components. *Nature Biotechnology* 2001;19(8):751-755.

Sun, X.Y., Yang, Q. and Xia, X.H. An Improved Implementation of Effective Number of Codons (N-c). *Molecular Biology and Evolution* 2013;30(1):191-196.

UniProt, C. The universal protein resource (UniProt). *Nucleic Acids Res* 2008;36(Database issue):D190-195.

Wang, M.*, et al.* PaxDb, a database of protein abundance averages across all three domains of life. *Mol Cell Proteomics* 2012;11(8):492-500.

Westbrook, J.*, et al.* The Protein Data Bank and structural genomics. *Nucleic Acids Research* 2003;31(1):489-491.

Zacharias, D.A.*, et al.* Partitioning of lipid-modified monomeric GFPs into membrane microdomains of live cells. *Science* 2002;296(5569):913-916.