CodonBERT: Large Language Models for mRNA design and optimization

Sizhen Li[*,†] **Saeed Moayedpour**[*,†] **Ruijiang Li**[‡] **Michael Bailey**[*] **Saleh Riahi**[*]

Lorenzo Kogler-Anele[*] **Milad Miladi**[§] **Jacob Miner**[§] **Dinghai Zheng**[§] **Jun Wang**[§]

Akshay Balsubramani[§] **Khang Tran**[§] **Minnie Zacharia**[§] **Monica Wu**[§] **Xiaobo Gu**[§]

Ryan Clinton[§] **Carla Asquith**[§] **Joseph Skaleski**[§] **Lianne Boeglin**[§] **Sudha Chivukula**[§]

Anusha Dias[§] **Fernando Ulloa Montoya**[¶] **Vikram Agarwal**[§]

Ziv Bar-Joseph[*,‖]
ziv.bar-joseph@sanofi.com

Sven Jager[*,‖]
sven.jager@sanofi.com

Abstract

mRNA-based vaccines and therapeutics are gaining popularity and usage across a wide range of conditions. One of the critical issues when designing such mRNAs is sequence optimization. Even small proteins or peptides can be encoded by an enormously large number of mRNAs. The actual mRNA sequence can have a large impact on several properties including expression, stability, immunogenicity, and more. To enable the selection of an optimal sequence, we developed CodonBERT, a large language model (LLM) for mRNAs. Unlike prior models, CodonBERT uses codons as inputs which enables it to learn better representations. CodonBERT was trained using more than 10 million mRNA sequences from a diverse set of organisms. The resulting model captures important biological concepts. CodonBERT can also be extended to perform prediction tasks for various mRNA properties. CodonBERT outperforms previous mRNA prediction methods including on a new flu vaccine dataset.

1 Introduction

mRNA vaccines have emerged as a high potency, fast production, low-cost, and safe alternative to traditional vaccines [1–4]. The expression level of a vaccine directly affects its potency, ultimate immunogenicity, and efficacy [5]. The higher the level of expression of the antigenic protein encoded by the mRNA sequence, the smaller amount of the vaccine is needed to achieve the desired immune response, which can make the vaccine more cost-effective and easier to manufacture [1].

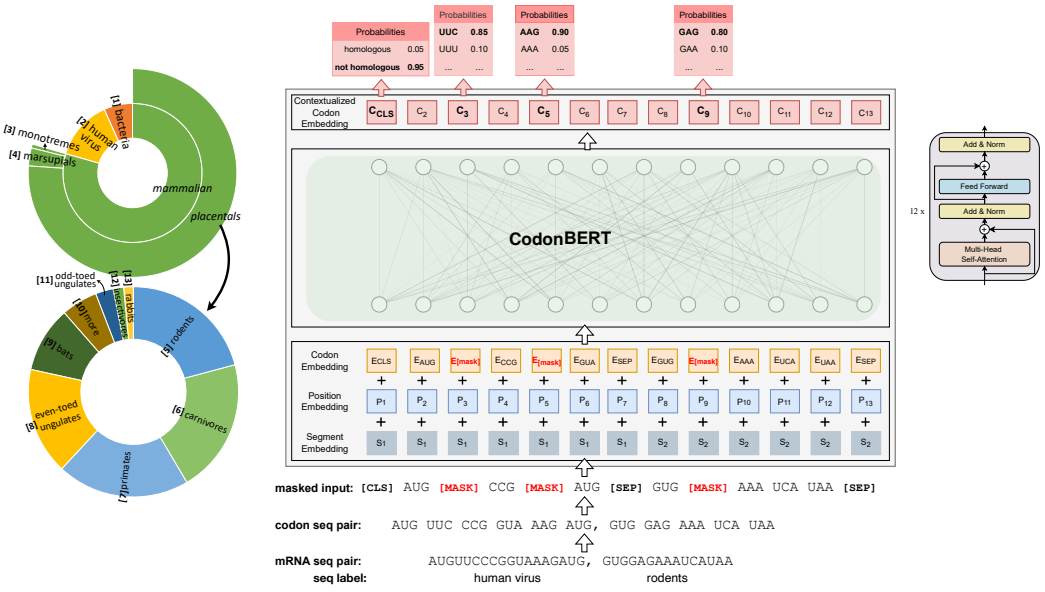
[*]Digital R&D, Sanofi, Cambridge, MA, USA
[†]Equally contributed
[‡]Digital Data, Sanofi, Shanghai, China
[§]mRNA Center of Excellence, Sanofi, Waltham, MA, USA
[¶]mRNA Center of Excellence, Sanofi, Marcy L'Etoile, France
[‖]Corresponding authors

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

A human protein with an average length of 500 amino acids can be encoded by roughly 3^{500} different codon sequences. While only one of those is encoded in the virus or DNA of interest, this is not necessarily the optimal sequence for a vaccine. The classical method to find the optimal mRNA sequence is codon optimization, which selects the most optimal codon for each amino acid using the codon bias in the host organism [6]. This method has been widely applied for optimizing recombinant protein drugs, nucleic acid therapies, gene therapy, mRNA therapy, and DNA/RNA vaccines [7–9]. However, codon optimization alone does not consider several key properties that impact protein expression [10]. For instance, RNA structural motifs (e.g., stem loops and pseudoknots) have been shown to play a major role for non-coding RNAs (such as riboswitches or aptamers) [11, 12].

Pre-training a large language model (LLM) has been scaled to biological sequences (protein, DNA, and RNA) [13–17]. However, as we show, such LLMs may not be ideal for predicting protein expression due to their focus on individual nucleotides and non-coding regions.

We developed CodonBERT, an LLM which extends the BERT model [18] and applies it to the language of mRNAs, which uses a multi-head attention transformer architecture framework. The pre-trained model can also be generalized to a diverse set of supervised-learning tasks. We pre-trained CodonBERT using 10 million mRNA coding sequences spanning an evolutionarily diverse set of organisms. Next, we used it to perform several mRNA prediction tasks, including protein expression and mRNA degradation prediction. As we show, both the pre-trained and fine-tuned version of models can learn biological information and improve on current state-of-the-art methods for mRNA vaccine design.



(a) (b) (c)

Figure 1: (a) Hierarchically classified mRNA sequences for pre-training. All the 13 leaf-level classes are numbered. The angle of each segment is proportional to the number of sequences belonging to this group. (b) Model architecture and training scheme deployed for two tasks of CodonBERT. (c) A stack of 12 transformer blocks employed in CodonBERT model.

2 Methods

2.1 Data for pre-training and evaluation

We collected mRNA sequences across diverse organisms for pre-training from NCBI [19]. The datasets included *mammalian* reference sequences [1], *bacteria* (*Escherichia coli*) reference sequences [2],

[1] https://www.ncbi.nlm.nih.gov/datasets/taxonomy/40674/
[2] https://www.ncbi.nlm.nih.gov/datasets/taxonomy/562/

and *homo sapiens virus* complete nucleotides [3]. To evaluate the prediction accuracy of the LLM, we collected several functional mRNA datasets. We also generated a new dataset consisting of mRNA sequences that encode the Influenza H3N2 A/Tasmania/503/2020 hemagglutinin protein. As illustrated in Fig. S5, sequences corresponding to these candidates were synthesized as gene fragments and PCR amplified to generate template DNA for high-throughput in vitro transcription reactions. HeLa cells were used to evaluate the expression of the protein encoded by different mRNA sequences.

2.2 Model architecture

As shown in Figure 1(b), CodonBERT takes a sequence pair as input and concatenates them using a separator token ([SEP]). It then adds a classifier token ([CLS]) and a separator token ([SEP]) at the beginning and end of the combined sequence, respectively. CodonBERT constructs the input embedding by concatenating codon, position, and segment embeddings.

The combined input embedding is fed into the CodonBERT model, which consists of a stack of 12 layers of bidirectional transformer encoders [20] as shown in Figure 1(c). Each transformer layer processes its input using 12 self-attention heads, and outputs a representation for each position with hidden size 768. In each layer, the multi-head self-attention mechanism captures the contextual information of the input sequence by considering all the other codons in the sequence. A key benefit of self-attention mechanism is the connection learned between all pairs of positions in an input sequence using parallel computation which enables CodonBERT to model not only short-range but also long-range interactions, which impact translation efficiency and stability [21]. Next a feed-forward neural network is added to apply a non-linear transformation to the output hidden representation from the self-attention network. A residual connection is employed around each of the multi-head attention and feed-forward networks. After processing the input sequence with a stack of transformer encoders, CodonBERT produces the final contextualized codon representations, which is followed by a classification layer to produce probability distribution over the vocabulary.

2.3 Pre-training CodonBERT

Pre-training CodonBERT performs two tasks: Masked Language Model (MLM) and Homologous Sequences Prediction (HSP). A fraction of input codons (15%) are randomly selected and replaced by the masking token ([MASK]). The self-training loop optimizes CodonBERT to predict the masked codons based on the remaining ones. A probability distribution over 64 possible codons is produced by CodonBERT for the masked positions. The average cross entropy loss of the masked language model $\mathcal{L}_{\text{MLM}}$ over the masked positions M is calculated by the optimization function:

$$\mathcal{L}_{\text{MLM}} = -\frac{1}{|X|}\frac{1}{|M|}\sum_{x\in X}\sum_{i\in M}\log p(x_i \mid x_M) \tag{1}$$

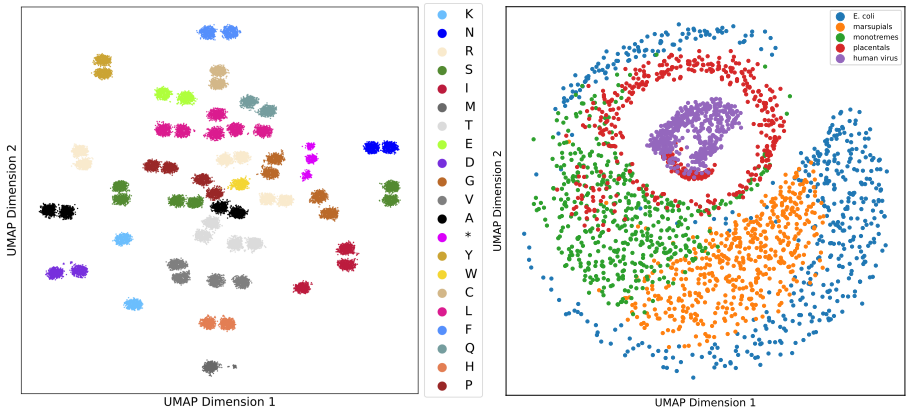
Where X represents a batch of sequences, x is one sequence and x_i is the original codon for the position i. x_M is the masked input with a set of positions M masked. $p(x_i \mid x_M)$ indicates the output probability of the real codon x_i given all the remaining codons in the masked sequence x_M.

Besides, The output embedding of the classifier token ([CLS]) is used for predicting whether these two input sequences belong to the same class, i.e., homologous sequences. The average cross entropy loss of the homologous sequences prediction task $\mathcal{L}_{\text{HSP}}$ is computed as:

$$\mathcal{L}_{\text{HSP}} = -\frac{1}{N}\sum_{n=1}^{N}[y_n \log p_n + (1-y_n)\log(1-p_n)] \tag{2}$$

Where N represents the number of sequence pairs. y_n is the expected value, which is 1 when two sequences are homologous and 0 when they are not. p_n indicates the predicted probability of two sequences belonging to the same category. Thus, the total loss is $\mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{HSP}}$.

CodonBERT was also applied to a wide range of downstream tasks. For this we can use either a single or a pair of sequences as input (Figure 1(b) and Figure S3(c)). To perform supervised analysis, the output embedding is followed by an output layer which is trained for the specific task.



(a) (b)

Figure 2: Genetic code and evolutionary homology information learned by pre-trained, unsupervised CodonBERT model. High-dimensional codon (a) and sequence (b) embeddings were projected into 2-dimensional space using UMAP [22]. (a) Each point represents a codon with different contexts, and its color corresponds to the type of amino acid accordingly. (b) Each point is a mRNA sequence, and its color represents the sequence label.

3 Results

3.1 Pre-trained Representation Model

We pre-trained CodonBERT using 10 million mRNA sequences (Methods). Pre-training on 4 A10G GPUs with 96 GB GPU memory and 192 GB memory took roughly two weeks. To assess the pre-trained CodonBERT model, we built a held-out dataset by randomly leaving out 1% of mRNA sequences for each category and trained the model with the remaining sequences. During the pre-training phase, the model performance on two tasks (MLM and HSP) is substantially improved on the losses and accuracies of the model prediction as shown in Fig. S1.

In addition to the quantitative evaluation of model predictions, we also performed several qualitative analyses of the embeddings provided by CodonBERT. To decipher what kind of biological information has been learned by the model and encoded in the codon and sequences representations. 2D projections of the codon and sequence embeddings for the held-out dataset is presented in Figure 2(a–b). As can be seen, codons that encode the same amino acid, i.e., synonymous codons, are spatially close to each other, which indicates that CodonBERT learns the genetic code from the large-scale training set. Figure 2(b) shows clusters for five high-level sequence categories: *E. coli*, human virus, and three subgroups of mammals. Homologous sequences are clustered together with clear boundaries between the homology classes. As the largest and most developed group within mammals, placentals are further split into eight specific categories (Figure 1(a)). Clustering of the embeddings corresponding to these eight subgroups is compact and well-separated. These clear cluster patterns implies that CodonBERT can learn the homologous information from the millions of mRNA sequences across diverse families and organisms.

3.2 Evaluating CodonBERT and comparison to prior methods on supervised learning tasks

CodonBERT can be extended to perform supervised learning for specific mRNA prediction tasks. To evaluate the use of our LLM for downstream tasks and to compare it to prior methods, we collected several mRNA property prediction datasets. Table S1 presents a diverse set of downstream tasks related to mRNA translation, stability, and regulation [23–28]. In addition, these datasets represent a range of molecules, including newly published data sets for recombinant protein, bio-computing, and SARS-CoV-2 vaccine design. Finally, we generated a new dataset to test CodonBERT in the context of mRNAs encoding the influenza hemagglutinin antigen for Flu vaccines (Fig. S5).

[3]https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/

	Model	Flu Vaccines	mRFP Expression	Fungal Expression	*E. coli* Proteins	mRNA Stability	Tc-Riboswitch	SARS-CoV-2 Vaccine Degradation
Nucleotide -based	plain TextCNN	0.72	0.62	0.53	0.39	0.01	0.41	0.55
	RNABERT	0.65	0.40	0.41	0.39	0.16	0.47	0.64
	RNA-FM	0.71	0.80	0.59	0.43	0.34	**0.58**	0.74
Codon -based	TF-IDF	0.68	0.57	0.68	0.44	**0.54**	0.49	0.69
	plain TextCNN	0.71	0.78	0.76	0.36	0.26	0.43	**0.80**
	Codon2vec	0.72	0.77	0.61	0.43	0.33	0.56	0.70
	CodonBERT	**0.78**	**0.88**	**0.89**	**0.57**	0.35	0.48	0.78

Table 1: Comparison of CodonBERT to prior methods on seven downstream tasks. For regression tasks, the corresponding Spearman's rank correlation values are listed. For the classification task (*E. coli* proteins data set), classification accuracy is calculated. The best values of correlation and accuracy for each task are in bold. The corresponding MSE loss and cross entropy loss is listed in Table S2.

To assess CodonBERT's performance on these tasks, we have also applied several other state-of-the-art methods that have been previously used for mRNA property prediction with different model complexities, including TF-IDF [29], TextCNN [30], Codon2vec [31], RNABERT [15], and RNA-FM [16]. Detailed training information for these models are shown in Fig. S3. Table 1 presents the performance of CodonBERT and other six methods on these downstream tasks. Note that the first three rows are nucleotide-based methods , while the rest are codon-based methods.

Overall, we see that codon-based methods outperform nucleotide-based methods on most tasks (Fig. S4). This is in part due to the critical role of codons on the protein expression. For example, and CodonBERT improved on RNABERT and RNA-FM, which are nuecleotide-based LLMs trained on non-coding RNAs, on protein expression tasks. Moreover, the codon-based variant of TextCNN also outperforms the original nucleotide implementation on most tasks. As for the detailed comparison, we observe that CodonBERT performed best on four of the seven tasks and second best (in most cases with very small difference) on two of the remaining three tasks.

4 Discussion

To enable the analysis and prediction of mRNA properties, we utilized 10 million mRNA coding sequences (CDS) from several species to train a large language model (CodonBERT), and to establish a foundational model. Projection of codon embedding obtained from CodonBERT produces distinct clusters that adhere to the amino acid types. In-depth analysis of CodonBERT representation of a set of genes from different organisms revealed that CodonBERT autonomously learns the genetic code and principles of evolutionary homology and aligns with our understanding of genetic evolution.

We also utilized CodonBERT to perform several supervised prediction tasks for mRNA properties. Our results indicate that CodonBERT is the top performing method overall and ranks first or second in performance for six of the seven tasks. The one exception in terms of performance was observed for the mRNA stability tasks. Stability is known to be structure-dependent, and stable structures such as stem-loops or hairpin structures can impede degradation enzymes, protecting the mRNA from rapid decay. A possible reason for the reduction in performance for these datasets is that structural properties are highly dependent on nucleotides whereas CodonBERT is a codon-based model. One possible solution for this is a model that combines codon and nucleotide representation. Similarly, mRNA modification events including capping at the $5'$ end and polyadenylation at the $3'$ end in eukaryotes are not currently encoded in our model but can also impact mRNA stability.

To conclude, our findings suggest that CodonBERT could serve as a versatile and foundational model for the development of new mRNA-based vaccines and the engineering and recombinant production of industrial and therapeutic proteins.

References

[1] Norbert Pardi, Michael J. Hogan, Frederick W. Porter, and Drew Weissman. mRNA vaccines — a new era in vaccinology. *Nature Reviews Drug Discovery*, 17(4):261–279, 2018.

[2] Norbert Pardi, Michael J Hogan, and Drew Weissman. Recent advances in mRNA vaccine technology. *Current opinion in immunology*, 65:14–20, 2020.

[3] Nicholas AC Jackson, Kent E Kester, Danilo Casimiro, Sanjay Gurunathan, and Frank DeRosa. The promise of mRNA vaccines: a biotech and industrial perspective. *npj Vaccines*, 5(1):11, 2020.

[4] Cuiling Zhang, Giulietta Maruggi, Hu Shan, and Junwei Li. Advances in mRNA vaccines for infectious diseases. *Frontiers in immunology*, page 594, 2019.

[5] Thomas Schlake, Andreas Thess, Mariola Fotin-Mleczek, and Karl-Josef Kallen. Developing mRNA-vaccine technologies. *RNA Biology*, 9(11):1319–1330, 2012. doi: 10.4161/rna.22269. URL `https://doi.org/10.4161/rna.22269`. PMID: 23064118.

[6] Vincent P. Mauro and Stephen A. Chappell. A critical analysis of codon optimization in human therapeutics. *Trends in Molecular Medicine*, 20(11):604–613, 2014. ISSN 1471-4914. doi: https://doi.org/10.1016/j.molmed.2014.09.003. URL `https://www.sciencedirect.com/science/article/pii/S1471491414001403`.

[7] Adnan B. Al-Hawash, Xiaoyu Zhang, and Fuying Ma. Strategies of codon optimization for high-level heterologous protein expression in microbial expression systems. *Gene Reports*, 9:46–53, 2017. ISSN 2452-0144. doi: https://doi.org/10.1016/j.genrep.2017.08.006. URL `https://www.sciencedirect.com/science/article/pii/S2452014417300614`.

[8] Gina R. Webster, Audrey Y.-H. Teh, and Julian K.-C. Ma. Synthetic gene design—the rationale for codon optimization and implications for molecular pharming in plants. *Biotechnology and Bioengineering*, 114(3):492–502, 2017. doi: https://doi.org/10.1002/bit.26183. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/bit.26183`.

[9] Vincent P. Mauro. Codon optimization in the production of recombinant biotherapeutics: Potential risks and considerations. *BioDrugs*, 32(1):69–81, 2018. doi: 10.1007/s40259-018-0261-x. URL `https://doi.org/10.1007/s40259-018-0261-x`.

[10] Annabel HA Parret, Hüseyin Besir, and Rob Meijers. Critical reflections on synthetic gene design for recombinant protein expression. *Current Opinion in Structural Biology*, 38:155–162, 2016.

[11] Michael Schmidt, Kay Hamacher, Felix Reinhardt, Thea S. Lotz, Florian Groher, Beatrix Suess, and Sven Jager. SICOR: Subgraph isomorphism comparison of rna secondary structures. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 17(6):2189–2195, 2020. doi: 10.1109/TCBB.2019.2926711.

[12] Florian Groher, Cristina Bofill-Bosch, Christopher Schneider, Johannes Braun, Sven Jager, Katharina Geißler, Kay Hamacher, and Beatrix Suess. Riboswitching with ciprofloxacin—development and characterization of a novel RNA regulator. *Nucleic Acids Research*, 46(4):2121–2132, 01 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx1319. URL `https://doi.org/10.1093/nar/gkx1319`.

[13] 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.

[14] Tristan Bepler and Bonnie Berger. Learning the protein language: Evolution, structure, and function. *Cell systems*, 12(6):654–669, 2021.

[15] Manato Akiyama and Yasubumi Sakakibara. Informative RNA base embedding for RNA structural alignment and clustering by deep representation learning. *NAR genomics and bioinformatics*, 4(1):lqac012, 2022.

[16] Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, Irwin King, and Yu Li. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *bioRxiv*, 2022. doi: 10.1101/2022.08.06.503062. URL https://www.biorxiv.org/content/early/2022/08/07/2022.08.06.503062.1.

[17] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 02 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab083. URL https://doi.org/10.1093/bioinformatics/btab083.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[19] David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 35 (suppl_1):D5–D12, 2007.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[21] Jong Ghut Ashley Aw, Yang Shen, Andreas Wilm, Miao Sun, Xin Ni Lim, Kum-Loong Boon, Sidika Tapsin, Yun-Shen Chan, Cheng-Peow Tan, Adelene YL Sim, et al. In vivo mapping of eukaryotic RNA interactomes reveals principles of higher-order organization and regulation. *Molecular cell*, 62(4):603–617, 2016.

[22] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction, 2020.

[23] Thijs Nieuwkoop, Barbara R Terlouw, Katherine G Stevens, Richard A Scheltema, Dick de Ridder, John van der Oost, and Nico J Claassens. Revealing determinants of translation efficiency via whole-gene codon randomization and machine learning. *Nucleic Acids Research*, 51(5):2363–2376, 01 2023. ISSN 0305-1048. doi: 10.1093/nar/gkad035. URL https://doi.org/10.1093/nar/gkad035.

[24] Rhondene Wint, Asaf Salamov, and Igor V Grigoriev. Kingdom-Wide Analysis of Fungal Protein-Coding and tRNA Genes Reveals Conserved Patterns of Adaptive Evolution. *Molecular Biology and Evolution*, 39(2), 01 2022. ISSN 1537-1719. doi: 10.1093/molbev/msab372. URL https://doi.org/10.1093/molbev/msab372. msab372.

[25] Zundan Ding, Feifei Guan, Guoshun Xu, Yuchen Wang, Yaru Yan, Wei Zhang, Ningfeng Wu, Bin Yao, Huoqing Huang, Tamir Tuller, et al. Mpepe, a predictive approach to improve protein expression in e. coli based on deep learning. *Computational and Structural Biotechnology Journal*, 20:1142–1153, 2022.

[26] Ann-Christin Groher, Sven Jager, Christopher Schneider, Florian Groher, Kay Hamacher, and Beatrix Suess. Tuning the performance of synthetic riboswitches using machine learning. *ACS Synthetic Biology*, 8(1):34–44, 2019. doi: 10.1021/acssynbio.8b00207. URL https://doi.org/10.1021/acssynbio.8b00207.

[27] Michay Diez, Santiago Gerardo Medina-Muñoz, Luciana Andrea Castellano, Gabriel da Silva Pescador, Qiushuang Wu, and Ariel Alejandro Bazzini. iCodon customizes gene expression based on the codon composition. *Scientific Reports*, 12(1):1–16, 2022.

[28] Hannah K Wayment-Steele, Wipapat Kladwang, Andrew M Watkins, Do Soon Kim, Bojan Tunguz, Walter Reade, Maggie Demkin, Jonathan Romano, Roger Wellington-Oguri, John J Nicol, et al. Deep learning models for predicting RNA degradation via dual crowdsourcing. *Nature Machine Intelligence*, pages 1–11, 2022.

[29] Anand Rajaraman and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, 2011.

[30] Yoon Kim. Convolutional neural networks for sentence classification, 2014.

[31] Rhondene Wint, Asaf Salamov, and Igor V Grigoriev. Kingdom-wide analysis of fungal protein-coding and trna genes reveals conserved patterns of adaptive evolution. *Molecular biology and evolution*, 39(2):msab372, 2022.

Supplemental Information

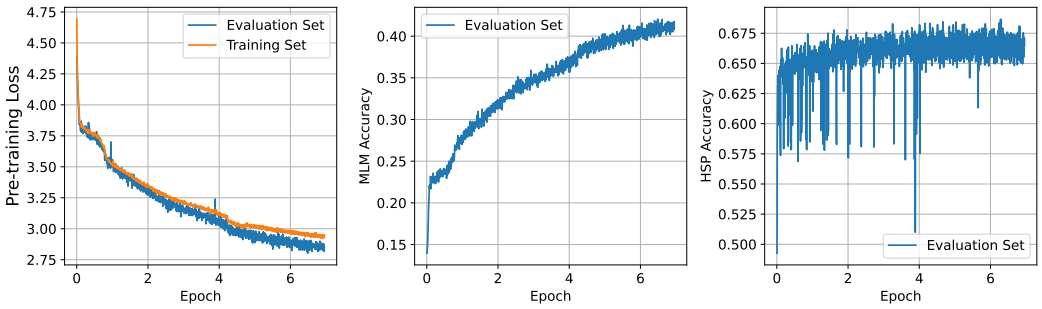
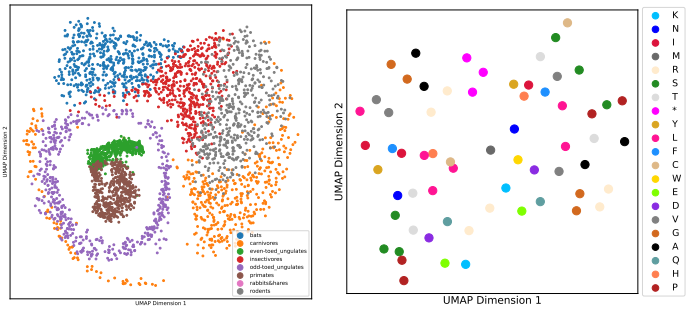


Figure S1: The pre-training curve of CodonBERT.



(a) **(b)**

Figure S2: (a) Projected sequence embedding from the pre-trained CodonBERT model. Each point is a mRNA sequence, and its color represents the sequence label. (b) Projected codon embedding from the pre-trained Codon2vec model. Each point shows a codon, and its color is the corresponding amino acid.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Data Set	Target	category	# mRNAs	seq length
MLOS Flu Vaccines (Sanofi-Aventis)	Expression	Regression	543	1698 – 1704
mRFP Expression	Expression	Regression	1459	678 – 678
Fungal expression	Expression	Regression	7056	150 – 3000
E. coli proteins	Expression	Classification	6348	171 – 3000
Tc-Riboswitches	Switching factor	Regression	355	67 – 73
mRNA stability	Stability	Regression	41123	30 – 1497
SARS-CoV-2 Vaccine Degradation	Degradation	Regression	2400	107 – 107

Table S1: The collection of the datasets with their corresponding mRNA source and property used for method evaluation. Each dataset is split into training, validation, and test with 0.7, 0.15, 0.15 ratio. All the methods were optimized on the same data split.

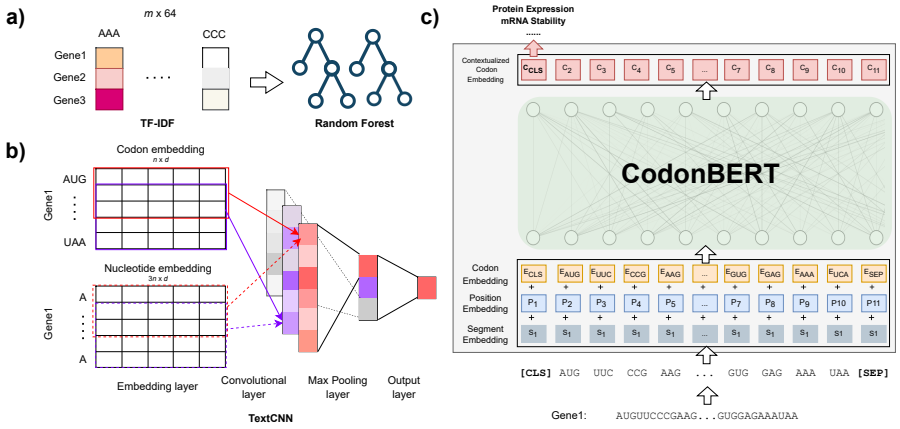


Figure S3: Comparison to prior methods (TF-IDF, Codon2vec, RNABERT and RNA-FM) and fine-tuning CodonBERT on downstream data sets. a) Given an input corpus with m mRNA sequences, TF-IDF is used to construct a feature matrix followed by a random forest regression model. b) Use a TextCNN model to learn task-specific nucleotide or codon representations. The model is able to fine-tune pre-trained representations by initializing the embedding layers with stacked codon or nucleotide embeddings extracted from pre-trained language models (Codon2vec, RNABERT, and RNA-FM). n is the number of codons in the input sequence and d is the dimension of the token embedding. As baseline, plain TextCNN initializes the embedding layer with a standard normal distribution. c) Fine-tune the pre-trained CodonBERT model on a given downstream task directly by keeping all the parameters trainable.

Model / Dataset	Flu Vaccines	mRFP Expression	Fungal Expression	*E. coli* 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	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 S2: Results of our CodonBERT model against other benchmarks on the test set of seven downstream tasks. For regression tasks, the MSE 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.

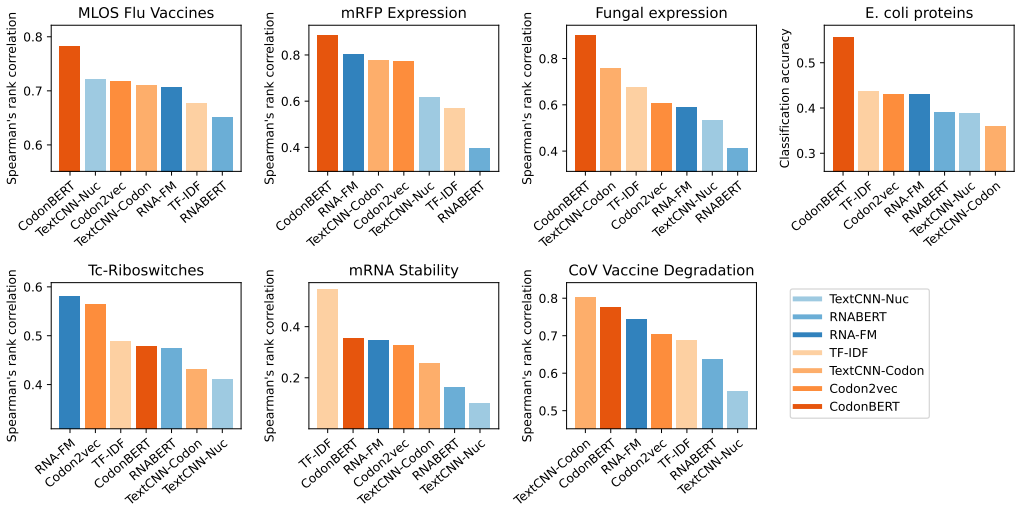


Figure S4: Ranked results (Spearman's rank correlation or classification accuracy, higher is better) for CodonBERT model and other benchmarks. Nucleotide-based and codon-based methods are in blue and orange colors, respectively. Shades represent model complexity.

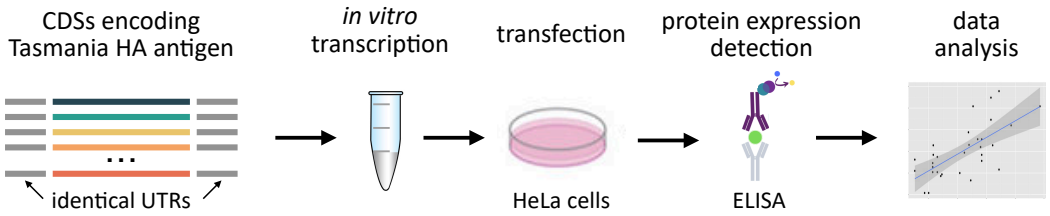


Figure S5: Experimental design for testing in-cell protein expression.