

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Simulation and modeling: All computer code developed in this work is available in the GitHub repositories https://github.com/obrien-lab/cg_simtk_protein_folding and <https://github.com/obrien-lab/Activation-Energy-Estimation-Workflow>, under the MIT license. MD simulations were performed via OpenMM v7.4.1 and Amber 17.
Enzyme activity measurements: UV-visible spectra were recorded on a Cary 60 spectrometer from Varian (Agilent Technologies, Santa Clara, CA) using the WinUV software package to control the instrument.
Limited Proteolysis experiments: No software was used in data collection. A Thermo Q-Exactive HF-x Orbitrap mass spectrometer was used to analyze protein digests.

Data analysis

Simulation and modeling: Computer code developed in this work for analyzing the simulation results is available in the GitHub repositories https://github.com/obrien-lab/cg_simtk_protein_folding and <https://github.com/obrien-lab/Activation-Energy-Estimation-Workflow>, under the MIT license. Visual Molecular Dynamics v1.9.1 was used for molecular/trajectory visualization and image/movie generation. Python v3.7, PyEMMA v2.5.7, ParmEd v3.4.1, Scipy, Numpy, Stride, Mdtraj v1.9.7, KnotPull v0.4.1, WHAM v2.0.11, and ChaperISM v1.0 were also used for analyzing simulation data.
Enzyme activity measurements: Data were analyzed by using Python v3.7.
Limited Proteolysis experiments: Proteome Discoverer (PD) Software Suite (v2.4, Thermo Fisher) and the Minora Algorithm were used to analyze mass spectra and perform Label Free Quantification (LFQ) of detected peptides. PD output files were outputted in a three-label hierarchy (protein > peptide group > consensus feature) and were further processed utilizing custom Python analyzer scripts that are available on GitHub: <https://github.com/FriedLabJHU/Refoldability-Tools/>.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw data for Figures 1 to 6 are provided. We cannot feasibly provide all ~5.3 TB of molecular dynamics trajectory data, but we provide the input data that was used to perform the simulations in the repository subdirectory https://github.com/obrien-lab/cg_simtk_protein_folding/blob/master/example/input_data.tar.xz. All the data that support the findings of this study, as well as the biological materials that were used for testing the enzymatic activity of DDLB and DHFR variants and for LiP-MS experiments, are available from the corresponding author upon reasonable request. The raw mass spectrometry data of DDLB and DHFR have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD031425. The PDB structures 3CLA, 4C5C, 1AQ2, 2ZCV, 1AKE, 1VHL, 3K6L, 3LBF, 2FMT, 1C2T, 1FDR, 3CW7, 2OFP, 4KKJ, 1PDA, 1RYK, 1LMB, 2ABD, 1CEI, 1IMQ, 256B, 1JO8, 1SHF, 1C9O, 1MJC, 1TEN, 1WIT, 1POH, 2QJL, 1SPR, 1E65, 3CHY, 2RN2, 5NWX, 5JTE, 6IOY, 3JBU, 4UY8, 6ENJ, 6ENU, 5JU8, 4V9D were used in this study. The databases EzCatDB (<https://ezcatdb.cbrc.pj.aist.go.jp/EzCatDB/>), UniProt (<https://www.uniprot.org/>), RCSB PDB (<https://www.rcsb.org/>) and CATH (<http://www.cathdb.info/>) were used in this study. A website (https://obrien-lab.github.io/visualize_entanglements/) was created to provide interactive visualization of the key misfolded entangled structures predicted in this study.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was conducted prior. The specific activity measurements of DDLB and DHFR were performed on five biological replicates, respectively. Each Limited Proteolysis experiment was conducted as biological triplicates to allow for statistical analysis. The numbers of simulation trajectories were chosen due to the limitations of our computational resources.
Data exclusions	No data were excluded.
Replication	The experiments for DDLB have been done for five biological replicates and all of them have the same trend of the enzymatic activity, which indicates the robustness and reproducibility of the our findings. The experiments for DHFR have been done for five biological replicates as well and exhibit no statistical difference in the specific activities of fast and slow variants. Limited Proteolysis experiments were conducted on two separate experiments. Experimental findings were consistent in both experiments. Simulations were replicated 50 (DHFR) and 100 (DDLB and CAT-III) times for co-translational folding, 500 (DHFR) and 1,000 (DDLB and CAT-III) times for post-translational folding, 30 times for disentangling timescale estimation and 100 times for DDLB refolding simulations, with different random seeds to allow for the calculation of statistics. Due to different random seeds, behavior is different between trajectories. All trajectories ran to completion. All attempts at replication were successful.
Randomization	E. coli lysates were divided into either being either a native or a refolded sample depending on if they were unfolded and refolded or not. Covariates are not relevant in our study as all bacteria are cultured simultaneously under identical growth conditions and simultaneously prepared under identical methods. Random allocation is not relevant to the other experiments/simulations because they were not grouped to get treatments.
Blinding	Investigators know which E. coli lysates are native samples and which ones are refolded samples as limited proteolysis experiments are conducted after different refolding time points. Our analysis compares the limited proteolysis peptide profile of our refolded samples to our native samples, so blinding is not possible. Blinding is also not possible to the other experiments/simulations for the same reason.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Antibodies used in the Time-course analysis of DDLB expression:
 Primary antibody: Mouse anti-His6 IgG, Clone: His.H8, Company: Invitrogen, catalog number: MA1-21315
 Secondary antibody: Goat anti-mouse-AP, Company: Millipore, Catalog number: 69266-3

Validation

<https://assets.thermofisher.com/TFS-Assets/LSG/Flyers/commitment-antibody-performance-flyer.pdf>