

1 Supplementary Information

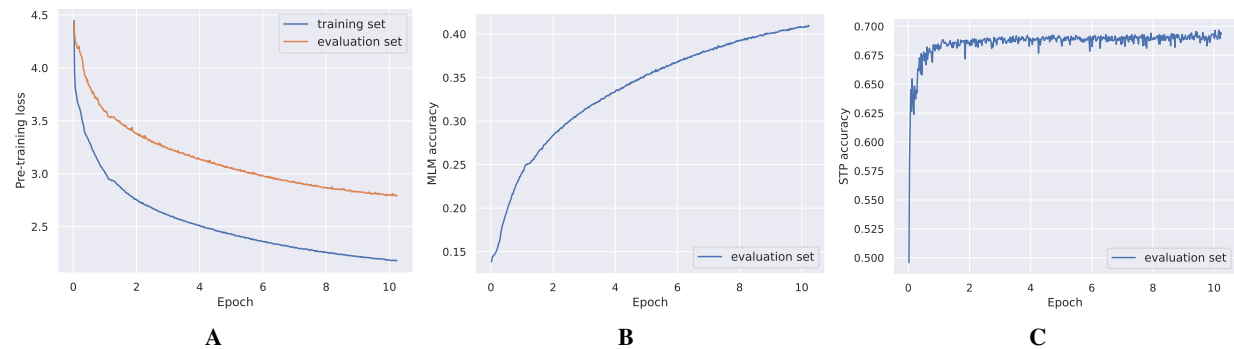


Figure S1: The pre-training curve of CodonBERT.

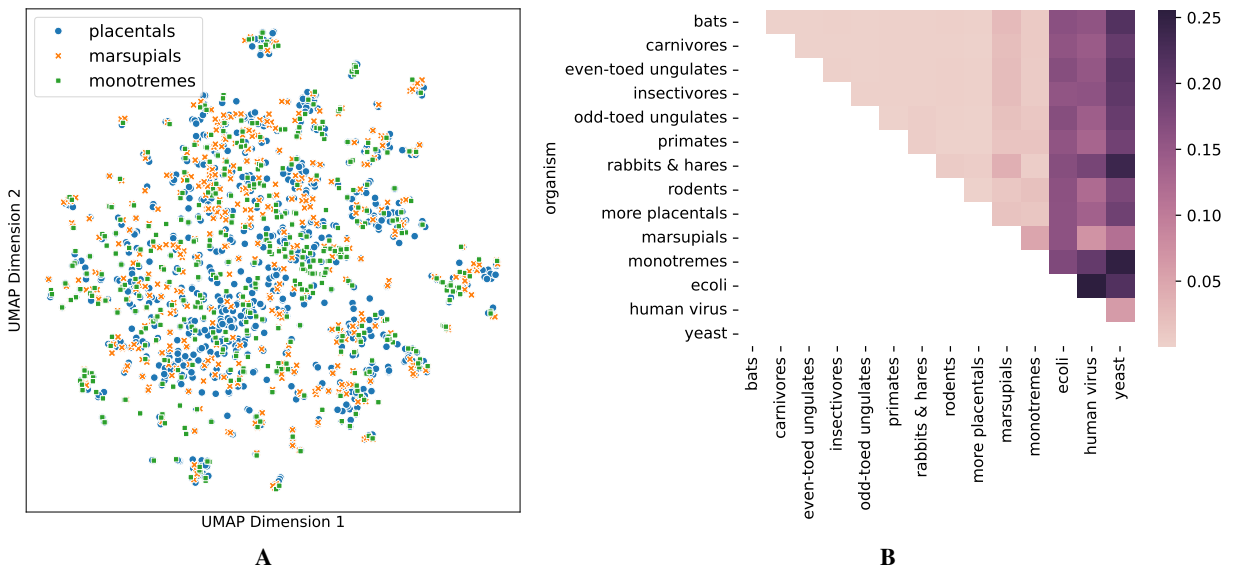


Figure S2: **A:** Projection of the latent space of the sequences from three taxonomic groups within mammals (placentals, marsupials, and monotremes) in the heldout dataset. **B:** Kullback-Leibler divergence of codon usage between different organisms.

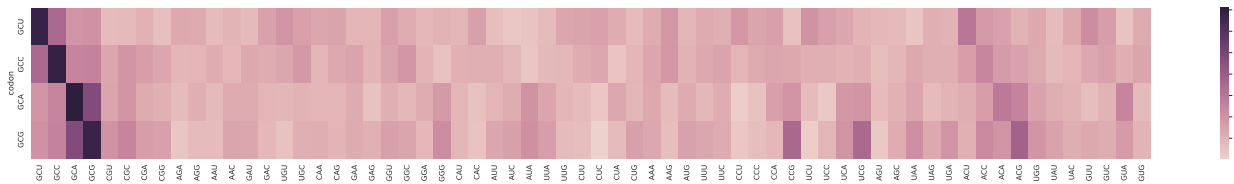


Figure S3: Average cosine similarity between four codons encoding Alanine and all 64 codons.

Model / Dataset	Flu Vaccines	mRFP Expression	Fungal Expression	<i>E. coli</i> Proteins	mRNA Stability	Tc-Riboswitch	CoV Vaccine Degradation
number of seqs	538	1459	7553	6348	41123	355	2400
plain TextCNN	0.26 / 0.72	0.35 / 0.62	3.35 / 0.53	1.09 / 0.39	1.01 / 0.01	0.53 / 0.41	0.017 / 0.55
RNABERT	0.45 / 0.65	0.47 / 0.40	3.85 / 0.41	1.09 / 0.39	0.98 / 0.16	0.44 / 0.47	0.017 / 0.64
RNA-FM	0.36 / 0.71	0.21 / 0.80	3.06 / 0.59	1.05 / 0.43	0.89 / 0.34	0.45 / 0.58	0.015 / 0.74
TF-IDF	0.37 / 0.68	0.43 / 0.57	2.59 / 0.68	- / 0.44	0.68 / 0.54	0.46 / 0.49	0.017 / 0.69
plain TextCNN	0.37 / 0.71	0.21 / 0.78	1.83 / 0.76	1.09 / 0.36	0.59 / 0.26	0.64 / 0.43	0.009 / 0.80
Codon2vec+TextCNN	0.30 / 0.72	0.28 / 0.77	3.04 / 0.61	1.06 / 0.43	0.91 / 0.33	0.43 / 0.56	0.016 / 0.70
CodonBERT	0.28 / 0.78	0.11 / 0.88	0.64 / 0.89	0.92 / 0.57	0.94 / 0.35	0.44 / 0.48	0.012 / 0.78

Table S1: Results of our CodonBERT model against other benchmarks on the test set of seven downstream tasks. For regression tasks, the root mean squared error loss and the corresponding Spearman’s rank correlation are listed. For the classification task (*E. coli* proteins data set), the cross entropy loss and classification accuracy are calculated. The best values of loss, correlation and accuracy for each task are in bold.

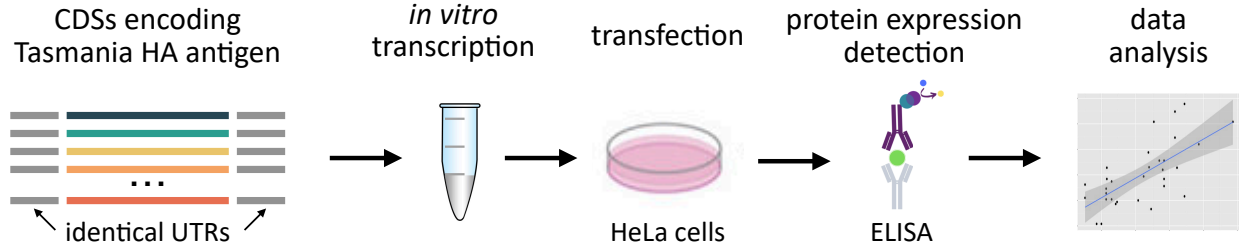


Figure S4: Experimental design for testing in-cell protein expression.

1.1 Comparisons to the original results for the different datasets

mRFP To ensure an apple-to-apple comparison, we fine-tuned the pre-trained CodonBERT model on the same training set and evaluated the model on the same testing set as reported in the paper. Specifically, 10% of the data points were reserved for the testing set, ensuring that both training and test subsets had the same proportions of the three libraries as the entire dataset. CodonBERT achieved a higher correlation (Pearson $r=0.840$ compared to the best model used in the original paper, which was a random forest regressor ($r=0.762$)).

***E. coli* Proteins** Following the original publication’s protocol, we conducted a binary classification using 10-fold cross-validation, achieving an Area Under the Curve (AUC) of 0.76. This result is similar to the performance reported in the initial study, though classification may be an easier task than regression for this data.

mRNA stability Utilizing the same training/testing split as the original study for our mRNA stability model, we achieved superior results. While the paper reported a spearman correlation of 0.366, using CodonBERT we improved the results to 0.395.

Degradataion A direct comparison with the degradation model presented in the referenced study was not feasible due to methodological differences in the level of analysis. The original model’s performance metrics are based on nucleotide-level predictions, whereas our model operates at the sequence-level.

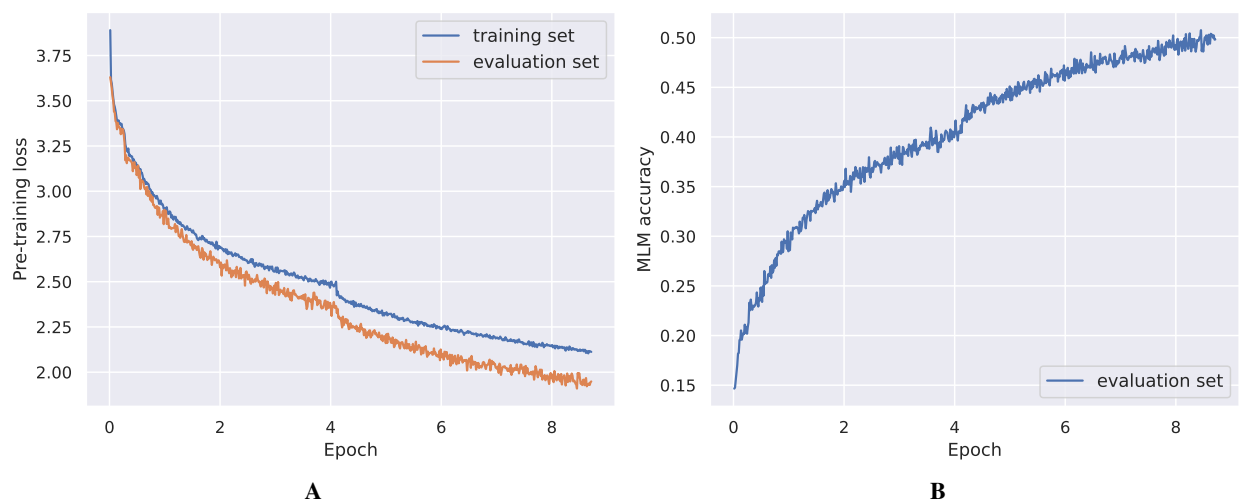


Figure S5: Pre-training CodonBERT model with masked language model task alone.

1.2 Ablation experiment for CodonBERT

We conducted further analysis by excluding the homology detection optimization and adhering to the same training protocol. The outcomes are illustrated in table S2. In comparing CodonBERT trained solely with Masked Language Modeling (MLM) to those trained with both MLM and Sequence Tagging Prediction (STP), we noted a modest improvements on the datasets for mRFP, Flu vaccines, Tc Riboswitch, and degradation. Additionally, significant enhancements were observed in the fungal expression dataset and the *E. coli* Protein dataset. The loss curves over the training period for the model trained with the MLM task alone is shown in Fig. S5.

Model	Flu Vaccines	mRFP Expression	Fungal Expression	<i>E. coli</i> Proteins	mRNA Stability	Tc- Riboswitch	SARS-CoV-2 Vaccine Degradation
MLM + STP	0.78	0.88	0.89	0.57	0.35	0.48	0.78
MLM	0.75	0.86	0.80	0.50	0.40	0.47	0.75

Table S2: Comparison of the performance of two models on seven downstream tasks. MLM and STP represent masked language model and sequence taxonomy prediction, respectively.