PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Protein stability prediction by fine-tuning a protein language model on a mega-scale dataset

Simon K. S. Chu[1], **Kush Narang**[2], **Justin B. Siegel**[3,4,5]*

1 Biophysics Graduate Program, University of California Davis, Davis, California, United States of America, **2** College of Biological Sciences, University of California Davis, Davis, California, United States of America, **3** Genome Center, University of California Davis, Davis, California, United States of America, **4** Department of Chemistry, University of California Davis, Davis, California, United States of America, **5** Department of Biochemistry and Molecular Medicine, University of California Davis, Davis, California, United States of America

* jbsiegel@ucdavis.edu

OPEN ACCESS

Citation: Chu SKS, Narang K, Siegel JB (2024) Protein stability prediction by fine-tuning a protein language model on a mega-scale dataset. PLoS Comput Biol 20(7): e1012248. https://doi.org/10.1371/journal.pcbi.1012248

Editor: Piero Fariselli, Universita degli Studi di Torino, ITALY

Received: December 11, 2023

Accepted: June 13, 2024

Published: July 22, 2024

Data Availability Statement: The code and materials are maintained on https://github.com/SimonKitSangChu/EsmTherm. Mutant-level predictions from our model and benchmark evaluation of state-of-the-art are available under Supplementary Materials.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Protein stability plays a crucial role in a variety of applications, such as food processing, therapeutics, and the identification of pathogenic mutations. Engineering campaigns commonly seek to improve protein stability, and there is a strong interest in streamlining these processes to enable rapid optimization of highly stabilized proteins with fewer iterations. In this work, we explore utilizing a mega-scale dataset to develop a protein language model optimized for stability prediction. $\text{ESM}_{\text{therm}}$ is trained on the folding stability of 528k natural and *de novo* sequences derived from 461 protein domains and can accommodate deletions, insertions, and multiple-point mutations. We show that a protein language model can be fine-tuned to predict folding stability. $\text{ESM}_{\text{therm}}$ performs reasonably on small protein domains and generalizes to sequences distal from the training set. Lastly, we discuss our model's limitations compared to other state-of-the-art methods in generalizing to larger protein scaffolds. Our results highlight the need for large-scale stability measurements on a diverse dataset that mirrors the distribution of sequence lengths commonly observed in nature.

Author summary

Research in Professor Justin Siegel's lab focuses on discovering and engineering enzyme catalysis. His work follows a design-build-test cycle, integrating computational protein modeling with wet-lab experiments. Key areas of his research include *de novo* enzyme design, enzyme therapeutics for celiac disease, and applications in food and renewable energy. Additionally, his lab has developed the Design2Data program, a multi-year, multi-campus effort to curate a high-quality dataset of enzymatic activity and stability for beta-glucosidase.

Under the supervision of Professor Justin Siegel, I am engaged in molecular modeling and machine learning in protein engineering. I have a background in molecular dynamics

simulations for protein and cell membrane permeability estimation and in using the Rosetta molecular modeling suite for protein structure modeling and enzyme-substrate interaction. Current research topics include the prediction of mutational effects on protein functions and protein language models for functional prediction and protein design.

Introduction

Protein stability is one of the foundations of protein engineering to design resilient proteins for industrial processes and therapeutic manufacturing [1–3]. Beyond protein engineering, destabilizing mutations are associated with pathogenicity, and stability predictors can help identify pathogenic mutations across human proteome [4–7]. Molecular modeling methods, including Rosetta [8, 9], FoldX [10], and molecular dynamics simulations [11], have been shown to predict the impact of mutation on protein stability. More recently, the use of machine learning models grounded in biophysical features and evolutionary statistics [12–20] has offered an alternative approach to stability and function prediction without the need for computationally intensive molecular modeling simulations. Fueled by the latest advances in deep learning, convolutional neural networks (CNNs) [21] and graph neural networks (GNNs) [22] are now being adopted to predict mutational impacts on stability by operating directly on the input protein structure [23, 24]. For example, RaSP is a CNN-based model trained on top of Rosetta [25], while ELASPIC-2, another stability predictor, operates on both sequence embedding from ESM and structural embedding from GNN [26–28].

Despite these advancements, the lack of a consistent and universal dataset remains an obstacle. While merging smaller datasets into a more comprehensive collection, such as ProTherm [29], ProtaBank [30] and ThermoMutDB [31], is a feasible approach, combined datasets often consist of closely related but distinct quantities accompanied by additional discrepancies in experimental conditions. While deep mutagenesis scanning (DMS) offers profound insights, these studies typically focus on a single protein target, limiting the broader applicability of the derived data and models subsequently trained on these datasets. In light of these challenges, Tsuboyama et al. introduced a mega-scale thermostability dataset, encompassing 776k short protein sequences derived from 479 small protein domains, all consistently evaluated using the same assay [32].

Utilizing this dataset, we fine-tuned a protein language model (pLM), named $\text{ESM}_{\text{therm}}$, from ESM-2 [33] to act as an end-to-end stability predictor. We observe that $\text{ESM}_{\text{therm}}$ performs comparably with state-of-the-art models and generalizes to small protein sequences distal to those of the training set. We also demonstrate that training on an ensemble of protein domains, instead of mutagenesis studies of a single domain, improves the performance of the fine-tuned protein language model for folding stability prediction. Lastly, we discuss the limitations of $\text{ESM}_{\text{therm}}$ and compare it to other state-of-the-art methods in the ability to generalize to longer protein sequences.

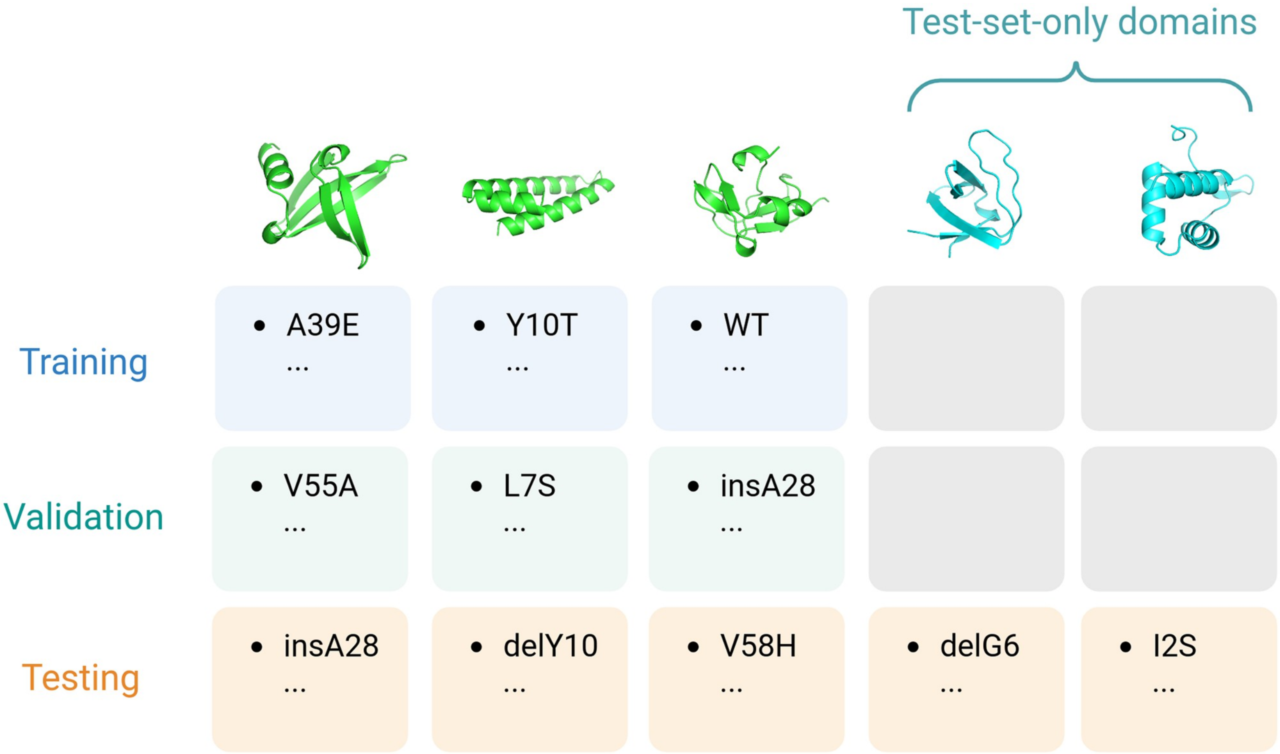
Results

Evaluating model generalizability on test-set-only domains

Protein stability prediction can be assessed on different scales of generalizability. Although machine learning algorithms are often trained and tested on different sets of non-overlapping samples, the definition of overlap is ambiguous in protein sequences. For example, assigning two point mutants from the same WW domain, one to the training set and another to the test set, can assess the generalizability of the model to sequences sharing the same protein domain. However, it fails to evaluate the generalizability of the model to a domain different from those in the training set, such as an SH2 domain. To benchmark our model on both scales, our test set sequences consist of two parts. The first part is formed by protein domains also found in training set, whereas the second part consists of protein domains exclusively found in test set only, denoted as test-set-only domains. We assess the model performance by Spearman's R, and its capability to generalize to these test-set-only sequences by the highest sequence identity to any domains in the training set. Given that domains are classified according to the wildtype definitions by Tsuboyama et al. [32], it is possible for domains exclusive to the test set to still share considerable sequence identity with those in the training set. This setup allows for an assessment of generalizability across varying degrees of sequence identity. The dataset-splitting scheme is illustrated in Fig 1 and further detailed in Methods and Materials.

$\text{ESM}_{\text{therm}}$ generalizes reasonably well to 47 test-set-only protein domains, illustrated in Fig 2. The Spearman's R evaluated on individual domains ranges from 0.2 to 0.9, except for the uncharacterized bacterial protein yahO (PDB code: 2MA4) [34]. Among all test-set-only domains, SH3-subunit of chicken alpha spectrin (PDB code: 6SCW) [35] has the highest sequence identity of 95.8% and scores a corresponding Spearman's R of 0.88. Going down the ladder to test-set-only domains in lower sequence identity, our model scores worse in Homo sapein J-domain protein HSJ1a (PDB code: 2LGW) [36] at 59% identity but still retains a Spearman's R of 0.52.

In the 13 cases where no alignment with the training set sequences passes e-value $< 10^{-3}$, $\text{ESM}_{\text{therm}}$ is capable of generalizing to both natural and *de novo* proteins. No training sequence



Training	• A39E ...	• Y10T ...	• WT ...		
Validation	• V55A ...	• L7S ...	• insA28 ...		
Testing	• insA28 ...	• delY10 ...	• V58H ...	• delG6 ...	• I2S ...

Fig 1. Dataset splitting scheme. Protein domains are first identified by their wildtype sequences and split into train-validation-test (green) and test-set-only partitions (cyan). Mutants are then randomly assigned to either training, validation and test sets or test set only according to their respective wildtype.

https://doi.org/10.1371/journal.pcbi.1012248.g001

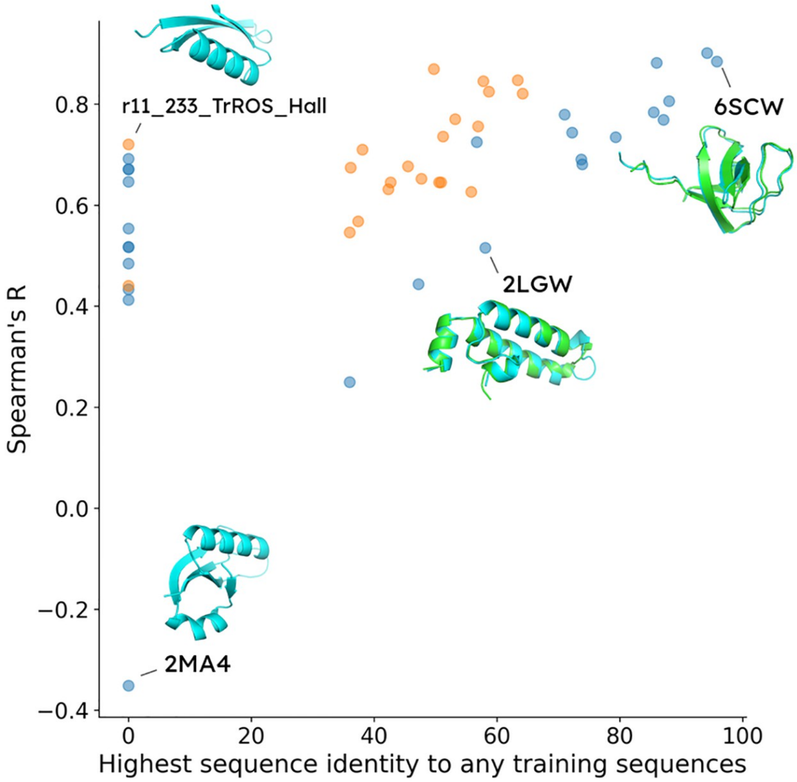


Fig 2. Spearman's R on test-set-only protein domains. Natural protein domains are labeled in blue and *de novo* domains are in orange. The x-axis is the highest sequence identity from the evaluated protein domain to those in the training set. In the case where no sequence alignment was found, 0% is assigned. The y-axis is the Spearman's R evaluated on all sequences from the corresponding domain. We highlighted some of the test-set-only AlphaFold2 models in cyan, and when possible, overlay them with the training-set protein domains of the highest sequence identity in green.

https://doi.org/10.1371/journal.pcbi.1012248.g002

can be aligned to *Escherichia coli* DNA-binding arginine repressor (PDB code: 1AOY) [37], and yet its Spearman's R evaluated is 0.69. For *de novo* designs, we highlight two protein domains from Baker Lab. αββα domain (HEEH_KT_rd6_0790) is a mini-protein from high-throughput computational design with Rosetta [38], whereas the trRosetta-hallucinated structure (r11_233_TrROS_Hall) was sampled with iterative sequence refinement to improve the confidence in the prediction of residue-residue distance map [39]. Spearman's R on these domains is 0.44 and 0.72, respectively.

Improving stability prediction by learning all domains collectively

Prior to the work by Tsuboyama et al. [32], DMS was often restricted to a single protein of interest. In the case where the target of interest is not thoroughly mapped, site-saturated mutagenesis studies from a homologous sequence(s) might provide insights into selecting the best mutation for the specific function of interest. However, direct cross-comparison between proteins is often complicated by the difference in measured quantities and experimental conditions between functional assays. This inconsistency makes it difficult to highlight the benefits of learning from multiple target proteins collectively in a systematic manner.

The mega-scale dataset addresses this difficulty by measuring folding stability across multiple protein domains in a uniform experimental condition, and it helps us compare two paradigms, i.e. transfer learning from homologous sequences and learning from all domains collectively. To contrast these approaches, we assess the generalizability of the model fine-tuned on these paradigms on test-set-only protein domains.

Extrapolating to test-set-only domains clearly benefits from learning all domains collectively. Collective training improves Spearman's R by 0.16 on average (p-value = 6x10^{-3}), as illustrated in Fig (3). CdnL protein (PDB code: 2LQK) [40] cannot be aligned with any training sequence and instead was matched with its closest structural alignment (PDB code: 2BTT) [41] with Foldseek [42]. Collective training increased CdnL's Spearman R from -0.25 to 0.65. Similarly, amino-terminal domain of phase 434 repressor (PDB code: 1R69) [43] was matched by structural alignment to a redesigned protein G (PDB code: 1EM7) [44] with a TM-score of 0.23, and gained 0.74 in Spearman's R from -0.22 to 0.52. Looking into the domains with sequence alignment to the training set, WW domain from APBB3 (PDB code: 2YSC) shares 47% identity with its training-set partner (PDB code: 1WR7) and yet still benefits from multi-domain training with an improvement of 0.32. In contrast, uncharacterized yahO protein remains a difficult target. Compared to training on its closest training-set domain (PDB code: 1IGV), learning on multiple domains only improves the correlation from -0.35 to -0.14. Overall, these results highlight the benefits of a protein stability dataset on a diverse collection of protein domains for generalization to previously understudied targets.

Although the improvement brought by collective training highlights the benefits of a consistent large-scale dataset on folding stability, it is still unclear whether the improvement originates from the shared knowledge on folding stability across multiple domains or the sheer number of samples. The discrepancy in dataset size is significant as an individual domain only constitutes up to 7k sequences, less than 2% of the training set on the collection of protein domains.

In addition to extrapolating to test-set-only protein domains, we conducted a similar comparison on the impact of training on a collection of protein domains on interpolation on previously observed protein domains. Overall, performance on sequences from training-set domains is marginally uplifted by learning from a multi-domain dataset. Illustrated in S2 Fig, learning from an ensemble of protein domains weakly outperforms models trained on the same domain by an average of 0.03 (p-value = 2x10^{-2}). However, the margin is slim. 72% of the domains have Spearman's R only change by 0.1.

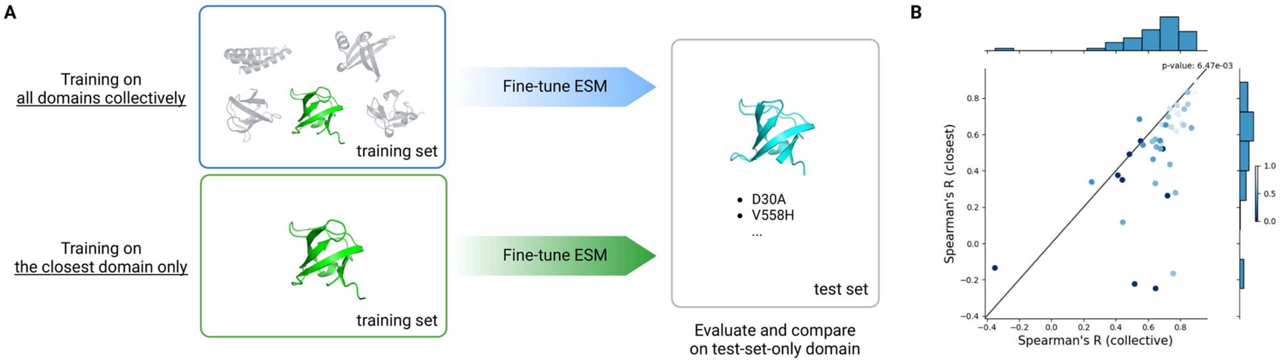


Fig 3. Comparison between transfer learning from the closest protein domain in training set and training on all domains collectively. (A) Schematic of the comparison. In the case of transfer learning, we match the test-set-only protein domain in cyan with the closest domain found in the training set in green. (B) Spearman's R in the test set-only protein domains (x-axis) by learning from all domains collectively and (y-axis) by learning from the closest training-set domain alone. Samples(s) located under the diagonal line indicate better performance by learning collectively. The closest training-set domains were identified primarily by sequence alignment using MMseqs2, then by structure alignment using Foldseek, or discarded when no match was found in either case. The color bar indicates the highest sequence identity to any training-set domains and 0% was assigned when no sequence alignment was found. Statistical significance is performed with Wilcoxon's rank sum test (p-value = 6 x 10^{-3}).

https://doi.org/10.1371/journal.pcbi.1012248.g003

Comparison with existing models on larger proteins

Although natural proteins often span between 200 and 400 residues [45], $\text{ESM}_{\text{therm}}$ is fine-tuned on sequences no longer than 72 residues in length. To explore its performance under this limitation, we benchmarked our model on seven stability-related datasets on larger proteins and compared our results with state-of-the-art covering different methodologies in Tables 1 and 2. These include Rosetta Cartesian ΔΔG for molecular modeling, MUPro for support vector machine (SVM) on traditional sequence features, RaSP for structure-based CNN, ELASPIC-2 which employs a machine-learning model based on both structure and sequence embedding, and unsupervised prediction from ESM-2 [46].

We observe comparable performance in predicting the thermostability of test-set-only protein domains across all models except MUPro. Our pLM achieves a Spearman's R of 0.65, compared to 0.64 from RaSP and ELASPIC-2, and 0.61 from Rosetta molecular modeling. MUPro finishes last by scoring 0.31. Drawing an interesting parallel between datasets, Huang et al. reported direct melting temperature measurements of beta-glucosidase active-site mutants (PDB code: 2JIE) manually selected based on biophysical knowledge [47, 48], while Romero et al. leveraged a log-enrichment value to gauge the stability for a similar beta-glucosidase (PDB code: 1GNX) in a site-saturated fashion [49]. The former closely resembles a smaller-scale study guided by domain knowledge in contrast to the latter dataset that leverages a parallelized assay. Despite an identical alpha-beta barrel scaffold and catalytic mechanism, and a shared sequence identity 48%, most models achieve Spearman's R above 0.4 on the Bgl3 dataset, and no method correlates with BglB dataset. This highlights the potential impact of sampling and assay through a comparative setting.

Trained specifically on small protein domains, $\text{ESM}_{\text{therm}}$ does not generalize to other datasets on larger protein sequences. In a collection of direct [50] and indirect [51–53] stability

Table 1. Comparison of Spearman's R across methods on individual DMS datasets. All evaluation is restricted to point mutations, except our pLM on the mega-scale dataset. We also report unsupervised prediction from pretrained ESM-2 to contrast with supervised approaches. While the mega-scale dataset from Tsuboyama et al. covers multiple protein domains [32], all other datasets studied only one target protein.

Dataset	Length	Supervised Prediction					Unsupervised Prediction	
		Rosetta	MUPro	RaSP	ELASPIC-2	$\text{ESM}_{\text{therm}}$	ESM-2 (35M)	ESM-2 (3B)
Mega-scale dataset	40–72	0.61	0.31	**0.64**	**0.64**	**0.65**	0.36	0.43
BglB dataset	445	-0.01	-0.02	0.02	-0.11	-0.11	-0.12	-0.12
Bgl3 dataset	510	0.49	0.10	0.44	**0.56**	0.06	0.43	0.53
Acetyltransferase dataset	177	0.33	0.16	0.30	**0.49**	0.03	0.25	**0.50**
Lipase EstA dataset	212	**0.48**	0.06	0.41	**0.47**	0.04	0.26	0.28
PTEN dataset	403	0.44	0.17	0.41	**0.47**	0.04	0.26	0.25
Methyltransferase dataset	245	0.48	0.21	0.40	**0.58**	0.03	0.42	0.46

https://doi.org/10.1371/journal.pcbi.1012248.t001

Table 2. Overview of benchmarked DMS datasets on protein stability.

Dataset name	Protein name	Protein length	Measured quantity	No. of sequences
Mega-scale dataset [32]	multiple	40–72	cDNA display proteolysis	100,794 (test set)
BglB dataset [47]	Beta-glucosidase (BglB)	445	melting temperature (T_m)	157
Bgl3 dataset [49]	Beta-glucosidase (Bgl3)	501	catalytic susceptibility to heat shock	2,999
Acetyltransferase dataset [51]	Gentamicin 3-N-acetyltransferase	177	chemical stability	1,801
Lipase EstA dataset [50]	Lipase EstA	212	melting temperature (T_{50})	2,172
PTEN dataset [52]	PTEN	403	protein abundance	5,083
Methyltransferase dataset [53]	Thiopurine S-methyltransferase	245	protein abundance	3,648

https://doi.org/10.1371/journal.pcbi.1012248.t002

measurements, state-of-the-art methods outperform our pLM convincingly. Cartesian ddG in Rosetta achieves generalizability through molecular modeling with a correlation between 0.33 and 0.48. Simultaneously, RaSP is built on top of Cartessian ddG and dramatically speeds up the protocol with marginal correlation setbacks. Overall, ELASPIC-2 ranks highest with a Spearman's R of 0.42–0.58 while our pLM correlates to none of these datasets.

Another intriguing observation is the performance of unsupervised predictions from pLM. While ESM-2 is less capable of predicting stability changes within the mega-scale dataset, it excels in datasets where indirect stability measurements correlate with function. These include log2-enrichment value which characterizes how catalytic activity reacts to heat shock in Bgl3 dataset and the intracellular abundance of the protein in the acetyltransferase dataset. Conversely, ESM-2 has a comparably weaker performance for proteolysis folding stability in the mega-scale dataset and chemical stability in the Lipase EstA dataset, where the assays measure stability directly. We also highlight the impact of fine-tuning by benchmarking the unsupervised prediction from 35M-parameter ESM-2 against our fine-tuned $\mathrm{ESM}_{\mathrm{therm}}$ of the same model size. Supervised prediction improves the correlation from 0.36 to 0.65.

Discussion

Although our model generalizes reasonably well to new small protein domains in the mega-scale thermostability dataset, it is substantially weaker on larger proteins. Studies have established a strong correlation between the parallelized assay and direct measurement of thermostability [54]. However, we cannot rule out that our language model is biased towards dataset-specific details, including experiment conditions and sampling distribution of protein sequences. One hypothesis is that our pLM is biased toward shorter sequences, while geometric learning do not suffer from the same pitfall and already performs better in unsupervised prediction [55]. The protein domains on which we trained are limited to 40 to 72 amino acids in length, a stark contrast to the 177- to 501-residue-long sequences in our additional DMS benchmark. This might suggest that fine-tuned pLM stability predictors would benefit from a large-scale folding stability dataset on longer sequences.

While most methods can rank ΔΔG between mutants successfully, predicting ΔG is still challenging. Our predictions often suffer from an offset and/or scale differently when compared to the experimental ΔG of the test-set-only domains (S3 Fig) and other methods might share the same problem. For example, Rosetta Cartesian ΔΔG follows a different energy unit (Rosetta Energy Unit), and it might not be suitable to be compared directly to kcal mol^{-1}. However, the misalignment can be easily resolved by a simple linear regression between model prediction and experiment. Upon recalibration per protein domain, the root mean square error from our model improved from 1.34 to 0.83 and R^2 from -0.85 to 0.45, averaged across all test-set-only domains. For instance, our model scores a negative R^2 on DNA-binding arginine repressor before recalibration and improves to 0.47 after rescaling, while Spearman's R remains the same at 0.69 regardless of any monotonic transformation (Fig 4).

Conclusion

In this work, we demonstrate that folding stability prediction is possible using a protein language model. Enabled by large-scale protein stability measurements, we fine-tuned ESM-2 on the absolute folding energy of small protein domains. This approach generalizes successfully to protein domains distal from the training set, showing the potential of transfer learning to reduce experimental burden. Furthermore, our result highlights the benefits of training collectively on all protein sequences instead of mutagenesis study on a single wildtype. Although its

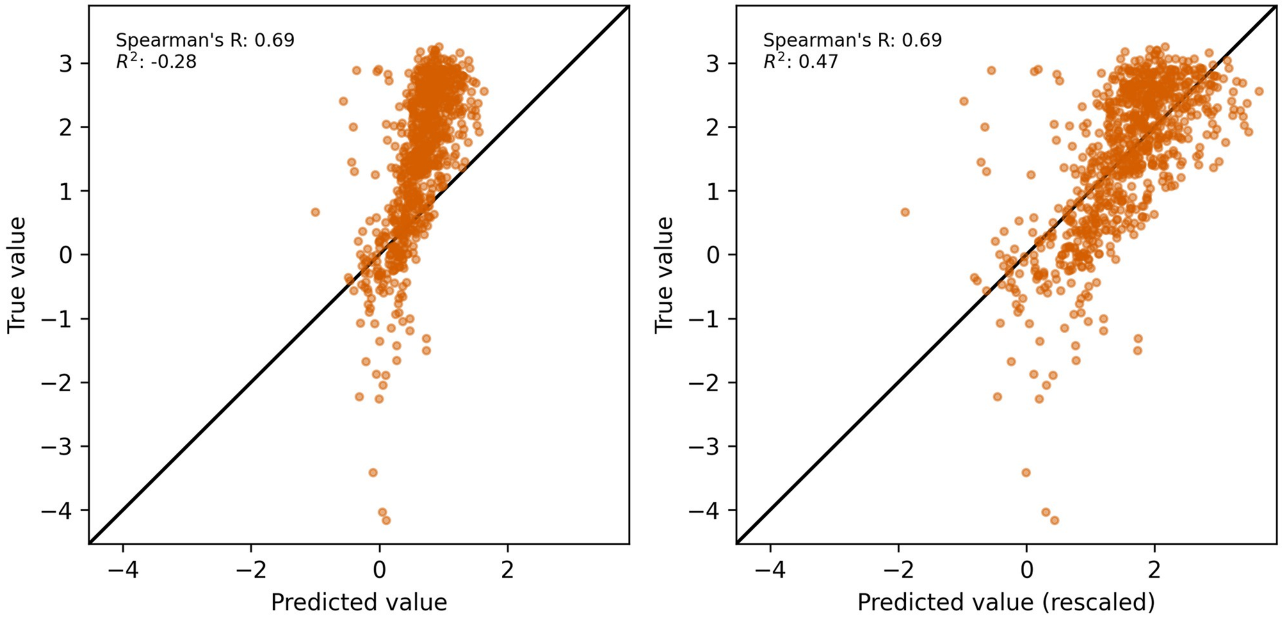


Fig 4. Impact of recalibration. (Left) Miscalibration between prediction and true value on the stability of DNA-binding arginine repressor. (Right) Recovered agreement between prediction and stability measurement through linear rescaling on the same set of data.

https://doi.org/10.1371/journal.pcbi.1012248.g004

performance on larger protein scaffolds is lagging behind state-of-the-art, a folding stability dataset of larger proteins might be vital to improving the generalizability of $\text{ESM}_{\text{therm}}$.

Methods and materials

Protein language model and fine-tuning protocol

ESM-2 is a transformer pLM pre-trained on masked-language-model (MLM) objective on UniRef50. We fine-tuned the model on whole-sequence regression task with a classification head on the starting token. All parameters were trainable in fine-tuning, and used a local batch size of 128 and a global batch size of 2048. We trained the model on A100 GPU at half precision with a patience of 500 steps. We report all test-set-only evaluations on the checkpoint with the best performance on the validation set.

We performed hyperparameter selection on model size (8, 35, 150, and 650 million parameters) (S1 Table), and selected the 35-million-parameter model to balance prediction performance and compute speed. In addition, we performed an ablation study on pretraining. Model with pretraining has a superior advantage over that with random initialization (S1 Fig).

Dataset construction

Tsuboyama et al. measured the folding stability of 1.8M measurements derived from 542 protein domains by cDNA display proteolysis [32]. We aggregated measurement(s) with the identical protein sequence, regardless of their DNA sequence(s), into a single entry. In cases where the DNA sequence was unique while sharing the same protein sequence, we evaluated the standard deviation of ΔG and log K_{50}. We removed measurements when the standard deviation of ΔG was greater than 2 kcal mol^{-1} or that of log K_{50} was greater than 0.5 and we kept only domains with at least 100 measurements by protein sequence. This reduced the number of entries from 851,552 protein sequences from their original criteria (K50_dG_Dataset1_Dataset2.csv) to 527,785 protein sequences and 258 natural and 203 *de novo* protein domains.

Under the hierarchical nature of this dataset, by which multiple domains are constituted and each domain holds a collection of multiple mutants, the definition of model generalizability has two layers. The first is the ability of the model to generalize to mutants on training protein domains, and the second is that on test-set-only domains. To evaluate the model on both training and test-set-only domains, we split our dataset into train, validation, and test sets by domains as illustrated in Fig 1. 10% of all domains, defined by wildtype by the authors, are randomly drawn and all of their mutants are assigned to the test set. Mutants are randomly assigned to train-validation-test sets in an 80–10-10 ratio for the remaining domains.

Sequence and structural alignment

We implemented sequence clustering and alignment through MMseqs2 [56]. For clustering, we clustered the domain wild-type sequences using a similar strategy in constructing the Uniclust database. We dropped prefiltering for all-to-all pairwise alignment. For Foldseek, we searched for the structural identity based on AlphaFold structures from Tsuboyama et al [32]. Unless otherwise specified, we used the default parameters in both MMseqs and Foldseek. The implementation details of alignment can be found src/esmtherm/alignment in the GitHub repository.

Matching test-set-only domains

We fine-tuned ESM-2 (esm2_t12_35M_UR50D) on each of the 416 protein domains in the training set as our independently learned models. We first matched each test-set-only domain to its closest partner in the training set by the highest sequence identity using MMseqs2. In the case where no sequence alignment is identified, we matched test-set-only domain by the highest structural identity by Foldseek. In the case that neither is identified, the test-set-only domain was not compared. Pairwise comparisons of interpolation and extrapolation are performed in Wilcoxon's rank sum test.

Benchmark protein dataset selection

Given the intensive computing resource required to benchmark Rosetta, we limited ourselves to six DMS datasets on direct and indirect stability measurements from ProteinGym [57, 58], and another independent mutational dataset (BglB) from Huang et al. [47] to cover a range of assays. Nutschel et al. reported the thermostability (ΔT_{50}) of *Bacillus subtilis* Lipase A [50], whereas Dandage et al. reports chemical stability on Gentamicin 3-N-acetyltransferase [51]. Contrary to direct stability measurements, PTEN and Methyltransferase datasets correlate with stability through enhancement or depreciation of intracellular abundance as an indirect indicator [52, 53]. The pair of Bglb and Bgl3 datasets was chosen for a comparative study on the impact of sampling and measurement assays. Bgl3 from Romero et al. and BglB datasets [47, 49] share homologous beta-glucosidase sequences but differ in log enrichment value and melting temperature (T_m) as indirect and direct thermostability measurements.

Supporting information

S1 Fig. Ablation of pretraining measured in Spearman's R. Each sample is a collection of mutants from a test-set-only domain. The x-axis is Spearman's R of a test-set-only domain with pretraining. The y-axis is that from randomly initialized model. The color bar on the right represents the closest sequence identity in the train and validation set domains. The

statistical assessment was performed using Wilcoxon's rank sum test.
(TIF)

S2 Fig. Comparison between learning from the same protein domain only and training on all domains collectively. (A) Schematic of the comparison. (B) Spearman's R on test mutants whose protein domains are also present in the training set. The x-axis represents learning from all domains collectively and the y-axis is learning from the same protein domain alone. Domain(s) located under the diagonal line indicate better performance when learning collectively. Statistical significance is performed with Wilcoxon's rank sum test.
(TIF)

S3 Fig. Offset in ΔG prediction on wildtype sequences. The x-axis is the ΔG prediction from $\text{ESM}_{\text{therm}}$ and the y-axis is the experimental ΔG label. The Spearman's R across all wildtypes in test-set-only protein domains is 0.39.
(TIF)

S1 Table. Performance evaluation on different model sizes on test set. Metrics are evaluated on each individual domain, and then aggregated into mean and standard deviation over all domains. All models have similar performance metrics with esm2_t12_35M_UR50D except esm2_t6_8M_UR50D on Spearman's R (p-value $< 5\text{x}10^{-2}$).
(PDF)

S1 Spreadsheet. Performance evaluation per protein domain.
(CSV)

S2 Spreadsheet. Model prediction per protein sequence.
(CSV)

Author Contributions

Conceptualization: Simon K. S. Chu, Justin B. Siegel.

Data curation: Simon K. S. Chu.

Formal analysis: Simon K. S. Chu.

Investigation: Simon K. S. Chu.

Methodology: Simon K. S. Chu.

Project administration: Justin B. Siegel.

Software: Simon K. S. Chu, Kush Narang.

Supervision: Justin B. Siegel.

Validation: Simon K. S. Chu, Kush Narang.

Visualization: Simon K. S. Chu.

Writing – original draft: Simon K. S. Chu, Justin B. Siegel.

Writing – review & editing: Simon K. S. Chu, Kush Narang, Justin B. Siegel.

References

1. Lv Y., Zheng S., Goldenzweig A., Liu F., Gao Y., Yang X., Kandale A., McGeary R.P., Williams S.J., Kobe B., Schembri M.A., Landsberg M.J., Wu B., Brück T.B., Sieber V., Bodén M., Rao Z., Fleishman S.J., Schenk G., Guddat L.W. Enhancing the Thermal and Kinetic Stability of Ketol-Acid Reductoisomerase, a

Central Catalyst of a Cell-Free Enzyme Cascade for the Manufacture of Platform Chemicals. Applied Biosciences.

2. Rennison A., Winther J.R., Varrone C. Rational Protein Engineering to Increase the Activity and Stability of IsPETase Using the PROSS Algorithm. Polymers, 13. https://doi.org/10.3390/polym13223884 PMID: 34833182
3. Hutchinson M., Ruffolo J.A., Haskins N., Iannotti M., Vozza G., Pham T., Mehzabeen N., Shandilya H., Rickert K., Croasdale-Wood R., Damschroder M., Fu Y., Dippel A., Gray J.J., Kaplan G. (2023). Enhancement of antibody thermostability and affinity by computational design in the absence of antigen. bioRxiv.
4. Gerasimavicius L., Liu X., Marsh J.A. Identification of pathogenic missense mutations using protein stability predictors. Scientific Reports. 2020; 10. https://doi.org/10.1038/s41598-020-72404-w PMID: 32958805
5. Cheng J., Novati G., Pan J., Bycroft C., Žemgulytė A., Applebaum T., et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023; 381. https://doi.org/10.1126/science.adg7492 PMID: 37733863
6. Stein A., Fowler D.M., Hartmann-Petersen R., Lindorff-Larsen K. Biophysical and Mechanistic Models for Disease-Causing Protein Variants. Trends in biochemical sciences, 2019; 44 7, 575–588. https://doi.org/10.1016/j.tibs.2019.01.003 PMID: 30712981
7. Yue P., Li Z., Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. Journal of molecular biology, 2005; 353 2, 459–73. https://doi.org/10.1016/j.jmb.2005.08.020 PMID: 16169011
8. Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins: Structure, Function and Bioinformatics. 2011; 79 (3):830–838. https://doi.org/10.1002/prot.22921 PMID: 21287615
9. Park H, Bradley P, Greisen P, Liu Y, Mulligan VK, Kim DE, et al. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. Journal of Chemical Theory and Computation. 2016; 12(12):6201–6212. https://doi.org/10.1021/acs.jctc.6b00819 PMID: 27766851
10. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: An online force field. Nucleic Acids Research. 2005; 33(SUPPL. 2):382–388. https://doi.org/10.1093/nar/gki387 PMID: 15980494
11. Wilson CJ, Chang M, Karttunen M, Choy WY. Keap1 cancer mutants: A large-scale molecular dynamics study of protein stability. International Journal of Molecular Sciences. 2021; 22(10). https://doi.org/10.3390/ijms22105408 PMID: 34065616
12. Dehouck Y, Kwasigroch JM, Gilis D, Rooman M. PoPMuSiC 2.1: A web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinformatics. 2011; 12. https://doi.org/10.1186/1471-2105-12-151 PMID: 21569468
13. Cao H, Wang J, He L, Qi Y, Zhang JZ. DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks. Journal of Chemical Information and Modeling. 2019; 59(4):1508–1514. https://doi.org/10.1021/acs.jcim.8b00697 PMID: 30759982
14. Witvliet DK, Strokach A, Giraldo-Forero AF, Teyra J, Colak R, Kim PM. ELASPIC web-server: Proteome-wide structure-based prediction of mutation effects on protein stability and binding affinity. Bioinformatics. 2016; 32(10):1589–1591. https://doi.org/10.1093/bioinformatics/btw031 PMID: 26801957
15. Worth CL, Preissner R, Blundell TL. SDM—A server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Research. 2011; 39(SUPPL. 2). https://doi.org/10.1093/nar/gkr363 PMID: 21593128
16. Masso M, Vaisman II. AUTO-MUTE 2.0: A portable framework with enhanced capabilities for predicting protein functional consequences upon mutation. Advances in Bioinformatics. 2014; 2014. https://doi.org/10.1155/2014/278385 PMID: 25197272
17. Strokach A., Corbi-Verge C., Teyra J., Kim P.M. Predicting the Effect of Mutations on Protein Folding and Protein-Protein Interactions. Methods in molecular biology, 2018; 1851, 1–17. https://doi.org/10.1007/978-1-4939-8736-8_1
18. Strokach A., Corbi-Verge C., Kim P.M. Predicting changes in protein stability caused by mutation using sequence-and structure-based methods in a CAGI5 blind challenge. Human Mutation, 40, 1414–1423. https://doi.org/10.1002/humu.23852 PMID: 31243847
19. Cheng J., Randall A., Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. Proteins: Structure, Function, and Bioinformatics, 2006; 62(4), 1125–1132. https://doi.org/10.1002/prot.20810 PMID: 16372356

20. Huang L., Gromiha M.M., Ho S. iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. Bioinformatics, 23 10, 1292–3. https://doi.org/10.1093/bioinformatics/btm100 PMID: 17379687

21. Lecun Y, Bottou E, Bengio Y, Haffner P. Gradient-Based Learning Applied to Document Recognition; 1998.

22. Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. arxiv. 2016;.

23. Wang S, Tang H, Shan P, Wu Z, Zuo L. ProS-GNN: Predicting effects of mutations on protein stability using graph neural networks. Computational Biology and Chemistry. 2023; 107. https://doi.org/10.1016/j.compbiolchem.2023.107952 PMID: 37643501

24. Chu SKS, Siegel J. Predicting single-point mutational effect on protein stability; 2021.

25. Blaabjerg LM, Kassem MM, Good LL, Jonsson N, Cagiada M, Johansson KE, et al. Rapid protein stability prediction using deep learning representations. eLife. 2023; 12. https://doi.org/10.7554/eLife.82593 PMID: 37184062

26. Strokach A, Lu TY, Kim PM. ELASPIC2 (EL2): Combining Contextualized Language Models and Graph Neural Networks to Predict Effects of Mutations. Journal of Molecular Biology. 2021; 433(11). https://doi.org/10.1016/j.jmb.2021.166810 PMID: 33450251

27. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. bioRxiv. 2019; 118(15):e2016239118.

28. Strokach A, Becerra D, Corbi-Verge C, Perez-Riba A, Kim PM. Fast and Flexible Protein Design Using Deep Graph Neural Networks. Cell Systems. 2020; 11(4):402–411. https://doi.org/10.1016/j.cels.2020.08.016 PMID: 32971019

29. Gromiha MM, An J, Kono H, Oobatake M, Uedaira H, Prabakaran P, et al. ProTherm, version 2.0: thermodynamic database for proteins and mutants; 2000. 1. Available from: http://www.rtc.riken.go.jp/protherm.html.

30. Wang CY, Chang PM, Ary ML, Allen BD, Chica RA, Mayo SL, et al. ProtaBank: A repository for protein design and engineering data. Protein Science. 2018; 27(6):1113–1124. https://doi.org/10.1002/pro.3406 PMID: 29575358

31. Xavier J.S., Nguyen T., Karmarkar M., Portelli S., Rezende P.M., Velloso J.P., et al. ThermoMutDB: a thermodynamic database for missense mutations. Nucleic Acids Research. 2020; 49, D475–D479. https://doi.org/10.1093/nar/gkaa925

32. Tsuboyama K, Dauparas J, Chen J, Laine E, Mohseni Behbahani Y, Weinstein JJ, et al. Mega-scale experimental analysis of protein folding stability in biology and design. Nature. 2023; 620(7973):434–444. https://doi.org/10.1038/s41586-023-06328-6 PMID: 37468638

33. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023; 379:1123–1130. https://doi.org/10.1126/science.ade2574 PMID: 36927031

34. Eletsky A, Michalska K, Houliston S, Zhang Q, Daily MD, Xu X, et al. Structural and Functional Characterization of DUF1471 Domains of Salmonella Proteins SrfN, YdgH/SssB, and YahO. PLoS ONE. 2014; 9:e101787. https://doi.org/10.1371/journal.pone.0101787 PMID: 25010333

35. Grohe K, Patel S, Hebrank C, Medina S, Klein A, Rovó P, et al. Protein Motional Details Revealed by Complementary Structural Biology Techniques. Structure. 2020; 28(9):1024–1034. https://doi.org/10.1016/j.str.2020.06.001 PMID: 32579946

36. Gao XC, Zhou CJ, Zhou ZR, Wu M, Cao CY, Hu HY. The C-terminal helices of heat shock protein 70 are essential for J-domain binding and ATPase activation. Journal of Biological Chemistry. 2012; 287 (8):6044–6052. https://doi.org/10.1074/jbc.M111.294728 PMID: 22219199

37. Sunnerhagen M, Nilges M, Otting G, Carey J. Solution structure of the DNA-binding domain and model for the complex of multifunctional hexameric arginine repressor with DNA; 1997. Available from: http://www.nature.com/nsmb.

38. Chevalier A., Silva DA., Rocklin G. et al. Massively parallel de novo protein design for targeted therapeutics. Nature, 2017; 550, 74–79. https://doi.org/10.1038/nature23912 PMID: 28953867

39. Anishchenko I, Chidyausiku TM, Ovchinnikov S, Pellock SJ, Baker D. De novo protein design by deep network hallucination. Nature. 2020; p. 547–552.

40. Gallego-García A, Mirassou Y, Elías-Arnanz M, Padmanabhan S, Jiménez MA. NMR structure note: N-terminal domain of Thermus thermophilus CdnL. Journal of Biomolecular NMR. 2012; 53(4):355–363. https://doi.org/10.1007/s10858-012-9648-z PMID: 22782235

41. Musi V, Birdsall B, Fernandez-Ballester G, Guerrini R, Salvatori S, Serrano L, et al. New approaches to high-throughput structure characterization of SH3 complexes: The example of Myosin-3 and Myosin-5 SH3 domains from S. cerevisiae. Protein Science. 2006; 15:795–807. https://doi.org/10.1110/ps.051785506 PMID: 16600966

42. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. Nature Biotechnology. 2023. https://doi.org/10.1038/s41587-023-01773-0 PMID: 37156916

43. Mondragbn A, Subbiah' S, Almolt SC, Drottar' M, Harrison SC. Structure of the Amino-terminal Domain of Phage 434 Repressor at 2.0 A Resolution. J Mol Hiol (1989). 1989; 205:189–200. https://doi.org/10.1016/0022-2836(89)90375-6

44. Strop P., Marinescu A., Mayo S.L. Structure of a protein G helix variant suggests the importance of helix propensity and helix dipole interactions in protein design. Protein Science, 2000; 9. https://doi.org/10.1110/ps.9.7.1391 PMID: 10933505

45. Nevers Y, Glover NM, Dessimoz C, Lecompte O. Protein length distribution is remarkably uniform across the tree of life. Genome Biology. 2023; 24(1). https://doi.org/10.1186/s13059-023-02973-2 PMID: 37291671

46. Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T. and Rives, A., 2021. Language models enable zero-shot prediction of the effects of mutations on protein function. Advances in Neural Information Processing Systems, 34 (2021): 29287-29303.

47. Huang P, Chu SKS, Frizzo HN, Connolly MP, Caster RW, Siegel JB. Evaluating Protein Engineering Thermostability Prediction Tools Using an Independently Generated Dataset. ACS Omega. 2020; 5 (12):6487–6493. https://doi.org/10.1021/acsomega.9b04105 PMID: 32258884

48. Isorna P, Polaina J, Latorre-García L, Cañada FJ, González B, Sanz-Aparicio J. Crystal Structures of Paenibacillus polymyxa *β*-Glucosidase B Complexes Reveal the Molecular Basis of Substrate Specificity and Give New Insights into the Catalytic Machinery of Family I Glycosidases. Journal of Molecular Biology. 2007; 371(5):1204–1218. https://doi.org/10.1016/j.jmb.2007.05.082 PMID: 17585934

49. Romero PA, Tran TM, Abate AR. Dissecting enzyme function with microfluidic-based deep mutational scanning. Proceedings of the National Academy of Sciences. 2015; 112(23):7159–7164. https://doi.org/10.1073/pnas.1422285112 PMID: 26040002

50. Nutschel C, Fulton A, Zimmermann O, Schwaneberg U, Jaeger KE, Gohlke H. Systematically Scrutinizing the Impact of Substitution Sites on Thermostability and Detergent Tolerance for Bacillus subtilis Lipase A. Journal of Chemical Information and Modeling. 2020; 60:1568–1584. https://doi.org/10.1021/acs.jcim.9b00954 PMID: 31905288

51. Dandage R, Pandey R, Jayaraj G, Rai M, Berger D, Chakraborty K. Differential strengths of molecular determinants guide environment specific mutational fates. PLOS Genetics. 2018; 14:e1007419. https://doi.org/10.1371/journal.pgen.1007419 PMID: 29813059

52. Matreyek KA, Stephany JJ, Ahler E, Fowler DM. Integrating thousands of PTEN variant activity and abundance measurements reveals variant subgroups and new dominant negatives in cancers. Genome Medicine. 2021; 13:165. https://doi.org/10.1186/s13073-021-00984-x PMID: 34649609

53. Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, Gray VE, et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. Nature Genetics. 2018; 50(6):874–882. https://doi.org/10.1038/s41588-018-0122-z PMID: 29785012

54. Rocklin GJ, Chidyausiku TM, Goreshnik I, Ford A, Houliston S, Lemak A, et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. Science. 2017; 357(6347):168–175. https://doi.org/10.1126/science.aan0693 PMID: 28706065

55. Paul, S., Kollasch, A., Notin, P., Marks, D. Combining Structure and Sequence for Superior Fitness Prediction. NeurIPS 2023 Generative AI and Biology (GenBio) Workshop.

56. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology. 2017; 35(11):1026–1028. https://doi.org/10.1038/nbt.3988 PMID: 29035372

57. Notin P, Dias M, Frazer J, Hurtado JM, Gomez AN, Marks D, et al. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In: International Conference on Machine Learning. PMLR; 2022. p. 16990–17017.

58. Notin P., Kollasch A.W., Ritter D., van Niekerk L., Paul S., Spinner H., et al. ProteinGym: Large-Scale Benchmarks for Protein Design and Fitness Prediction. bioRxiv, 2023. https://doi.org/10.1101/2023.12.07.570727 PMID: 38106144