SCALING WORKS IN BIOLOGY IF YOU USE LATENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling is a dominant paradigm in the field of artificial intelligence. Despite the advantages of scaling in computer vision and large language models, recent studies have shown that scaling models trained on genome and protein sequences plateau in performance when evaluating mutational effects in the zero-shot setting. This zero-shot setting compares the likelihood of a variant sequence and its wild-type sequence, under the assumption that likelihood is a proxy for general fitness. The likelihood is based solely on the probability that a sequence, $p(sequence)$, occurs in the distribution of the data on which the model is trained. This largely ignores the biological context of the variant, leading to poor prediction of mutational effects on phenotypes such as enzymatic catalytic activity. Here we show that simple supervised probes fitted to frozen model representations reveal scaling gains in phenotype prediction that zero-shot likelihood misses. By examining the latent space of protein and genome language models across different scales and deep mutational scanning benchmarks, we restore the gains that should be expected from scaling.

1 INTRODUCTION

The scaling hypothesis has been a reliable engine for progress in artificial intelligence: increasing model parameters and training data predictably improves downstream capabilities (Hoffmann et al., 2022; Kaplan et al., 2020). In sequence biology, this paradigm has driven the development of massive protein and genomic language models (pLMs and gLMs) trained on billions of evolutionary sequences Brixi et al. (2026); Lin et al. (2023a); Bhatnagar et al. (2025); Penić et al. (2025).

Yet, a strange divergence occurs when applying these scaled models to one of biology's most critical tasks: predicting the functional effects of mutations. While larger models consistently yield better overall perplexity, their zero-shot predictions of biological fitness worsen with increasing mutational depth, failing to capture complex functional motifs. These models plateau in performance as they scale (Hou et al., 2026; Weinstein et al., 2022; Muir et al., 2025; Chungyoun & Gray, 2025).

The traditional zero-shot prediction paradigm computes the log-likelihood ratio 1 (LLR) of a mutant sequence x_j^{mut} and its wildtype sequence x_j^{wt} (Meier et al., 2021; Brixi et al., 2026):

$$LLR = \log p_\theta(x_j^{\text{mut}}) - \log p_\theta(x_j^{\text{wt}}) \tag{1}$$

LLR estimate assumes that fitness can be identified from sequence likelihood alone. However, in practice, this assumption does not hold true. Likelihood estimates are influenced by biases in the training datasets, favoring sequences from species that are better represented in the training data (Ding & Steinhardt, 2024). LLR tends to predict mutational effects that alter protein activity poorly, such as antibody binding and enzymatic activity Chungyoun & Gray (2025); Muir et al. (2025).

Pugh et al. address the perplexity issue with an inference method that averages likelihoods across sequences that share similar fitness landscapes. While this improves fitness assessments over pure LLR estimates and supports the notion of scaling models, this metric does not explicitly incorporate phylogeny or phenotype selection criteria, and still requires multiple sequence alignment, a computationally-intensive process. As LLR is an unconditioned estimate that does not explicitly include important biological context, such as phylogeny and phenotype, we argue that LLR is not a universal biological fitness function. Instead, we demonstrate that fitness assessments should directly account for biological context, such as phenotype selection and homology, through phenotype-specific prediction from latent representations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To take context into account, we harness the latent space of protein and genome language models. Language models transform high dimensional data into latent spaces, which are numerical representations of the input in context to the whole dataset. Latent spaces of larger language models have been shown to encode phylogenetic relationships between sequences (Ding et al., 2019; Ziegler et al., 2023). pLM latent spaces have also been used to extract steering vectors to perturb models towards generating sequences with properties of interest, such as low and high stability (Huang et al., 2025). The ability to steer the latent space with vectors implies that these language models linearly represent certain properties within their latent spaces (Park et al., 2024). Moreever, latents have been directly used for fitness assessments and protein design schemes through variational autoencoders, regression models, and transformers, and generally, they outperform LLR (Notin et al., 2023b; Tang et al., 2026; Jiang et al., 2025; Marquet et al., 2024; Ding et al., 2019).

We further explore the geometry of pLM and gLM latent spaces for phenotype-specific prediction across several model families, benchmarks, and scales. As models scale, they undergo performance gains for predicting fitness across multiple mutational depths, and can be used to better extrapolate to more extreme activity or stability values in a deep mutational scanning assay.

2 Latents reveal mutational effects missed by LLR

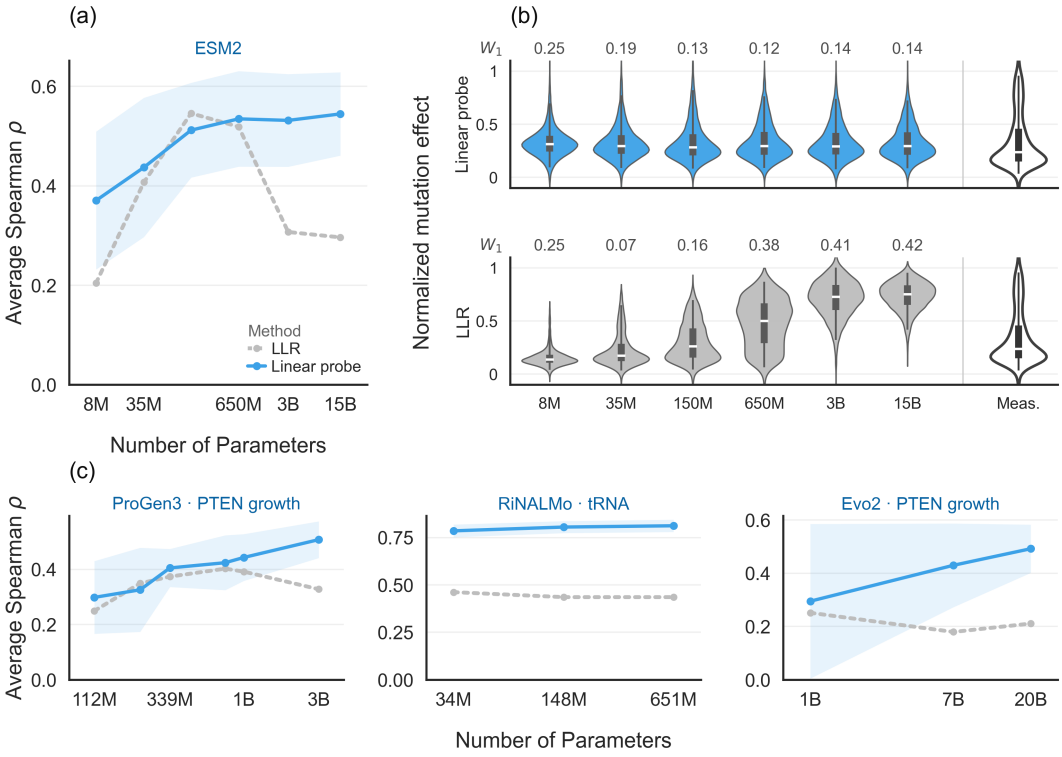


Figure 1: **Latent probes reveal scaling gains missed by sequence likelihood.** (a) PTEN growth prediction across ESM2 sizes, comparing latent probes with zero-shot log-likelihood ratios (LLR). Performance is Spearman correlation with measurements (mean ± standard deviation across folds). (b) Z-scored distributions of measured growth, linear-probe predictions, and LLR scores. W_1 is the 1-Wasserstein distance between predicted and measured distributions (lower is better). (c) Corresponding comparisons for ProGen3 (PTEN), RiNALMo (tRNA), and Evo2 (PTEN).

Phosphatase and tensin homolog (PTEN) is one example where larger protein language models computed LLR poorly predict mutational effects on growth and stability (Hou et al., 2026). In this case, larger protein language models tend to assign higher probability to the wildtype sequence. Performance across ESM2 sizes, as assessed by distributions of mutational effects and Spearman correlation between predicted and experimentally measured growth, peaks with ESM2-150M and 650M. When using latents from ESM2, however, we demonstrate that scaling does improve the

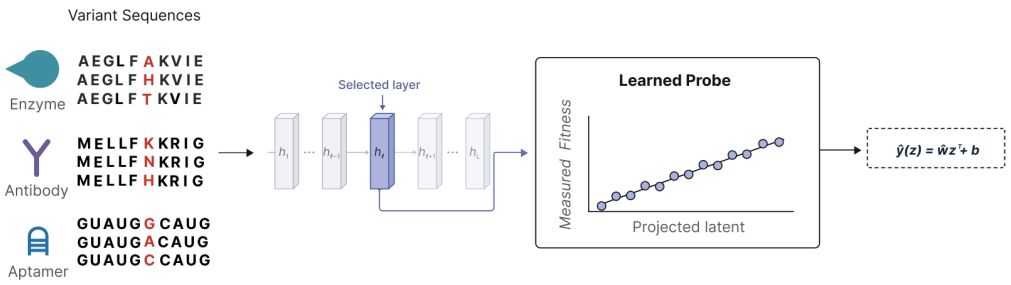


Figure 2: **Predicting biological properties from sequence latents.** A probe maps sequence representations from a selected language-model layer to the property of interest, using experimental measurements for supervision.

model's ability to recapitulate PTEN variants' growth. Scaling also helps recover the bimodal nature of the dataset, as quantified by Wasserstein distance W_1. We also see improvements in Spearman correlations with ProGen3 and Evo2 when evaluated on ProteinGym's PTEN growth folds, and with RiNALMo when evaluated on a tRNA benchmark from Li et al..

3 Methods for using latents

For each desired property, we fit probes on latents extracted from sequence language models of different biological modalities, namely, ESM2, ProGen3, Evo2, and RiNALMo to experimental measurements (Bhatnagar et al., 2025; Brixi et al., 2026; Lin et al., 2023b; Penić et al., 2025). For simplicity, we select ridge regression as our linear probe and a multilayer perceptron (MLP) as our nonlinear probe. More complex models optimized for each assay could also be used. Unless otherwise stated, we follow reported folds for training and evaluation splits from several fitness benchmarks, namely, ProteinGym, Flab2, and NABench for the genome language models (See Appendix A.2) (Notin et al., 2023a; Li et al., 2025; Chungyoun & Gray, 2025). We also include the sequence-catalysis landscape of the enzyme adenylate kinase (ADK) sampled across bacterial and archaeal phylogeny (Muir et al., 2025).

For an input sequence x containing N tokens, where each token is an amino acid or nucleotide, a protein or genome language model maps each token to a D-dimensional representation in each layer ℓ, resulting in a hidden-state matrix where $h_{j,i}^{(\ell)} \in \mathbb{R}^D$ is the latent representation of token i in variant x_j, and D is the hidden width of the model.

We conduct a layer, pooling, and normalization sweep to examine where biological information is best represented in each model and how to harness it. In general, the last hidden states of ESM2 and ProGen3 result in probes with higher performance. For Evo2, we see highest performance around 80% layer depth. For each mutant sequence within a deep mutational scanning assay, we pool token representations within a window centered on a mutant site $\mathcal{C}$. We denote this mutation-centered, variant-level representation as $z_j^{(\ell,r)} \in \mathbb{R}^D$, where r is the pooling radius that we sweep (Appendix A.3.2).

$$\tilde{z}_j = (z_j - \mu_{\text{fold}}) \oslash \sigma_{\text{fold}} \tag{2}$$

Here, μ_{fold} and σ_{fold} are the per-feature mean and standard deviation computed on the training fold only. Each $z_j^{(\ell,r)}$ is therefore paired with the experimentally measured phenotype y_j of variant j for downstream fitness values. To normalize across the latent space, broadly, we compare the following methods: subtracting the mean of the feature vector associated with the wildtype, z-scoring with respect to the wildtype, subtracting the mean of the sequences across the training fold, and z-scoring across the training fold (Appendix A.3.5). Broadly, z-scoring across the training fold results in better probe performance, evaluated by Spearman correlation between ground truth and predicted values.

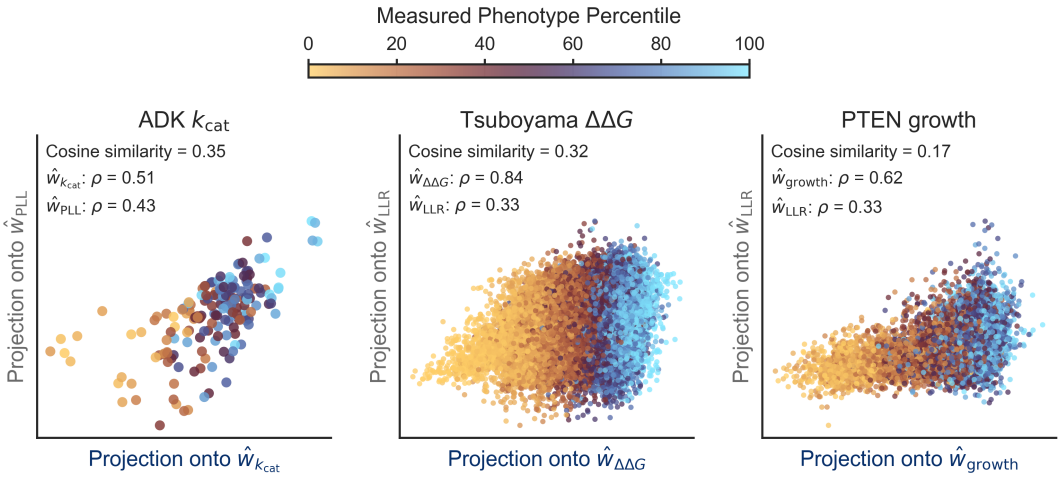


Figure 3: **Sequence likelihood aligns poorly with phenotype directions.** Each point is a sequence projected onto ESM2-3B directions fitted to measured phenotype (horizontal) and zero-shot likelihood (vertical). Color indicates phenotype percentile within each assay. Cosine similarity measures alignment between directions; ρ is the Spearman correlation of each projection with measured phenotype.

4 Sequence likelihood aligns poorly with phenotype directions

We find a vector $\hat{w}_{property}$ fitted on experimental measurements for a given property of interest with the linear probe. We also find a direction $\hat{w}_{LLR}$ within the latent space of ESM2-3B that corresponds with the computed zero-shot likelihoods with Muir et al.'s ADK k_{cat} assay, Tsuboyama et al.'s $\Delta\Delta G$, and Mighell et al.'s PTEN growth (Appendix Fig. 16). To compute the zero-shot log-likelihoods for ADK's ortholog sequences, we use the standard LLR formula (Equation 3) where natural sequences serve as the wildtype. For sequences without any counterparts, we use a sequence's own model likelihood rather than a log-likelihood ratio relative to a wild type (Equation 6).

The cosine similarities of the 2 fitted directions are under 0.35, showing limited alignment between the two directions (Figure 3). Projecting each sequence onto the fitted phenotype and LLR directions reveals a substantially stronger correlation between the measured phenotype and the fitted phenotype direction than with the LLR direction within the latent space.

5 Scaling improves overall mutational effect assessments

Next, we test whether these gains extend across assays, phenotypes, and model families. With all of the benchmarks described in appendix A.2, we compare the average Spearman correlations of the latent-based probes with those of the baseline zero-shot LLR. This spans over 300 assays for the protein language models and multiple properties: thermostability, growth, expression, enzymatic catalytic activity, binding affinity, and general protein activity. We survey fewer assays for RiNALMo, since it is trained on noncoding sequences and instead benchmarked on NABench (Li et al., 2025). Each Spearman correlation is computed as the average of the latent-based methods evaluated on held-out folds, per assay. We examine the kernel density of the average Spearman correlations per model size for each model family (Fig. 4). The Spearman correlations with LLR span a wider range than those of the linear probe, and are negative for certain assays. The linear probe recovers the negative Spearman correlations and generally achieves higher performance than the LLR. Increasing model size also improves the median Spearman correlation, although the increase is minimal between two similarly-sized models (Table 6). The same trends are observed with the nonlinear probe (Appendix Fig. 17). The range of average Spearman correlations varies depending on the phenotype and representation of an assay's sequences within a model's training data distributions. OpenGenome2 is not constructed to explicitly represent experimentally selected aptamer libraries and thus aptamer binding affinities are not as well predicted with Evo2 (Brixi et al., 2026).

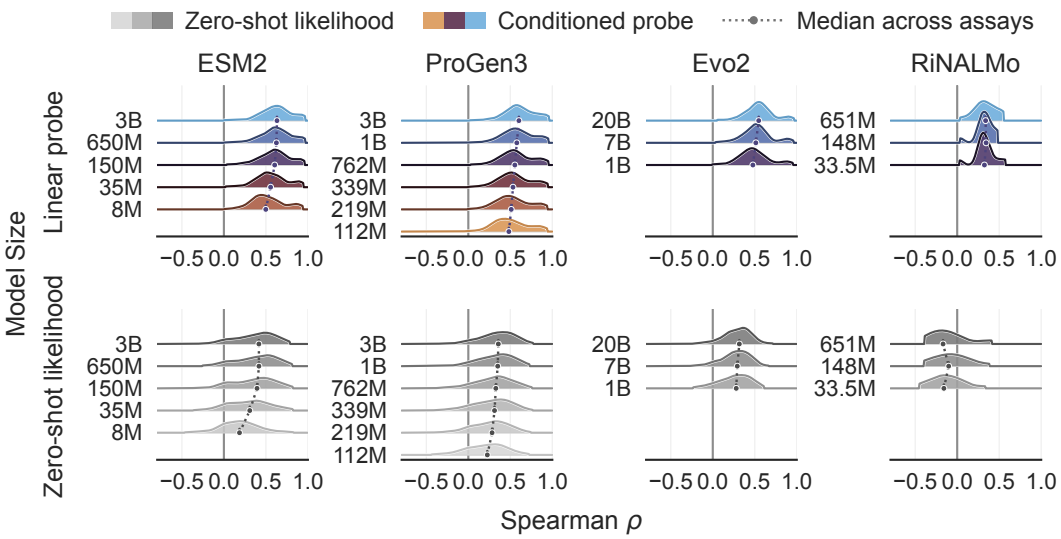


Figure 4: **Latent probes outperform zero-shot likelihood across model families.** Each curve shows the distribution of per-assay Spearman correlations, averaged over held-out folds, for linear probes (top) or zero-shot likelihood (bottom). Dotted lines connect medians across model sizes; the assay set is fixed within each family.

The probe still has a higher Spearman correlation compared to zero-shot likelihood (Appendix Fig. 19), but it is not as high performing compared to a probe fit on a better represented protein, like PTEN.

6 Probes recover local geometry of fitness landscapes

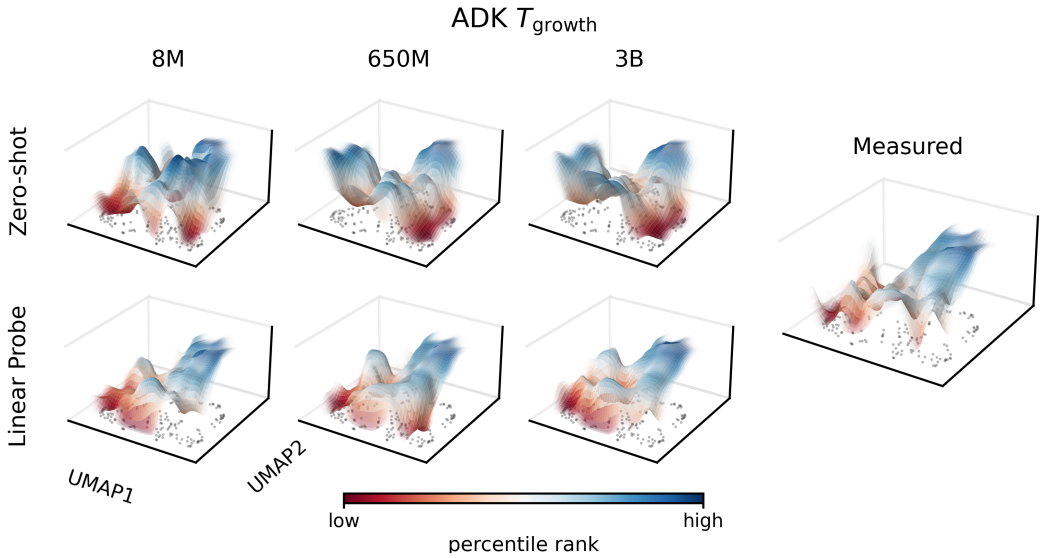
The majority of the phenotypes examined in figure 4 are from deep mutational scans that examine variant effects from single or double point substitutions. We further investigate the fitness landscapes of Muir et al.'s ADK dataset, which attains higher sequence diversity compared to single-point mutational scans. Muir et al.'s assay consists of 5149 curated ADK enzymes from bacteria and archaea, as well as 175 ortholog sequences. The overall sequence similarity, based on `MMSeq2` is no more 42% (Kallenborn et al., 2025).

We visualize these fitness landscapes as 3D surfaces following a similar approach as Diplock et al., with implementation differences as described in Appendix A.7. Likelihood-scoring tends to overestimate the temperatures of sequences with low T_{growth} temperatures, as seen in figure 5. To quantitatively compare the ADK landscapes, we examine the standardized landscapes with respect to unit variance and compare each landscape's standardized roughness as the mean absolute difference in fitness between neighboring sequences. We also examine the gradient similarity as the Spearman correlation between each surface local gradients.

In the case of the ADK T_{growth} landscapes, the zero-shot likelihoods overestimate roughness and have low gradient similarity with the true landscape of the measured T_{growth} values. Both the nonlinear and linear probes more closely capture the roughness of the measured T_{growth} landscape, and have nearly 2.5-fold improvement in gradient similarity over the likelihood-based T_{growth} landscape. Scaling from the 8M to 3B ESM2 model improves the overall geometry of the T_{growth}, with roughness ratio improving from 0.914 to 0.979 respectively. The nonlinear probe and the other nine measured landscapes are in Appendix A.7.

7 Latent probes improve prediction across mutational depths

On ProteinGym's multiple mutational depth benchmark, which includes assays with increasing numbers of substitutions away from a wildtype sequence, latent-based probes from ESM2, Evo2, and



ADK T_{growth} landscape metrics

	Roughness ratio			Gradient similarity		
Model	**Measured**	**Zero-shot**	**Linear**	**Measured**	**Zero-shot**	**Linear**
8M	1.000	0.914	0.914	1.000	0.367	0.538
650M	1.000	1.098	0.963	1.000	0.272	**0.620**
3B	1.000	1.113	**0.979**	1.000	0.259	0.615

Figure 5: **Latent probes better recover the shape of the ADK T_{growth} landscape.** Surfaces compare zero-shot likelihood (top row), linear probes (bottom row), and measurements (right) across ESM2 sizes. Height and color indicate within-surface percentile rank. Roughness ratio compares local variation with the measured landscape after unit-variance scaling (ideal: 1). Gradient similarity is the Spearman correlation between predicted and measured changes across neighboring sequences (higher is better). Bold marks the closest match at each model size. Probe predictions are out of fold across 30 random splits.

ProGen3 outperform LLR nearly 3-fold Notin et al. (2023a). Scale improves the linear probe performance, although the improvements are incremental (Fig. 6). Performance is generally lower at greater mutational depths, but latent probes retain an advantage over LLR. This is also seen with ProGen3 and Evo2 (Appendix Fig. 20).

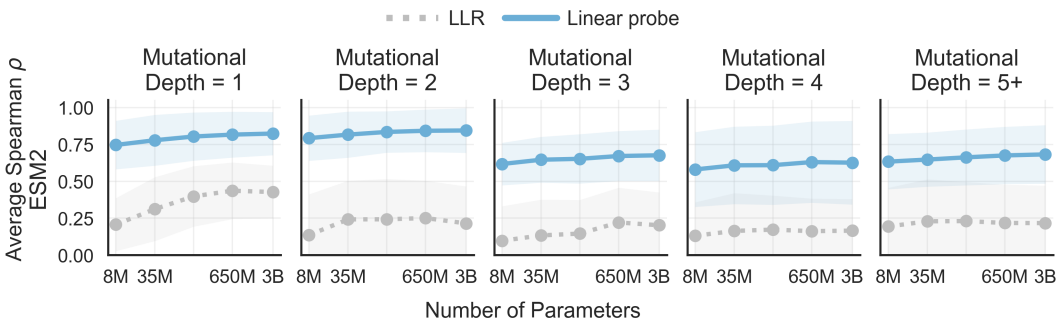


Figure 6: **Latent probes better predict the effects of multiple mutations.** Panels group ProteinGym variants by the number of amino-acid substitutions from wild type (1, 2, 3, 4, or $\geq$ 5). Curves compare linear probes with zero-shot log-likelihood ratios (LLR) across ESM2 sizes. Performance is the mean Spearman correlation with experimental measurements (higher is better).

8 LATENTS IMPROVE EXTRAPOLATION BEYOND A MEASURED RANGE

The folds we use so far are split at random or by mutation sites. Here, we explore how well latent-based probes perform when training and held-out folds are split by the range of experimental measurements. Specifically, we ask: how well do probes extrapolate to measurements that are beyond the range of labels seen during training? We examine Mighell et al.'s PTEN growth once more, as well as Sarkisyan et al.'s fluorescence, and Koenig et al.'s antibody binding and expression assay (Appendix Fig. 28). Across the board, the more labels included during supervision, the better the probe performance (Appendix Fig. 27). The probe performance is also influenced by positional encodings: certain variants held out, such as the antibodies with highest binding affinities, have more mutations within the light chain. Probes that do not fit on as many sequences with higher variability in the light chain may not extrapolate as well to the held-out sequences.

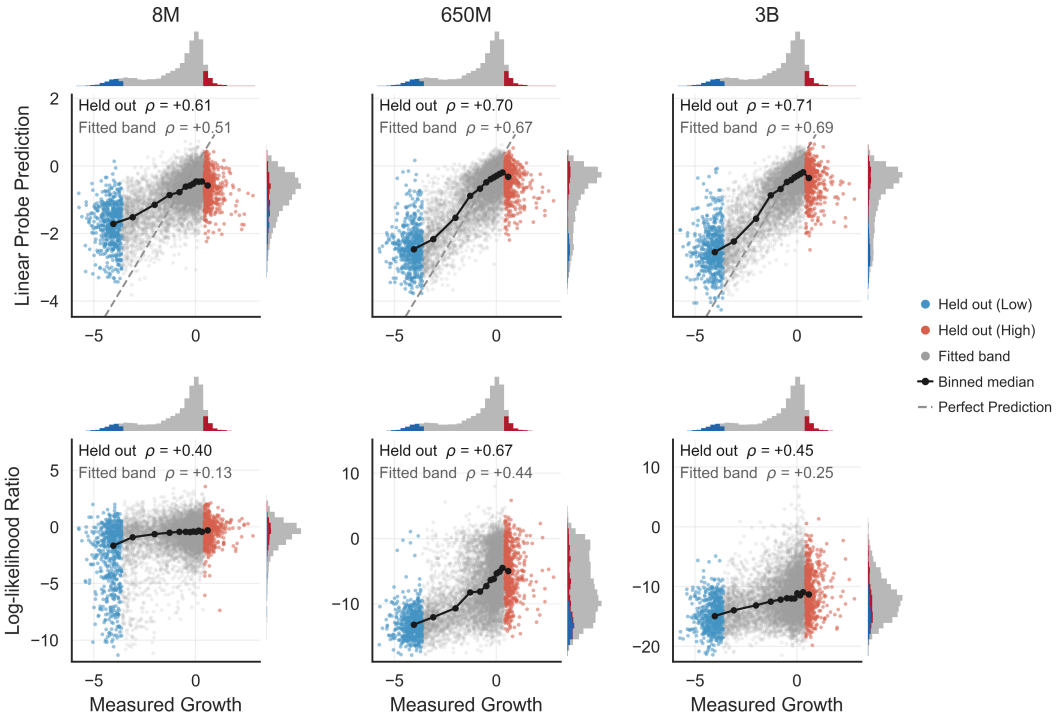


Figure 7: **Probes extrapolate to held-out measurements beyond the range of what was fitted.** 47.5% of the total range is within the fitted band (6,170 of 7,260 variants fitted, held out 545 from each extreme).

Within one fitted band, probes extrapolate better to more extreme measurements compared to LLR, as assessed by the Spearman correlation of the held out and predicted measurements, and this improves with increasing model size. The improvement in the held-out Spearman correlation is more extreme from ESM2-8M to ESM2-3B, with a change in Spearman correlation of $\Delta\rho$ of 0.10. The improvement from ESM2-650M and ESM2-3B is smaller, with a $\Delta\rho$ of 0.01. In contrast, the Spearman correlation of LLR computed with the ESM2-3B model is lower than that of ESM2-650M.

9 HOW MUCH SUPERVISION IS NEEDED?

We establish that the latent-based probes perform well across benchmarks consisting of well-characterized sequences. Each assay ranges from several hundreds to thousands of sequences and experimental measurements. Thus far, we typically fit probes on assays with 80-20 splits of a whole assay, typically consisting of thousands of sequences and measurements. One advantage of supervised models trained on latent spaces of known sequences is the reduction of the quantity of data to attain high performance. We further investigate how much experimental data is required in order

Table 1: **Probes achieve the same performance as LLR with a fraction of labels.** Overall, the same Spearman correlation as that of LLR can be reached with a couple hundred variants across assays. Assays were selected based on those with a Spearman correlation $\rho_{\mathrm{LLR}} \geq 0.45$.

Assay	Assay size	ρ_{LLR}	n_{LLR}	Range spanned	Reduction
HXK4	6,860	0.51	128	0.28	54×
KKA2	3,974	0.49	96	0.60	41×
β_2-adrenergic receptor	6,244	0.47	256	0.33	24×
RNC	3,424	0.57	192	0.53	18×
TEM-1 β-lactamase	4,002	0.66	256	0.75	16×
Tsuboyama $\Delta\Delta G$	737	0.51	64	0.60	10×
DYR	1,895	0.46	192	0.35	10×
IF1	1,098	0.55	256	0.69	4×

for fitted probes to attain high performance when splits are randomly sampled by sampling varied fraction of sequences and labels (Appendix A.9).

For each assay and model size, we partition variants into five folds across six independent shuffles, yielding 30 train–test splits with approximately 80% of variants available for training and 20% held out for evaluation. Within each split, we subsampled n $\in \{8, 16, 32, 64, 128, 256\}$ variants from the training pool and fit a ridge regression probe using only these measurements. We evaluate all measurement budgets and selection strategies within a split on the same held-out test variants. We define n_{LLR} as the smallest budget at which the probe's mean Spearman correlation over the 30 splits matches or exceeds that of the model's zero-shot likelihood, scored on the same held-out variants.

Overall, only a fraction of a whole deep mutational scanning assay is required for latent-based probes to attain the same Spearman correlation of LLR, with over a factor of 50 decrease for assays like glucokinase (HXK4). Only a couple hundred of sequences are required. Position coverage of substitutions and indels matter more for probe performance, as seen with worsened performance when folds are split contiguously as opposed to at random (Appendix A.3.6).

10 Do the probes generalize?

To test how generalizable fitted probes are, we examine how well a probe fit on one assay predicts the phenotypes of another assay. We also compute the cosine similarity between each assay's $\hat{w}$ vector. To keep phenotype constant, we use Tsuboyama et al.'s $\Delta\Delta G$ measurements across single-point mutations of 28 protein domains. Each probe is fit per protein domain.

Probes with higher cosine similarities transfer well when predicting phenotypes of the other assay. Probe cosine similarity increases with sequence similarity between the corresponding proteins, as quantified by `MMseqs2`. To examine transfer between probes fit on the same sequences, different property, we use Koenig et al.'s binding and expression measurements of vascular endothelial growth factor (VEGF) antibodies. We see that probes fit on the same sequences, but different phenotypes, still predict the other phenotype, although with worse performance as seen in figure 8. Low sequence similarity and no phenotype overlap results in the lowest cosine similarities, and worst probe performance when evaluating on other assays.

Averaging 2 separate fitted vectors for homolog proteins improves overall Spearman correlation when assessed on each homolog. Pooling labeled training examples from homologous proteins improves probe performance further. Generally, increasing the number of training examples improves the Spearman correlation between predicted and experimentally measured properties (Muir et al., 2025). To isolate the effect of protein relatedness, we keep the total number of training examples fixed across pooling conditions. Pooling latents of training sequences with high structural similarity, as quantified by `Foldseek` TM-score, but low sequence similarity produces little change in probe performance van Kempen et al. (2024). Likewise, combining latents for assays with low structural and sequence similarity similarly yields only small changes in predictive performance.

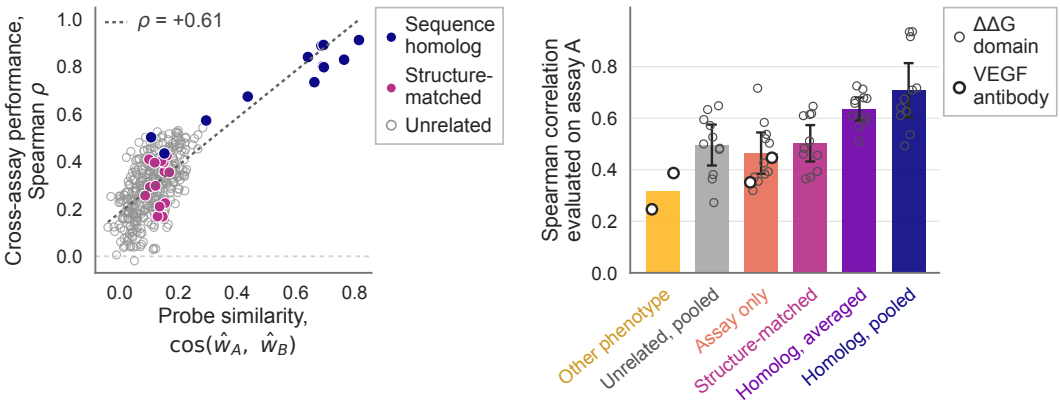


Figure 8: **Examination of assay-combination strategies, orthogonality of fitted probes, and property-separation.** Left: Scatterplot of cosine similarity versus Spearman correlation. A and B refer to 2 different protein domains from Tsuboyama et al.'s $\Delta\Delta G$ measurements. $\hat{w}_A$ and $\hat{w}_B$ refer to vectors fitted on assays A and B, respectively. Cross-assay performance is the Spearman correlation of linear probe with vector $\hat{w}_A$ evaluated on assay B's sequences and phenotypes. Sequence homologs refer to proteins with high sequence similarity. Structure-match refers to proteins with high structural similarity and low sequence similarity. Right: Held-out Spearman correlations for different training and transfer strategies: probes fit to the same sequences for a different phenotype; pooled latent representations from unrelated domains; assay-specific probes; pooled latent representations from structurally matched domains, defined by TM-score; averaged probe vectors $\hat{w}$ from homologous assays; and pooled latent representations from homologous proteins.

11 Limitations

Probes are dependent on the positions of mutations seen during training. The probes have the worst performance with the contiguous folds, where they are not fit on variants within an entire segment of a sequence. The generalization of probes depends on the sequence similarity of the sequences represented in the latents and the held-out sequence. This positional and sequence dependency may be due to the positional-encodings of these language models. This enables models to learn a notion of distance between nucleotides or amino acids in each sequence, but also reinforces positional biases in the latent space (Su et al., 2023). Furthermore, for this study, we choose simple methods of harnessing latents. Other methods, such as logistic regressions, variational autoencoders, cross-attention heads, and optimized MLPs, could be used for each phenotype or assay. We moreso seek to demonstrate simple ways of using latent spaces to evaluate sequences across different phenotypes and scale.

12 Conclusion

We establish that scaling laws do apply to protein and genome language models. As long as compute, model size, and datasets scale, models become more capable and develop stronger representations of biological information, such as thermostability, growth, and expression. With advancements in methods for evaluating sequences, our ability to engineer proteins for therapeutics, develop aptamers for targeting, and understand genetic diseases improves.

AI use statement

We used generative AI tools to execute code and to assist with literature review. LLM-generated code was double checked and tested. We conduct our own literature search as well, and read suggested papers from LLMs to cross-check LLM claims. We only use generative AI tools to assist with Latex formatting, suggestions for wording, and clarity. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced

with the aid of generative AI. The rest of the required disclosure tasks are not applicable to this work.

Reproducibility statement

The datasets and protein and genome language models described in this paper are all public. The probes are off-the-shelf models from `scikit-learn` or implemented as described in the methods on `pytorch`.

References

Ali Bashir, Qin Yang, Jinpeng Wang, Stephan Hoyer, Wenchuan Chou, Cory McLean, Geoff Davis, Qiang Gong, Zan Armstrong, Junghoon Jang, Hui Kang, Annalisa Pawlosky, Alexander Scott, George E. Dahl, Marc Berndl, Michelle Dimon, and B. Scott Ferguson. Machine learning guided aptamer refinement and discovery. *Nature Communications*, 12(1):2366, 2021. doi: 10.1038/s41467-021-22555-9. URL https://doi.org/10.1038/s41467-021-22555-9.

Aadyot Bhatnagar, Sarthak Jain, Joel Beazer, Samuel C. Curran, Alexander M. Hoffnagle, Kyle S. Ching, Michael Martyn, Stephen Nayfach, Jeffrey A. Ruffolo, and Ali Madani. Scaling unlocks broader generation and deeper functional understanding of proteins. In *Advances in Neural Information Processing Systems 38*, pp. 51531–51567, 2025. doi: 10.52202/085713-1536.

Garyk Brixi, Matthew G. Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, Alison Fanton, Gabriel A. Gonzalez, et al. Genome modelling and design across all domains of life with Evo 2. *Nature*, 652(8112):1349–1361, 2026. doi: 10.1038/s41586-026-10176-5.

Michael Chungyoun and Jeffrey Gray. Fitness landscape for antibodies 2: Benchmarking reveals that protein ai models cannot yet consistently predict developability properties. *bioRxiv*, 2025. doi: https://doi.org/10.64898/2025.12.27.696706.

Frances Ding and Jacob Steinhardt. Protein language models are biased by unequal sequence sampling across the tree of life. In *ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design*, 2024. URL https://openreview.net/forum?id=gVwiYMo4by.

Xinqiang Ding, Zhengting Zou, and Charles L. Brooks III. Deciphering protein evolution and fitness landscapes with latent space models. *Nature Communications*, 10(1):5644, 2019. doi: 10.1038/s41467-019-13633-0. URL https://doi.org/10.1038/s41467-019-13633-0.

Nathan Diplock, Thomas Desautels, Riyaz Haque, Micaela Finney, Daniel J. Wackelin, Christina S. Kang-Yun, Dan M. Park, and Sean P. Leonard. Unobserved sequence space has many functional proteins. *bioRxiv*, 2026. doi: 10.64898/2026.08.24.746828. URL https://www.biorxiv.org/content/early/2026/08/26/2026.08.24.746828.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

Chao Hou, Di Liu, Aziz Zafar, and Yufeng Shen. Understanding language model scaling for protein fitness prediction. *Nature Computational Science*, 6(7):778–788, 2026. doi: 10.1038/s43588-026-01010-z.

Long-Kai Huang, Rongyi Zhu, Bing He, and Jianhua Yao. Steering protein language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 26247–26260, 2025.

Kaiyi Jiang, Zhaoqing Yan, Matteo Di Bernardo, Samantha R. Sgrizzi, Lukas Villiger, Alisan Kayabolen, B. J. Kim, Josephine K. Carscadden, Masahiro Hiraizumi, Hiroshi Nishimasu, Jonathan S. Gootenberg, and Omar O. Abudayyeh. Rapid in silico directed evolution by a protein language model with evolvepro. *Science*, 387(6732):eadr6006, 2025. doi: 10.1126/science.adr6006. URL https://www.science.org/doi/abs/10.1126/science.adr6006.

Felix Kallenborn, Alejandro Chacon, Christian Hundt, Hassan Sirelkhatim, Kieran Didi, Sooyoung Cha, Christian Dallago, Milot Mirdita, Bertil Schmidt, and Martin Steinegger. Gpu-accelerated homology search with mmseqs2. *Nature Methods*, 22(10):2024–2027, 2025. doi: 10.1038/s41592-025-02819-8. URL https://doi.org/10.1038/s41592-025-02819-8.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Patrick Koenig, Chingwei V. Lee, Benjamin T. Walters, Vasantharajan Janakiraman, Jeremy Stinson, Thomas W. Patapoff, and Germaine Fuh. Mutational landscape of antibody variable domains reveals a switch modulating the interdomain conformational dynamics and antigen binding. *Proceedings of the National Academy of Sciences*, 114(4):E486–E495, 2017. doi: 10.1073/pnas.1613231114.

Zhongmin Li, Runze Ma, Jiahao Tan, Chengzi Tan, and Shuangjia Zheng. Nabench: Large-scale benchmarks of nucleotide foundation models for fitness prediction, 2025. URL https://arxiv.org/abs/2511.02888.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023a. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023b. doi: 10.1126/science.ade2574.

Céline Marquet, Julius Schlensok, Marina Abakarova, Burkhard Rost, and Elodie Laine. Expert-guided protein language models enable accurate and blazingly fast fitness prediction. *Bioinformatics*, 40(11), Nov 2024. ISSN 1367-4811 (Electronic); 1367-4803 (Print); 1367-4803 (Linking). doi: 10.1093/bioinformatics/btae621.

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

Taylor L. Mighell, Sara Evans-Dutson, and Brian J. O'Roak. A saturation mutagenesis approach to understanding pten lipid phosphatase activity and genotype-phenotype relationships. *The American Journal of Human Genetics*, 102(5):943–955, 2018. ISSN 0002-9297. doi: https://doi.org/10.1016/j.ajhg.2018.03.018. URL https://www.sciencedirect.com/science/article/pii/S0002929718301071.

Duncan F. Muir, Garrison P. R. Asper, Pascal Notin, Jacob A. Posner, Debora S. Marks, Michael J. Keiser, and Margaux M. Pinney. Evolutionary-scale enzymology enables exploration of a rugged catalytic landscape. *Science*, 388(6752):eadu1058, 2025. doi: 10.1126/science.adu1058.

Pascal Notin, Aaron W. Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Hansen Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora S. Marks. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems 36*, pp. 64331–64379, 2023a. doi: 10.52202/075280-2810.

Pascal Notin, Ruben Weitzman, Debora S. Marks, and Yarin Gal. ProteinNPT: Improving protein property prediction and design with non-parametric transformers. In *Advances in Neural Information Processing Systems 36*, pp. 33529–33563, 2023b. doi: 10.52202/075280-1458.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666, 2024.

Rafael Josip Penić, Tin Vlašić, Roland G. Huber, Yue Wan, and Mile Šikić. Rinalmo: general-purpose rna language models can generalize well on structure prediction tasks. *Nature Communications*, 16(1):5671, 2025. doi: 10.1038/s41467-025-60872-5. URL https://doi.org/10.1038/s41467-025-60872-5.

Charles W J Pugh, Paulina G. Nuñez Valencia, Mafalda Dias, and Jonathan Frazer. From likelihood to fitness: Improving variant effect prediction in protein and genome language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 130835–130866. Curran Associates, Inc., 2025. doi: 10.52202/085713-4359. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/bdb30687f1c2255c29b11b0b45204ebe-Paper-Conference.pdf.

Karen S. Sarkisyan, Dmitry A. Bolotin, Margarita V. Meer, Dinara R. Usmanova, Alexander S. Mishin, George V. Sharonov, Dmitry N. Ivankov, Nina G. Bozhanova, Mikhail S. Baranov, Onuralp Soylemez, Natalya S. Bogatyreva, Peter K. Vlasov, Evgeny S. Egorov, Maria D. Logacheva, Alexey S. Kondrashov, Dmitry M. Chudakov, Ekaterina V. Putintseva, Ilgar Z. Mamedov, Dan S. Tawfik, Konstantin A. Lukyanov, and Fyodor A. Kondrashov. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603):397–401, 2016. doi: 10.1038/nature17995. URL https://doi.org/10.1038/nature17995.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

Michelle Tang, Gareth A. Cromie, Anowarul Kabir, Martin S. Timour, Julee Ashmead, Russell S. Lo, Nathaniel Corley, Frank DiMaio, Hiroki Morizono, Ljubica Caldovic, Nicholas Ah Mew, Andrea Gropman, Amarda Shehu, and Aimée M. Dudley. Predicting epistasis across proteins by structural logic. *Proceedings of the National Academy of Sciences*, 123(3):e2516291123, 2026. doi: 10.1073/pnas.2516291123. URL https://www.pnas.org/doi/abs/10.1073/pnas.2516291123.

Kotaro Tsuboyama, Justas Dauparas, Jonathan Chen, Elodie Laine, Yasser Mohseni Behbahani, Jonathan J. Weinstein, Niall M. Mangan, Sergey Ovchinnikov, and Gabriel J. Rocklin. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, 620(7973):434–444, 2023. doi: 10.1038/s41586-023-06328-6.

Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, 42(2):243–246, 2024. doi: 10.1038/s41587-023-01773-0. URL https://doi.org/10.1038/s41587-023-01773-0.

Eli N. Weinstein, Alan N. Amin, Jonathan Frazer, and Debora S. Marks. Non-identifiability and the blessings of misspecification in models of molecular fitness. In *Advances in Neural Information Processing Systems 35*, 2022.

Cheyenne Ziegler, Jonathan Martin, Claude Sinner, and Faruck Morcos. Latent generative landscapes as maps of functional diversity in protein sequence space. *Nature Communications*, 14(1):2222, 2023. doi: 10.1038/s41467-023-37958-z. URL https://doi.org/10.1038/s41467-023-37958-z.

A Appendix

A.1 Fitness assessment details

A.1.1 Zero-shot log-likelihood ratio (LLR)

Zero-shot fitness computed through LLR measures a model's unconditional preference for a variant relative to its wild type.

For ESM2 and RiNALMo, we use the masked-marginal log-likelihood ratio. If $x^{\mathrm{wt}}_{\setminus i}$ is the wild-type sequence with position i masked, then

$$LLR = \sum_{(a_i^{\mathrm{wt}},i,a_i^{\mathrm{mut}})\in\mathcal{M}_j} \left[\log p_\theta(a_i^{\mathrm{mut}} \mid x^{\mathrm{wt}}_{\setminus i}) - \log p_\theta(a_i^{\mathrm{wt}} \mid x^{\mathrm{wt}}_{\setminus i})\right]. \tag{3}$$

Inference is run once per mutated position, and the ESM2 masked position is read at token index i under the convention in Table 3.

ProGen3 and Evo2 are autoregressive models, so their baselines are full-sequence log-likelihood differences:

$$LLR = \log p_\theta(x_j^{\mathrm{mut}}) - \log p_\theta(x_j^{\mathrm{wt}}), \qquad \log p_\theta(x) = \sum_{i=1}^{T} \log p_\theta(x_i \mid x_{<i}) \tag{4}$$

For Evo2, x^{mut} and x^{wt} are equal-length nucleotide windows and we report the mean token log likelihood (nats per nucleotide) before taking the difference. $x_{<i} = (x_1, \ldots, x_{i-1})$ denotes all preceding tokens. For ProGen3, the score uses the model's bidirectional N-to-C/C-to-N evaluation protocol.

A.1.2 Zero-shot likelihood

Certain assays may not have reference wildtype sequences. In this scenario, for masked language models, we compute the pseudo-log-likelihood (PLL) by masking each canonical amino-acid position in turn and summing the log-probability assigned to the observed residue,

$$\mathrm{PLL}\theta(x) = \sum i \in \mathcal{A}(x) \log p_\theta(a_i \mid x_{\setminus i}), \tag{5}$$

where $\mathcal{A}(x)$ contains positions with canonical amino acids and $x_{\setminus i}$denotes the sequence with position i masked.

Because sequences can vary in length, summed likelihoods can be confounded by sequence length. We therefore use the mean per-residue pseudo-log-likelihood,

$$\bar{\ell}\theta(x) = \frac{\mathrm{PLL}\theta(x)}{|\mathcal{A}(x)|}, \tag{6}$$

as the zero-shot score. For autoregressive models, we analogously use the mean per-token log-likelihood, $\log p_\theta(x)/T$. Zero-shot phenotype prediction is evaluated by correlating these sequence-level scores with the experimentally measured phenotype across sequences.

A.2 Benchmarks

The cross-validation folds follow the canonical splits from ProteinGym, FLAb2, and NABench, when provided (Notin et al., 2023a; Chungyoun & Gray, 2025; Li et al., 2025). The exception is the ADK dataset, which does not follow the same fold schemes, and instead uses 30 random seeds, consistent with Muir et al.. Bashir et al.'s aptamer for neutrophil gelatinase-associated lipocalin (NGAL) was included as an additional aptamer benchmark for RiNALMo and Evo2.

Fold	Description
Random	Splits created at random
Modulo	Splits are created by modulo operator
Contiguous	Strictest of 3 folds, divides by sequence sites to reduce mutation-position leakage

To compare the assays with one another, we use `MMseq2` on A40 GPUs to assess the sequence similarity (Kallenborn et al., 2025). We also examine structural similarity quantified as TM-score with Foldseek van Kempen et al. (2024).

Assays that contain non-canonical amino acids, non-canonical nucleotides, and stop codons were excluded. This filters out 3 assays from Flab2.

A.3 Represented models

Model	Number of Parameters N (total / active)	Tokens D	FLOPs $\approx 6ND$
ESM-2			
ESM-2 8M	8M	1.0 T	4.8×10^{19}
ESM-2 35M	35M	1.0 T	2.1×10^{20}
ESM-2 150M	150M	1.0 T	9.0×10^{20}
ESM-2 650M	650M	1.0 T	3.9×10^{21}
ESM-2 3B	3B	1.0 T	1.8×10^{22}
ESM-2 15B	15B	0.86 T	7.8×10^{22}
ProGen3			
ProGen3 112M	112M / 32M	n/p[a]	—
ProGen3 219M	219M / 63M	n/p[a]	—
ProGen3 339M	339M / 98M	200 B	1.2×10^{20}
ProGen3 762M	762M / 221M	n/p[a]	—
ProGen3 1.4B	1.4B / 393M	500 B	1.2×10^{21}
ProGen3 3.0B	3.0B / 866M	500 B	2.6×10^{21}
Evo 2			
Evo2 1B base	1.1B	1.0 T	6.6×10^{21}
Evo2 7B	6.5B	2.4 T	9.4×10^{22}
Evo2 20B	20B	n/p[b]	—
Evo2 40B	40.3B	9.3 T	2.25×10^{24}
RiNALMo			
rotary, dense RiNALMo micro	33.5M	n/p[c]	—
RiNALMo mega	148M	n/p[c]	—
RiNALMo giga	651M	n/p[c]	—

Table 2: Models examined in this study and how they fit within scaling laws (Hoffmann et al., 2022).

Model Family	Input sequence	Token layout
ESM2	amino acids	$\langle\text{cls}\rangle, x_1, \ldots, x_L, \langle\text{eos}\rangle$
ProGen3	amino acids	$\langle\text{bos}\rangle, 1, x_1, \ldots, x_L, 2, \langle\text{eos}\rangle$
Evo2	nucleotides (DNA)	$x_1, \ldots, x_L$
RiNALMo	nucleotides (RNA)	$\langle\text{cls}\rangle, x_1, \ldots, x_L, \langle\text{eos}\rangle$

Table 3: Tokenization details of selected models.

A.3.1 Protein sequence to coding DNA sequence conversion for Evo2

Evo2 is tokenized by nucleotides, so a protein assay with amino acid sequences must be converted to coding DNA sequences. To do so, we obtain a wild-type CDS is with UniProt/NCBI homology lookup or built from seeded-random synonymous codons. For each mutant, we swaps one codon from the wild-type sequence for a seeded-random codon of the mutant amino acid.

For feature engineering, we use windows of size W centered on the mutated codon's middle nucleotide,

$$\mathrm{C} \;=\; 3(\mathrm{pos}-1)+1, \tag{7}$$

against a wild-type CDS per assay. This may introduce some further biases in Evo2's fitness estimates, as certain codons may occur more frequently within the OpenGenome2's dataset compared

to other codons. We do not codon optimize for each species. The same CDS is used across all of the presented fitness metrics, so any biases from the seeded-random codon selection would persist in all of the fitness estimations.

A.3.2 Full Layer sweeps

To determine which latents to use, layer and pooling sweeps were completed for ESM2, Evo2, ProGen3, and RiNALMo.

To conduct pooling sweeps, we vary r where $I(C_j, r)$ contains all valid sequence positions within r tokens of any mutated site in C_j. Overlapping positions are counted once.

$$I(\mathcal{C}_j, r) = \bigcup_{c \in \mathcal{C}_j} \{t \in \mathbb{Z} : \max(\text{lo}, c - r) \le t \le \min(\text{hi}, c + r)\}, \qquad z_j = \frac{1}{|I(\mathcal{C}_j, r)|} \sum_{t \in I(\mathcal{C}_j, r)} h_{j,t}^{(\ell)} \tag{8}$$

$$z_j^{(\ell, r)} = \underset{t \in I(C_j, r)}{\text{mean}} \, h_{j,t}^{(\ell)}. \tag{9}$$

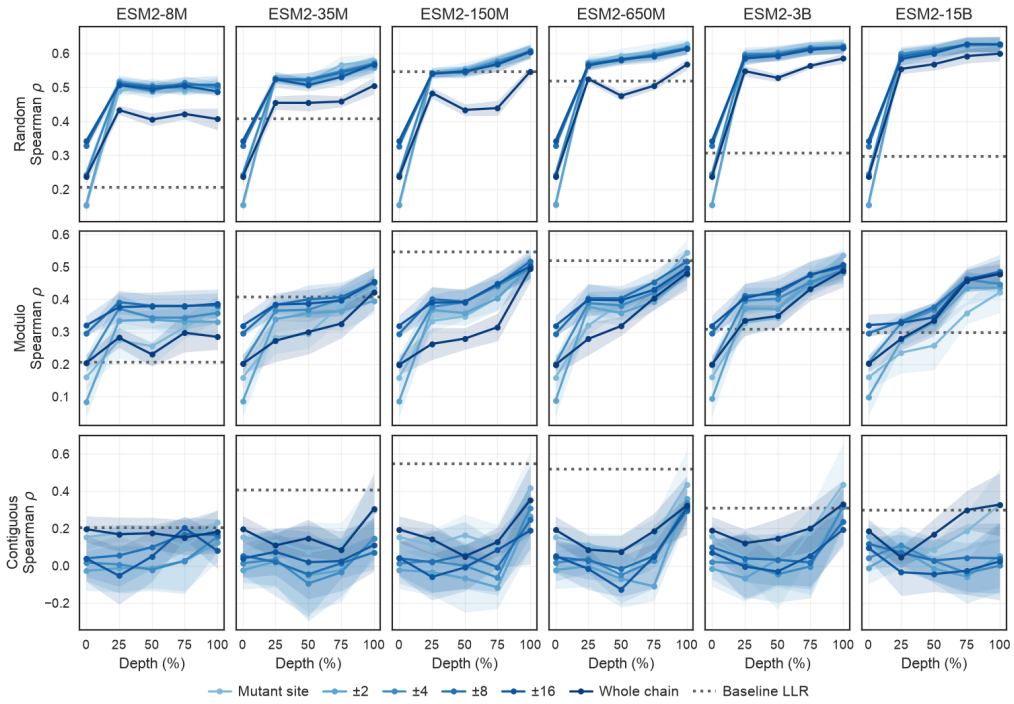


Figure 9: Layer and pooling sweep across ESM2 models and cross-validation folds.

With random folds, whole chain pooling consistently resulted in lower Spearman correlations. Varying window sizes up to r=16 does not significantly change the Spearman correlation. With larger model sizes, the gap between the whole chain pooling and the smaller window sizes decrease. The whole chain pooling with the modulo fold results in a similar pattern as the random fold: it overall underperforms other pooling schemes but improves with larger model sizes. In contrast, the effects of different window sizes is more prominent. Larger pooling windows around the mutant size improves the overall Spearman correlation. The contiguous fold has the most variance in probe performance and is the hardest benchmark out of the three folds. On average, the larger pooling window sizes improves the Spearman correlations, with the whole chain pooling resulting in better probe performance.

For Evo2, generally, the best layers were around 75 to 80% depths, and had a steep decrease in property encoding in the final layers.

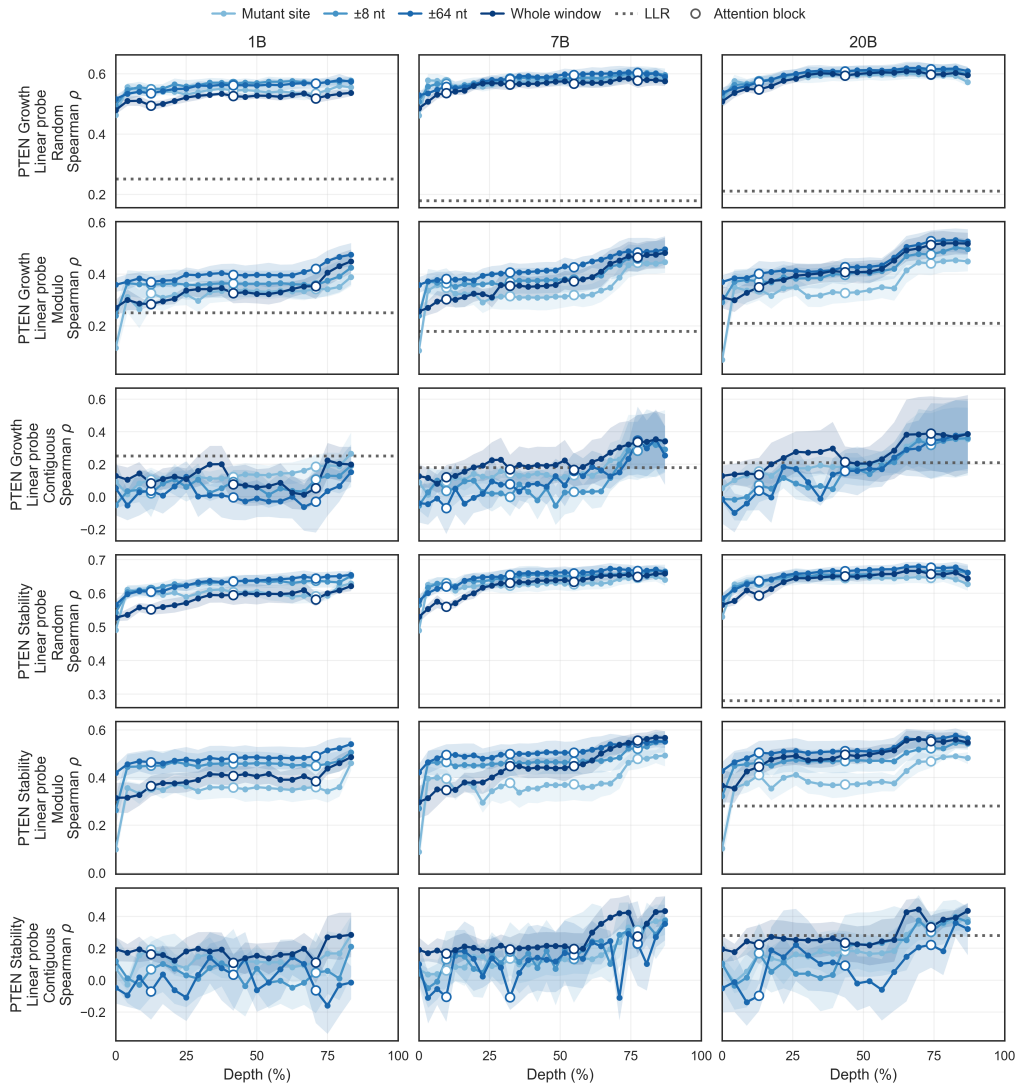


Figure 10: Evo2 layer and pooling sweep.

A.3.3 Linear Probe Fit (Ridge Regression)

We use `RidgeCV` from `scikit-learn` to implement the ridge regression. Fitting a ridge regression estimates $\widehat{w}$, the standardized direction describing the phenotype of interest, and $\widehat{b}$, the intercept. An example of ridge regression fits on different folds is shown in Appendix A.3.6.

$$(\widehat{w}, \widehat{b}) = \arg\min_{w,b} \sum_{j \in \mathcal{I}_{\text{tr}}} \left(y_j - w^\top \widetilde{z}_j - b\right)^2 + \alpha \|w\|_2^2. \tag{10}$$

The penalty α is selected over a sweep over a 19-value log-spaced grid required for the higher dimensionality of larger models.

$$\alpha \in \{10^{-2}, 10^{-1.5}, \ldots, 10^{7}\} \tag{11}$$

On the evaluation fold, we take the standardized $\widehat{w}$ and obtain the "raw" direction in the original activation coordinates and estimate the new $\widehat{y}(x)$.

$$w = \widehat{w} \oslash \sigma_{\text{fold}}, \qquad \widehat{y}(x) = z(x)^\top w + b' \tag{12}$$

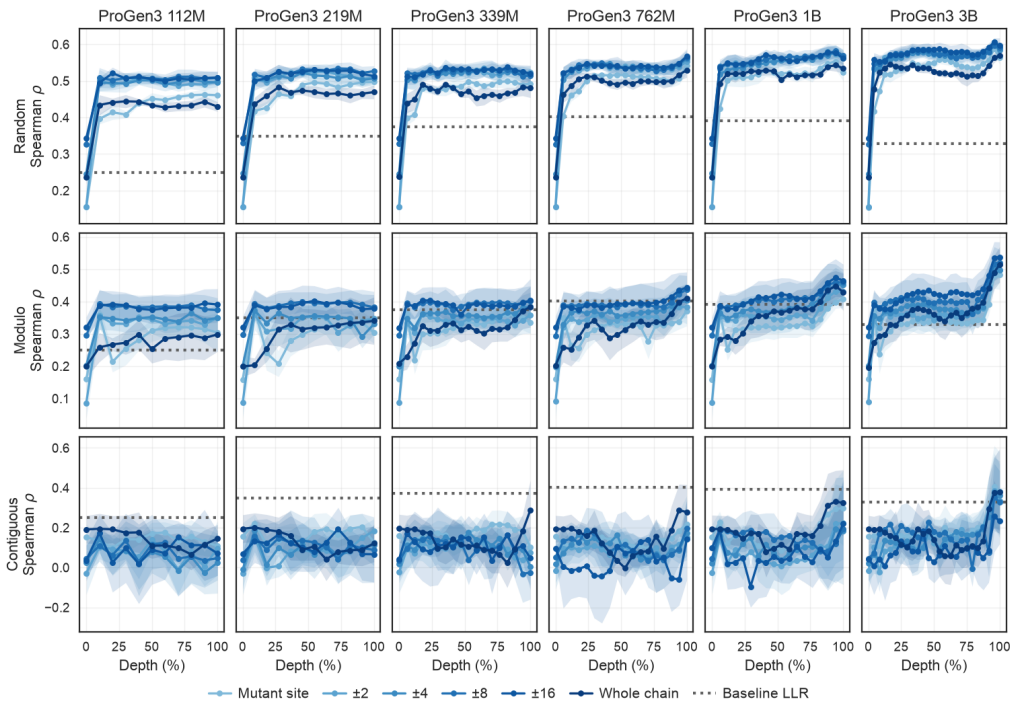


Figure 11: ProGen3 layer and pooling sweep.

A.3.4 Nonlinear Probe Fit (Multilayer Perceptron)

The Multilayer Perceptron (MLP) maps standardized features through

$$\widehat{y}_j^{\mathrm{MLP}} = W_3 \,\mathrm{Drop}_{0.1}[\mathrm{GELU}(W_2 \,\mathrm{Drop}_{0.1}[\mathrm{GELU}(W_1 \widetilde{z}_j + b_1)] + b_2)] + b_3. \tag{13}$$

with widths $D \to 512 \to 128 \to 1$. We standardize labels on the training fold for MLP optimization and map predictions back to the assay scale before reporting errors or correlations. The model is trained with AdamW, weight decay, mean-squared-error loss, batch size of 256, and at most 200 epochs. Early stopping uses a random 10% split of the outer training fold. For simplicity, the architecture and optimization settings are fixed for each assay, model size, layer, pooling radius, and split scheme.

A.3.5 Feature normalization details

The following normalizations were applied to the latent representation h_i:

$$\begin{aligned}
\text{none} &: \tilde{h}_i = h_i \\
\text{fold } z\text{-score} &: \tilde{h}_i = (h_i - \mu_{\text{fold}}) \oslash \sigma_{\text{fold}} \\
\text{WT-centred} &: \tilde{h}_i = h_i - \mu_{\text{WT}} \\
\text{WT } z\text{-score} &: \tilde{h}_i = (h_i - \mu_{\text{WT}}) \oslash \sigma_{\text{WT}} \\
\text{WT-position-centred} &: \tilde{h}_i = h_i - H_{\text{pos}(i)} \\
\text{WT-position } z\text{-score} &: \tilde{h}_i = (h_i - H_{\text{pos}(i)}) \oslash \sigma_{\text{WT}} \\
\text{lowest } z\text{-score} &: \tilde{h}_i = (h_i - \mu_{\text{lowest}}) \oslash \sigma_{\text{lowest}}
\end{aligned} \tag{14}$$

Here, $\oslash$ denotes elementwise division. The WT-position z-score uses the whole-sequence σ_{WT}, rather than a position-specific spread. Because ridge regression includes an intercept, a constant

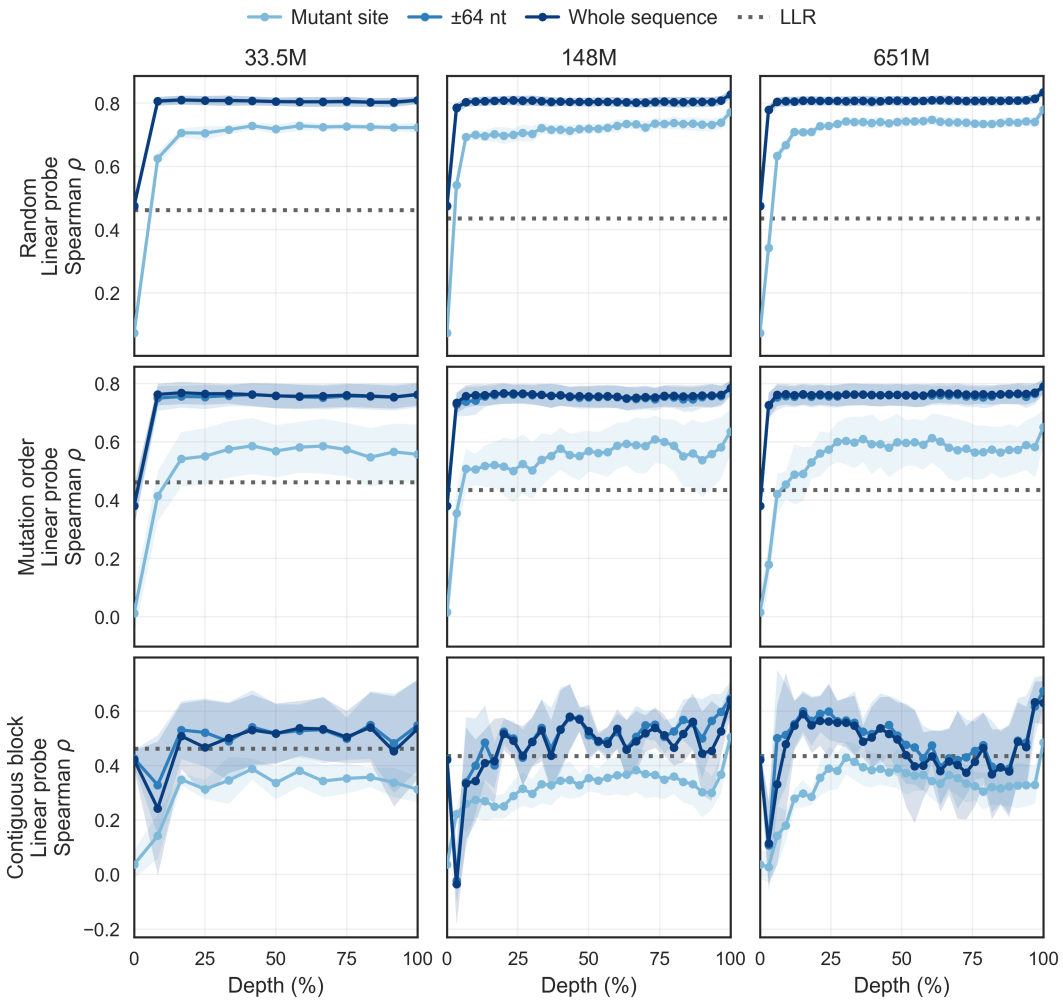


Figure 12: RiNALMo ridge regression layer and pooling sweep.

shift (WT-centred or lowest-centred) does not change the fitted predictions; only rescaling or a row-dependent shift, such as $H_{\mathrm{pos}(i)}$ alters the probe performance.

Symbol	Meaning
μ_{fold}	Per-feature mean over all of the sequences in the training fold
σ_{fold}	Per-feature standard deviation over all of the sequences in the training fold
$H_{\mathrm{pos}(i)}$	The reference's state at variant i's own mutated residue
μ_{WT}	Per-feature mean of H over all L residues
σ_{WT}	Per-feature standard deviation of H over all L residues
μ_{lowest}	Mean the sequence with the known lowest measurement
σ_{lowest}	standard deviation of the sequence with the known lowest measurement
$\Delta\rho$	Change in Spearman correlation from latent without further normalization

Table 4: Change in Spearman correlation ($\Delta\rho$) of PTEN growth probes under each feature-normalization strategy, against probes fit on latents with no feature normalization. Mean $\pm$ s.d. over the six ESM2 sizes (8M–15B); ρ is the mean over the five folds of each ProteinGym scheme. Bold marks the largest $\Delta\rho$ in each column. Ridge's α is selected on scheme-matched inner folds.

Normalization	**PG random**	**PG modulo**	**PG contiguous**
Ridge			
Fold center	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
Fold scale	$\mathbf{+0.0275 \pm 0.0090}$	$\mathbf{+0.0372 \pm 0.0117}$	$\mathbf{+0.0225 \pm 0.0233}$
Fold z-score	$\mathbf{+0.0275 \pm 0.0090}$	$\mathbf{+0.0372 \pm 0.0117}$	$\mathbf{+0.0225 \pm 0.0233}$
WT center	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
WT scale	$\mathbf{+0.0275 \pm 0.0090}$	$+0.0371 \pm 0.0117$	$+0.0186 \pm 0.0272$
WT z-score	$\mathbf{+0.0275 \pm 0.0090}$	$+0.0371 \pm 0.0117$	$+0.0186 \pm 0.0272$
WT position center	-0.0902 ± 0.0347	-0.0583 ± 0.0530	-0.0165 ± 0.0367
WT position z-score	-0.0337 ± 0.0401	-0.0261 ± 0.0633	$+0.0132 \pm 0.0445$
Minimum center	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
Minimum z-score	$\mathbf{+0.0275 \pm 0.0089}$	$+0.0370 \pm 0.0117$	$+0.0185 \pm 0.0272$
MLP			
Fold center	-0.0039 ± 0.0044	$+0.0001 \pm 0.0031$	$\mathbf{+0.0085 \pm 0.0167}$
Fold scale	-0.0041 ± 0.0072	$+0.0064 \pm 0.0125$	-0.0011 ± 0.0165
Fold z-score	-0.0050 ± 0.0038	$+0.0028 \pm 0.0126$	-0.0077 ± 0.0189
WT center	-0.0017 ± 0.0022	-0.0043 ± 0.0043	-0.0055 ± 0.0139
WT scale	-0.0040 ± 0.0096	$\mathbf{+0.0078 \pm 0.0114}$	-0.0052 ± 0.0129
WT z-score	-0.0055 ± 0.0040	$+0.0043 \pm 0.0106$	-0.0093 ± 0.0237
WT position center	-0.0410 ± 0.0257	-0.0661 ± 0.0412	-0.0076 ± 0.0244
WT position z-score	-0.0401 ± 0.0190	-0.0487 ± 0.0221	-0.0035 ± 0.0209
Minimum center	$\mathbf{-0.0013 \pm 0.0027}$	-0.0042 ± 0.0043	-0.0090 ± 0.0068
Minimum z-score	-0.0052 ± 0.0035	$+0.0038 \pm 0.0112$	-0.0085 ± 0.0232

Several entries are algebraic identities rather than measurements, because ridge fits an intercept and therefore cannot see a constant per-feature shift: the three centring arms reproduce the unnormalized fit exactly (0.0000), fold scale equals fold z-score, and WT scale equals WT z-score. Ties are bolded together.

Table 5: Change in Spearman correlation ($\Delta\rho$) of ADK probes under each feature-normalization strategy, against probes fit on latents with no feature normalization, for the four measured phenotypes. Mean $\pm$ s.d. over the six ESM2 sizes (8M–15B); within each size ρ is the mean over the 30 stratified splits of Pinney et al. at $n_{\text{train}} = 140$. Bold marks the largest $\Delta\rho$ in each column.

Normalization	k_{cat}	k_{cat}/K_M	K_M	T_{growth}
Ridge				
Fold center	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
Fold scale	$+0.0002 \pm 0.0096$	-0.0008 ± 0.0083	$+0.0027 \pm 0.0100$	$\mathbf{+0.0020 \pm 0.0032}$
Fold z-score	$+0.0002 \pm 0.0096$	-0.0008 ± 0.0083	$+0.0027 \pm 0.0100$	$\mathbf{+0.0020 \pm 0.0032}$
WT center	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
WT scale	$\mathbf{+0.0020 \pm 0.0088}$	$+0.0011 \pm 0.0074$	$\mathbf{+0.0077 \pm 0.0070}$	$+0.0006 \pm 0.0030$
WT z-score	$\mathbf{+0.0020 \pm 0.0088}$	$+0.0011 \pm 0.0074$	$\mathbf{+0.0077 \pm 0.0070}$	$+0.0006 \pm 0.0030$
WT position center	–	–	–	–
WT position z-score	–	–	–	–
Minimum center	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
Minimum z-score	-0.0002 ± 0.0072	$\mathbf{+0.0031 \pm 0.0063}$	$+0.0059 \pm 0.0080$	$+0.0004 \pm 0.0031$
MLP				
Fold center	-0.0236 ± 0.0267	-0.0009 ± 0.0339	-0.0097 ± 0.0426	-0.0107 ± 0.0237
Fold scale	$\mathbf{+0.0070 \pm 0.0100}$	$+0.0156 \pm 0.0227$	$\mathbf{+0.0255 \pm 0.0485}$	-0.0090 ± 0.0243
Fold z-score	-0.0068 ± 0.0252	$\mathbf{+0.0249 \pm 0.0371}$	$+0.0221 \pm 0.0407$	$\mathbf{+0.0074 \pm 0.0232}$
WT center	-0.0228 ± 0.0158	-0.0029 ± 0.0234	-0.0116 ± 0.0344	-0.0110 ± 0.0218
WT scale	-0.0205 ± 0.0204	$+0.0098 \pm 0.0201$	$+0.0047 \pm 0.0204$	-0.0046 ± 0.0165
WT z-score	-0.0367 ± 0.0265	$+0.0041 \pm 0.0301$	$+0.0018 \pm 0.0263$	-0.0103 ± 0.0250
WT position center	–	–	–	–
WT position z-score	–	–	–	–
Minimum center	-0.0070 ± 0.0173	-0.0006 ± 0.0242	$+0.0033 \pm 0.0391$	-0.0181 ± 0.0195
Minimum z-score	-0.0258 ± 0.0352	$+0.0202 \pm 0.0331$	$+0.0050 \pm 0.0327$	-0.0234 ± 0.0225

ADK is evaluated on the source paper's own 30 splits rather than on ProteinGym fold schemes, so it has phenotype columns instead of scheme columns and its α is selected by leave-one-out GCV in every arm. The two position-matched arms are undefined here: the rows are 175 natural orthologs rather than point mutants of a reference, so a mutated position does not exist.

A.3.6 PROBE FITS

Below are examples of the linear probe fits across ProteinGym's cross-validation schemes.

A.4 LLR LATENT DIRECTION

Below are the fits of computed zero-shot likelihood fitness assessments within the last hidden state of ESM2-3B.

A.5 SPEARMAN CORRELATIONS ACROSS ASSAYS BY MODEL FAMILY

A.6 PROGEN3 AND EVO2 MULTIPLE MUTATIONAL DEPTH BENCHMARK

A.7 FITNESS LANDSCAPE DETAILS

We use SciPy's 1D `Akima1DInterpolator` independently along each axis of the 2D projection from the UMAP. Each reconstructed surface is converted to within-surface percentile ranks for visualization. Percentile rank determines both surface height and color, allowing the topography of the three surfaces to be compared directly without conflating variations in the phenotype ranges.

A.8 EXTRAPOLATION ADDITIONAL ASSESSMENTS

We include extrapolation assessments across different fitted ranges for PTEN growth, and across different assays. The more labels included during training, the better the probe performance. The probes extrapolate better to certain phenotypes, such as growth, stability, expression, and $\Delta\Delta G$, over others more tied to protein activity, such as antibody binding.

Table 6: Spearman ρ across n assays, by model family and checkpoint, for the linear conditioned probe and the model's own zero-shot likelihood.

			Linear Probe ρ		Zero-shot ρ	
Family	n	Model	Median	Mean	Median	Mean
ESM2	303	8M	0.499	0.526	0.186	0.189
		35M	0.556	0.567	0.308	0.293
		150M	0.605	0.606	0.395	0.352
		650M	0.625	0.624	0.417	0.381
		3B	0.632	0.625	0.417	0.373
ProGen3	321	112M	0.483	0.514	0.226	0.213
		219M	0.511	0.537	0.283	0.257
		339M	0.532	0.554	0.312	0.279
		762M	0.554	0.575	0.328	0.295
		1B	0.577	0.594	0.350	0.318
		3B	0.602	0.618	0.357	0.330
Evo2	207	1B	0.478	0.495	0.279	0.261
		7B	0.513	0.523	0.292	0.271
		20B	0.547	0.552	0.316	0.288
RiNALMo	14	33.5M	0.324	0.329	−0.159	−0.121
		148M	0.344	0.331	−0.102	−0.067
		651M	0.338	0.339	−0.168	−0.094

Table 7: Spearman ρ across assays, by model family and checkpoint, for the nonlinear conditioned probe and the model's own zero-shot likelihood. Rows are ordered smallest to largest within a family.

			Nonlinear Probe ρ		Zero-shot ρ	
Family	n	Model	Median	Mean	Median	Mean
ESM2	316	8M	0.526	0.552	0.188	0.189
		35M	0.577	0.591	0.309	0.294
		150M	0.627	0.623	0.402	0.355
		650M	0.649	0.639	0.425	0.386
		3B	0.651	0.639	0.424	0.378
ProGen3	319	112M	0.527	0.557	0.226	0.214
		219M	0.547	0.571	0.283	0.258
		339M	0.557	0.584	0.313	0.280
		762M	0.574	0.596	0.328	0.297
		1B	0.593	0.609	0.351	0.321
		3B	0.613	0.625	0.359	0.332
Evo2	205	1B	0.476	0.491	0.279	0.262
		7B	0.492	0.505	0.292	0.272
		20B	0.500	0.519	0.320	0.289
RiNALMo	14	33.5M	0.338	0.350	−0.159	−0.121
		148M	0.357	0.372	−0.102	−0.067
		651M	0.383	0.410	−0.168	−0.094

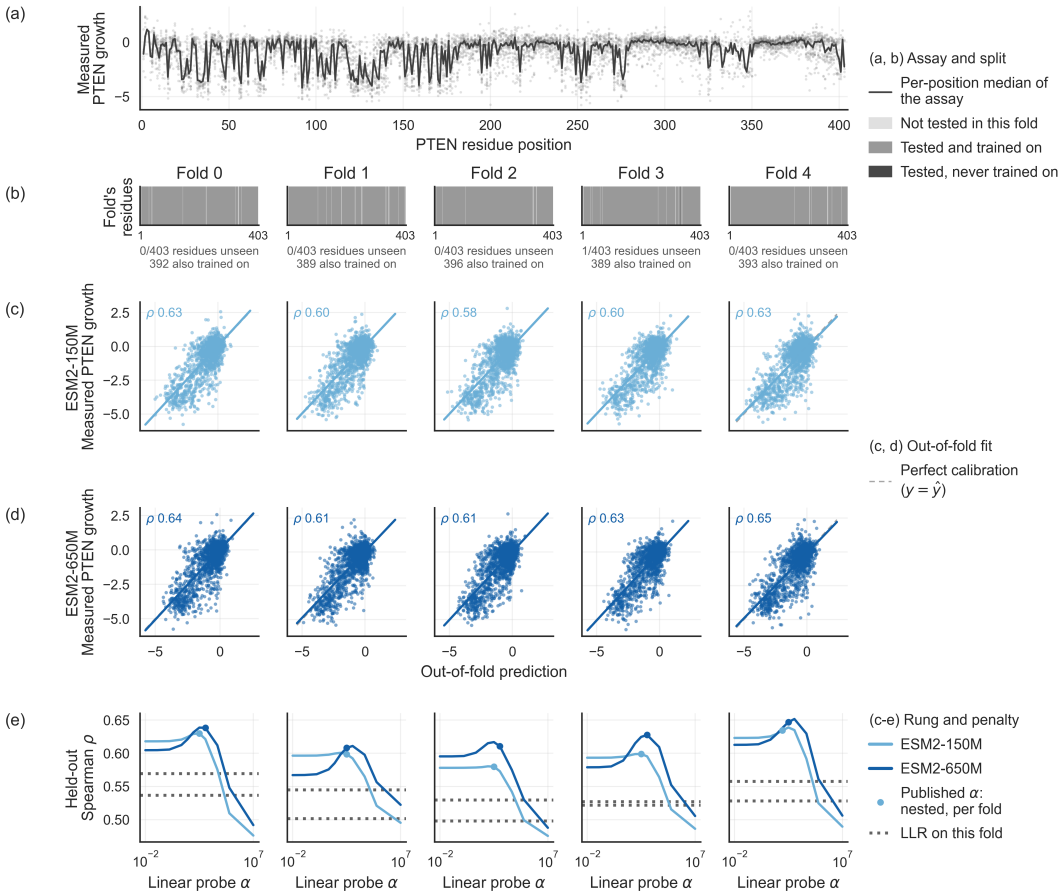


Figure 13: PTEN growth random fold linear probe fits. (a) Measured growth values and site of mutation along PTEN's sequence. (b) 5 folds within ProteinGym and which positions are included in probe training. (c, d) Out-of-fold predictions with ESM2-150 and ESM2-650M models and the Spearman correlation for each fold. (e) Regularization term α sweep per fold.

A.9 Measurement efficiency analysis detailed methods

We assessed measurement efficiency using repeated held-out cross-validation schemes. For each checkpoint, we generated 30 train–test splits (5 folds × 6 shuffles), with 80% of observations assigned to the training pool and 20% held out for evaluation. Within each split, training observations were ranked by measured fitness, and the outermost 5% at each end of the fitness distribution were excluded. From the remaining observations, we selected a measurement set of size n comprising the $n/2$ lowest- and $n/2$ highest-fitness observations. For odd n, the additional observation was taken from the high-fitness end.

A ridge regression model was fit using these n selected observations and evaluated by Spearman correlation ρ on the split's held-out 20% test fold. The test fold was strictly disjoint from the training pool and was not used for ranking, trimming, or measurement-set selection.

Performance at each measurement-set size n was averaged across the 30 splits. We defined n_{LLR} as the smallest tested n for which the mean held-out Spearman correlation equaled or exceeded the corresponding ρ_{LLR}, evaluated for the same checkpoint on the same held-out test observations. Finally, measurement efficiency for each assay was summarized as the median n_{LLR} across its (size, scheme) conditions.

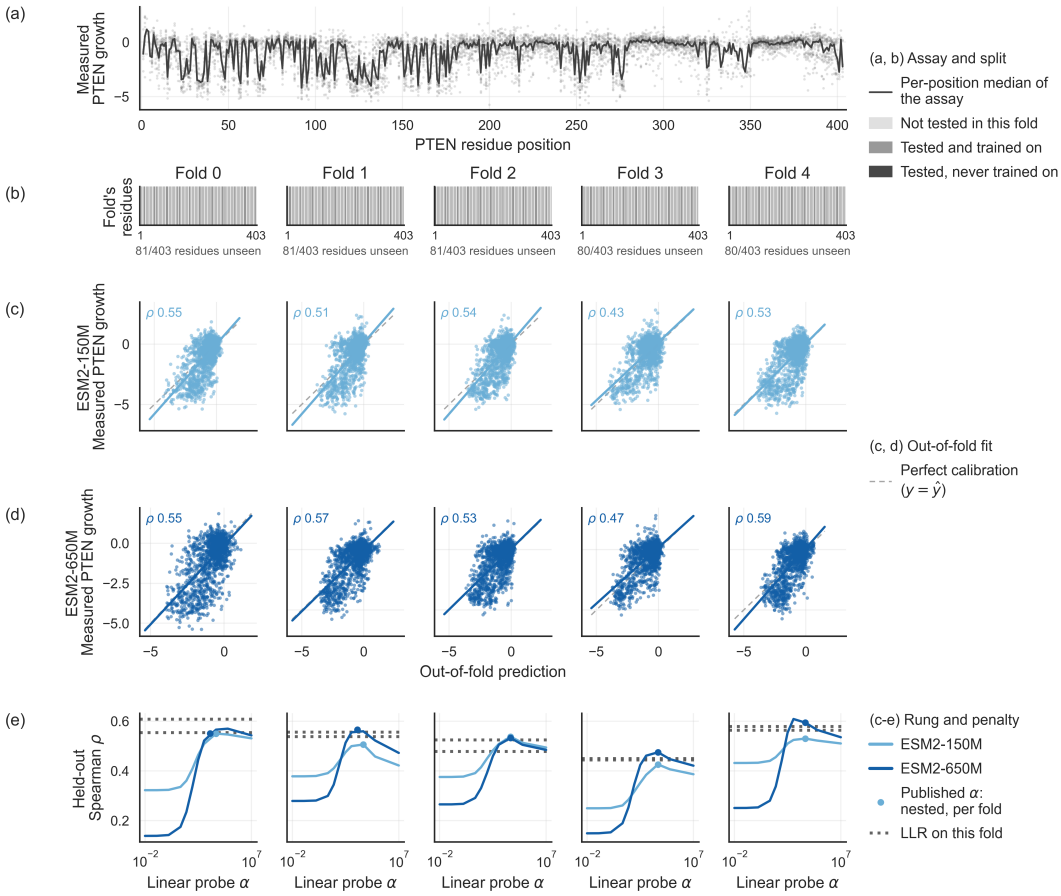


Figure 14: PTEN growth modulo folds linear probe fits and α value selection.

A.10 Protein domain similarity for the pooled-assay arms

Sequence identity is quantified by MMSeq2 percent identity (`easy-search -s 7.5 -e 1e5 --exhaustive-search 1`), symmetrised by the maximum over the two directed rows. TM-score was assessed with TMalign $\min(\mathrm{qtm}, \mathrm{ttm})$ from Foldseek `easy-search --alignment-type 1 --exhaustive-search 1`.

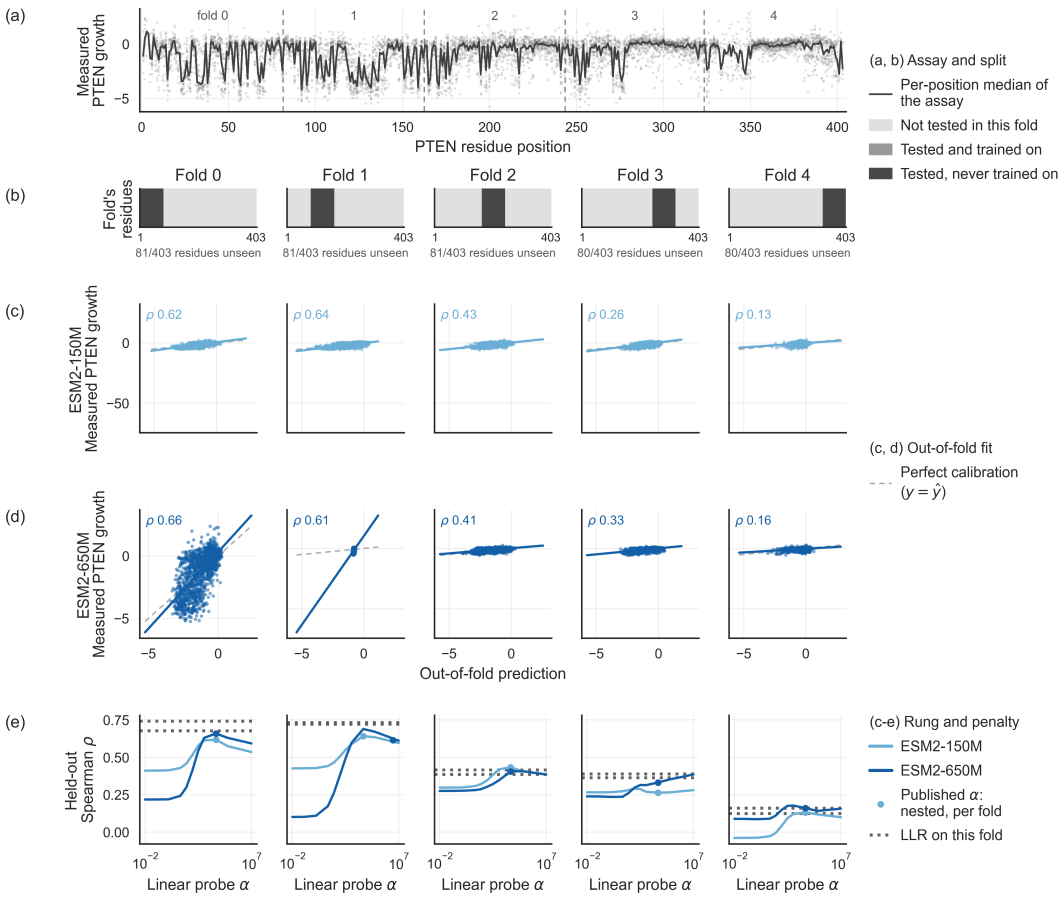


Figure 15: PTEN growth contiguous folds linear probe fits and α value selection.

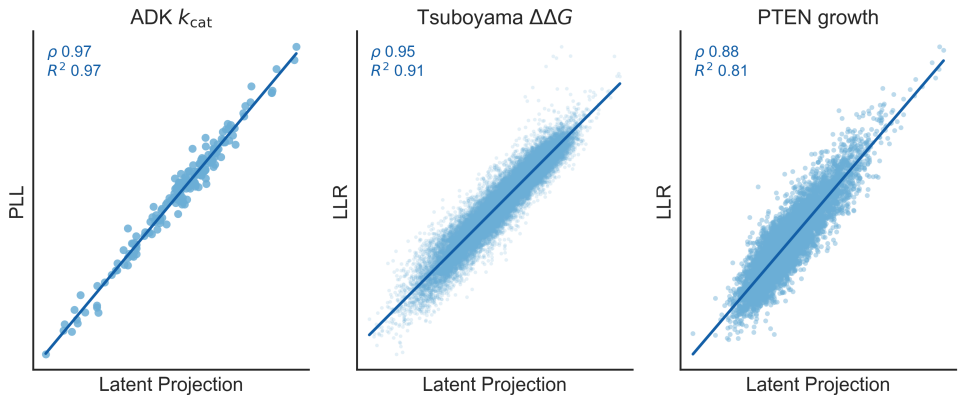


Figure 16: Scatter plots of zero-shot likelihood assessments of ADK, varied protein domains, and PTEN and the last hidden state of ESM2-3B.

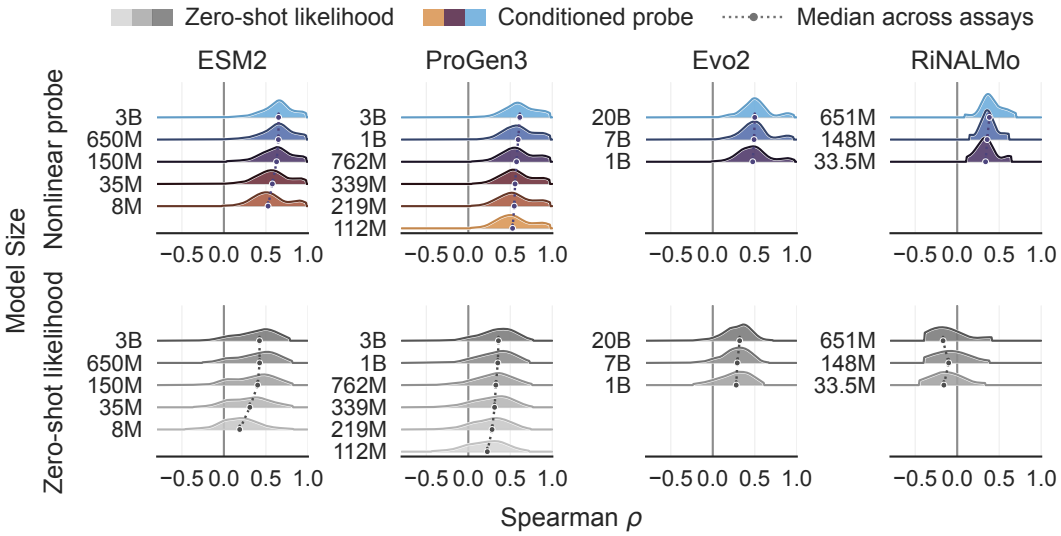


Figure 17: Distribution of per-assay Spearman ρ across each model family, for the nonlinear probe (top row) and the model's zero-shot likelihood (bottom row). Each band is a kernel density estimate over assays at one checkpoint. Medians and means are in Table 7.

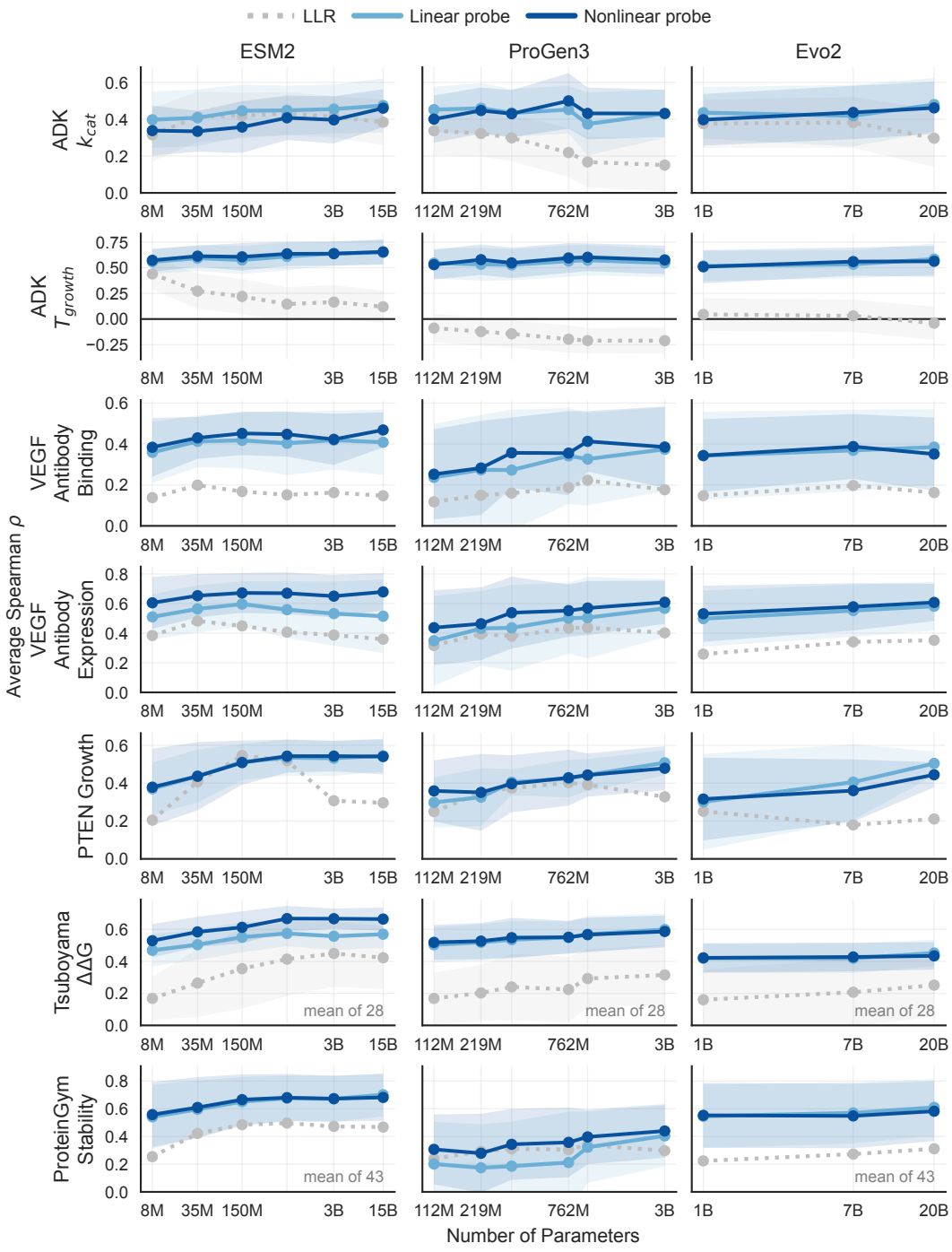


Figure 18: Average Spearman correlation across different model sizes for varied assays.

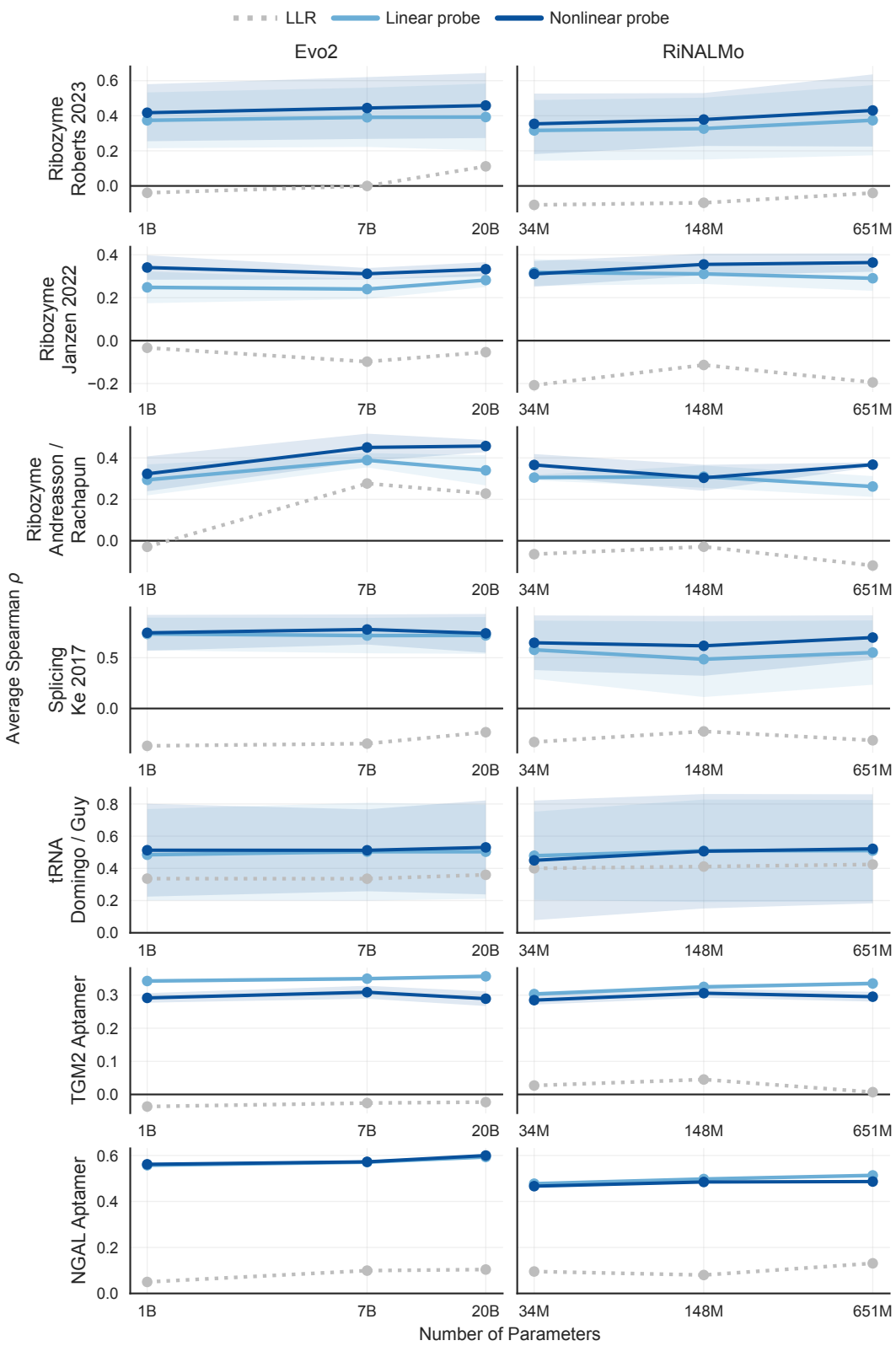


Figure 19: Average Spearman correlation across RiNALMo and Evo2 for assays from NABench(Li et al., 2025).

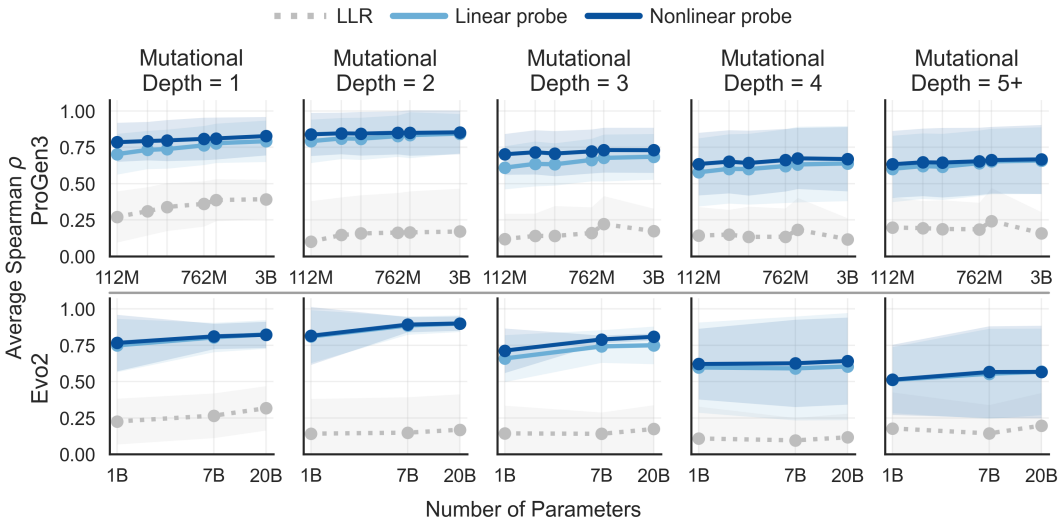


Figure 20: Average Spearman correlations of latent-based and LLR fitness estimates across 5 different mutational depths assessed with ProGen3 and Evo2.

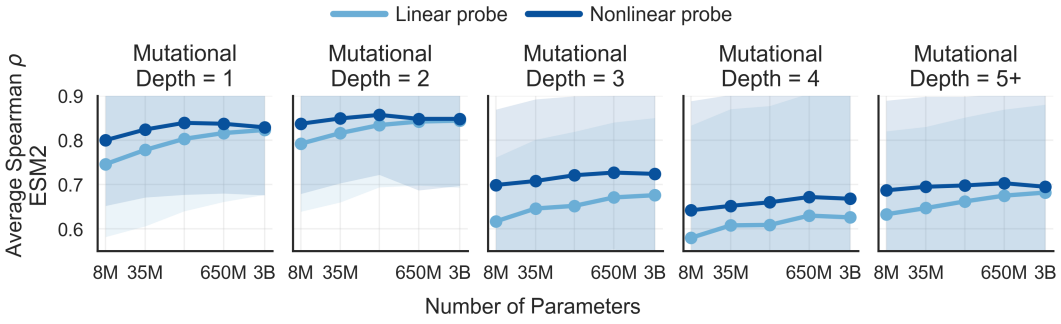


Figure 21: ESM2 multiple mutational depth assessments with y-axis rescaled to examine effects from scaling model size.

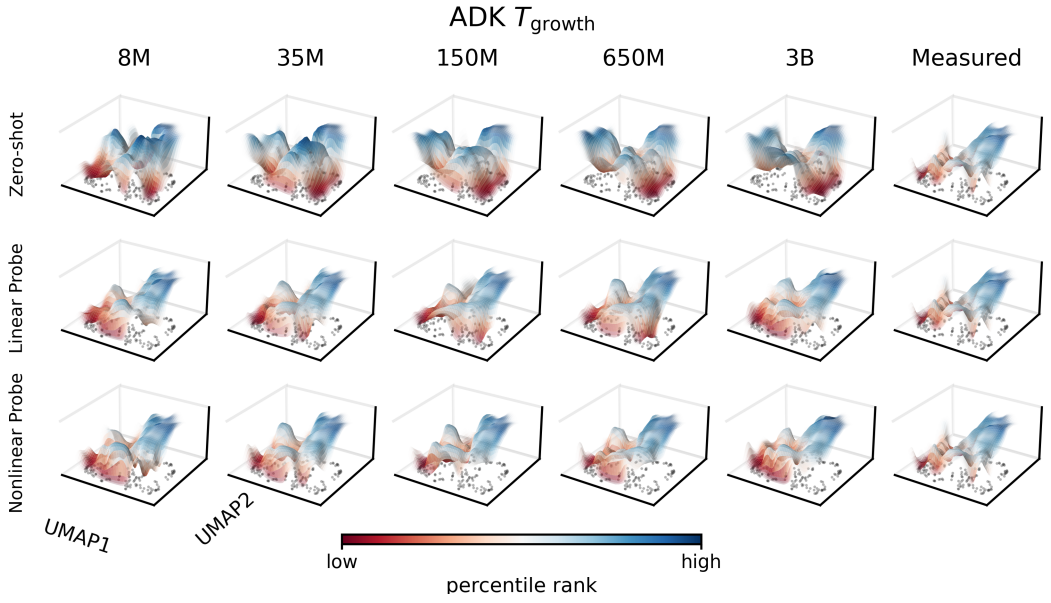


Figure 22: 3D T_{growth} reconstructions of ADK across different methods and ESM2 sizes.

Table 8: **Roughness ratio across measured fitness landscapes and model surfaces, by model size.** Landscape geometry at every ESM2 model size, for every measured ADK phenotype from Muir et al.'s dataset. Roughness ratio is the mean absolute difference between neighbouring variants with both surfaces standardised to unit variance, so the measured landscape is 1.000 by construction and a value above it is a surface rougher than the phenotype. Rows are the three surfaces read off the same encoder: the model's own zero-shot likelihood, the conditioned linear probe and the conditioned nonlinear probe, the last two out of fold under each arm's held-out split. The measured landscape is the reference and is 1.000 in every cell of both metrics by construction, so it is not given a row. Bold marks the surface whose value is closest to 1.000 at that rung, i.e. the one that reproduces how rugged the measured landscape is. Neighbours are defined on an explicit sequence-space graph and never on an embedding: same position, different amino acid for the saturation scans, and k-NN ($k = 10$) over global-alignment identity for the ADK ortholog panels.

Landscape	Surface	8M	35M	150M	650M	3B	15B
ADK T_{growth}	Zero-shot	0.914	**1.001**	1.024	1.098	1.113	—
	Linear Probe	0.914	0.893	0.944	0.963	**0.979**	—
	Nonlinear Probe	**0.958**	0.974	**0.983**	**1.008**	0.955	—
ADK $\log_{10} k_{\mathrm{cat}}$	Zero-shot	0.849	0.931	0.927	0.962	**1.002**	—
	Zero-shot, designed tier[§]	0.820	0.903	0.870	0.877	0.872	—
	Linear Probe	**0.990**	0.910	0.937	0.941	0.830	—
	Nonlinear Probe	0.981	**0.936**	**0.973**	**0.964**	0.934	—
ADK $\log_{10} K_{\mathrm{M}}$	Zero-shot	0.746	0.817	0.813	0.844	0.880	—
	Zero-shot, designed tier[§]	0.686	0.756	0.728	0.734	0.730	—
	Linear Probe	**0.909**	0.763	0.806	0.781	0.782	—
	Nonlinear Probe	0.790	**0.827**	**0.827**	**0.852**	**0.939**	—
ADK $\log_{10}(k_{\mathrm{cat}}/K_{\mathrm{M}})$	Zero-shot	0.945	1.036	1.031	1.070	1.116	—
	Zero-shot, designed tier[§]	0.826	0.910	0.876	0.883	0.878	—
	Linear Probe	1.050	**1.015**	**0.986**	1.056	**0.980**	—
	Nonlinear Probe	**0.982**	0.941	1.054	**1.042**	0.973	—
ADK urea tolerance	Zero-shot	0.877	0.960	0.982	1.051	1.065	—
	Linear Probe	0.817	0.748	0.891	0.881	0.856	—
	Nonlinear Probe	**1.038**	**1.039**	**1.004**	**0.995**	**0.998**	—

[§] Scored on the designed tier only. An ortholog is scored by whole-sequence pseudo-log-likelihood and a design by a point-mutant log-likelihood ratio, two scales that are never pooled, so this row covers 463 of the arm's 2,324 edges. The measured landscape is not equally rugged there, so the row is reported and held out of the bold comparison.

Table 9: **Gradient similarity across measured fitness landscapes and model surfaces, by model size.** Landscape geometry at every ESM2 rung, for every measured phenotype in the reconstruction arm. Gradient similarity is the Spearman correlation between a surface's per-edge gradients and the measured landscape's, so it asks whether a surface is steep in the same places rather than by the same amount; the measured landscape is 1.000 against itself. Rows are the three surfaces read off the same encoder: the model's own zero-shot likelihood, the conditioned linear probe and the conditioned nonlinear probe, the last two out of fold under each arm's held-out split. The measured landscape is the reference and is 1.000 in every cell of both metrics by construction, so it is not given a row. Bold marks the highest value at that rung, i.e. the surface that is steep in the same places as the measured landscape. Neighbours are defined on an explicit sequence-space graph and never on an embedding: same position, different amino acid for the saturation scans, and k-NN ($k = 10$) over global-alignment identity for the ADK ortholog panels.

Landscape	Surface	8M	35M	150M	650M	3B	15B
ADK T_{growth}	Zero-shot	0.367	0.269	0.267	0.272	0.259	—
	Linear Probe	0.538	0.564	0.554	**0.620**	0.615	—
	Nonlinear Probe	**0.565**	**0.588**	**0.595**	0.609	**0.651**	—
ADK $\log_{10} k_{\mathrm{cat}}$ ‡	Zero-shot	0.267	0.313	0.305	0.330	0.270	—
	Zero-shot, designed tier§	0.194	0.271	0.320	0.281	0.242	—
	Linear Probe	0.362	0.398	0.366	0.420	0.431	—
	Nonlinear Probe	**0.370**	**0.403**	**0.415**	**0.432**	**0.453**	—
ADK $\log_{10} K_{\mathrm{M}}$ ‡	Zero-shot	0.224	**0.274**	**0.232**	**0.272**	0.259	—
	Zero-shot, designed tier§	0.275	0.314	0.427	0.384	0.403	—
	Linear Probe	**0.256**	0.202	0.206	0.221	**0.264**	—
	Nonlinear Probe	0.194	0.182	0.163	0.184	0.174	—
ADK $\log_{10}(k_{\mathrm{cat}}/K_{\mathrm{M}})$ ‡	Zero-shot	0.203	0.224	0.217	0.235	0.184	—
	Zero-shot, designed tier§	0.164	0.192	0.259	0.221	0.235	—
	Linear Probe	0.239	0.326	**0.339**	**0.347**	**0.432**	—
	Nonlinear Probe	**0.393**	**0.382**	0.299	0.306	0.281	—
ADK urea tolerance	Zero-shot	0.072	0.101	0.154	0.184	0.151	—
	Linear Probe	0.075	0.136	0.182	0.218	0.221	—
	Nonlinear Probe	**0.257**	**0.186**	**0.280**	**0.318**	**0.347**	—

§ Scored on the designed tier only.

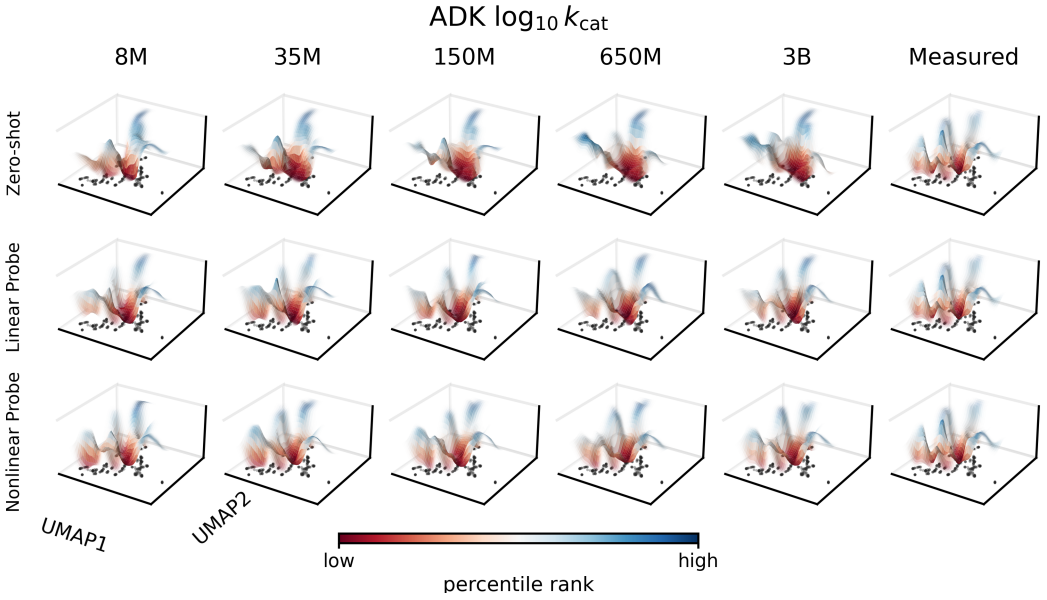


Figure 23: 3D k_{cat} reconstructions of ADK across different methods and ESM2 sizes.

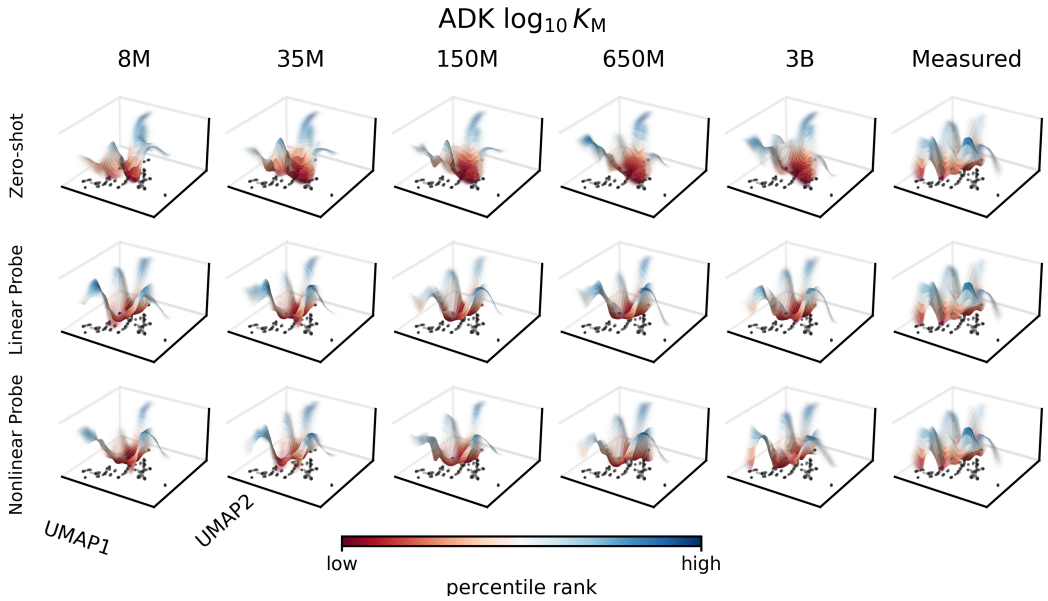


Figure 24: 3D k_M reconstructions of ADK across different methods and ESM2 sizes.

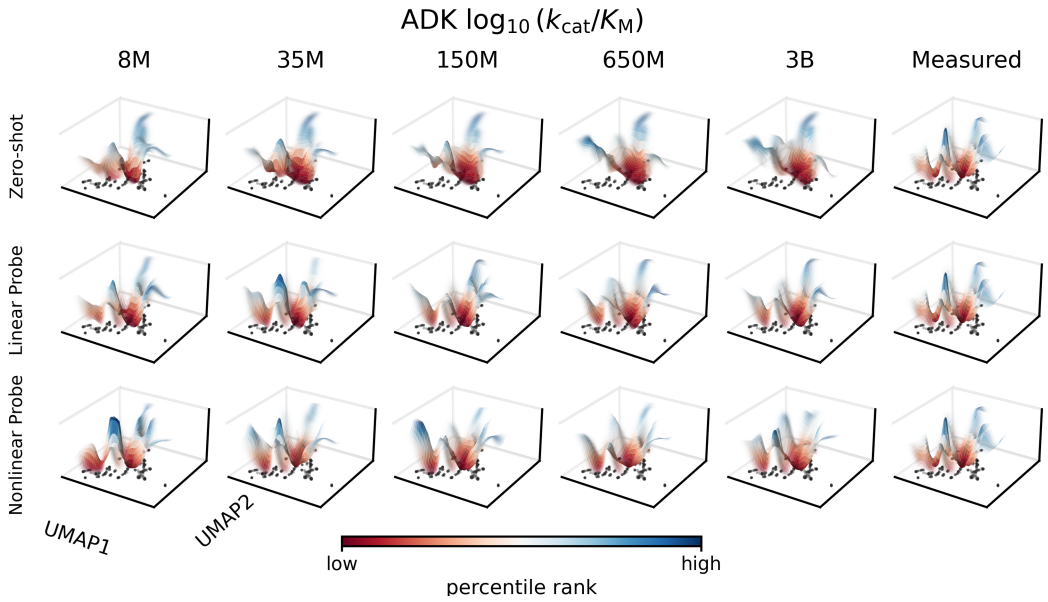


Figure 25: 3D $\frac{k_{cat}}{k_M}$ reconstructions of ADK across different methods and ESM2 sizes.

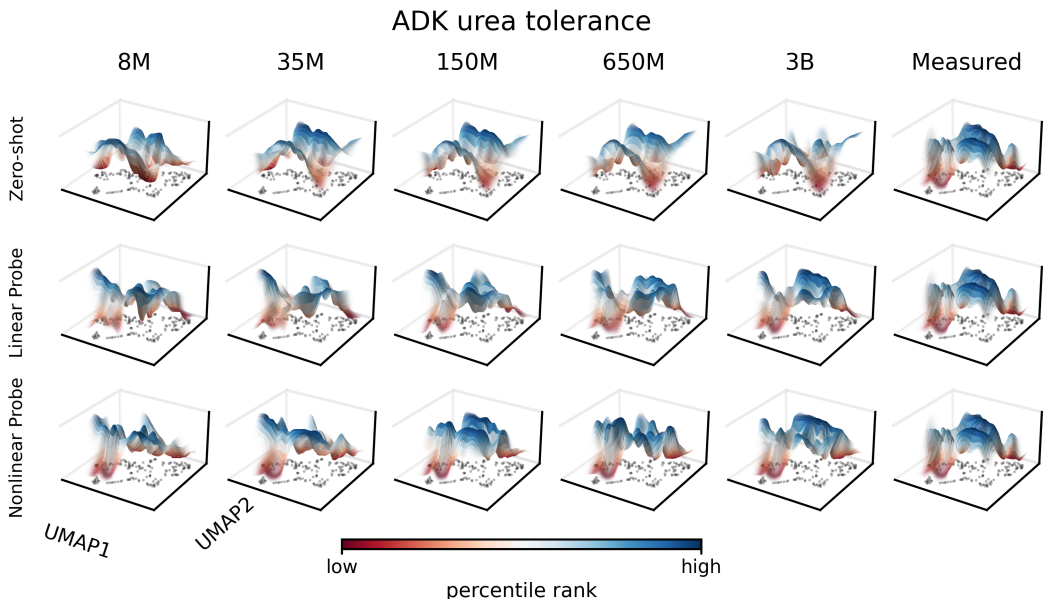


Figure 26: 3D urea tolerance reconstructions of ADK across different methods and ESM2 sizes.

Table 10: A dash in an identity column means `MMSeq2` reports no alignment for that pair.

Sequence homolog				*Similar structure, low sequence identity*			
Target	**Partner**	**Id. (%)**	**TM**	**Target**	**Partner**	**Id. (%)**	**TM**
1PSE	1QP2	69.2	0.861	1PSE	2LQK	–	0.610
1QP2	1PSE	69.2	0.861	1PSE	3DKM	–	0.626
1TUC	2KXD	100.0	0.809	1QP2	2LQK	–	0.668
1W4H	2BTH	97.7	0.999	1QP2	3DKM	–	0.639
1W4H	2WXC	97.7	0.976	1TUC	2K5H	–	0.549
2BTH	1W4H	97.7	0.999	1W4H	HHH_0142	–	0.535
2BTH	2WXC	95.4	0.975	1W4H	HHH_0244	–	0.516
2K1B	2K28	76.0	0.942	2BTH	HHH_0142	–	0.530
2K28	2K1B	76.0	0.942	2BTH	HHH_0244	–	0.509
2KXD	1TUC	100.0	0.809	2K5H	1TUC	–	0.549
2WXC	1W4H	97.7	0.976	2K5H	2KXD	–	0.504
2WXC	2BTH	95.4	0.975	2KXD	2K5H	–	0.504
6EWS	6EWT	71.4	0.978	2LQK	1PSE	–	0.610
6EWS	6EWU	78.5	0.992	2LQK	1QP2	–	0.668
6EWT	6EWS	71.4	0.978	2LQK	3DKM	–	0.584
6EWT	6EWU	73.2	0.977	2WXC	HHH_0142	–	0.532
6EWU	6EWS	78.5	0.992	2WXC	HHH_0244	–	0.512
6EWU	6EWT	73.2	0.977	3DKM	1PSE	–	0.626
HHH_0142	HHH_0244	52.9	0.913	3DKM	1QP2	–	0.639
HHH_0244	HHH_0142	52.9	0.913	3DKM	2LQK	–	0.584
r11_1081	r18_3	48.9	0.612	HHH_0142	1W4H	–	0.535
r12_757	r18_3	63.8	0.739	HHH_0142	2BTH	–	0.530
r18_3	r11_1081	48.9	0.612	HHH_0142	2WXC	–	0.532
r18_3	r12_757	63.8	0.739	HHH_0244	1W4H	–	0.516
				HHH_0244	2BTH	–	0.509
				HHH_0244	2WXC	–	0.512
				r11_1081	r12_757	–	0.636
				r12_757	r11_1081	–	0.636

Both measures are properties of the sequences and structures alone and do not depend on the model rung. The 28 structural pairs are disjoint from the 24 sequence pairs by construction. Alignment length, E-value, lDDT, `alntmscore` and both chain lengths are in the accompanying CSV.

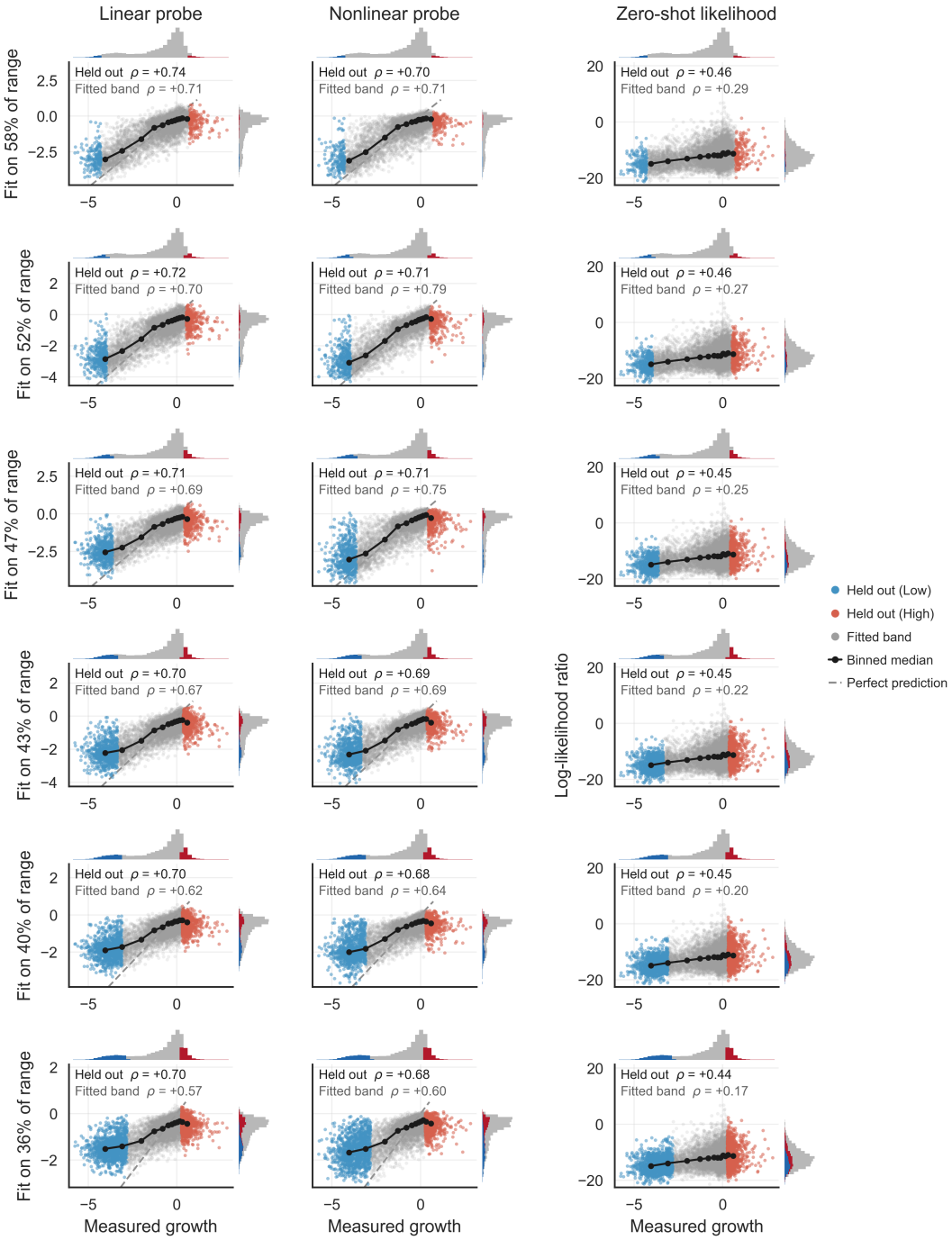


Figure 27: Extrapolation assessment with ESM2-3B, varying fitted label range.

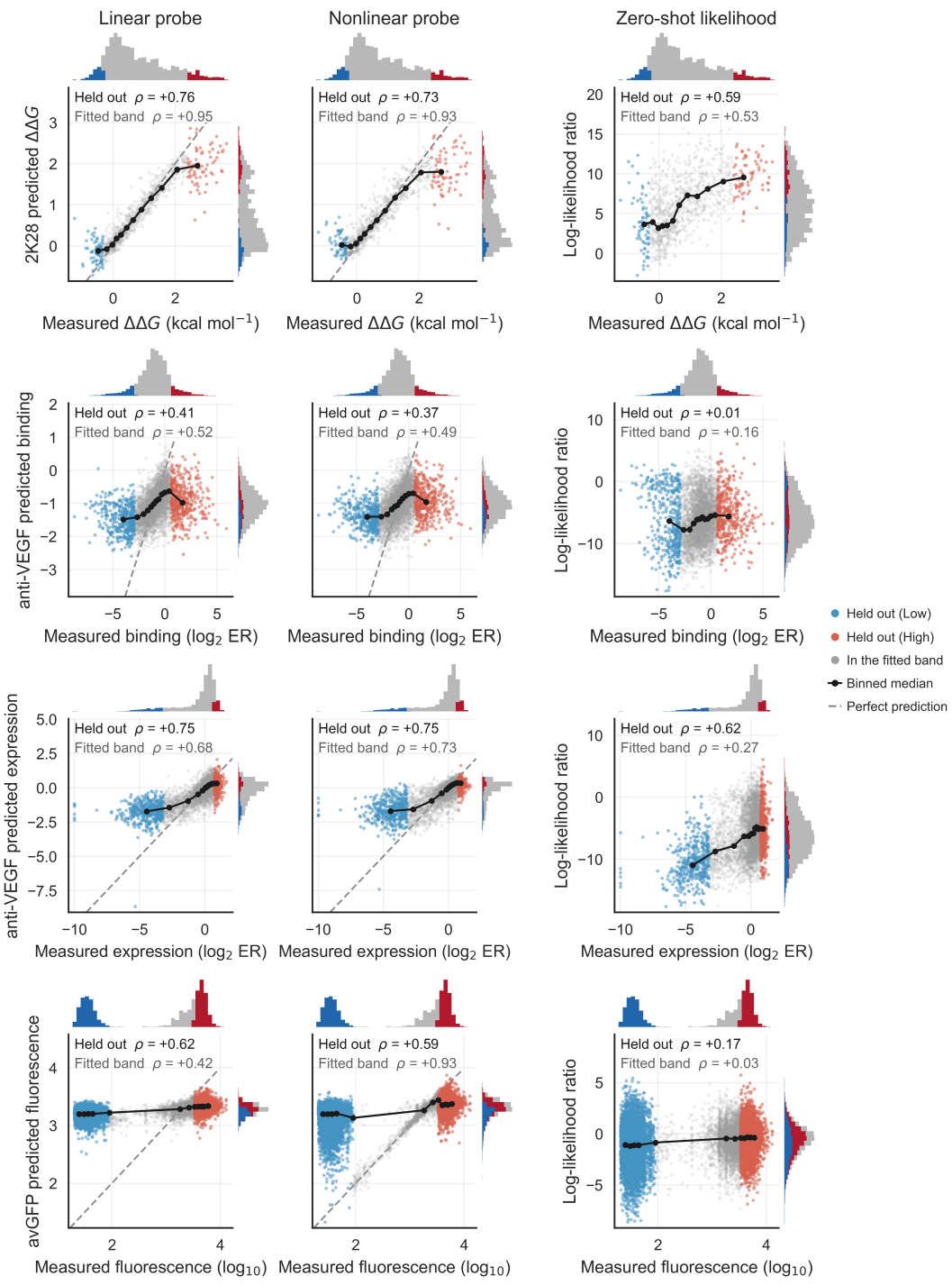


Figure 28: Extrapolation assessment across different assays with ESM2-3B.