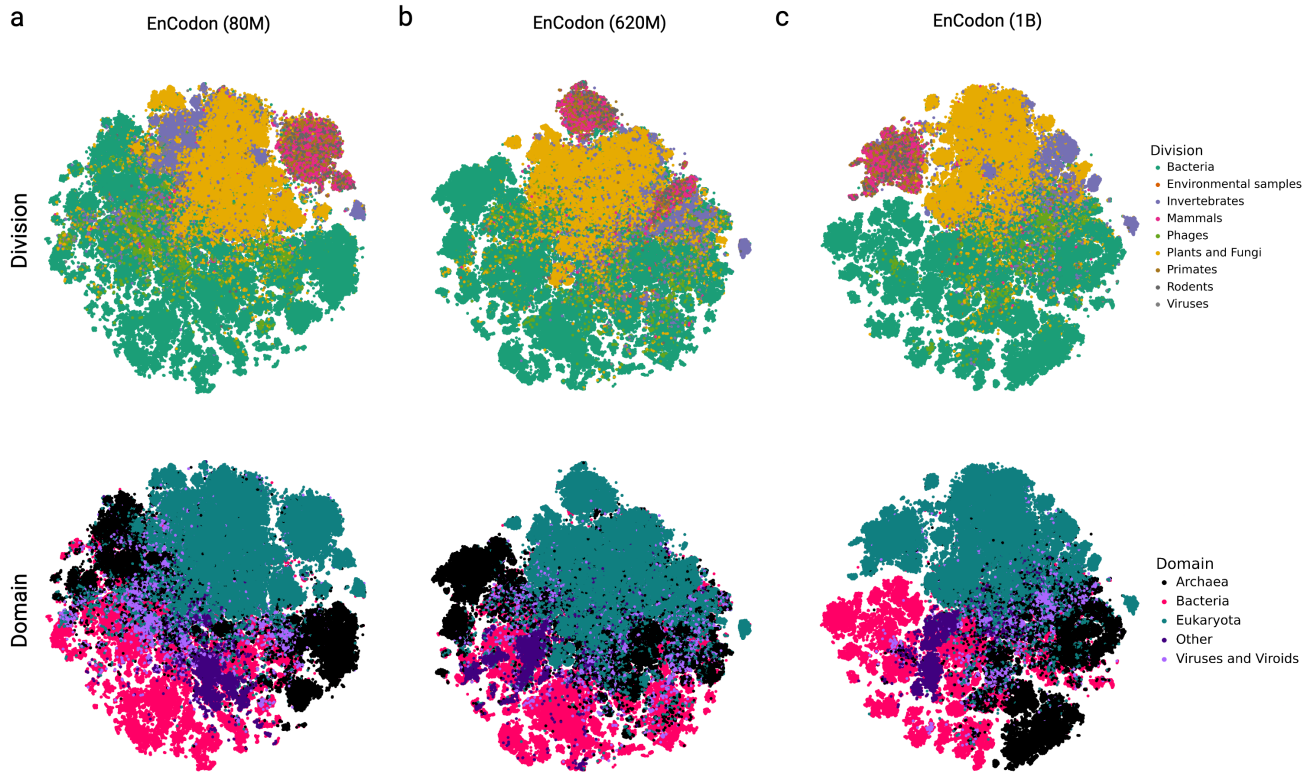
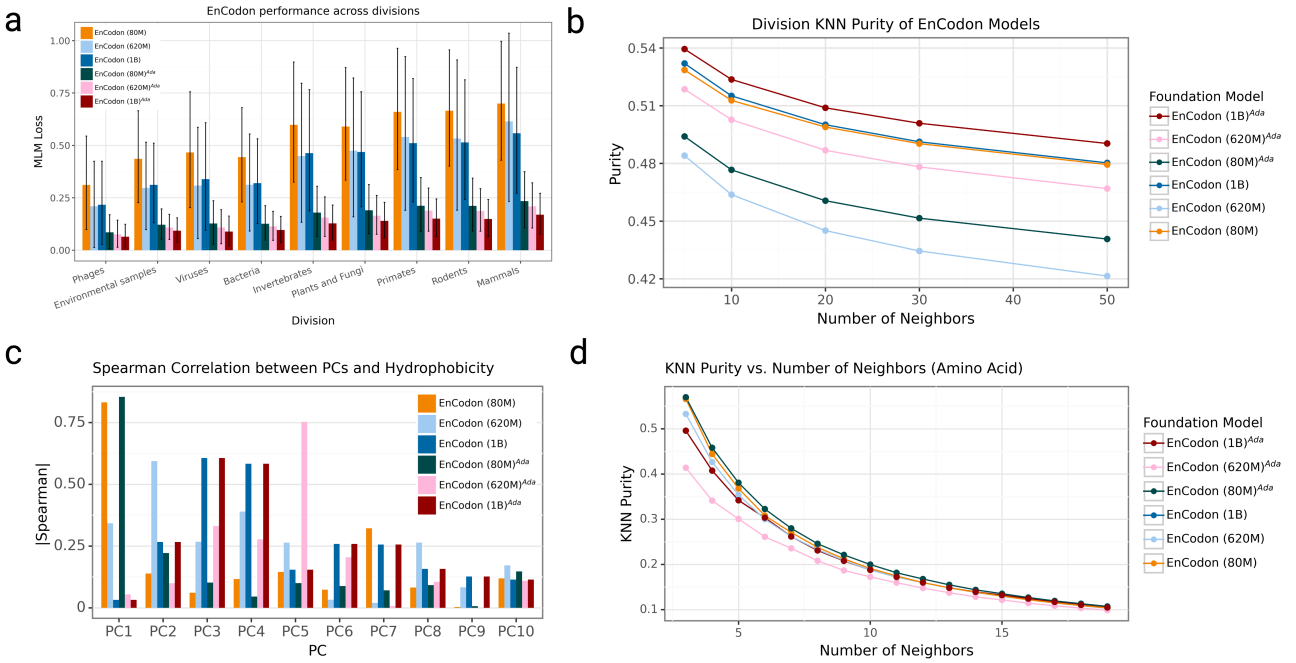
Foundation Model	# layers	Emb. dim.	Inter. dim.	LR	WD	warmup steps
EnCodon (80M)	12	1024	2048	1e-4	1e-2	10,000
EnCodon (620M)	12	2048	8192	5e-5	1e-2	10,000
EnCodon (1B)	18	2048	8192	1e-5	1e-2	10,000
DeCodon (200M)	12	1024	2048	1e-4	1e-2	10,000
DeCodon (1B)	18	2048	8192	1e-5	1e-2	10,000

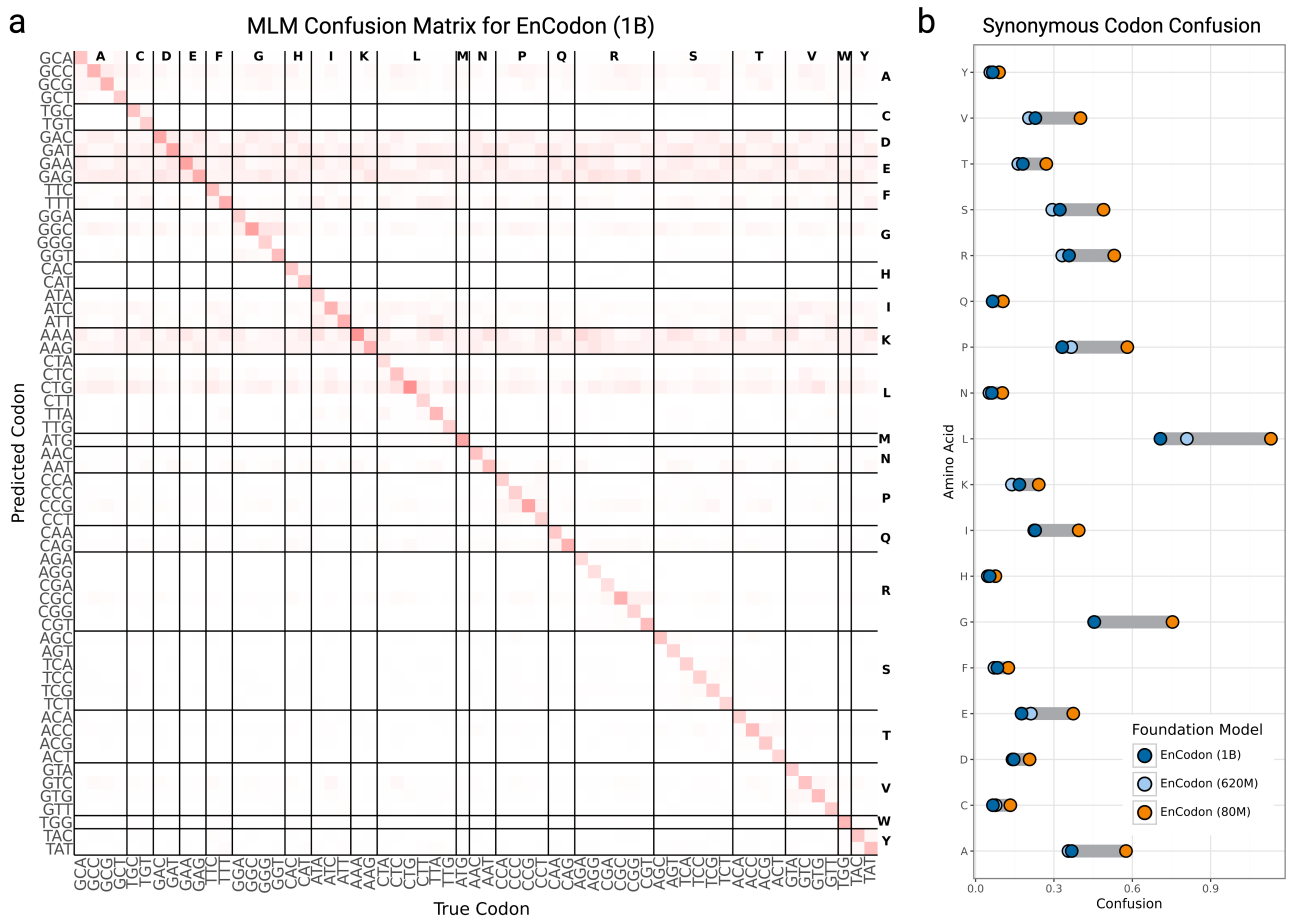
Supplementary Table 1 | Hyperparameters used for each of our pre-trained codon foundation models (cdsFMs). Emb. dim.: codon-level embedding dimensionality, Inter. dim.: Intermediate layers' dimensionality, LR: learning rate, WD: weight decay



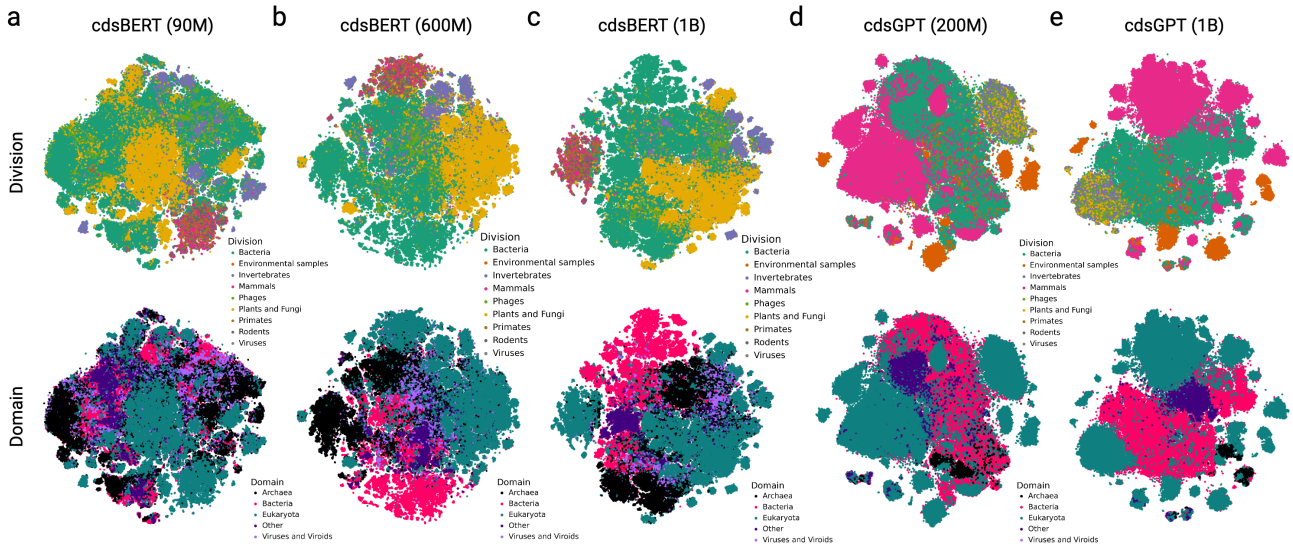
Supplementary Figure 1: T-SNE visualization of sequence embedding space learned by a) EnCodon (80M), b) EnCodon (620M), and c) EnCodon (1B) where each dot is a sequence and they are colored by sequence's organism division (top row) and domain (bottom row).



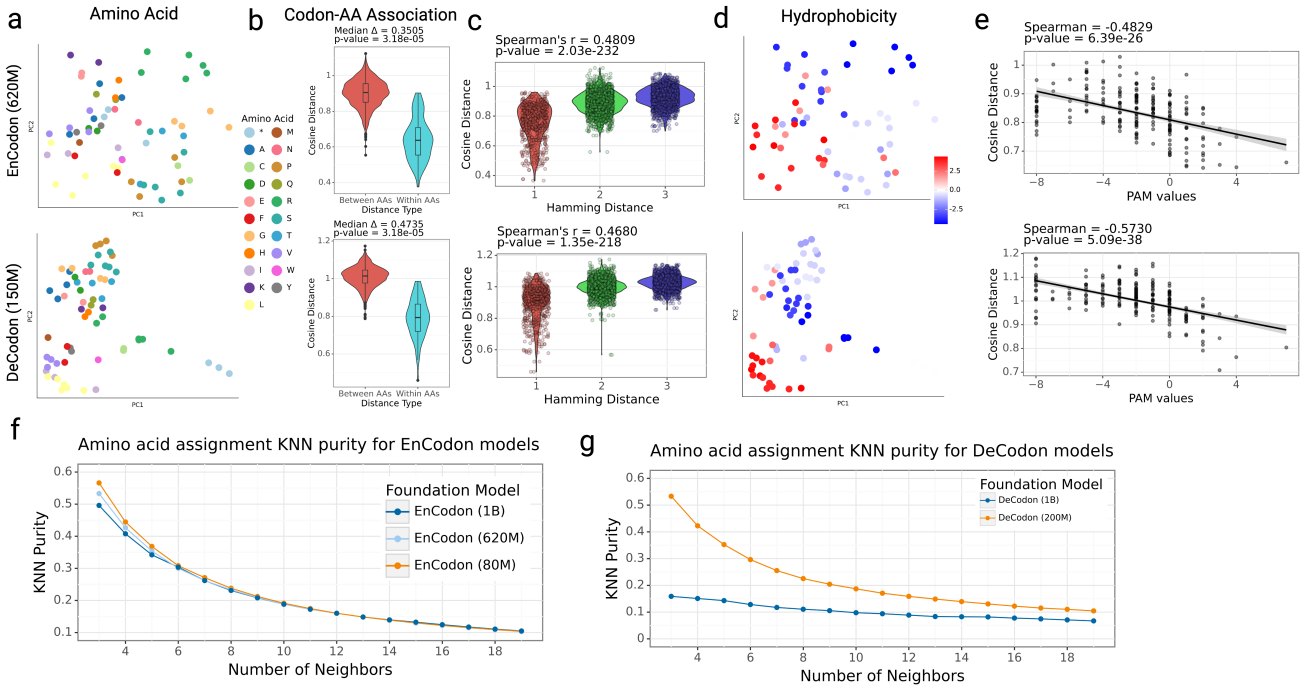
Supplementary Figure 2: **a)** Representation of MLM loss distribution for pre-trained and adapted EnCodon models across taxonomy divisions with mean bars and standard error lines. **b)** Scatter plots of KNN Purity scores against numbers of nearest neighbors, using organisms' Division as clustering labels. **c)** Spearman correlations bar plot between the top 10 principal components (PC) of the pre-trained/adapted EnCodons and the hydrophobicity index of codon's amino acid. **d)** KNN Purity scores of the codon embedding space of EnCodons with amino acid labels against the number of neighbors (K).



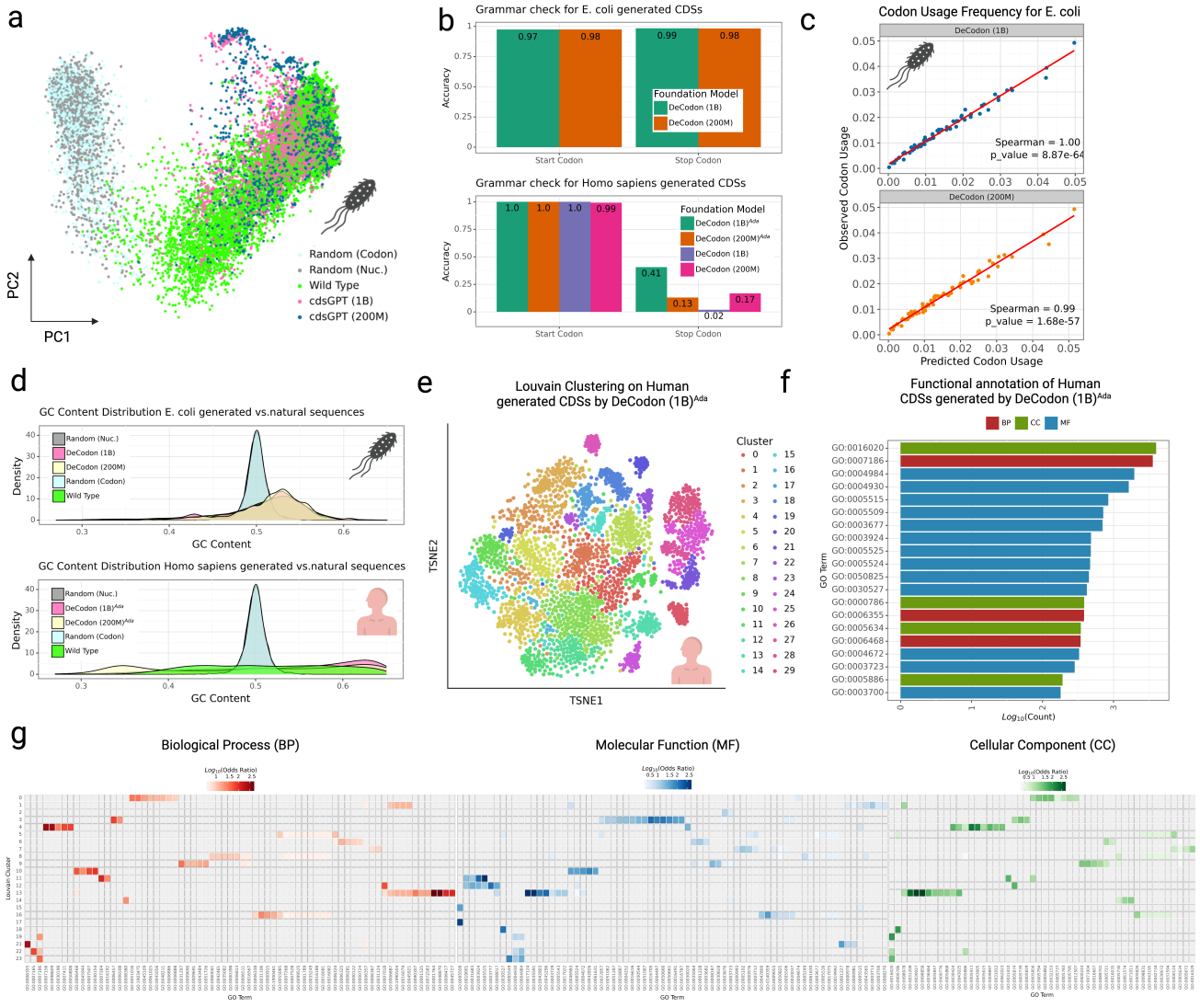
Supplementary Figure 3: **a)** Masked language modeling confusion matrix of pre-trained EnCodon (1B) model. We use sequences in the pre-training test split and randomly masked each codon in the sequence with 0.15 probability. The shown confusion matrix is computed from EnCodon's prediction on the masked positions. **b)** Difference plot of synonymous codon confusion per amino acid is shown for the purpose of comparing pre-trained EnCodons – EnCodon (80M), EnCodon (620M), and EnCodon(1B).



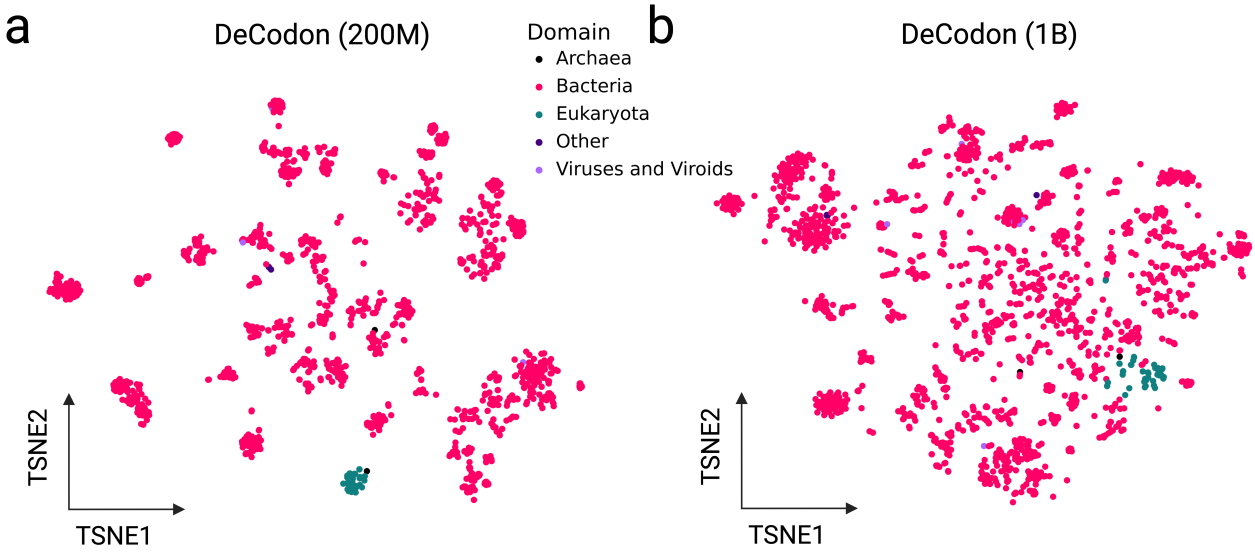
Supplementary Figure 4: T-SNE visualization of sequence embedding space learned performing eukaryotic adaptation on **a)** EnCodon (80M), **b)** EnCodon (620M), **c)** EnCodon (1B), **d)** DeCodon (200M), and **e)** DeCodon (1B) where each dot is a sequence and they are colored by sequence's organism division (top row) and domain (bottom row).



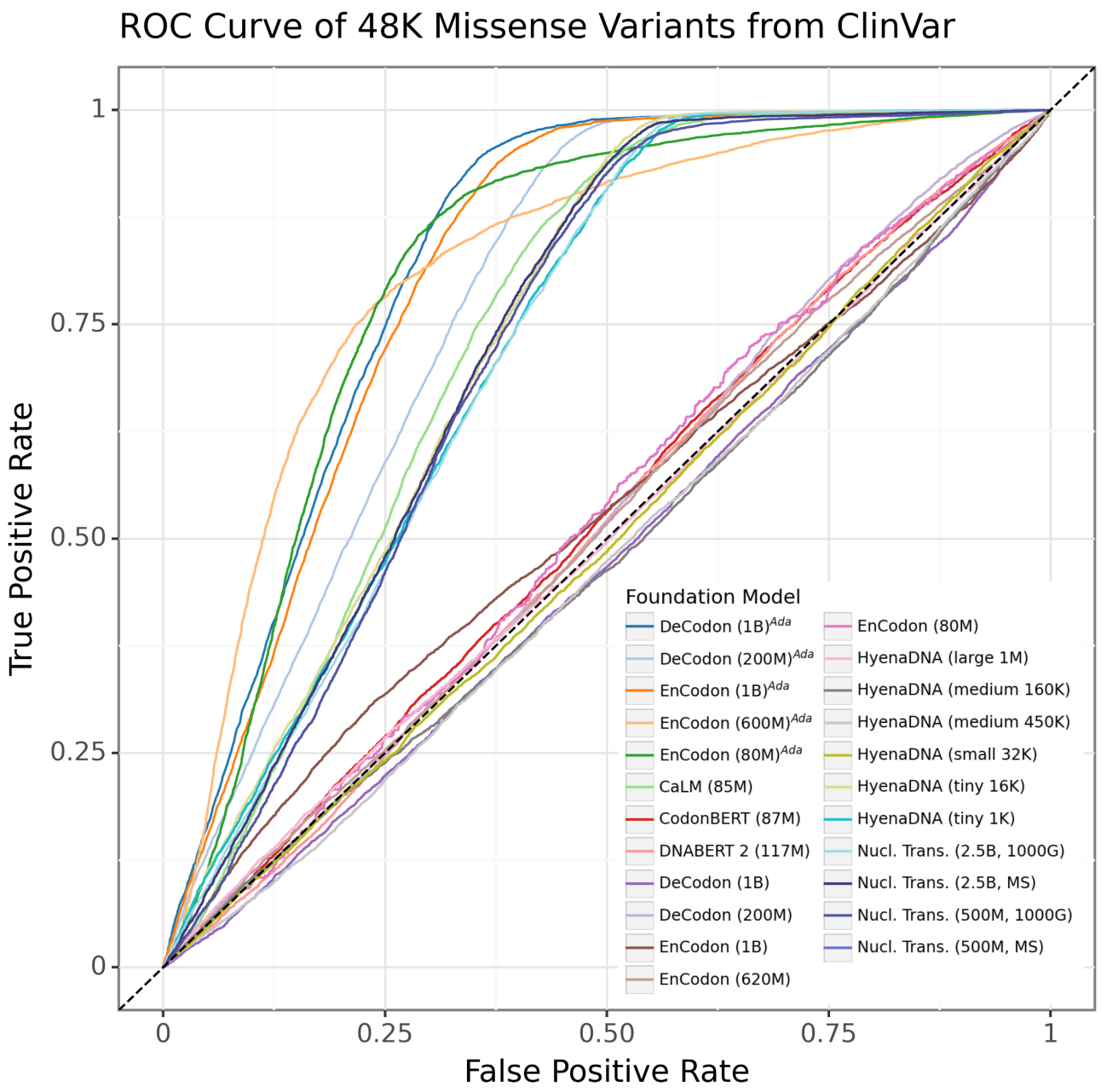
Supplementary Figure 5: **Codon Embedding Space Analysis for pre-trained EnCodons and DeCodons**: **a)** PCA visualization of codon embeddings learned by EnCodon (620M) and DeCodon (150M) colored by Amino Acid. **b)** Violin plots of two cosine distance between pairwise synonynous against non-synonymous codons. **c)** Violin plot of two codon distance metrics i.e. cosine distance in learned embedding space and hamming distance between codon sequences for all possible pairs of codons annotated with spearman correlation between the two metrics. **d)** PCA visualization codon embeddings colored by amino acid's Hydrophobicity Index. **e)** Scatter-plot of pair-wise cosine distance between amino acids and their corresponding PAM250 entry score for pre-trained models. Scatter plot of KNN purity scores of clusters of synonymous codons in learned codon embedding space by **f)** EnCodon and **g)** DeCodon models against different numbers of neighbors.



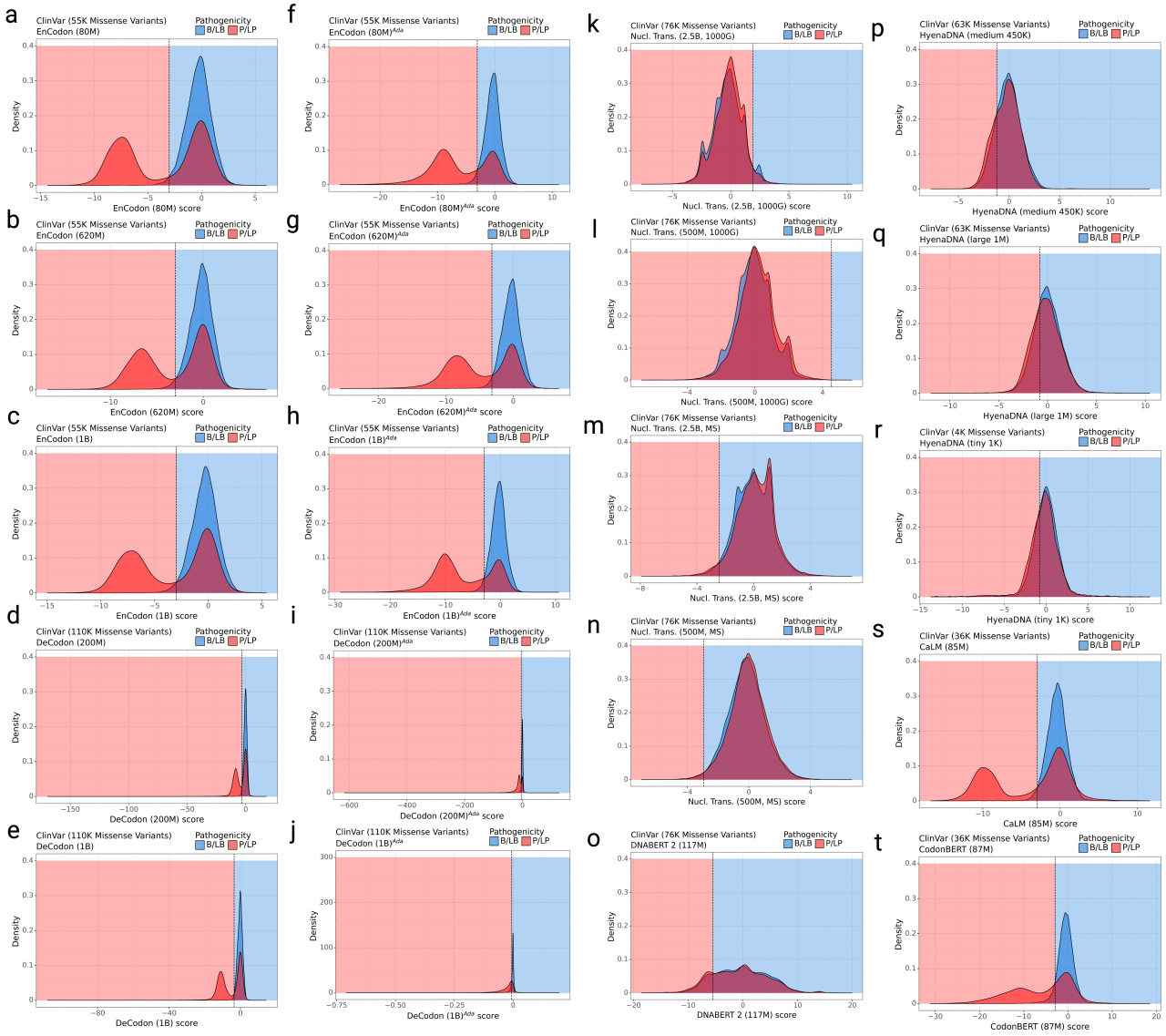
Supplementary Figure 6: **Analysis of DeCodon-generated coding sequences: a)** PCA visualization of EnCodon (1B)^{Ada}'s sequence embedding space to compared generated coding sequences with wild-type and random cohorts in E. coli. **b)** Accuracy bar plots of start and stop codon grammar checks on the generated coding sequences for human and E. coli. **c)** Scatter plot of observed codon usage (in wild-type sequences) against codon usage in generated coding sequences (x-axis) for E. coli. **d)** Comparison of the GC content distribution between generated sequences and natural coding sequences from E. coli (top) and human (bottom), where each distribution is compared with two sets of 10K randomly generated sequences. **e)** Louvain clustering performed on 10,000 sequences generated by DeCodon (1B)^{Ada}, with a t-SNE visualization colored by cluster ID. **f)** Functional region prediction of generated sequences using InterPro and PANTHER, highlighting the top 20 most common Gene Ontology (GO) terms as bars representing log-transformed number of annotated sequences colored by their namespace. **g)** Fisher's Exact Test for GO term enrichment across Louvain clusters, with a heatmap showing significant enrichments ($p\_adjusted < 0.05$) based on the GO namespaces: biological process (BP), molecular function (MF), and cellular component (CC).



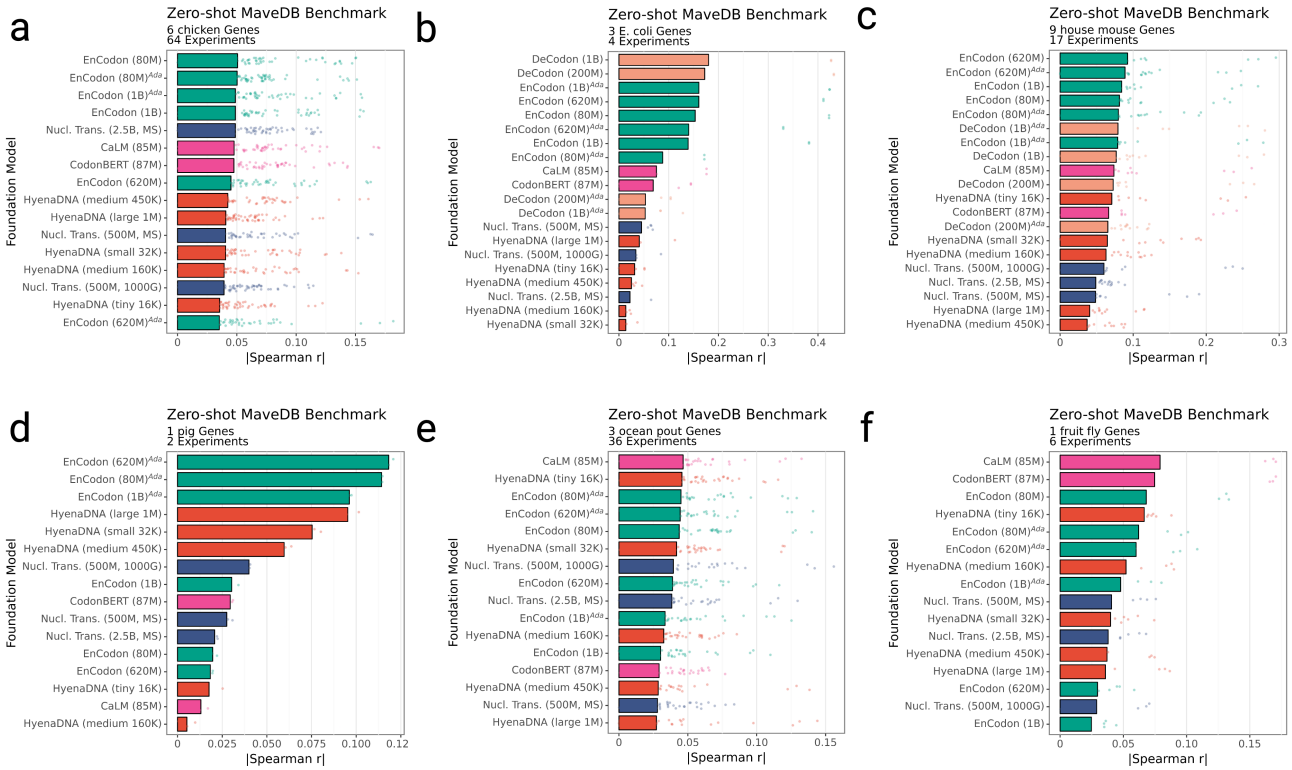
Supplementary Figure 7: DeCodon organism embedding space: PCA visualization of pre-trained DeCodon's organism embedding space for **a)** DeCodon (200M) and **b)** DeCodon (1B)



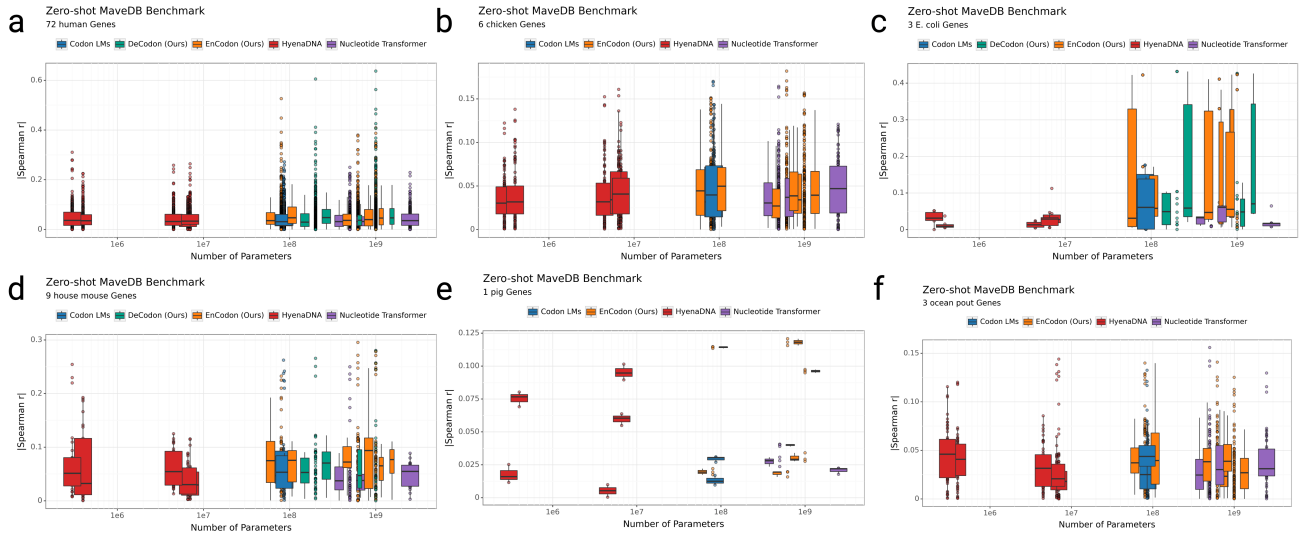
Supplementary Figure 8: Receiver Operating Characteristic (ROC) curve depicting the True Positive Rate (TPR, y-axis) versus the False Positive Rate (FPR, x-axis) for the foundation models evaluated in predicting ClinVar variant pathogenicity. The comparison was standardized by calculating the TPR and FPR on a common set of 48,000 shared missense variants across all models.



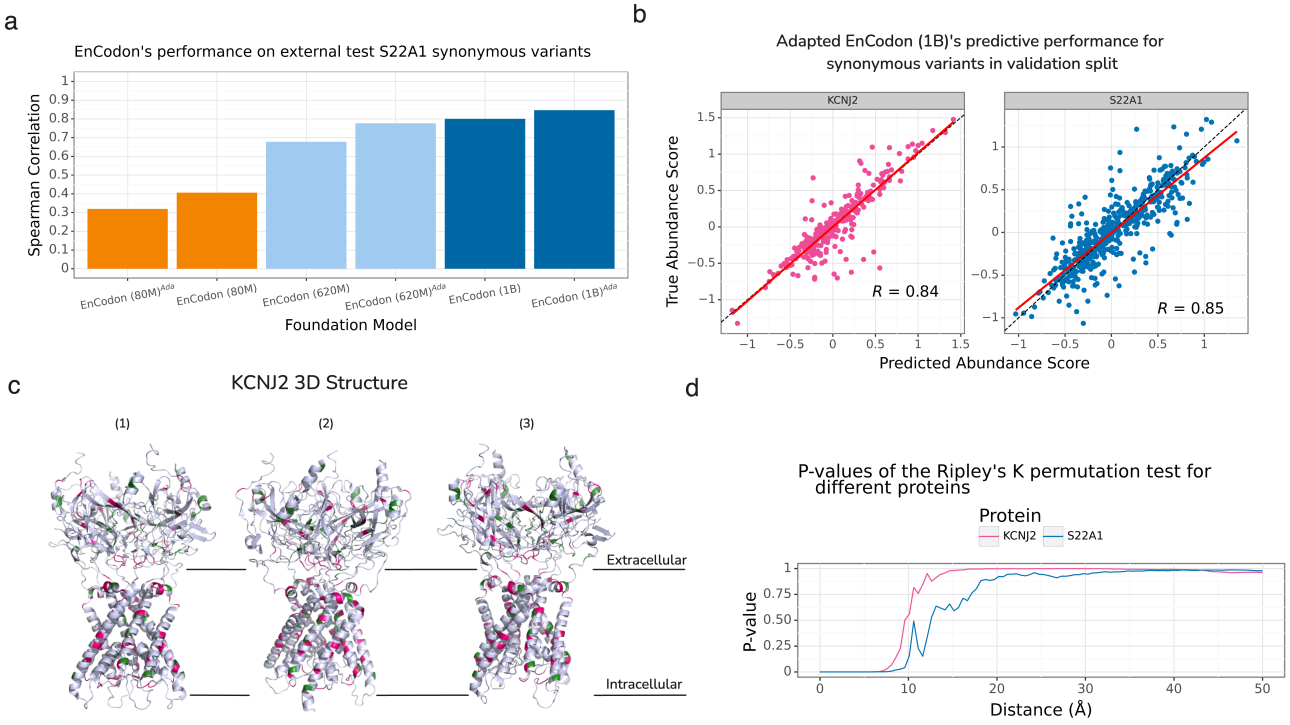
Supplementary Figure 9: Distribution of missense variant scores for all the models used in the zero-shot ClinVar benchmark. Missense Variant scores distribution colored by the consequence of the variant where P/LP and B/LB represents Pathogenic/Likely Pathogenic and Benign/Likely Benign variants. The score distribution is shown for **a)** EnCodon (80M), **b)** EnCodon (620M), **c)** EnCodon (1B), **d)** DeCodon (200M), **e)** DeCodon (1B), **f)** EnCodon (80M)^{Ada}, **g)** EnCodon (620M)^{Ada}, **h)** EnCodon (1B)^{Ada}, **i)** DeCodon (200M)^{Ada}, **j)** DeCodon (1B)^{Ada}, **k)** Nucleotide Transformer (2.5B, 1000G), **l)** Nucleotide Transformer (500M, 1000G), **m)** Nucleotide Transformer (2.5B, MS), **n)** Nucleotide Transformer (500M, MS), **o)** DNABERT 2 (117M), **p)** HyenaDNA (medium 450K), **q)** HyenaDNA (large 1M), **r)** HyenaDNA (tiny 1K), **s)** CaLM (85M), and **t)** CodonBERT (87M).



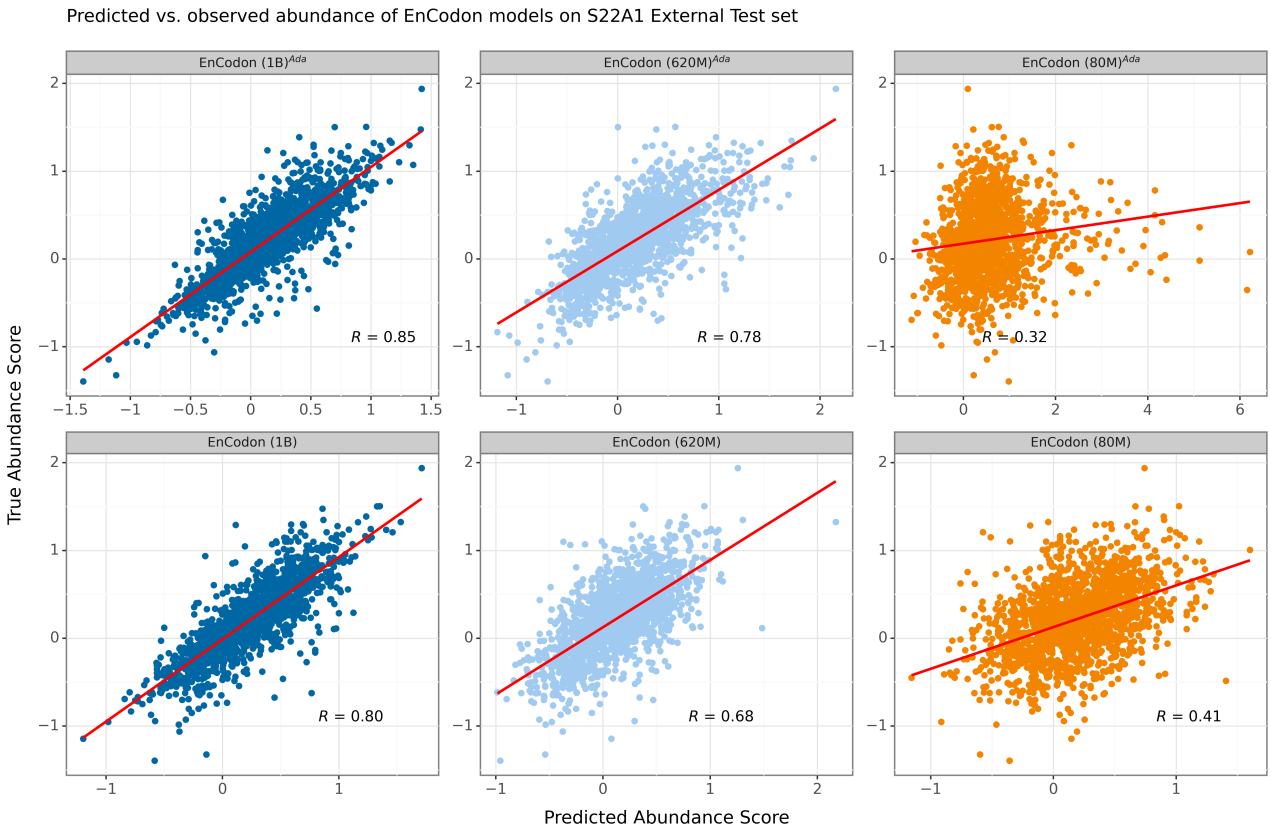
Supplementary Figure 10: Distribution of absolute Spearman correlations (x-axis) for the tested foundation models (y-axis) across different organisms in the Zero-shot MaveDB benchmark: **a)** Chicken, **b)** E. coli, **c)** House mouse, **d)** Pig, **e)** Ocean pout, and **f)** Fruit fly.



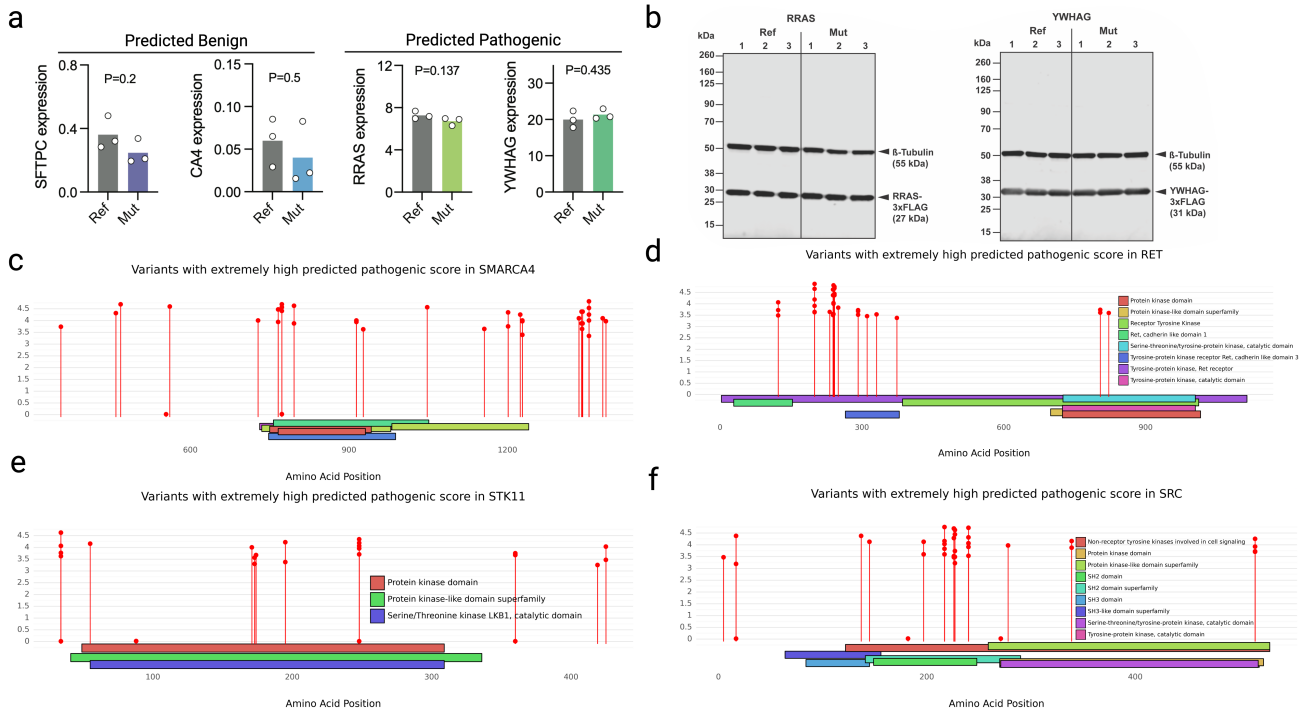
Supplementary Figure 11: Relationship between model size (log-scaled number of trainable parameters, x-axis) and Zero-shot MaveDB performance, reported as the distribution of absolute Spearman correlations (y-axis) for each organism: **a)** Chicken, **b)** E. coli, **c)** House mouse, **d)** Pig, **e)** Ocean pout, and **f)** Fruit fly.



Supplementary Figure 12: **a)** 3D structure visualization of KCNJ2 protein from 3 different angles where critical variants identified by EnCodon (1B) are colored in pink and green. **b)** Performance of EnCodon $(1B)^{Ada}$ in abundance prediction for KCNJ2 (left) and SLC22A1 (right) variants in the validation set. **c)** Bar plot of Speraman correlations of fine-tuned EnCodon models on the external set of SLC22A1 variants. **d)** Line plot of computed p-values at different distance cut-offs for KCNJ2 and SLC22A1 proteins.



Supplementary Figure 13: Scatter plot of all 6 fine-tuned EnCodon models on the external test set of SLC22A1 variants. 3 pre-trained (bottom row) and 3 eukaryotic adapted EnCodon models (top row) were fine-tuned.



Supplementary Figure 14: **a)** Barplot of observed gene expression levels of 4 other tested synonymous variants (2 controls and 2 predicted as pathogenic). **b)** Immunoblots of FLAG-tagged RRAS and YWHAG protein variants expressed in HEK293T cells, showing three biological replicates for both wild-type (Ref) and mutated (Mut) sequences. β-tubulin signal is used as a loading control. Lollipop plots showing potential synonymous variants with extremely pathogenic score for **c)** SMARCA4, **d)** RET, **e)** STK11, and **f)** SRC.

Version	# Variants	# P/LP	# B/LB	# VUS
v0.1	76,051	27,760	48,291	0
v0.2	1,120,127	81,465	98,755	939,907

Supplementary Table 2 | Summary statistics of the different versions of the preprocessed ClinVar dataset. P/LP: Pathogenic/Likely Pathogenic; B/LB: Benign/Likely Benign; VUS: Variant of Uncertain Significance.

HGVS	Gene	PHRED	Prediction
NM_006412.4(AGPAT2):c.702C>T	AGPAT2	0.034	-3.207 (pathogenic)
NM_006270.5(RRAS):c.333C>T	RRAS	1.505	-3.057 (pathogenic)
NM_002872.5(RAC2):c.501C>T	RAC2	0.069	-2.928 (pathogenic)
NM_012479.4(YWHAG):c.564C>T	YWHAG	0.295	-2.667 (pathogenic)
NM_001317778.2(SFTPC):c.228G>C	SFTPC	0.012	0.097 (benign)
NM_000717.5(CA4):c.492G>A	CA4	0.061	0.000 (benign)

Supplementary Table 3 | Nominated synonymous variants, including detailed PHRED scores and predictions from our codon-based model (showing pathogenicity scores and corresponding predicted labels)