PROTEIN LANGUAGE MODELS ARE BIASED BY UNEQUAL SEQUENCE SAMPLING ACROSS THE TREE OF LIFE

Frances Ding and Jacob Steinhardt
Departments of Statistics and Electrical Engineering and Computer Sciences
University of California, Berkeley
{frances, jsteinhardt}@berkeley.edu

ABSTRACT

Protein language models (pLMs) trained on large protein sequence databases have been used to understand disease and design novel proteins. In design tasks, the likelihood of a protein sequence under a pLM is often used as a proxy for protein fitness, so it is critical to understand what signals likelihoods capture. In this work we find that pLM likelihoods unintentionally encode a species bias: likelihoods of protein sequences from certain species are systematically higher, independent of the protein in question. We quantify this bias and show that it arises in large part because of unequal species representation in popular protein sequence databases. We further show that the bias can be detrimental for some protein design applications, such as enhancing thermostability. These results highlight the importance of understanding and curating pLM training data to mitigate biases and improve protein design capabilities in under-explored parts of sequence space.

1 INTRODUCTION

Proteins are the building blocks and workhorses of life, performing essential roles in human and ecosystem health. Inspired by advances in natural language processing, many different protein language models (pLMs) have been trained to model the distribution of naturally occurring protein sequences (Alley et al., 2019; Rives et al., 2021; Elnaggar et al., 2021; Madani et al., 2023; Lin et al., 2023; Alamdari et al., 2023). pLMs have been successfully used to predict protein 3D structure (Lin et al., 2023), catalytic activity (Eom et al., 2024), and other biophysical properties (Brandes et al., 2022; Jagota et al., 2023), generally with additional supervision for fine-tuning. Excitingly, without needing additional supervision, *likelihoods* from pLMs have been shown to correlate well with protein fitness, i.e. desirable qualities such as catalytic activity, stability, and binding affinity (Meier et al., 2021; Notin et al., 2023; Nijkamp et al., 2023).

Because of this correlation with fitness, pLM likelihoods are increasingly used in protein design. They have been used to screen for potentially beneficial mutations (Johnson et al., 2023), to design libraries of protein candidates with higher hit rates than previously state-of-the-art synthetic libraries (Shin et al., 2021), and to efficiently evolve human antibodies without any additional supervision (Hie et al., 2023).

In this work we find that likelihoods from popular pLMs have a species bias: likelihoods of naturally occurring protein sequences are systematically higher in certain species, which can be detrimental for some protein design applications.

In Section 2 we show that across the many different proteins we study, certain species almost always have higher pLM likelihoods for their protein sequences than other species. For example, in the data we collect, fruit fly proteins have higher likelihoods than the *C. elegans* (roundworm) versions of the same proteins 92% of the time, even though there is no biological reason for fruit fly proteins to be uniformly "fitter" or more canonical. Next, in Section 3 we show that the bias can be largely explained by species representation in protein databases, combined with careful accounting of evolutionary relationships. Finally, in Section 4 we show that in protein design applications, the bias causes designs to gravitate towards sequences from favored species. In some cases, this leads the design

process to produce worse outcomes. For example, many thermophilic species are under-represented in databases, and we find that pLM-guided designs of their heat-stable proteins have significant decreases in predicted thermostability.

Looking forward, these results suggest that protein designers should use pLM likelihoods carefully and consider whether the species bias should be corrected for a given application. Looking forwards, we believe the protein design field would benefit from more deliberate curation of training data for pLMs, potentially tailored to different applications.

2 PLM LIKELIHOODS ARE HIGHER FOR SEQUENCES FROM CERTAIN SPECIES

We first empirically investigate what factors affect pLM likelihoods. We collect a dataset of orthologous sequences (different species' versions of the same proteins) (Appendix A) and compute pLM likelihoods for each sequence. Unsurprisingly, some protein types have much higher overall likelihoods than others (due to intrinsic disorder, conservation, etc.), but surprisingly, we find that some *species* also have much higher likelihoods than others (across proteins), and that this generalizes across pLMs with different training objectives and data sampling.

PLMs we study. We focus on two families of pLMs in this work: the Progen2 suite (Nijkamp et al., 2023) (in 5 sizes: xlarge, BFD90, large, base, and medium) and the ESM2 suite (Lin et al., 2023) (in 3 sizes: 15B, 3B, and 650M). These models are among the most popular for downstream use and achieve the best performance among pLMs on many benchmark tasks in ProteinGym (Notin et al., 2023), particularly metrics geared towards protein design (see Appendix B for more details).

Variance explained by species identity. We first investigate what factors explain pLM likelihoods in our dataset. We compute linear regressions of pLM likelihood against species only, protein type only, and both variables and report the R^2 value for each setting in Table 1. We also compute the fraction of variance explained by species, after controlling for protein type. We find that protein type explains some of the variance, as expected, since proteins vary in prevalence, conservation, and other factors that intuitively affect likelihood. Surprisingly, species identity also explains a significant amount of the variance in pLM likelihoods; for example, for Progen2-xlarge likelihoods, species accounts for 50% of the variance by itself, and 67% of the variance after controlling for protein type. This suggests that likelihoods have a consistent species bias across the diverse universe of proteins.

Table 1: **Variance in likelihood explained by species and protein type.** $R^2_{\text{Species|Protein}}$ is the fraction of variance explained by species identity, after controlling for protein type[2].

MODEL	R^2_{SPECIES}	R^2_{PROTEIN}	R^2_{BOTH}	$R^2_{\text{SPECIES|PROTEIN}}$
PROGEN2-XLARGE	0.50	0.42	0.81	0.67
PROGEN2-BFD90	0.49	0.51	0.85	0.69
PROGEN2-LARGE	0.46	0.60	0.87	0.67
PROGEN2-BASE	0.25	0.64	0.84	0.55
PROGEN2-MEDIUM	0.44	0.59	0.86	0.66
ESM2-15B	0.25	0.42	0.60	0.32
ESM2-3B	0.26	0.46	0.63	0.32
ESM2-650M	0.19	0.62	0.72	0.26

Quantifying species bias via Elo. We next quantify the bias associated with each species without assuming a linear model of likelihoods. Since each protein is only found in a subset of species, species cannot be fairly compared by a simple average likelihood score. To solve this, we use the Elo rating system to summarize how often one species has higher likelihoods than another (Appendix C).

Figure 1 plots Elo ratings for each species in our dataset, annotated by phylogenetic taxa. We find that Elo ratings vary widely across species. Using Progen2-xlarge likelihoods, the 25th percentile species (*A. baylyi*) has an Elo rating of 1235 while the 75th percentile species (*S. glossinidius*) has an Elo rating of 1745. This Elo difference of 510 implies that *S. glossinidius* has a higher likelihood for its homologs (versions) of proteins 95% of the time. Using ESM2-15B pseudo-likelihoods, there is a

[2]Example $R^2_{\text{Species|Protein}}$ derivation: 0.67 = (0.81 - 0.42)/(1 - 0.42)

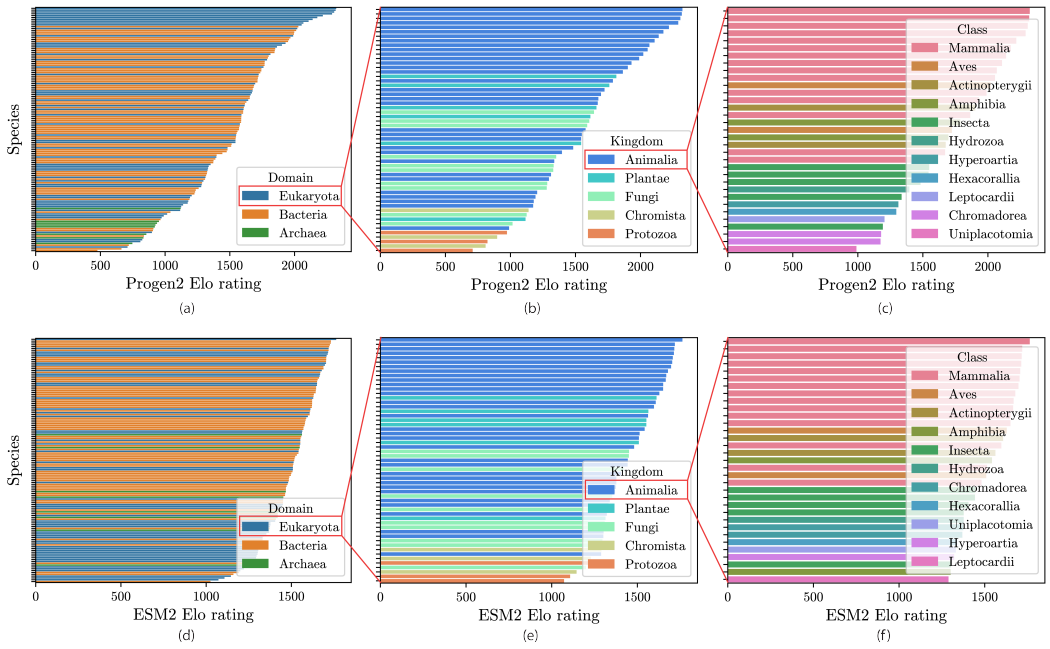

Figure 1: **Elo ratings for different species.** Elo ratings computed from Progen2-xlarge (top) and ESM2-15B (bottom).

220 Elo difference between the 25th and 75th percentiles, which implies an 80% chance of a higher likelihood. Both models thus have a significant species bias, with Progen2's being somewhat larger.

Progen2 and ESM2 also show largely similar biases: Elo ratings from Progen2-xlarge and ESM2-15B have a Pearson correlation of 0.83 (see Appendix D for correlations for all pairs of pLMs). Figure 1 further shows that the species bias has some interpretable trends across both models: within eukaryotes, animals have the highest Elo ratings, and within animals, mammals do. This species bias motivates understanding how it arises, which we study in the next section.

3 PLM BIAS IS LARGELY EXPLAINED BY SPECIES REPRESENTATION IN SEQUENCE DATABASES

We investigate what factors explain the species bias and find that the species representation in popular sequence databases plays a major role, once we appropriately account for evolutionary structure.

We test an initial hypothesis that Elo ratings will correlate with the number of sequences a species has in a database. Figure 2 (a) and (c) plots each species' Elo rating against its sequence counts and we see that although a few species, such as *H. sapiens*, *M. musculus*, and *E. coli* show the expected trend of high sequence counts and high Elo ratings, most other species are not fit well.

One factor not captured by the initial hypothesis is sequence similarity due to evolution. We conjecture that the sequence count from a given species contributes to the "effective" sequence count for another species in proportion to the sequence similarity between the two species' orthologs. Assuming a Poisson model of mutations accumulated over time, sequence similarity between two species' orthologs is directly related to their evolutionary closeness. Thus we posit our second hypothesis that Elo ratings will correlate with an evolution-weighted sequence count, $n^{\text{evo-weighted}}$, which we define as follows:

$$n_i^{\text{evo-weighted}} = \sum_j n_j e^{-\frac{d(i,j)}{\alpha}},$$

where n_j is the raw sequence count for species j, $d(i,j)$ is the time to last common ancestor between species i and j collected from the TimeTree of Life resource (Kumar et al., 2022), and $\alpha \in \mathbb{R}_{\geq 0}$ is a hyperparameter used to scale d appropriately. Under the assumption that mutations occur at a fixed

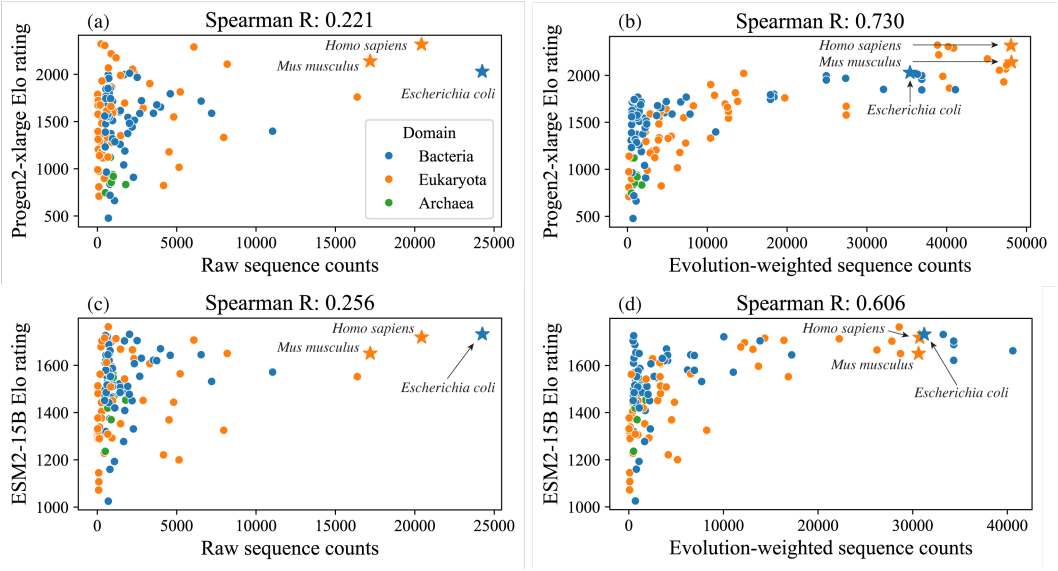

Figure 2: **Species Elo ratings plotted against their Swiss-Prot sequence counts and evolution-weighted sequence counts.** Top: Elo ratings computed from Progen2-xlarge likelihoods. Bottom: Elo ratings computed from ESM2-15B pseudolikelihoods. Correlation using raw sequence counts (left) is low, while correlation using evolution-weighted sequence counts (right) is high.

rate, $e^{-\frac{d(i,j)}{\alpha}}$ gives the expected overlap in sequence between two species' orthologs, to approximate the effective sequence counts they contribute to each other[3].

Figure 2(b, d) plot Elo rating against evolution-weighted sequence counts, and we see that this achieves a much higher Spearman correlation of 0.73 and 0.606 for Progen2-xlarge and ESM2-15B Elo ratings, respectively. Thus species representation, with the addition of evolutionary distance, explains a large fraction of the species bias in both pLMs.

4 PLM SPECIES BIAS AFFECTS PROTEIN DESIGN

Finally, we investigate the implications for protein design. Since sequences from high Elo species have higher likelihoods, they may be basins of attraction when designing proteins to optimize likelihood. We test this prediction by simulating a simple protein design workflow and find that designs indeed systematically drift towards sequences from high Elo species. Of the species with the lowest Elo ratings, many are extremophiles. Despite their low Elo, these species' proteins often have unique and useful properties, suggesting that the typical use of likelihood for design will be detrimental when trying to enhance these proteins' properties. We test this, finding that sequences that started as thermostable proteins have much lower predicted stability after design, and sequences that started as salt-tolerant proteins have lower predicted tolerance after design.

Simulated design. We follow the design methodology in Zhu et al. (2024) and Fannjiang et al. (2022) to generate sets of protein designs that achieve optimal tradeoffs between fitness and diversity, using various species' orthologs as starting points and Progen2-xlarge to guide sampling (Appendix G).

We run three tranches of design, each with a different species focus:

1. *Species representing the full range of Elo ratings.* 20 different proteins, each with 15–20 different species' homologs as starting points.
2. *Thermophilic (heat-loving) species.* 13 proteins; 3–7 species' homologs as starting points.
3. *Halophilic (salt-loving) species.* 18 proteins; 1–3 species' homologs as starting points.

[3]Mutation rates in fact vary across species and proteins, so as more species' proteomes are annotated, one could compute a more precise effective sequence count by directly calculating the average sequence similarity between each species' proteomes, and using this in place of the $e^{-\frac{d(i,j)}{\alpha}}$ term.

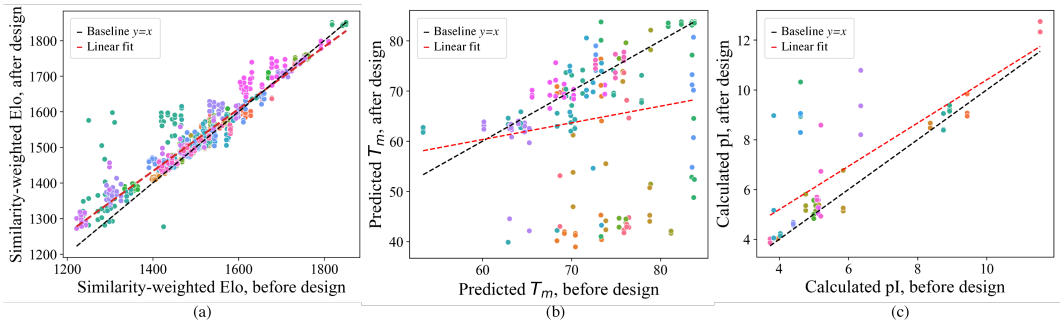

Figure 3: **Protein properties before and after design.** (a) Similarity weighted Elo from design is higher after design. (b) Predicted melting temperature (T_m) is lower (i.e. proteins are less stable) after design. (c) Calculated isoelectric point (pI) is higher (i.e. proteins are less tolerant to salt) after design. Each dot represents one protein design, and color indicates protein type (see Appendix G).

Analysis of species drift. To assess whether designed sequences systematically become more similar to sequences from high-Elo species, we define the *similarity-weighted Elo* of a sequence to reflect the average Elo of species who have homologs similar to that sequence. Formally:

$$s_{\text{sim-weighted Elo}}(x) = \frac{1}{\sum_j s(x, x_j)} \sum_j \big(s(x, x_j) \cdot \text{Elo}(j)\big),$$

where x_j is the homolog of x from species j, $s(x, x_j)$ is the sequence similarity between x and x_j, and $\text{Elo}(j)$ is the Elo rating of species j. We use $s(x, x_j) = (1 - \text{Levenshtein}(x, x_j))^2$; other choices of s lead to similar results (Appendix G).

Figure 3a plots the similarity-weighted Elo of a sequence before and after design. We see that final design sequences tend to increase their similarity-weighted Elo, and we find that this increase is statistically significant under a paired sample t-test ($t = 15.2$, $p = 7.7\text{e-}47$). This drift holds across the spectrum of Elo ratings and is most prominent for low Elo starting points. This is consequential because many low Elo species are rich sources of useful proteins, as we discuss further in the following section.

Analysis of extremophile properties. Many of the species with the lowest Elo ratings are extremophiles. Despite their low Elo, these species' proteins are a valuable resource for developing novel biotechnology. For instance, species that thrive in extreme heat have evolved proteins with high thermostability (the ability to remain stably folded at high temperatures), which is necessary for many industrial use-cases of engineered proteins (Modarres et al., 2016). Similarly, species that thrive in salty environments have evolved proteins to be more acidic (negatively charged) to prevent aggregation, and this salt tolerance is crucial to bioremediation strategies such as biofuel production (Daoud & Ben Ali, 2020).

We evaluate the results of design *in silico* (i.e. with a predictive model). We assess the thermostability of a sequence with the protein melting temperature (T_m) predictor DeepSTABp (Jung et al., 2023). We assess the salt tolerance of a sequence with the isoelectric point (pI) calculator (Kozlowski, 2016), as proteins in halophilic species have lower pI than other species' homologs in order to remain stable at high salt concentrations (Gunde-Cimerman et al., 2018).

Figure 3b plots the predicted melting temperature after design vs. before design. We see that designs tend to decrease their predicted melting temperatures, and we find that this decrease is statistically significant under a paired sample t-test ($t = -7.1$, $p = 3.5\text{e-}11$). 63% of designs have lower predicted melting temperatures than their starting sequence, and in one-third of those cases, the predicted decrease is over 20°C. The magnitude of this thermostability decrease is consequential for many applications. For example, engineering strains to successfully ferment bioethanol at 15°C higher temperatures makes a much wider set of raw materials economically feasible for biofuel production (Miah et al., 2022).

Figure 3c plots the calculated isoelectric point (pI) after design vs. before design. We see that designs tend to increase their pI, and we find that this increase is statistically significant under a paired sample t-test ($t = 4.6$, $p = 2.0\text{e-}5$). 83% of designs have higher pI than their starting sequence, and the average pI increase of 4.6 may be consequential–the difference in pH between vinegar and neutral water is approximately 4-5.

5 CONCLUSION

In this work we identify and quantify a species bias in pLM likelihoods, trace its origins to uneven sequence sampling across the tree of life, and document its effect of pushing protein designs toward sequences from favored species. This design bias is most likely to be detrimental when the starting point comes from an organism under-represented in sequence databases, and we demonstrate that likelihood-guided design can reduce the thermostability of proteins from heat tolerant species and reduce the salt-tolerance of proteins from species that thrive at high salinity.

It will be important to try to mitigate this species bias, both during pre-training and with post-hoc adjustments. During pre-training, up-weighting sequences from under-represented branches of the tree of life could help reduce bias. There may also be a need to develop novel algorithms for protein design that mitigate impacts of the bias, for example by adjusting the acceptance rate of proposed mutations based on whether the mutation moves towards a common or uncommon ortholog.

However, it is also possible that in some settings the bias will happen to align with design goals. For example, antibody therapeutics are often produced from non-human sources and can generate immunogenic responses in humans (Marks et al., 2021). It would be interesting to test pLM capabilities for humanizing antibodies such that they do not elicit an immune response and thus become safe for therapeutic use.

We focus on likelihoods, but pLM embeddings are also increasingly used in protein design, especially when supervision is available to fine-tune the model. It will be interesting to examine whether embeddings are affected by similar biases and whether they remain after fine-tuning. Future research could also study whether other types of protein models inherit related biases, such as generative models of protein sequences based on 3D structure (Strokach et al., 2020; Hsu et al., 2022; Dauparas et al., 2022) or generative models of protein structures (Ingraham et al., 2023; Watson et al., 2023). The training data for these models is often even more limited in coverage than the sequence databases used for pLM training.

More broadly, this work highlights the importance of data curation for biological datasets. Training pLMs has only been possible due to decades-long efforts from scientists to standardize sequence information in huge, public databases. While the databases at first primarily served as a resource for protein information queries, today they are additionally treated as defining *distributions* over natural protein sequences. As databases and models continue to grow, it is critical to understand biases present in the data collection process, evaluate whether mitigation of these biases is warranted, and leverage the rich annotations and meta-data in these databases to curate training data with downstream use-cases of models in mind.

ACKNOWLEDGMENTS

We thank Stephan Allenspach, James Bowden, Moritz Hardt, Milind Jagota, Hanlun Jiang, Erik Jones, Jennifer Listgarten, Thi Nguyen, Hunter Nisonoff, Alex Pan, Jeremy Reiter, Yun Song, and Junhao Xiong for helpful discussions. FD is supported by the NSF Graduate Research Fellowship Program under Grant No. DGE 1752814 and the Open Philanthropy AI Fellowship Program. JS is supported by the NSF SaTC CORE Award No. 1804794 and the Simons Foundation.

REFERENCES

Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Alex Xijie Lu, Nicolo Fusi, Ava Pardis Amini, and Kevin K Yang. Protein generation with evolutionary diffusion: sequence is all you need. *bioRxiv*, pp. 2023–09, 2023.

Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322, 2019.

Amos Bairoch and Rolf Apweiler. The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic acids research*, 28(1):45–48, 2000.

Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. Elo uncovered: Robustness and best practices in language model evaluation. *arXiv preprint arXiv:2311.17295*, 2023.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 11 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac1052. URL https://doi.org/10.1093/nar/gkac1052.

Lobna Daoud and Mamdouh Ben Ali. Chapter 5 - halophilic microorganisms: Interesting group of extremophiles with important applications in biotechnology and environment. In Richa Salwan and Vivek Sharma (eds.), *Physiological and Biotechnological Aspects of Extremophiles*, pp. 51–64. Academic Press, 2020. ISBN 978-0-12-818322-9. doi: https://doi.org/10.1016/B978-0-12-818322-9.00005-8. URL https://www.sciencedirect.com/science/article/pii/B9780128183229000058.

J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022. doi: 10.1126/science.add2187. URL https://www.science.org/doi/abs/10.1126/science.add2187.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

Arpad E Elo. *The Rating of Chess Players: Past and Present*. Ishi Press International, 1978.

Hyunuk Eom, Kye Soo Cho, Jihyeon Lee, Stephanie Kim, Sukhwan Park, Hyunbin Kim, Jinsol Yang, Young-Hyun Han, Juyong Lee, Chaok Seok, et al. Discovery of highly active kynureninases for cancer immunotherapy through protein language model. *bioRxiv*, pp. 2024–01, 2024.

Clara Fannjiang, Micah Olivas, Eric R Greene, Craig J Markin, Bram Wallace, Ben Krause, Margaux M Pinney, James Fraser, Polly M Fordyce, Ali Madani, et al. Designing active and thermostable enzymes with sequence-only predictive models. In *NeurIPS 2022 Workshop on Learning Meaningful Representations of Life*, 2022.

Nina Gunde-Cimerman, Ana Plemenitaš, and Aharon Oren. Strategies of adaptation of microorganisms of the three domains of life to high salt concentrations. *FEMS microbiology reviews*, 42(3):353–375, 2018.

Brian L Hie, Varun R Shanker, Duo Xu, Theodora UJ Bruun, Payton A Weidenbacher, Shaogeng Tang, Wesley Wu, John E Pak, and Peter S Kim. Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology*, 2023.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8946–8970. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/hsu22a.html.

John B Ingraham, Max Baranov, Zak Costello, Karl W Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M Lord, Christopher Ng-Thow-Hing, Erik R Van Vlack, et al. Illuminating protein space with a programmable generative model. *Nature*, pp. 1–9, 2023.

Milind Jagota, Chengzhong Ye, Carlos Albors, Ruchir Rastogi, Antoine Koehl, Nilah Ioannidis, and Yun S Song. Cross-protein transfer learning substantially improves disease variant prediction. *Genome Biology*, 24(1):182, 2023.

Sean R Johnson, Xiaozhi Fu, Sandra Viknander, Clara Goldin, Sarah Monaco, Aleksej Zelezniak, and Kevin K Yang. Computational scoring and experimental evaluation of enzymes generated by neural networks. *bioRxiv*, pp. 2023–03, 2023.

Felix Jung, Kevin Frey, David Zimmer, and Timo Mühlhaus. Deepstabp: A deep learning approach for the prediction of thermal protein stability. *International Journal of Molecular Sciences*, 24(8): 7444, 2023.

Lukasz P Kozlowski. Ipc–isoelectric point calculator. *Biology direct*, 11(1):1–16, 2016.

Sudhir Kumar, Michael Suleski, Jack M Craig, Adrienne E Kasprowicz, Maxwell Sanderford, Michael Li, Glen Stecher, and S Blair Hedges. Timetree 5: an expanded resource for species divergence times. *Molecular Biology and Evolution*, 39(8):msac174, 2022.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, pp. 1–8, 2023.

Claire Marks, Alissa M Hummer, Mark Chin, and Charlotte M Deane. Humanization of antibodies using a machine learning approach on large-scale repertoire data. *Bioinformatics*, 37(22):4041–4047, 2021.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34:29287–29303, 2021.

Roni Miah, Ayesha Siddiqa, Udvashita Chakraborty, Jamsheda Ferdous Tuli, Noyon Kumar Barman, Aukhil Uddin, Tareque Aziz, Nadim Sharif, Shuvra Kanti Dey, Mamoru Yamada, et al. Development of high temperature simultaneous saccharification and fermentation by thermosensitive saccharomyces cerevisiae and bacillus amyloliquefaciens. *Scientific Reports*, 12(1):3630, 2022.

H Pezeshgi Modarres, MR Mofrad, and AJRA Sanati-Nezhad. Protein thermostability engineering. *RSC advances*, 6(116):115252–115270, 2016.

Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978, 2023.

Pascal Notin, Aaron W Kollasch, Daniel Ritter, Lood Van Niekerk, Steffan Paul, Han Spinner, Nathan J Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Jung-Eun Shin, Adam J Riesselman, Aaron W Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew C Kruse, and Debora S Marks. Protein design and variant prediction using autoregressive generative models. *Nature communications*, 12(1):2403, 2021.

Alexey Strokach, David Becerra, Carles Corbi-Verge, Albert Perez-Riba, and Philip M. Kim. Fast and flexible protein design using deep graph neural networks. *Cell Systems*, 11(4):402–411.e4, 2020. ISSN 2405-4712. doi: https://doi.org/10.1016/j.cels.2020.08.016. URL https://www.sciencedirect.com/science/article/pii/S2405471220303276.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Danqing Zhu, David H. Brookes, Akosua Busia, Ana Carneiro, Clara Fannjiang, Galina Popova, David Shin, Kevin C. Donohue, Li F. Lin, Zachary M. Miller, Evan R. Williams, Edward F. Chang, Tomasz J. Nowakowski, Jennifer Listgarten, and David V. Schaffer. Optimal trade-off control in machine learning–based library design, with application to adeno-associated virus (aav) for gene therapy. *Science Advances*, 10(4):eadj3786, 2024. doi: 10.1126/sciadv.adj3786. URL https://www.science.org/doi/abs/10.1126/sciadv.adj3786.

A Details on dataset creation

To create our orthologous protein sequence dataset, we started with the top 100 most sequenced species in the UniProt database (Consortium, 2022), filtered for redundancy, then augmented this list with additional model organisms that had whole genomes sequenced, resulting in 133 species total. Next we collected all protein sequences in the Swiss-Prot database (the human-annotated subset of UniProt (Bairoch & Apweiler, 2000)) associated with any of the species in our list. Based on their annotations, we divide the proteins into orthologous sets The vast majority of sequences were bacterial, so to create a balanced dataset with many points of comparison between eukaryotes and bacteria, we restricted our attention to proteins with at least 15 eukaryotic orthologs, resulting in 203 distinct protein types, and a total of 7545 sequences in our dataset, 40% being eukaryotic.

B Details on Progen2 and ESM2 models

Progen2 is an autoregressive transformer trained with next-token prediction on the UniRef90 database (a curated subset of UniProt clustered at 90% sequence identity), and we can compute exact sequence likelihoods. ESM2 has a bidirectional transformer architecture and is trained with masked language modeling on data collected in a two-tiered sampling scheme: first randomly select a UniRef50 database member, and then sample a training data point from the UniRef90 cluster that member belongs to. For ESM models, we compute a pseudo-likelihood by masking each token in the sequence, as in Lin et al. (2023).

C Details on Elo rating system

The Elo rating system was developed to calculate the relative skill levels of players in zero-sum games (Elo, 1978). The difference in two players' Elo ratings directly translates to the probability of one player winning in a match against the other; for example, the 400 Elo difference between a chess grandmaster and a candidate master implies that the grandmaster is expected to win 90% of matches. In our setting, each time two species have different sequences of the same protein type, we count this as a "match", where the winner is the species with the higher likelihood for their sequence. If a species has multiple sequences of the same protein, its median likelihood is used to determine the match result. All species start with a baseline Elo rating of 1500, and each pair-wise matchup updates the winner's rating upwards and the loser's downwards in a stochastic gradient descent-like step. We use the standard Elo update algorithm with $K = 32$ and average results over 50 permutations of the matchups to ensure results are robust (Boubdir et al., 2023).

D ELO RATINGS FROM DIFFERENT PLMS

Figure 4 plots the correlation between Elo ratings computed from different pLMs. We see that pLMs within the same family have nearly identical Elo ratings, and pLMs across families also correlate highly.

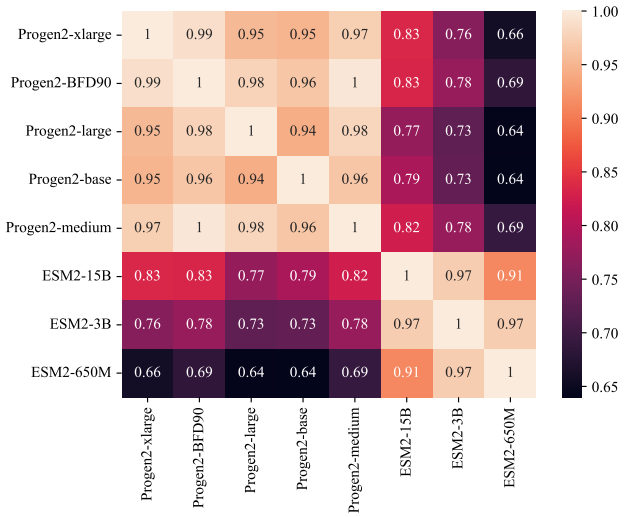

Figure 4: **Heatmap of the Pearson correlation between Elo scores from different pLMs.**

E PHYLOGENETIC TREE

Figure 5 displays the phylogenetic tree connecting the species in our dataset, annotated with sequence counts and Elo ratings. A few model organisms have extremely large sequence counts, while Elo ratings are spread more diffusely across species. Many species with few sequence counts nonetheless have a high Elo rating, often when the species shares a recent common ancestor with one of the model species with a large sequence count. This motivated our hypothesis that Elo ratings are influenced not just by a single species' sequence counts, but also by sequence counts from evolutionarily close species.

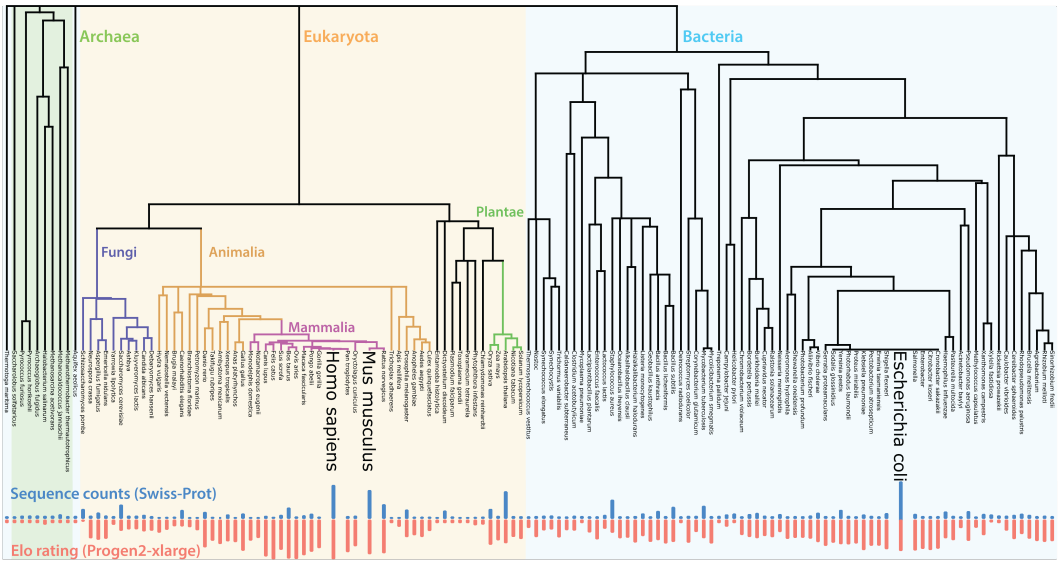

Figure 5: **Phylogenetic tree annotated with sequence counts and Elo ratings.** The tree's vertical axis represents time: number of years to the last common ancestor between two species is proportional to the distance from the leaves to the last common node between them.

F ROBUSTNESS CHECKS WITH DIFFERENT DATABASE SEQUENCE COUNTS

In the main text, Figure 2 shows the correlation between Elo scores and raw sequence counts, and between Elo scores and evolution-weighted sequence counts, using the number of SwissProt entries per species as the raw sequence count. Here we show that we find similar results by using other choices of raw sequence counts. Figure 6 shows results with UniRef90 sequence counts. Figure 7 shows results with sequence counts tallied from two-tiered sampling from UniProt: first sample a representative member of UniRef50, then sample a protein sequence from the UniRef90 cluster that the member belongs to. In all cases, correlation with Elo ratings is significantly higher with evolution-weighted sequence counts.

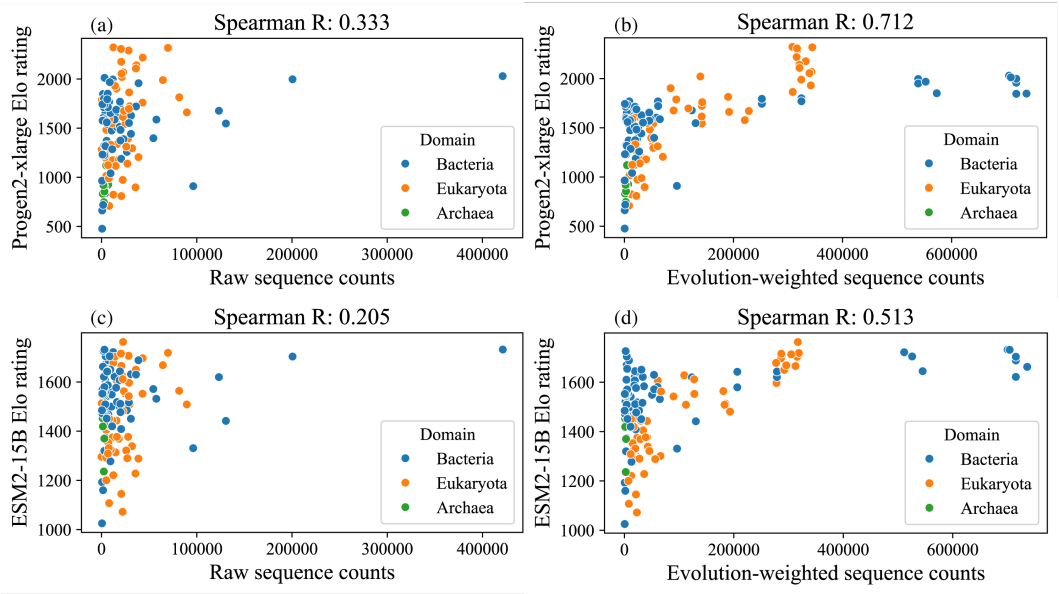

Figure 6: **Species Elo ratings plotted against their UniRef90 sequence counts and evolution-weighted sequence counts.** Top: Elo ratings computed from Progen2-xlarge likelihoods. Bottom: Elo ratings computed from ESM2-15B pseudolikelihoods. Correlation using raw sequence counts (left) is low, while correlation using evolution-weighted sequence counts (right) is higher.

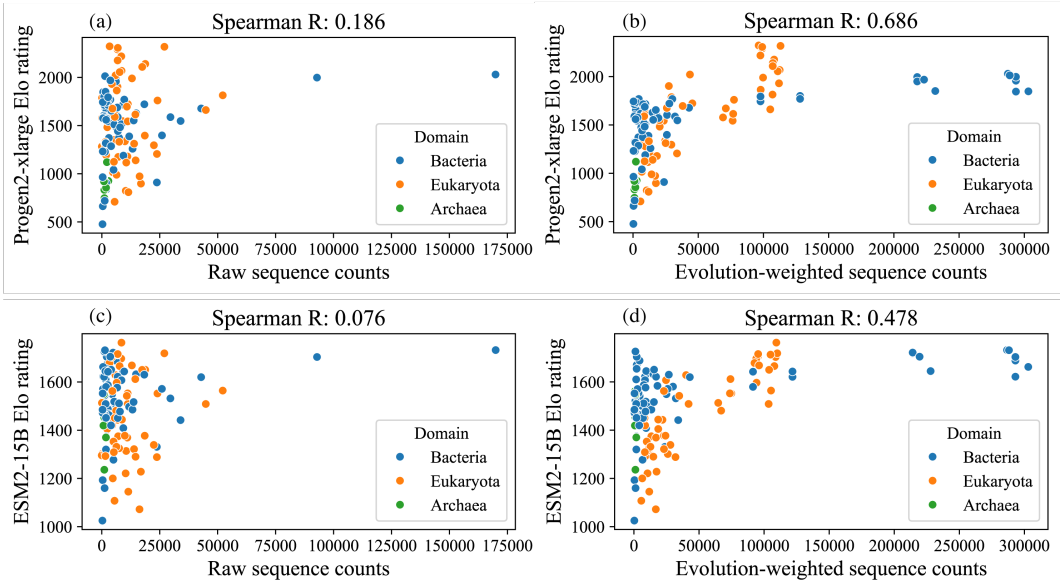

Figure 7: **Species Elo ratings plotted against sequence counts from two-tiered sampling: first sample a representative from UniRef50 and then sample a sequence from the UniRef90 cluster of the representative.** Top: Elo ratings computed from Progen2-xlarge likelihoods. Bottom: Elo ratings computed from ESM2-15B pseudolikelihoods. Correlation using raw sequence counts (left) is low, while correlation using evolution-weighted sequence counts (right) is higher.

G PROTEIN DESIGN DETAILS

In most protein design applications, a set of candidate designs is proposed, with the goal of both high protein fitness and diversity in the set. We follow the design methodology in Zhu et al. (2024) and Fannjiang et al. (2022), who show that optimal tradeoffs between fitness and diversity are achieved by sampling designs from sequence distributions that maximize entropy while satisfying constraints on mean fitness.

Formally, let $\mathcal{X}$ denote the set of all amino acid sequences of length L and let $\mathcal{P}$ denote set of all distributions over $\mathcal{X}$. We aim to sample from the sequence distribution, $p^\star$, that solves the following optimization problem:

$$\begin{aligned} \arg\max_{p \in \mathcal{P}} \quad & H(p) \\ \text{subject to} \quad & \mathbb{E}_p[f(x)] \geq \tau, \\ & \text{support}(p) \subseteq \textsc{HomologyRegion}(x_0) \end{aligned} \tag{1}$$

where $H(p)$ is the entropy of p, $f(x)$ is the sequence log-likelihood under some pLM, $\tau \in \mathbb{R}$ is a predicted fitness target threshold, and $\textsc{HomologyRegion}(x_0) \subseteq \mathcal{X}$ denotes the region of sequence space that is plausibly homologous to the starting sequence x_0 (to ensure we are designing sequences that still have similar function).

The distribution, $p^\star$, that solves equation 1 has a likelihood of the following form:

$$p^\star(x) \propto \begin{cases} \exp[\lambda f(x)] & \text{if } x \in \textsc{HomologyRegion}(x_0) \\ 0 & \text{otherwise,} \end{cases}$$

for Lagrange multiplier λ which has a one-to-one correspondence with the threshold τ in equation 1. Higher λ increases the average predicted fitness at the cost of lower diversity.

The distribution $p^\star$ is intractable to compute directly, so we use MCMC techniques to sample from it. From the starting sequence x_0, we iteratively propose mutations with Gibbs sampling for 10,000 steps, using Progen2-xlarge likelihoods for the predicted fitness $f(x)$. We set $\lambda = 1$ and sample 3 designs with different random seeds for each x_0.

G.1 EFFECTS OF DESIGN ON SIMILARITY-WEIGHTED ELO

For the results in the main text, we compute similarity-weighted Elo with $s(x, x_j) = (1 - \text{Levenshtein}(x, x_j))^2$. The Levenshtein distance between two sequences is the minimum number of single-character edits required to change one sequence into another. Here we show that results are robust to other choices of s. We use the Bio.Align package to score the optimal alignment between sequences with the Smith-Waterman algorithm, using the BLOSUM62 matrix to score the penalties for each amino acid substitution. Similar amino acids receive a smaller edit penalty compared to more chemically distinct amino acids, and we use the maximum alignment score as the similarity s. In contrast, the Levenshtein distance assigns an equal penalty to all substitutions. Figure 8 plots similarity-weighted Elo after design vs. before design. We see that for both similarity metrics, similarity-weighted Elo increases after design.

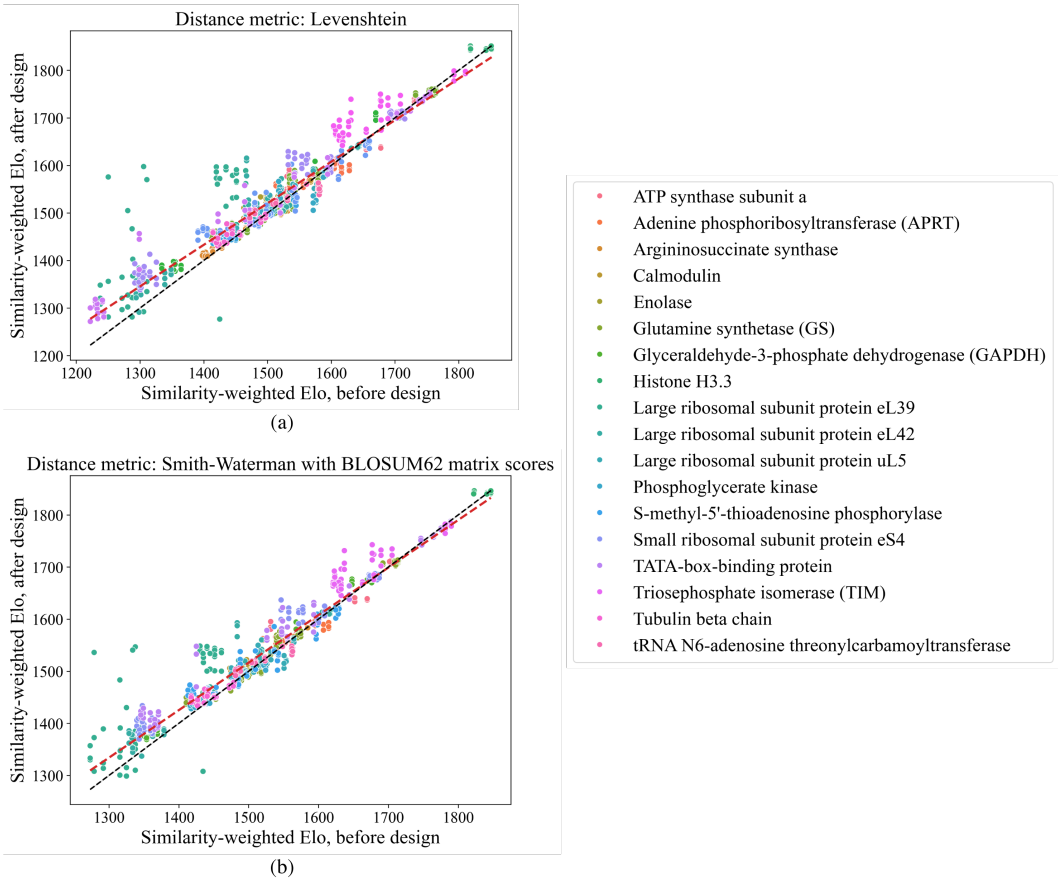

Figure 8: **Similarity-weighted Elo before and after design.** Top: similarity-weighted Elo computed using Levenshtein distance. Bottom: similarity-weighted Elo computed using the Smith-Waterman algorithm to compute a local alignment under BLOSUM62 matrix scores for amino acid substitutions.

G.2 Effects of design on thermostability

The 7 thermophilic species we study are *Methanocaldococcus jannaschii*, *Archaeoglobus fulgidus*, *Methanothermobacter thermautotrophicus*, *Methanosarcina acetivorans*, *Pyrococcus furiosus*, *Pyrococcus horikoshii*, and *Saccharolobus solfataricus*. Figure 9 plots the same data as Figure 3b with the legend added.

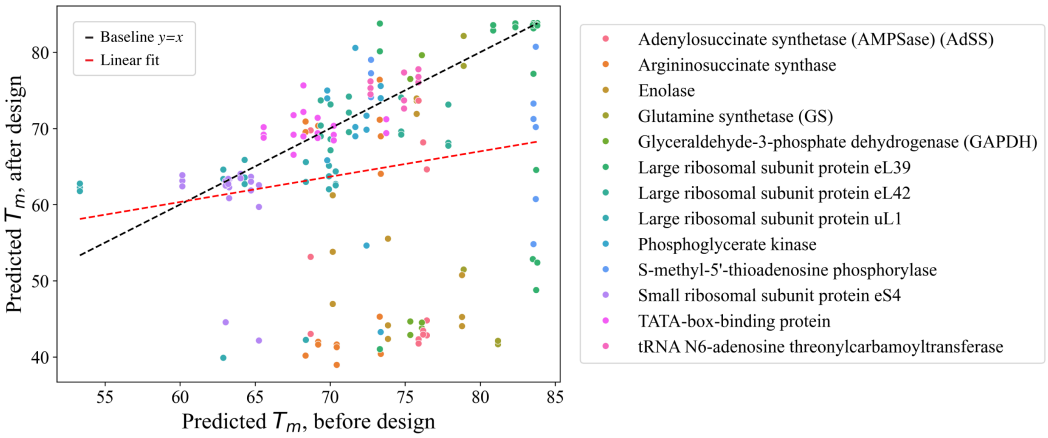

Figure 9: **Predicted melting temperature (T_m) after design vs. before design.**

G.3 Effects of design on salt tolerance

The 3 halophilic species we study are *Halobacterium salinarum*, *Alkalihalobacillus clausii*, and *Halalkalibacterium halodurans*. Figure 10 plots the same data as Figure 3c with the legend added.

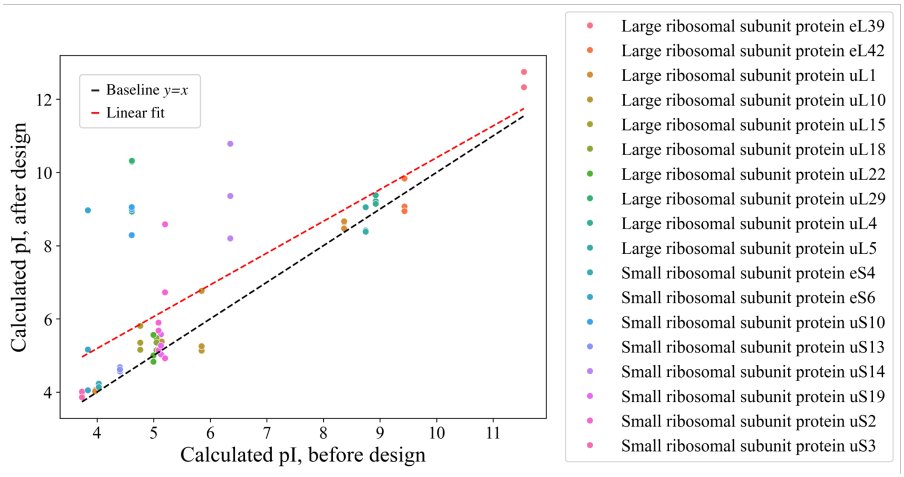

Figure 10: **Calculated isoelectric point (pI) after design vs. before design.**

G.4 CONVERGENCE TO HIGH ELO HOMOLOGS

We find that some protein designs result in sequences nearly identical to homologs from high Elo species. To quantify this, we compute the fraction of designs that "converge" to a homolog in our dataset, when we set the threshold for convergence to be 90%, 95%, and 98% sequence identity. Figure 11 plots these convergence frequencies. We see that the convergence rate varies significantly between different proteins, and that convergence happens much more often to a higher Elo species' homolog compared to a lower Elo species' homolog.

Examining the designs with the greatest changes in their predicted thermostability or salt-tolerance, we find that a number of them have extremely high sequence similarity (>90%) to mammalian homologs of the original protein. These homologs have been experimentally verified to have lower melting temperatures and lower salt tolerance. Thus, even though we only have access to predicted properties, these results suggest that protein design often pushes unique sequences into a basin of attraction around sequences from well-represented species, causing designs to diminish their unique properties.

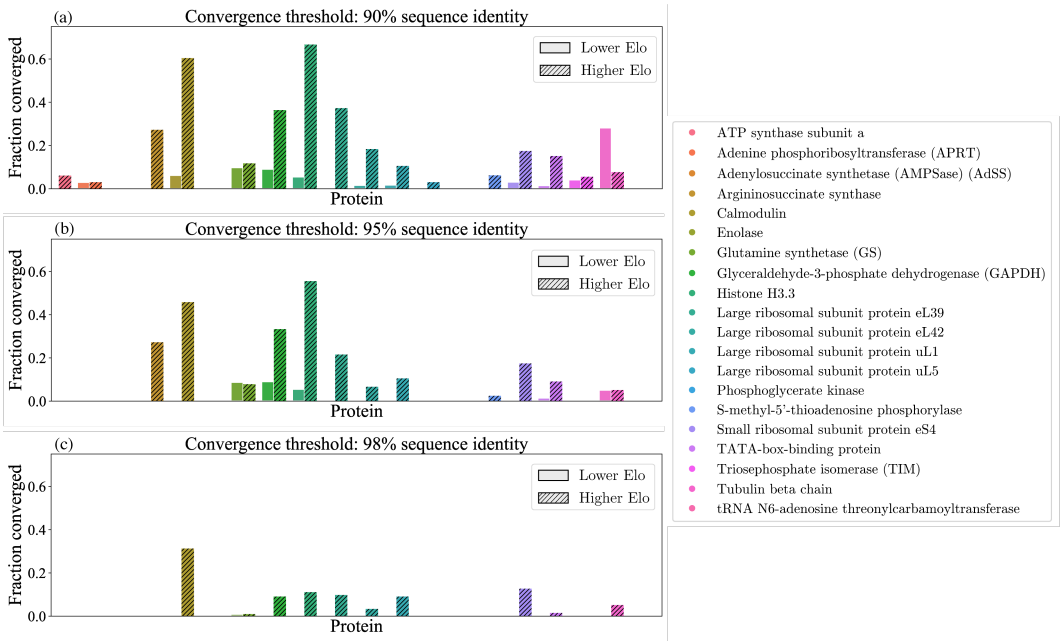

Figure 11: **Frequency of convergence to naturally-occurring homologs from a different species.** Convergence to a lower Elo species' homolog is represented by the solid bars, and convergence to a higher Elo species' homolog is represented by the shaded bars. The threshold for convergence is 90% sequence identity, 95% sequence identity, and 98% sequence identity for the top, middle, and bottom, respectively.