FINE-TUNING PROTEIN LANGUAGE MODELS WITH DEEP MUTATIONAL SCANNING IMPROVES VARIANT EFFECT PREDICTION

Aleix Lafita[1,†,*] **Ferran Gonzalez**[2,†,*] **Mahmoud Hossam**[2] **Paul Smyth**[2]

Jacob Deasy[2] **Ari Allyn-Feuer**[2] **Daniel Seaton**[1] **Stephen Young**[2]

[1] GSK, Human Genetics and Genomics
[2] GSK, Artificial Intelligence and Machine Learning
[†] equal contribution
[*]{aleix.x.lafita,ferran.x.gonzalez}@gsk.com

ABSTRACT

Protein Language Models (PLMs) have emerged as performant and scalable tools for predicting the functional impact and clinical significance of protein-coding variants, but they still lag experimental accuracy. Here, we present a novel fine-tuning approach to improve the performance of PLMs with experimental maps of variant effects from Deep Mutational Scanning (DMS) assays using a Normalised Log-odds Ratio (NLR) head. We find consistent improvements in a held-out protein test set, and on independent DMS and clinical variant annotation benchmarks from ProteinGym and ClinVar. These findings demonstrate that DMS is a promising source of sequence diversity and supervised training data for improving the performance of PLMs for variant effect prediction.

1 INTRODUCTION

The pace of discovery of new human genetic variants is rapidly increasing, led by genome sequencing of large human cohorts (Lek et al., 2016; Backman et al., 2021). However, the functional characterisation of these human variants has not scaled at the same pace, limiting their impact for clinical diagnosis and drug target discovery, and ultimately hampering our ability to understand and treat human diseases (Landrum et al., 2018). Missense variants are the most common type of coding variant, causing single amino acid changes in proteins that can have a wide range of consequences, from severe disease-causing protein function disruptions to no significant effect. Predicting the impact of missense mutations on protein function and the downstream clinical consequences remains a crucial challenge (Karczewski et al., 2020).

Over the last few years, advancements in deep learning have led to significant improvements for tackling the missense variant effect prediction challenge (Frazer et al., 2021; Cheng et al., 2023; Gao et al., 2023). Protein Language Models (PLMs) have demonstrated state-of-the-art (SOTA) performance in accuracy and generalisability at various protein variant effect prediction tasks (Brandes et al., 2023; Lin et al., 2023b; Rives et al., 2021), but there are still gaps in their performance for clinical variant classification and correlation with experimental assays (Livesey & Marsh, 2023).

In this study, we improve the performance of PLMs for variant effect prediction using experimental scores from Deep Mutational Scanning (DMS). We first introduce a rescaling and normalisation pipeline to integrate DMS assays from multiple proteins into a common functional scale. We then present a novel lightweight fine-tuning approach for PLMs named Normalised Log-odds Ratio (NLR) that can efficiently learn from DMS data by adding parameter-free layers on top of the language modelling head of PLMs. We finally evaluate the performance improvements of our approach on held-out test proteins and independent DMS and clinical annotation benchmarks, while ensuring low sequence similarity between training and evaluation proteins to assess model generalisation.

2 BACKGROUND

Deep Mutational Scanning (DMS) is an experimental technique that leverages high-throughput DNA sequencing and fitness selection assays to exhaustively measure the effect of variants in a protein region (Fowler & Fields, 2014), providing comprehensive maps of protein variant effects that can be used to understand the clinical relevance of human genetic variants (Findlay et al., 2018; Radford et al., 2023). Recognising the rapidly expanding volume of DMS data and the challenges with data compilation and reproducibility, the Atlas of Variant Effects (AVE) alliance (Fowler et al., 2021) developed MaveDB, an open-source repository of DMS assays (Esposito et al., 2019; Rubin et al., 2021). More recently, ProteinGym has emerged as an independent collection of manually curated DMS assays, providing a standardised framework for benchmarking protein fitness prediction and design (Notin et al., 2023). Despite the wide adoption of DMS datasets for benchmarking variant effect prediction models (Livesey & Marsh, 2023), one of the remaining challenges in combining DMS scores across assays and proteins is that their scale is highly dependent on experimental methods and selection assays, requiring rescaling and normalisation (Dunham & Beltrao, 2021).

Protein Language Models (PLMs) are pre-trained on large corpora of naturally evolved protein sequences using self-supervision and have shown great promise predicting the impact of missense variants without additional supervision (i.e. zero-shot) (Meier et al., 2021). The Evolutionary Scale Modelling (ESM) family of PLMs (Rives et al., 2021; Meier et al., 2021; Lin et al., 2023b) pre-trained with masked language modelling on the UniRef database (UniProt Consortium, 2023) demonstrated the ability to encode functional and structural patterns crucial for variant effect and structural predictions. More recently, AlphaMissense (Cheng et al., 2023), pre-trained on the Protein Data Bank (PDB) (wwPDB Consortium, 2019) and fine-tuned on population frequency data, achieved SOTA results in predicting the clinical pathogenicity of human missense variants. The success of PLMs in zero-shot variant effect predictions has led to the use of fine-tuning approaches to improve performance on specific tasks, including protein stability (Umerenkov et al., 2023), pathogenicity (Lin et al., 2023a), protein-protein interactions (Sledzieski et al., 2023), secondary structure and sub-cellular location (Schmirler et al., 2023), and protein fitness (Rives et al., 2021; Hsu et al., 2022; Jagota et al., 2023).

3 METHODS

3.1 TRAINING AND EVALUATION DATASETS

Normalisation of DMS datasets from MaveDB. We downloaded a total of 308 score-sets from 113 experiments in MaveDB (https://www.mavedb.org) in July 2023. We selected experiments of type *Protein coding*, manually mapped targets to 80 unique gene names and reassigned variants to UniProt sequence positions. We then categorised variants into *nonsense*, *missense* and *synonymous* types using the *hgvs_pro* column, and filtered out indels (insertions and deletions) and multiple amino acid variants. We filtered out viral proteins and selected datasets with at least one synonymous, one nonsense, and over 50 missense variants, resulting in 103 datasets for 30 proteins (Table S3). To normalise functional scores across all datasets, we converted them to log-scales and rescaled the distribution so that the mean score of synonymous variants (S_{syn}) was 0 and the mean score for nonsense variants ($S_{nonsense}$) was -1, using equation 1 (Figure 1A, Figure S1). Values were further capped in the [-2, 2] range to limit outliers.

$$S_{norm} = \frac{S_{raw} - mean(S_{syn})}{mean(S_{syn}) - mean(S_{nonsense})} \tag{1}$$

We then selected scores for missense variants and aggregated assays for each protein. In the case of duplicated scores for the same protein variant in multiple assays, we calculated the mean score. The final dataset contained 142,696 missense variants covering 8,636 unique protein positions in the 30 genes. We clustered the 30 protein sequences using MMseqs2 (version 14) (Steinegger & Söding, 2017) with 20% coverage and 20% sequence identity thresholds, yielding 29 unique clusters (only CCR5 and CXCR4 were clustered together). We selected the 5 proteins with most ClinVar labels (TP53, GCK, CBS, HMBS, and BRCA1) for model testing (Table S3). The remaining clusters were used to select variants in 25 proteins for model training and cross-validation.

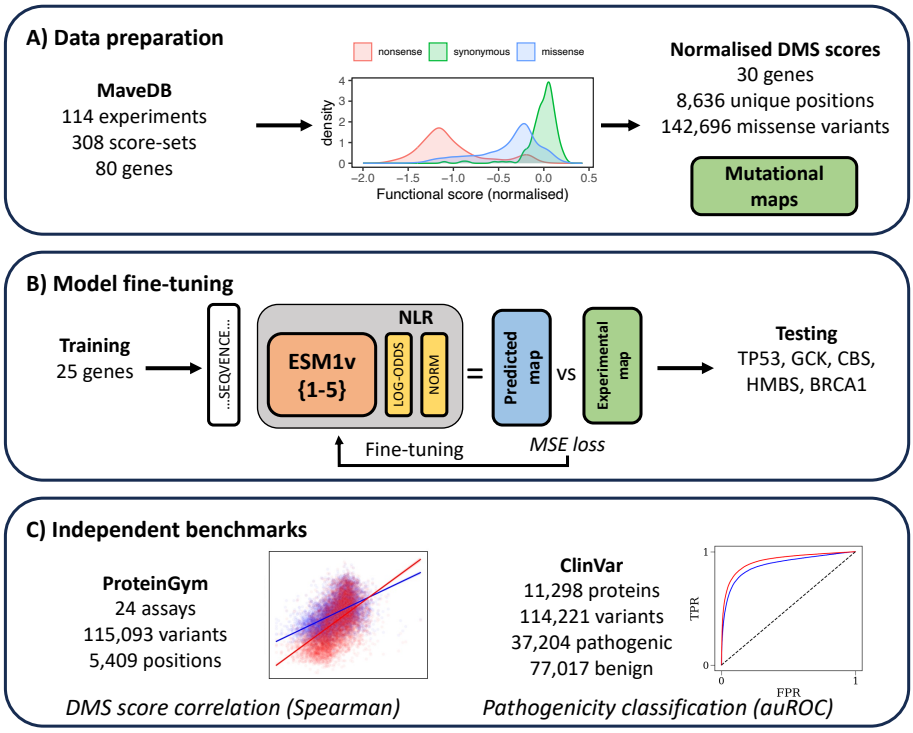


Figure 1: Methods overview. A) Preparation of normalised DMS functional scores from a subset of MaveDB experiments. The mean scores of synonymous and nonsense variants are used to create a common scale across assays and proteins. B) Fine-tuning pipeline for ESM-1v models using the Normalised Log-odds Ratio (NLR) head. C) Performance evaluation on two independent benchmarks: DMS assays from ProteinGym, and pathogenic and benign missense variants from ClinVar.

Benchmarking DMS datasets from ProteinGym. ProteinGym provides functional scores for two types of variants: amino acid substitutions and indels. Since ProteinGym does not provide scores for synonymous or nonsense variants, the normalisation approach needed for model training that we propose here was not possible. However, ProteinGym assays can still be used for benchmarking using correlation metrics. We downloaded ProteinGym substitution scores for a subset of 43 DMS assays with open licenses from the ProteinGym website (https://proteingym.org) in November 2023. We then removed 19 assays with sequence similarity to our 25 MaveDB training proteins using the same procedure described above and selected variants with single amino acid substitutions, resulting in a final benchmark dataset of 24 assays and 115,093 missense variants (Figure 1C).

ClinVar pathogenic and benign variants. ClinVar is a public archive of human genetic variants and interpretations of their significance to disease (Landrum et al., 2018). We downloaded the complete set of 862,666 variant annotations in the *variant_summary.txt* file (version 2023_04) from ClinVar's FTP website (https://ftp.ncbi.nlm.nih.gov/pub/clinvar). We then selected missense variants mapped to protein transcripts (*HGVSp* column), with at least one reviewer star (*ReviewStatus* column), and with clinical significance (*CLNSIG* column) labels as *Benign* and *Likely benign* (considered as *benign* in this study), and variants labelled as *Pathogenic* and *Likely pathogenic* (considered as *pathogenic*), similar to previous work (Cheng et al., 2023; Lin et al., 2023a). We then removed 430 proteins with sequence similarity to our 25 MaveDB training proteins using the same procedure described above. The final ClinVar dataset contains 114,221 missense variants in 11,298 proteins (Figure 1C). To evaluate per-protein performance of our models, we constructed a balanced dataset by selecting proteins with at least 10 benign and 10 pathogenic labels, similar to Cheng et al. (2023), resulting in 37,142 variant annotations for 361 proteins.

3.2 NORMALISED LOG-ODDS RATIO

We present a Normalised Log-odds Ratio (NLR) framework (Figure 1B), which allows efficient fine-tuning on DMS data by adding parameter-free layers on top of PLMs. The key components of the NLR architecture are shown in Figure S2 and include:

1. **Training instances**: Individual amino acid substitutions with corresponding DMS scores.
2. **Variant representation**: For each training instance we provide the wildtype protein as input, allowing efficient computation of all missense variant effects within a protein in a single forward pass during inference. We also experimented masking input tokens to encourage model regularisation, but simply using the raw wildtype sequence outperformed the masking strategies explored (Appendix A.2.1).
3. **Encoder**: We chose ESM as our baseline models and initialised the encoder weights from their corresponding checkpoints. In contrast to most fine-tuning approaches (Schmirler et al., 2023; Lin et al., 2023a), NLR fine-tunes both the pre-trained transformer encoder blocks and the masked language modelling head. We used ESM-1v (esm1v_t33_650M_UR90S_[1-5]) for most experiments but also benchmarked NLR when using ESM-2 (esm2_t33_650M_UR50D) and ESM-1b (esm1b_t33_650M_UR50S) as pre-trained models.

4.1 **NLR - Log-odds Ratio computation**: NLR follows the inference approach of existing zero-shot variant effect prediction methods, scoring all possible amino acid substitutions in a sequence by calculating the log-odds ratio between reference and alternate amino acid probabilities at each mutated position (e.g. Meier et al. (2021); Cheng et al. (2023)). However, NLR performs this computation during both, fine-tuning and inference phases. This is facilitated by the NLR head which, given a wildtype sequence, computes the matrix of log-odds ratios in a differentiable form. As visualised in Figure S2, this computation results in a matrix of sequence length by vocabulary size after each forward pass, providing a score for all possible amino acid substitutions.

4.2 **NLR - Normalisation**: To address the distinct scale between log-odds ratios and DMS scores, NLR applies a normalisation layer to each log-odds ratio matrix during training. After exploring the effect of different approaches, such as instance normalisation (Ulyanov et al., 2016), min-max normalisation between -2 and 2 proved superior to other methods tested (Appendix S2).

5. **Loss and output** During training, the predicted score for a specific variant is retrieved by indexing the corresponding position in the normalised log-odds ratio matrix. This score is then compared to the ground-truth DMS score to compute the mean squared error (MSE) loss for backpropagation.

Training and evaluation. We used the subset of 25 MaveDB proteins for model training and explored hyperparameters and architecture choices with 5-fold cross-validation due to varied zero-shot correlation across proteins (Appendix A.1). We fine-tuned all model parameters as early experiments exhibited larger performance gains than freezing transformer encoder blocks (Appendix A.2.2). After conducting cross-validation experiments, the final models were trained using the entire dataset, consisting of DMS data for 109,215 variants from 25 proteins. Final runs were trained for 2,000 optimization steps to prevent overfitting, which was observed during the cross-validation experiments (Appendix A.1). Since ESM-1v is a five-model ensemble, five independent runs were performed, each starting with a different ESM-1v checkpoint. At inference time, new variants were scored by averaging the log-odds ratio scores across the five model predictions.

4 RESULTS

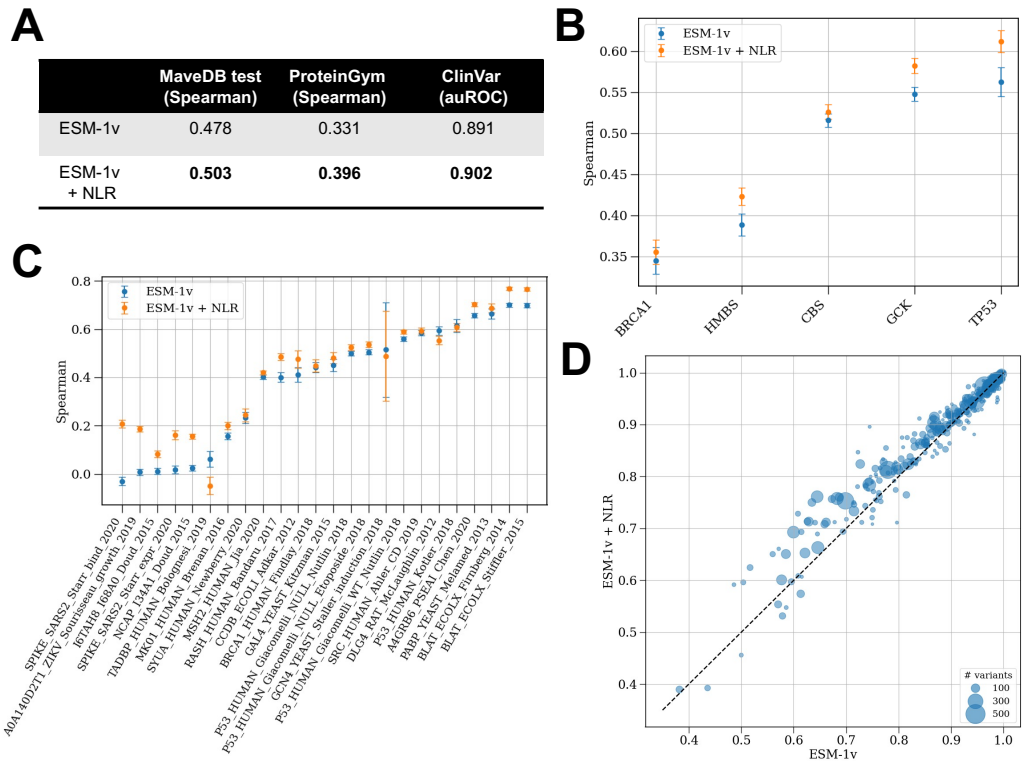
We ran inference with pre-trained and NLR fine-tuned models on the five MaveDB test proteins, the 24 ProteinGym DMS assays and the 114,221 ClinVar pathogenic and benign variants. As shown in Figure 2A, NLR fine-tuning of ESM-1v improves the performance of missense variant effect predictions across all benchmarks.

Improved MaveDB DMS predictions. Our MaveDB test set was composed of five human genes (TP53, GCK, CBS, HMBS and BRCA1) with associated DMS data, which went through the filtering and standardisation procedure presented in this study. From the results in Figure 2A, we find that NLR fine-tuning of ESM-1v ensemble improves the micro-averaged[1] Spearman correlation

[1]Micro-averaging of metrics applies weights per protein proportional to their relative number of variants, while macro-averaging applies equal weight per protein.

across proteins in our MaveDB test set from 0.478 to 0.503 (+5.2%, Figure 2A). In Table S4, the performance is stratified by each of the five ESM-1v model checkpoints, displaying improvements from 5.4% to 9.4% across all model versions. These results indicate that the fine-tuning improvements on individual model checkpoints are larger than the ensemble of ESM-1v models. Figure 2B displays the performance of ESM-1v per protein, exhibiting improvements after fine-tuning on all five proteins in the test set. We also assessed the impact of the number of proteins used for training (Appendix A.3). We observed an upward trend in model performance with increasing number of training proteins, indicating that NLR fine-tuning can scale with more available DMS data.

Improved ProteinGym DMS predictions. Our ProteinGym benchmark was composed of 24 DMS assays which differed from the five proteins in the MaveDB test set in two ways: they included DMS data for non-human proteins and we did not apply our MaveDB pre-processing pipeline. From the results in Figure 2A we find that NLR fine-tuning of ESM-1v improves the average Spearman correlation across assays in ProteinGym from 0.331 to 0.396 (+19.6%, Figure 2A). Figure 2C reveals that this improvement extends to nearly all DMS studies, both human and non-human. Notably, the improvements were larger for viral proteins, which were excluded from the training set and had the lowest zero-shot correlation.



	MaveDB test (Spearman)	ProteinGym (Spearman)	ClinVar (auROC)
ESM-1v	0.478	0.331	0.891
ESM-1v + NLR	**0.503**	**0.396**	**0.902**

Figure 2: Results after NLR fine-tuning of ESM-1v models across benchmarks. A) Performance in the five MaveDB test proteins. ProteinGym DMS assays and ClinVar pathogenic variants. B) Spearman correlation in MaveDB test proteins. Mean ± standard deviation (std) of 50 bootstrapped samples. C) Spearman correlation in ProteinGym DMS assays. Mean ± std of 50 bootstrapped samples. D) Per-protein auROC for ClinVar proteins with over 10 benign and 10 pathogenic variants.

Improved ClinVar pathogenicity classification. We analysed 114,221 variants from 11,298 proteins, calculating the area under the receiver operator characteristic curve (auROC) between predicted log-odds ratios and ClinVar labels. Fine-tuning increased auROC from 0.891 to 0.902 (+1.23%, Figure 2A). As seen in Figure S3B, after fine-tuning, log-odds ratio scores for pathogenic variants became more negative, and benign scores more positive, resulting in improved pathogenic-benign variant separation. We further used the balanced ClinVar subset with proteins having at least 10 benign and 10 pathogenic labels to examine the fine-tuning effect on individual proteins. As

shown in Figure 2D, fine-tuning consistently improved variant classification for most proteins. Notably, the improvements were larger for proteins with lower baseline auROCs (below 0.8), while proteins with strong zero-shot performance saw minimal gains. This analysis revealed a micro-averaged auROC increase from 0.860 to 0.877 (+1.98%), and a macro-averaged auROC rise across proteins from 0.880 to 0.891 (+1.25%).

NLR fine-tuning improves other ESM models. We further assessed the impact of NLR fine-tuning with ESM-1b and ESM-2. ESM-1b shares the same architecture as ESM-1v but was trained on different corpora and not directly optimised for variant effect predictions (Rives et al., 2021). ESM-2 introduced improvements in architecture, number of training parameters and pre-training data, outperforming previous ESM models on protein structure prediction (Lin et al., 2023b). Across all benchmarks, ESM-1b and ESM-2 exhibited performance improvements after NLR fine-tuning (Table 1). Notably, performance gains were higher for ESM-1b and ESM-2 compared to the ensemble of ESM-1v models on our DMS benchmarks, with up to 25.6% increase in average Spearman correlation on ProteinGym for ESM-1b. This indicates that NLR's benefits extend beyond specific architectures and may be more significant in single-model settings.

Table 1: Performance of ESM models in each evaluation benchmark before (Zero-Shot) and after NLR fine-tuning (+ NLR ft).

		MaveDB (avg Spearman)	ProteinGym (avg Spearman)	ClinVar (auROC)
ESM-1v ensemble	Zero-Shot	0.478	0.331	0.891
	+ NLR ft	**0.503**	**0.396**	**0.902**
ESM-1b	Zero-Shot	0.451	0.309	0.913
	+ NLR ft	**0.498**	**0.388**	**0.919**
ESM-2 (650M)	Zero-Shot	0.482	0.317	0.884
	+ NLR ft	**0.509**	**0.393**	**0.894**

5 DISCUSSION

We have presented a novel approach to enhance variant effect predictions from PLMs by fine-tuning on DMS datasets. We applied score normalisation across DMS assays to tackle challenges in data integration and introduced a novel lightweight fine-tuning Normalised Log-odds Ratio head that allows PLMs to efficiently learn from DMS data without adding task-specific parameters. After NLR fine-tuning, we observed moderate but consistent improvements across variant effect prediction benchmarks, proteins and ESM models, highlighting the robustness of the approach. Despite fine-tuning with DMS data from a limited set of 25 proteins, we observed accuracy improvements in the classification of the clinical significance of variants from 11,298 proteins in ClinVar, and in the correlation with experimental measurements from independent DMS assays in ProteinGym and MaveDB. The improvements were more pronounced for proteins with a lower zero-shot performance and lesser representation in the pre-training datasets (for example, viral proteins). These findings demonstrate that our fine-tuning approach improves the performance of PLMs beyond self-supervision on naturally selected protein sequences.

This study focused on the ESM family of PLMs, which utilises single protein sequences as input. Other models like MSA Transformer, EVE, and AlphaMissense (Rao et al., 2021; Frazer et al., 2021; Cheng et al., 2023) leverage Multiple Sequence Alignments (MSAs) in the input space. Alignment-based PLMs can perform better in proteins with a limited number of homologs in protein sequence databases, such as viral proteins, overcoming some of the performance gaps of ESM models. Notably, AlphaMissense has recently achieved SOTA performance on multiple missense variant effect prediction tasks, with reported ClinVar accuracy and DMS correlations superior to the NLR fine-tuned models in this study. Adapting NLR fine-tuning to MSA-based PLMs is a promising avenue for future work. Although we used DMS data from a limited set of 25 proteins, we observed that NLR fine-tuning can continue to scale model performance with more DMS data. We believe that NLR fine-tuning provides a lightweight and efficient approach to improve PLMs variant effect predictions as the volume, quality, and standards of DMS data continue to grow.

REFERENCES

Joshua D Backman, Alexander H Li, Anthony Marcketta, Dylan Sun, Joelle Mbatchou, Michael D Kessler, Christian Benner, Daren Liu, Adam E Locke, Suganthi Balasubramanian, et al. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature*, 599(7886):628–634, 2021.

Nadav Brandes, Grant Goldman, Charlotte H Wang, Chun Jimmie Ye, and Vasilis Ntranos. Genome-wide prediction of disease variant effects with a deep protein language model. *Nature Genetics*, 55(9):1512–1522, 2023.

Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664):eadg7492, 2023.

Alistair S Dunham and Pedro Beltrao. Exploring amino acid functions in a deep mutational landscape. *Molecular systems biology*, 17(7):e10305, 2021.

Daniel Esposito, Jochen Weile, Jay Shendure, Lea M Starita, Anthony T Papenfuss, Frederick P Roth, Douglas M Fowler, and Alan F Rubin. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome biology*, 20:1–11, 2019.

Gregory M Findlay, Riza M Daza, Beth Martin, Melissa D Zhang, Anh P Leith, Molly Gasperini, Joseph D Janizek, Xingfan Huang, Lea M Starita, and Jay Shendure. Accurate classification of BRCA1 variants with saturation genome editing. *Nature*, 562(7726):217–222, 2018.

DM Fowler, M Hurles, DJ Adams, AL Gloyn, WC Hahn, DS Marks, JT Neal, F Roth, AF Rubin, LM Starita, et al. The Atlas of Variant Effects (AVE) Alliance: understanding genetic variation at nucleotide resolution. *Zenodo*, 2021.

Douglas M Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature methods*, 11(8):801–807, 2014.

Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly Brock, Yarin Gal, and Debora S Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.

Hong Gao, Tobias Hamp, Jeffrey Ede, Joshua G Schraiber, Jeremy McRae, Moriel Singer-Berk, Yanshen Yang, Anastasia SD Dietrich, Petko P Fiziev, Lukas FK Kuderna, et al. The landscape of tolerated genetic variation in humans and primates. *Science*, 380(6648):eabn8153, 2023.

Chloe Hsu, Hunter Nisonoff, Clara Fannjiang, and Jennifer Listgarten. Learning protein fitness models from evolutionary and assay-labeled data. *Nature biotechnology*, 40(7):1114–1122, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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.

Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809): 434–443, 2020.

Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth R Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic acids research*, 46(D1):D1062–D1067, 2018.

Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O'Donnell-Luria, James S Ware, Andrew J Hill, Beryl B Cummings, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, 2016.

Weining Lin, Jude Wells, Zeyuan Wang, Christine Orengo, and Andrew CR Martin. VariPred: Enhancing Pathogenicity Prediction of Missense Variants Using Protein Language Models. *bioRxiv*, pp. 2023–03, 2023a.

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, 2023b.

Benjamin J Livesey and Joseph A Marsh. Updated benchmarking of variant effect predictors using deep mutational scanning. *Molecular Systems Biology*, pp. e11474, 2023.

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.

Pascal Notin, Aaron W Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Hansen Spinner, Nathan Rollins, Ada Shaw, Ruben Weitzman, Jonathan Frazer, et al. ProteinGym: Large-Scale Benchmarks for Protein Design and Fitness Prediction. *bioRxiv*, pp. 2023–12, 2023.

Elizabeth J Radford, Hong-Kee Tan, Malin HL Andersson, James D Stephenson, Eugene J Gardner, Holly Ironfield, Andrew J Waters, Daniel Gitterman, Sarah Lindsay, Federico Abascal, et al. Saturation genome editing of DDX3X clarifies pathogenicity of germline and somatic variation. *Nature Communications*, 14(1):7702, 2023.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA transformer. In *International Conference on Machine Learning*, pp. 8844–8856. PMLR, 2021.

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.

Alan F Rubin, Joseph K Min, Nathan J Rollins, Estelle Y Da, Daniel Esposito, Matthew Harrington, Jeremy Stone, Aisha Haley Bianchi, Mafalda Dias, Jonathan Frazer, et al. MaveDB v2: a curated community database with over three million variant effects from multiplexed functional assays. *bioRxiv*, pp. 2021–11, 2021.

Robert Schmirler, Michael Heinzinger, and Burkhard Rost. Fine-tuning protein language models boosts predictions across diverse tasks. *bioRxiv*, pp. 2023–12, 2023.

Samuel Sledzieski, Meghana Kshirsagar, Minkyung Baek, Bonnie Berger, Rahul Dodhia, and Juan Lavista Ferres. Democratizing Protein Language Models with Parameter-Efficient Fine-Tuning. *bioRxiv*, pp. 2023–11, 2023.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

Dmitriy Umerenkov, Fedor Nikolaev, Tatiana I Shashkova, Pavel V Strashnov, Maria Sindeeva, Andrey Shevtsov, Nikita V Ivanisenko, and Olga L Kardymon. PROSTATA: a framework for protein stability assessment using transformers. *Bioinformatics*, 39(11):btad671, 2023.

UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2023.

wwPDB Consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic acids research*, 47(D1):D520–D528, 2019.

A APPENDIX

A.1 FINE-TUNING DETAILS

Experimental setup. The base model for hyperparameter search and ablations was the first ESM-1v model checkpoint (esm1v_t33_650M_UR90S_1). Experiments ran for a maximum of 10 epochs, saving checkpoints based on peak micro-averaged Spearman correlation on the validation fold. We used an effective batch size of 128 samples across 8 V100s (32GB), training in Distributed Data Parallel mode, with two samples per device and accumulating gradients for 8 steps. All experiments performed learning rate warmup from 0 to the peak learning rate over 200 steps, and the peak learning rate was chosen via grid search (1e-5, 2e-5, 5e-5, and 1e-4). Cross-validation experiments revealed an optimal learning rate of 2e-5, exhibiting overfitting after 2,000 optimization steps.

Cross-validation. Five-fold cross-validation was performed on the training set (25 proteins from MaveDB) to compare model architectures and hyperparameter settings. DMS data for 5 proteins was kept in each validation fold while 20 proteins were used for model training. The cross-validation scheme grouped variant instances by protein cluster to assess generalisation to unseen and dissimilar proteins. In each run, the absolute improvement in Spearman correlation over the zero-shot model was computed at each step (t):

$$improvement(t) = Spearman(t) - Spearman(t = 0) \tag{S1}$$

where $Spearman(t{=}0)$ represents the correlation before starting model fine-tuning and $Spearman(t)$ represents the correlation after t optimization steps. We save the model checkpoint at the step (t_{max}) with peak Spearman improvement.

Log-odds ratio matrix normalisation. During NLR fine-tuning, the log-odds ratio matrix is computed for each training instance, and min-max normalisation is then independently applied to each matrix:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}(range_{max} - range_{min}) + range_{min} \tag{S2}$$

where X is the log-odds ratio matrix, X_{min} and X_{max} are its minimum and maximum values (masking out padded positions), and $range_{min}$ and $range_{max}$ define the desired normalised range, -2 and 2, respectively, aligning with DMS scores. Normalisation was only performed during fine-tuning but not at inference time.

A.2 REGULARISATION EXPERIMENTS

Training instances are defined by the wildtype protein sequence and a single DMS variant score. Each DMS variant score corresponds to a specific amino acid substitution at a certain position, and its corresponding prediction is sampled from the log-odds ratio matrix. Given the large number of DMS scores per protein sequence, the model is repeatedly exposed to the same input sequence during training which could make it prone to overfitting. To test for this, we explored the effect of two regularisation strategies: input masking and layer freezing.

A.2.1 INPUT MASKING

Randomly masking a percentage of input tokens for each variant instance has the potential to prevent the model relying on local, noise-based correlations between the amino acid sequence and the DMS output distribution. To test this we experimented with masking 5, 15 and 30% of input tokens for each instance during training, while at inference time, only the variant position was masked and the remaining context was shown to the model. Despite the potential for regularisation, no positive effects were observed when masking the input sequences (see Table S1). This result may be attributed to (1) a potential mismatch between the training and validation input distribution, since masked tokens are seen in a higher proportion during training than inference, when only the variant position was masked, and (2) strong dependencies between the effects of single nucleotide variants and specific motifs within the context of those variants. The latter scenario could result in masking of crucial information during training that prevents effective learning.

Table S1: Peak Spearman correlation improvement across folds in MaveDB (mean ± std) when randomly masking input tokens.

Inputed tokens masked (%)	$improvement(t_{max})$
0	0.033 ± 0.008
5	0.022 ± 0.009
15	0.019 ± 0.004
30	0.017 ± 0.007

A.2.2 LAYER FREEZING

Layer freezing has the potential to regularise the model by preserving the information learned during pre-training in the frozen layers while fine-tuning the remaining encoder blocks. However, while progressively freezing ESM-1v we did not observe any clear improvements (see Table S2). It is worth noting that the high variance across folds might hinder more general conclusions about the potential optimal number of layers frozen. However, three important conclusions were taken from this analysis:

1. As a result of the experiment freezing 33 layers, it was clear that tuning transformer encoder blocks is essential to obtain higher improvements on Spearman correlation, and exclusively tuning the language modelling head only provides a minor improvement.
2. Only unfreezing one encoder block performed worse than the rest of the experiments, suggesting that modelling this task requires a higher number of trainable parameters and learning higher-order token interactions.
3. From the results above, it was not clear that any number of frozen transformer encoder block experiments outperformed the fully unfrozen model. However, future experiments with parameter-efficient fine-tuning strategies (Hu et al., 2021) could reveal an optimal number of parameters to tune for this task.

Table S2: Peak Spearman correlation improvement across folds in MaveDB (mean ±std) when freezing layers.

Layers frozen	$improvement(t_{max})$
0	0.033 ± 0.008
12	0.027 ± 0.015
16	0.027 ± 0.014
20	0.027 ± 0.014
24	0.028 ± 0.012
28	0.031 ± 0.015
32	0.022 ± 0.019
33[†]	0.001 ± 0.007

[†] Only the Language Modelling head was fine-tuned.

A.3 DATA SCALING

We conducted experiments to evaluate how performance improvement scales with the number of training proteins that have DMS data. Analogous to regularisation experiments, we performed five-fold cross-validation on the training set. However, in each run, we randomly selected a subset of n proteins from the training folds, and conducted experiments with n being 2, 5, 10, 15 or 20 and ran all the experiments for five epochs. The results, which are displayed in Figure S4, demonstrate that the improvements of NLR fine-tuning scale with the number of proteins with DMS data. Furthemore, when increasing the amount of DMS data, we observed a consistent upward trend in improvement across all folds. Overall, these results indicate that NLR fine-tuning can effectively scale the performance of Protein Language Models (PLMs) in variant effect prediction as the number of DMS assays openly available contines to grow.

A.4 TABLES AND FIGURES

Table S3: List of proteins and DMS assays from MaveDB used in this study.

Gene	Uniprot ID	Organism	MaveDB URN	Data split
BRCA1	P38398	Homo sapiens	00000003-a-2,00000097-e-1,00000097-f-1,00000081-a-1,00000097-g-1,00000097-h-1,00000097-0-1,00000097-d-1,00000097-a-1,00000097-b-1,00000097-c-1,00000097-i-1,00000097-j-1,00000003-a-1,00000097-l-1,00000097-m-1,00000097-n-1,00000097-o-1,00000097-p-1,00000097-q-1,00000097-k-1,00000097-t-1,00000097-u-1,00000097-v-1,00000097-w-1,00000097-x-1,00000097-r-1,00000097-s-1,00000097-y-1,00000097-z-1	Test
CBS	P35520	Homo sapiens	00000005-a-2,00000005-a-1,00000005-a-3,00000005-a-4,00000005-a-5,00000005-a-6	Test
GCK	P35557	Homo sapiens	00000096-a-1,00000096-b-1	Test
HMBS	P08397	Homo sapiens	00000108-a-1,00000108-a-2,00000108-a-3	Test
TP53	P04637	Homo sapiens	00000059-a-1	Test
ACE2	Q9BYF1	Homo sapiens	00000069-a-1,00000069-a-2	Train
CALM1	P0DP23	Homo sapiens	00000001-c-1	Train
CCR5	P51681	Homo sapiens	00000047-c-1,00000047-a-1,00000047-b-1	Train
CD86	P42081	Homo sapiens	00000046-a-1	Train
CXCR4	P61073	Homo sapiens	00000048-c-1,00000048-b-1,00000048-a-1	Train
CYP2C9	P11712	Homo sapiens	00000095-b-1,00000095-a-1	Train
DHFR	P0ABQ4	Escherichia coli	00000063-a-1,00000063-b-1	Train
HMGCR	P04035	Homo sapiens	00000035-a-1,00000035-a-3	Train
HSP90	P02829	Saccharomyces cerevisiae	00000074-a-1,00000039-a-4,00000039-a-5,00000011-a-1,00000039-a-6,00000039-a-7,00000040-a-1,00000040-a-4,00000040-a-3,00000040-a-2,00000039-a-1,00000039-a-3,00000039-a-2	Train
LDLRAP1	Q5SW96	Homo sapiens	00000036-a-1,00000036-a-2	Train
LamB	P02943	Escherichia coli	00000064-a-1,00000064-b-1	Train
MTHFR	P42898	Homo sapiens	00000049-a-1,00000049-a-3,00000049-a-4,00000049-a-2,00000049-a-8,00000049-a-5,00000049-a-6,00000049-a-7	Train
NUDT15	Q9NV35	Homo sapiens	00000055-a-1,00000055-b-1	Train
PRKN	O60260	Homo sapiens	00000114-a-1	Train
PTEN	P60484	Homo sapiens	00000013-a-1,00000054-a-1	Train
SCN5A	Q14524	Homo sapiens	00000098-a-1	Train
SUMO1	P63165	Homo sapiens	00000001-b-2	Train
TPK1	Q9H3S4	Homo sapiens	00000001-d-1	Train
TPMT	P51580	Homo sapiens	00000013-b-1	Train
UBE2I	P63279	Homo sapiens	00000001-a-3,00000001-a-4	Train
UBE4B	Q9ES00	Mus musculus	00000004-a-2,00000004-a-3	Train
UBI4	P0CG63	Saccharomyces cerevisiae	00000037-a-1	Train
VKOR	Q9BQB6	Homo sapiens	00000078-a-1,00000078-b-1	Train
YAP1	P46937	Homo sapiens	00000002-a-1,00000002-a-2	Train
avGFP	P42212	Aequorea victoria	00000080-a-1,00000080-a-2	Train

Table S4: Micro-averaged Spearman correlation across MaveDB test proteins stratified by ESM-1v model versions. Improvement is reported as % of increase over Zero-Shot baseline.

ESM-1v checkpoint	Zero-Shot	NLR fine-tuned	% improvement
1	0.444	**0.476**	7.2
2	0.438	**0.479**	9.4
3	0.456	**0.481**	5.5
4	0.456	**0.487**	6.8
5	0.452	**0.490**	8.4
Ensemble 1-5	0.478	**0.503**	5.2

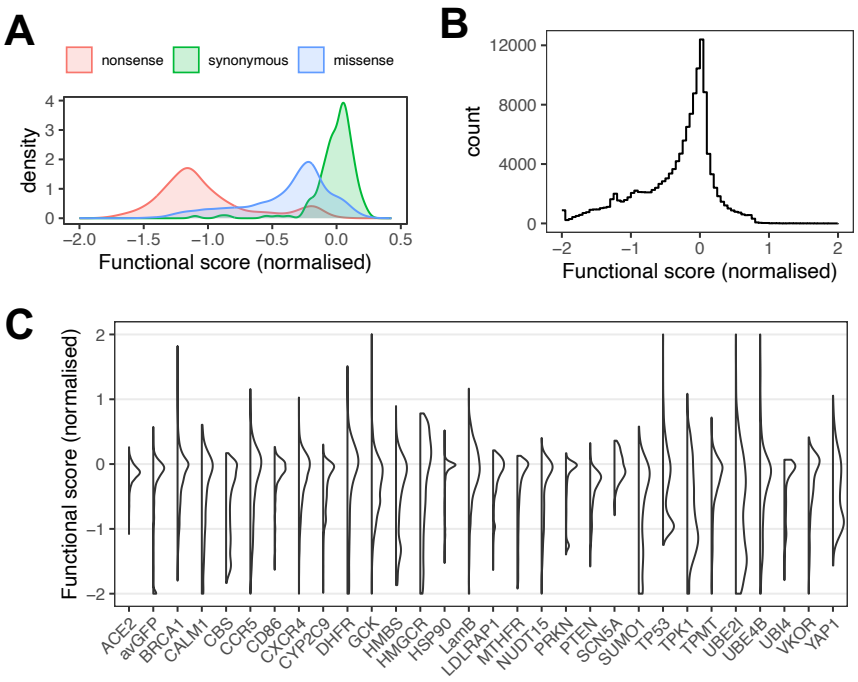


Figure S1: Normalisation of DMS functional scores. A) Distribution of DMS functional scores for missense, nonsense, and synonymous variants in MaveDB for the PTEN protein after score rescaling and normalisation, shown as density functions. B) Stacked histogram of DMS functional scores for all missense variants in the 30 proteins of the final normalised MaveDB dataset. C) Distributions of normalised DMS functional scores for each of the 30 proteins in the MaveDB dataset, shown as half-violin plots.

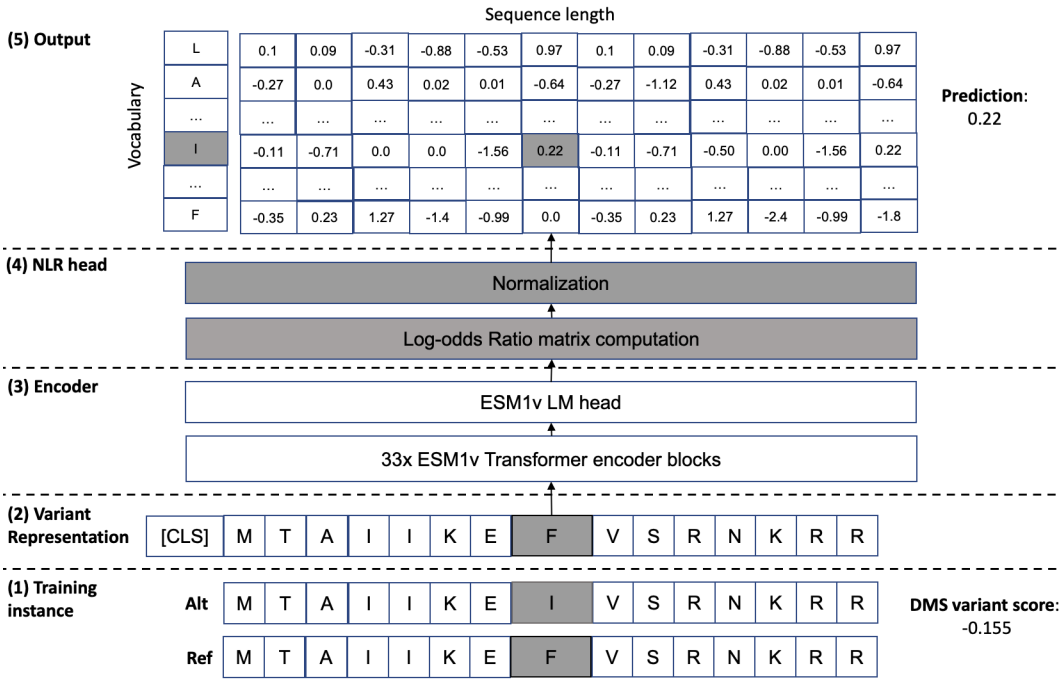


Figure S2: Diagram of the fine-tuning ESM-1v architecture with Normalised Log-odds Ratio (NLR) head. (1) Training instance with a single amino acid swap and its DMS label. (2) Input token representation for the wild-type sequence. (3) ESM-1v pre-trained encoder blocks + Language Modelling head. (4) Fine-tuning blocks, including log-odds ratio matrix calculation and normalisation layers. (5) Output matrix with the chosen cell reflecting the predicted score for the input variant.

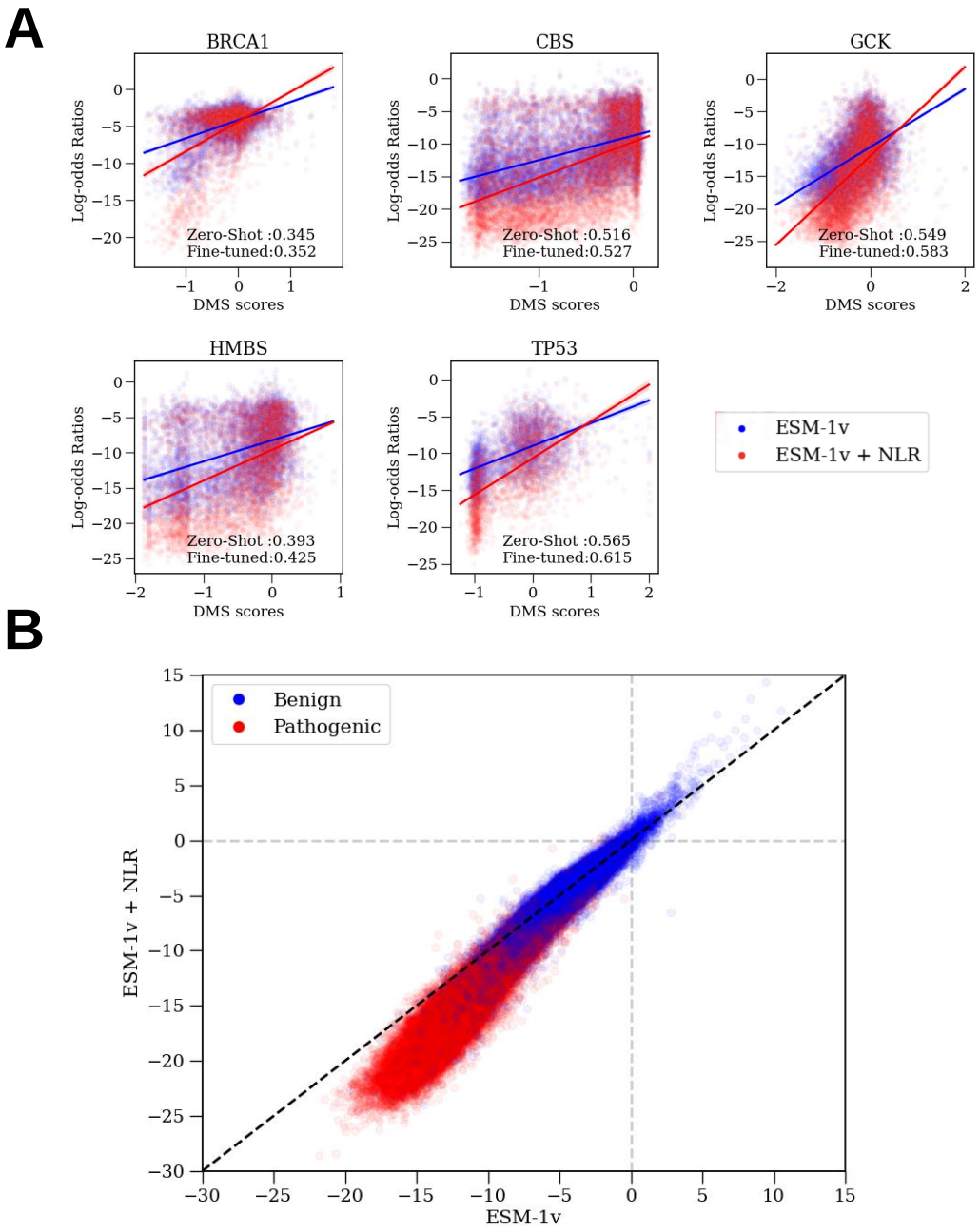


Figure S3: A) DMS vs. predicted log-odds ratios for each variant in the MaveDB test set, stratified by gene. In blue are the scores from the zero-shot ESM-1v, while in red are the NLR fine-tuned scores. A linear fit is displayed together with the score distribution and the Spearman correlation is shown for each distribution. B) Predicted log-odds ratios from ESM-1v vs. NLR-finetuned ESM-1v, for each variant in the ClinVar benchmark. In blue and red are the benign and pathogenic variants, respectively.

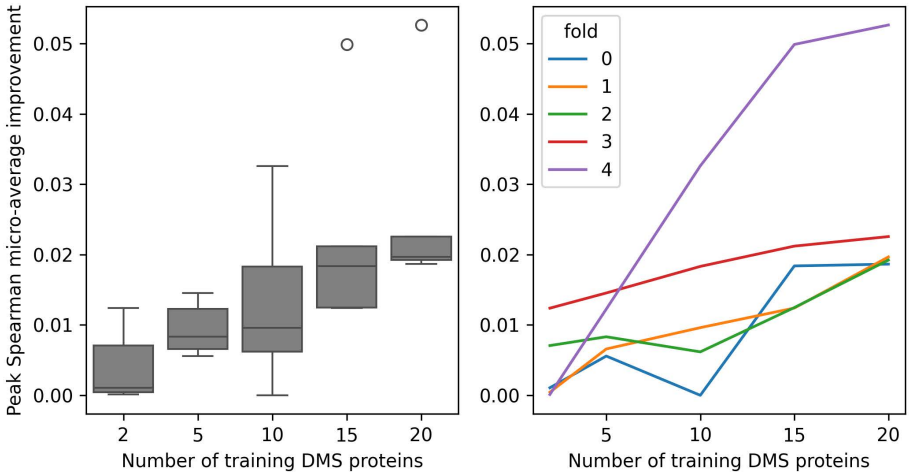


Figure S4: Peak Spearman correlation improvement across validation folds in MaveDB as a function of the number of training proteins. Left-side panel exhibits the improvement as a box-plot distribution across validation folds. The right-side panel shows the improvement stratified by validation fold.