Learning the Language of Codon Translation with CodonFM

Sajad Darabi [1,+], Fan Cao [1,+], Mohsen Naghipourfar [2,3,+], Sara Rabhi [1], Ankit Sethia [1], Kyle Gion [1], Jasleen Grewal [1], Jonathan Cohen [1], William J. Greenleaf [1,6], Hani Goodarzi [3,4,5,*], Laksshman Sundaram [1,*]

[1]NVIDIA, Santa Clara, CA, USA
[2]Molecular Cell Biomechanics Laboratory, Departments of Bioengineering and Mechanical Engineering, University of California, Berkeley, Berkeley, CA, USA
[3]Arc Institute, Palo Alto, CA, USA
[4]Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA, USA
[5]Department of Urology, University of California, San Francisco, San Francisco, CA, USA
[6]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA

[+] *equal contribution*

[*] **Correspondence to:** Laksshman Sundaram lsundaram@nvidia.com; Hani Goodarzi hani.goodarzi@arcinstitute.org

Abstract

The canonical genetic code is degenerate, with most amino acids encoded by multiple synonymous codons whose choice can influence translation, RNA stability, and protein expression. Despite this complexity, the underlying rules linking codon usage to molecular phenotypes remain poorly captured by existing models. Here, we introduce the EnCodon model series within CodonFM, a family of large foundation models trained on more than 130 million coding sequences spanning over 22,000 species, designed to learn the contextual grammar of codon usage directly from sequence. EnCodon models exhibit clear scaling behavior, with larger models showing lower normalized confusion scores across synonymous codons, revealing an emergent understanding of synonymous codon grammar. In zero-shot settings, EnCodon achieves state-of-the-art performance across diverse benchmarks, including prediction of *de novo* missense mutation pathogenicity, clinical missense mutation classification, and ClinVar synonymous variant discrimination. EnCodon generalizes to downstream mRNA design tasks, accurately predicting translation efficiency and protein expression from sequence context. Together, these results demonstrate that learning the intrinsic grammar of codon usage is sufficient to infer a broad spectrum of biological and clinical effects, establishing EnCodon as a scalable foundation for modeling translation and RNA-driven gene regulation.

Introduction

The canonical genetic code—the blueprint that links nucleic acid instructions to proteins—is highly degenerate, with 18 of the 20 amino acids encoded by more than one codon. These synonymous codons were once thought to be interchangeable, since they specify the same amino acid [1]. However, codon usage bias—the unequal use of synonymous codons—has long been recognized and extensively studied [1–7]. Increasing evidence shows that codon usage patterns are not random but exhibit organism-specific [4–6] and even tissue-specific [8,9] signatures across species. Numerous studies have demonstrated that codon usage can regulate gene expression and protein folding through multiple mechanisms [10–14]. For instance, synonymous codons are influenced by cognate tRNA abundance and tRNA gene copy numbers, affecting translation elongation rates [1,4,6]. Furthermore, there are known synonymous variants associated with diseases and tumors [15–17]. This underscores the fact that nucleotide sequences encode rich layers of information beyond their amino acid content.

This complexity has profound implications for applications such as mRNA therapeutics and vaccines, where the expression level of the encoded protein directly influences potency, immunogenicity, and efficacy [18]. Higher antigen expression enables lower vaccine doses, improving safety, reducing reactogenicity [19], and extending immune durability [20]. However, achieving such high expression levels depends on the choice of codons within the mRNA sequence. Despite each protein being encodable by thousands of possible synonymous sequences, the one found in nature or in viral genomes is not always optimal for expression in a given host cell. Classical codon optimization methods attempt to improve expression by aligning codon usage with host bias [21–24]. Yet, such approaches often overlook key biophysical factors—particularly RNA structural features such as stem-loops and pseudoknots—that play critical roles in RNA stability and translation efficiency [25–31]. Even a single synonymous substitution can alter local base-pairing patterns and disrupt or stabilize nearby structural motifs [32], highlighting that true optimization must account for the interplay between codon choice, RNA structure, and expression outcomes.

Over evolutionary timescales, the choice of synonymous codons has been finely tuned to optimize translation efficiency, accuracy, and resource allocation within each organism [7]. The specific pattern of codon usage for a given gene reflects a

balance among multiple selective pressures—including tRNA availability, GC content, RNA structure, and expression requirements—resulting in a context-dependent "grammar" of synonymous usage. This grammar encodes evolutionary solutions for maintaining translational fidelity and regulating expression in response to cellular environment. Learning these patterns directly from natural coding sequences offers a unique opportunity to model how evolution has encoded context-specific optimization into the structure of the genetic code itself.

To address the challenge of learning codon usage and its contextual grammar, we train a family of large foundation models collectively termed CodonFM. These transformer-based models are pre-trained on more than 130 million coding sequences spanning over 22,000 species from the NCBI Genomes database [33], encompassing the full phylogenetic diversity of the genetic code. We envision CodonFM as a growing family of codon foundation models designed to learn the contextual grammar of codon usage. For the first of these models, EnCodon, we develop three parametric variants of the model at increasing scales to systematically study how model capacity influences biological representation learning [34,35]. The models are then evaluated across a suite of downstream tasks probing both protein-related variant effects and synonymous codon changes relevant to disease genetics and mRNA design [16,36–38]. Across these diverse benchmarks, EnCodon demonstrates robust and generalizable performance, capturing latent features that link sequence composition to translation and expression outcomes [11–13]. Moreover, we observe a clear scaling effect with larger parameter models consistently outperforming their smaller counterparts—highlighting that model capacity enhances the ability to capture the nuanced, context-dependent rules governing synonymous codon usage and its biological impact [39].

Results

Overview of EnCodon Data, Architecture, and Pretraining

To train our models, we assemble a large-scale dataset of >130 million coding sequences (CDS) spanning >22,000 species from the NCBI RefSeq/Genomes database [33] (**Figure 1A**). Each sequence is tokenized at the codon level, allowing the

model to directly capture codon-context relationships within open reading frames (ORFs). The training corpus spans all major phylogenetic groups—including bacteria, archaea, fungi, plants, protozoa, and metazoans—representing the full evolutionary spectrum of the genetic code (**Figure 1B**). We exclude the human-affecting pathogen sequences for biosafety purposes. This diverse distribution ensures that the CodonFM models learn generalizable representations of codon redundancy, GC bias, and translation-associated sequence features across kingdoms [1,2,6]. Most CDS entries are under 1,000 codons, with a right-skewed length distribution characteristic of natural genes; hence, nearly all sequences are completely represented within the 2,046-codon context of the model (**Figure 1C)**.

EnCodon employs a transformer-based encoder architecture that operates directly on codon tokens, beginning with a *<CLS>* and ending with a *<SEP>* token (**Figure 1D**). The model is trained using a masked-codon prediction objective, analogous to masked language modeling [40]. To examine how scaling and masking influence biological representations, we train three variants—EnCodon 80M, EnCodon 600M, and EnCodon 1B—alongside a fourth variant, EnCodon 1B-CDWT, which leverages a Codon-frequency Weighted Masking Strategy (**Figure 1E**). This frequency-weighted approach is inspired by distributional token weighting in natural-language models [41,42], emphasizing underrepresented synonymous codons, encouraging the model to capture the interplay among codon bias, translation efficiency, and tRNA adaptation [6,10,13]. All variants show smooth convergence with larger models achieving lower validation loss and improved generalization, reflecting a clear scaling relationship between model capacity and its ability to encode codon-level structure and redundancy across diverse genetic contexts (**Figure 1E**) [35].

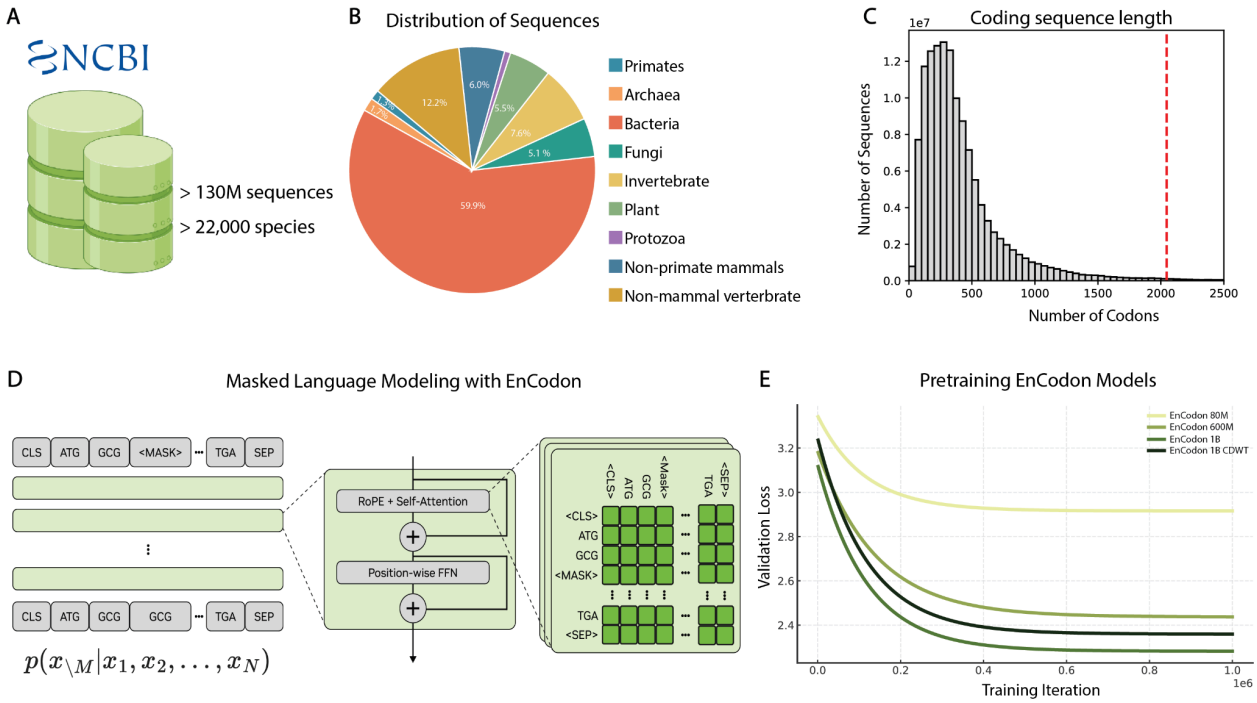

Figure 1: Dataset composition, model architecture, and training of EnCodon.
(A) EnCodon models were trained on >130 M coding sequences from >22,000 species in the NCBI RefSeq database, tokenized at the codon level within ORFs. (B) Sequence-grouped composition of the dataset. (C) CDS length distribution of ORF sequences. (D) Schematic of the EnCodon model's transformer encoder trained with a masked language modeling objective. (E) Validation loss across EnCodon models—80M, 600M, 1B, and 1B-CDWT—showing improved convergence with larger models.

Interpreting the EnCodon models

To better understand the representations learned by EnCodon across scales, we analyze both codon confusion patterns and embedding structures (**Figure 2A–C**). The synonymous codon confusion matrices (**Figure 2A**) reveal a progressive decrease in normalized confusion scores from the 80M to the 1B model—indicating that larger and frequency-aware models more accurately distinguish synonymous codons and capture their regulatory specificity.

To visualize the models' embedding space, we use UMAP projections (**Figure 2B**) on the PCA-reduced embeddings. Quantitatively, larger EnCodon models achieve significantly lower (better) masked language modeling (MLM) losses across phylogenetic groupings (**Figure 2C**, *left*), indicating improved predictive accuracy and contextual understanding of codon relationships. Consistent with this, larger models also show higher K-nearest neighbour (KNN) purity across taxonomic divisions (**Figure 2C**, *middle*), demonstrating that their embeddings organize sequences more coherently with respect to biological groupings. The 1B-CDWT

model achieves the highest purity, indicating that the codon-weighted masking strategy enables the model to focus more effectively on domain-specific codon usage patterns compared to random masking models. The correlation between the principal components (PCs) and amino acid hydrophobicity (**Figure 2C**, *right*) also show that the smallest model tend to have higher correlation with basic properties like hydrophobicity than the larger ones across top 10 PCs, indicating that higher-capacity models balance biochemical features with additional contextual signals related to codon usage. Together, these results indicate that larger EnCodon variants encode richer, context-dependent representations that extend beyond simple physico-chemical or genetic mappings, enabling more nuanced modeling of synonymous codon usage.

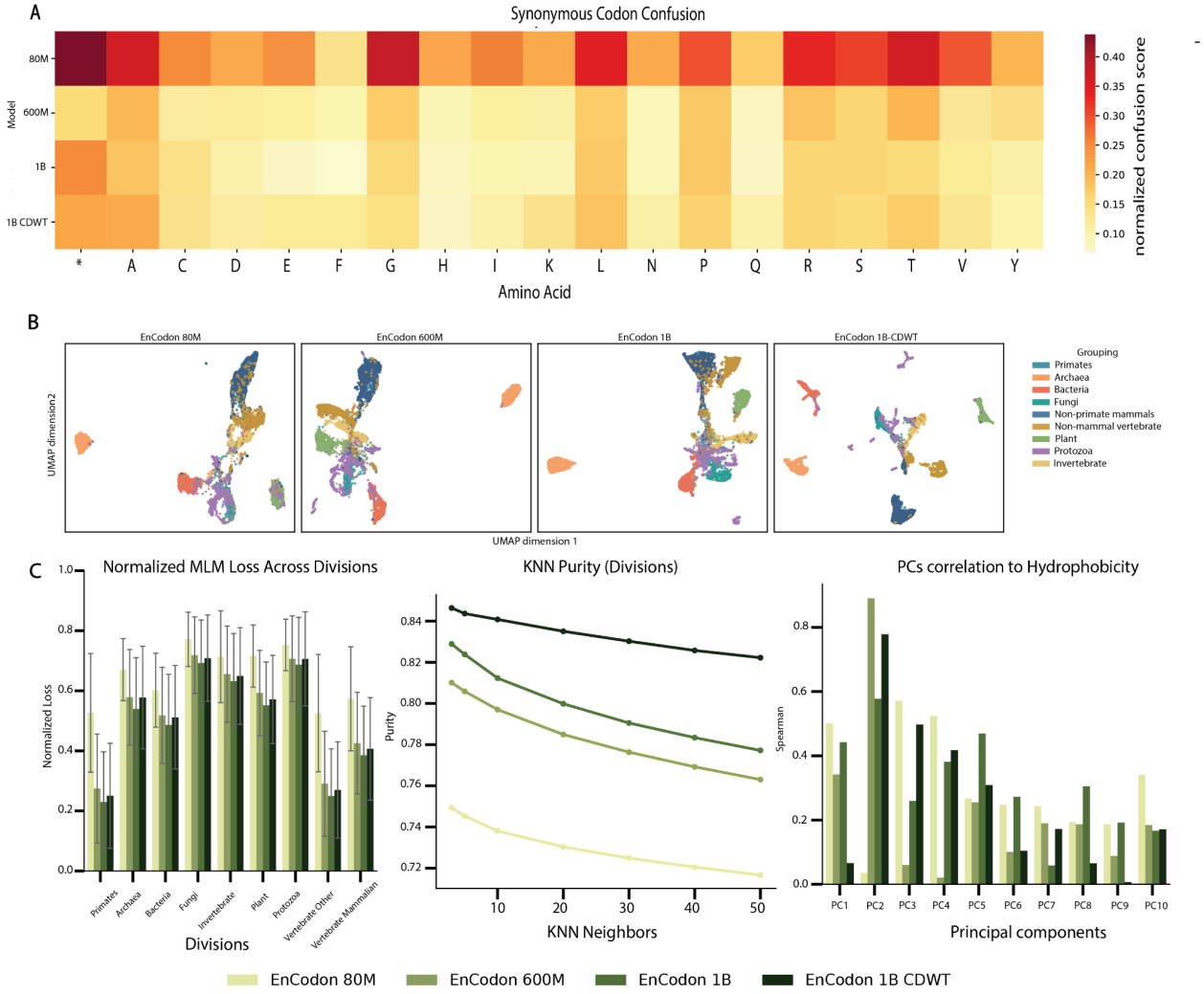

Figure 2: Emergent codon structure and phylogenetic organization across model scales*. (A) Normalized codon confusion matrices of synonymous codons across EnCodon model variants. (B) UMAP projections of sequence embedding space learned by EnCodon models, colored by sequence grouping. (C) (left) Representation of MLM loss distribution for pre-trained EnCodon models across taxonomy divisions, (middle) KNN purity scores across nearest neighbors in the embedding space. (right) Correlation between the top 10 PCs of the pre-trained EnCodon models and the hydrophobicity index of the codon's amino acids.*

Interpreting missense variants with EnCodon

We benchmark the EnCodon models' performance on multiple missense mutation prediction tasks to evaluate whether these models capture biologically meaningful information about coding variation (**Figure 3A–D**). In the zero-shot setting, each variant is scored using the log-likelihood difference between the reference and mutated codons, without any task-specific training. On two large *de novo* mutation datasets—the Deciphering Developmental Disorders (DDD) and Autism Spectrum Disorder (ASD) cohorts [37]— EnCodon demonstrates the strongest separation between case and control variants, compared to other unsupervised protein[43] and RNA sequence models [44–51]. We also observe robust improvement in model performance across different scales. The superior zero-shot performance suggests that EnCodon, despite being trained purely on codon sequences, implicitly learns representations that encode protein sequence constraint and functional relevance, capturing context beyond synonymous codon structure.

We next assess EnCodon on classifying ClinVar missense variants [36] and somatic missense variants in cancer hotspots [52,53] (**Figure 3C–D**). We observe that the EnCodon models are highly performant and are second best to the ESM2 models trained on amino acid sequences [43] while outperforming all other RNA-based baselines [44–51]. Finally, to examine how pre-training transfers to fine-tuned models, we fine-tune EnCodon 1B on gnomAD [54] missense variants and evaluate its (EnCodon 1B-FT) performance on the DDD and ASD cohorts [37] (**Figure 3F–G**). The fine-tuned EnCodon is similarly performant or better than AlphaMissense [52], a supervised protein pathogenicity model, even in the absence of explicit structural priors.

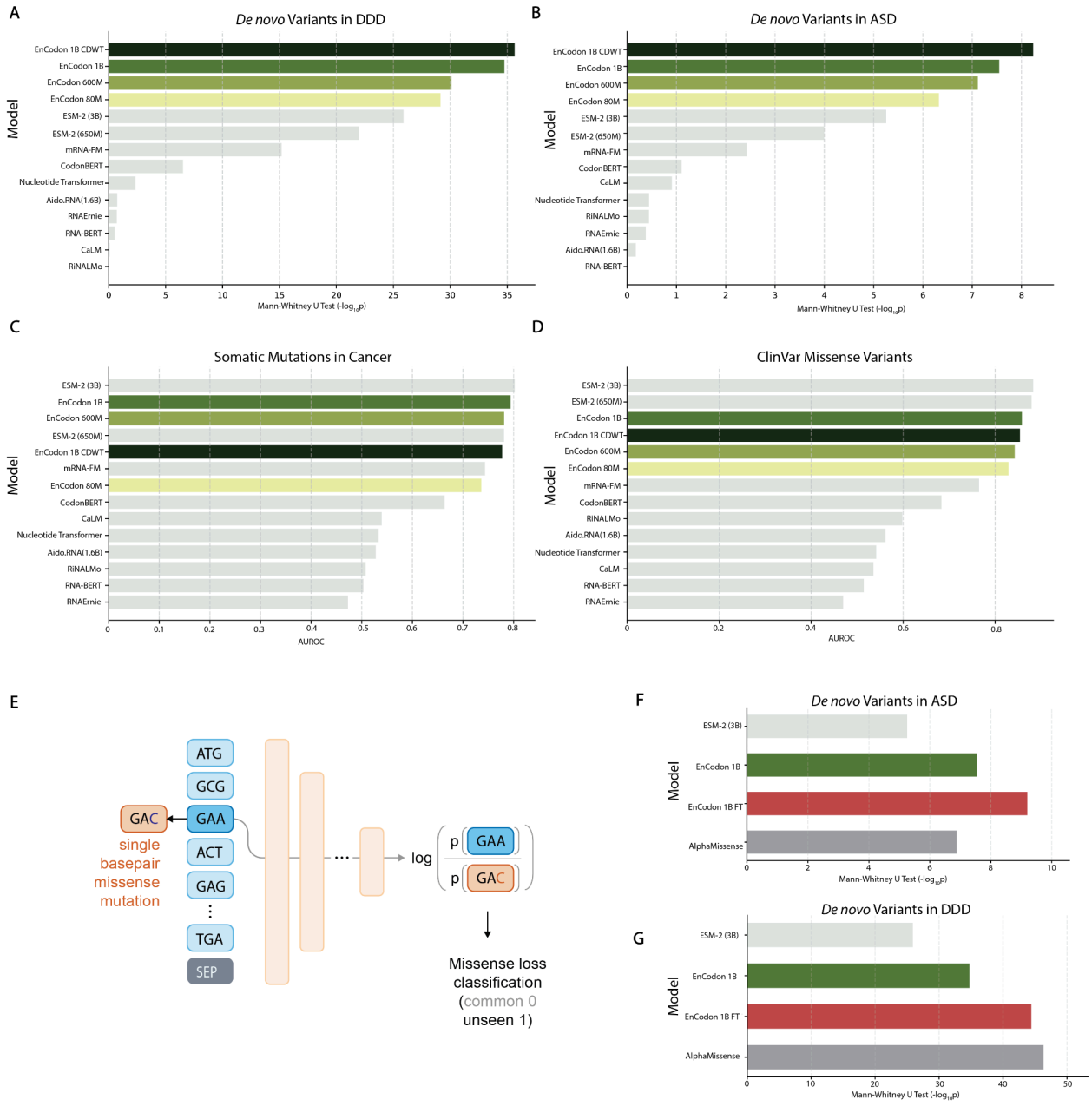

Figure 3: Benchmarking EnCodon on missense mutation prediction tasks.

*(**A-D**) Zero-shot performance comparison of classifiers across multiple missense mutation tasks. (A-B) Mann-Whitney U Test (two-sided) p-values ($-\log_{10}$ p-value) are shown for DDD and ASD case vs control variants, and (C–D) AUROC on ClinVar and Cancer Hotspot mutation datasets. (**E–G**) EnCodon fine-tuning procedure with common and unseen missense variants from gnomAD (E), Fine-tuned performance comparison of EnCodon 1B to other supervised and zero-shot models on ASD (F) and DDD (G).*

Interpreting synonymous variants with CodonFM

To assess whether EnCodon generalizes beyond missense mutations to synonymous changes, we evaluate its ability to distinguish pathogenic versus benign synonymous variants from the ClinVar database [36] (**Figure 4A)**. Given that the two lists of variants are confounded by codon composition, gene context, and mutation-rate biases, we perform 50 stratified subsampling iterations in which each pathogenic variant is compared against benign variants matched for reference and alternate codon, position in gene, gene-level probability of loss-of-function intolerance (pLI) [54], and local mutation rate. Model performance is measured by the Mann–Whitney U test across iterations (**Figure 4B**). EnCodon models consistently outperform all RNA- and mRNA-based baselines [44–51], with the 1B-CDWT variant achieving the highest median significance across replicates, demonstrating robustness to covariate control. Because the biological effects of synonymous substitutions have historically been the most difficult to resolve, this task highlights a distinct strength of EnCodon: its ability to infer functional and clinical consequences from the subtle grammar of codon choice. Importantly, EnCodon attains state-of-the-art performance relying solely on its learned understanding of the underlying grammar of codon usage to infer the clinical effects of synonymous variants.

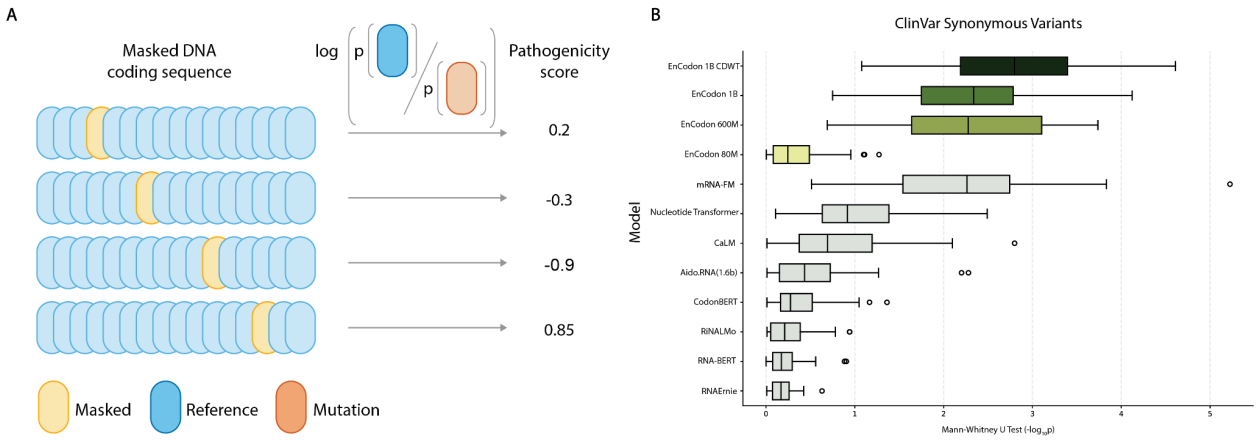

Figure 4: Evaluation of EnCodon on ClinVar synonymous variants.
(A) Schematic of the zero-shot evaluation of synonymous variants with EnCodon models. (B) Zero-shot performance (Mann-Whitney U Test, uncorrected) of models comparing pathogenic and benign synonymous variants from ClinVar using 50 matched subsampling iterations, controlling for codon, position in gene, gene-level pLI, and local mutation rate.

Evaluating EnCodon on mRNA Expression and Translation Efficiency

Finally, we assess whether EnCodon embeddings capture quantitative features relevant to mRNA design and functional features (**Figure 5A–B**). Using zero-shot embeddings from each model, we train a random forest regressor to predict experimental readouts from two independent datasets on mRNA translation efficiency [38] and mRFP protein expression [48,55]. These tasks reflect key determinants of mRNA design, such as codon composition, local structure, and translational control. Across both benchmarks, EnCodon models achieve the highest predictive performance among all nucleotide-level baselines, with the 1B parameter model showing the strongest correlation and explained variance. The EnCodon-1B-CDWT embeddings are less influenced by simple sequence features such as GC content and therefore perform slightly worse than the random masking models on tasks where these features have a stronger effect. By learning the contextual grammar of codon usage, EnCodon captures position-dependent effects on expression and stability, providing biophysically meaningful embeddings useful for zero-shot mRNA sequence optimization.

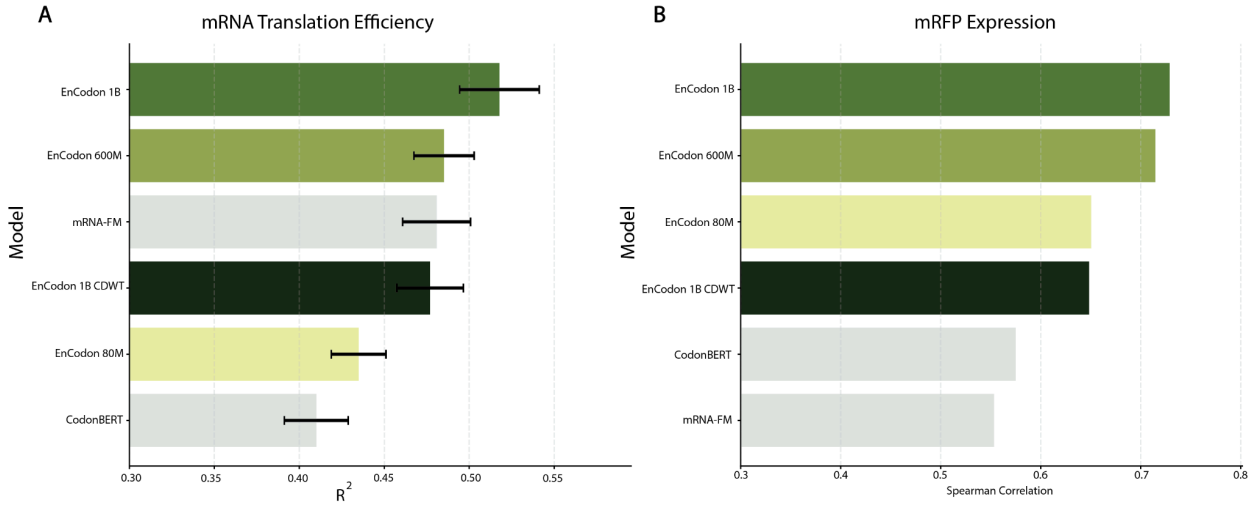

Figure 5: Evaluation of EnCodon on mRNA design features. *Comparison of random forest performance on pretrained model embeddings, (A) on the translation efficiency task, showing the mean 10-fold cross-validation R^2 scores (± standard deviation). (B) on the mRFP protein expression task, showing Spearman correlation between predicted and observed expression levels.*

Discussion

EnCodon demonstrates that large-scale, codon-level language modeling can learn the underlying grammar of the genetic code, linking sequence composition to translation efficiency, pathogenicity, and mRNA expression. Trained on over 130 million coding sequences from 22,000 species, the model captures evolutionary and functional constraints, with larger variants revealing organized structure among synonymous codons. The EnCodon 1B-CDWT model performs best overall, as codon-weighted masking encourages attention to rare and functionally constrained codons, enabling finer contextual understanding. In the use case of pathogenicity prediction for missense variants (ClinVar missense mutation prediction, cancer hotspot variant prediction), we observe that EnCodon models are marginally outperformed by the ESM family of protein language models. The marginal performance gap suggests that refinement in training strategies for codon-level models could help bridge this gap, which is likely due to explicit encoding of protein structure information and amino-acid substitution patterns in protein language models compared to nuanced codon grammar influencing said patterns in codon language models.

Beyond its predictive power, EnCodon highlights that the same protein can be encoded by an immense number of synonymous coding sequences, each capable of producing distinct translational and regulatory outcomes. This redundancy transforms the genetic code into a programmable substrate—one that can be tuned for context-dependent activity, expression, and stability. The hidden layer of functional variability in synonymous mutations has likely been shaped by evolution to fine-tune gene regulation, yet remains difficult to quantify or predict. EnCodon provides a computational framework to expose this latent space of synonymous effects—revealing that what was once considered silent variation can, in fact, encode nuanced regulatory information. As the CodonFM framework evolves, new architectures and objectives can be added as modular components within the same foundational ecosystem, supporting increasingly rich representations of codon-level biology.

Despite these advances, several limitations remain. EnCodon's results are based on computational inference and require experimental validation to confirm their biological relevance. The number of synonymous variants with experimentally verified functional effects is still small, constraining validation datasets and

necessitating cautious interpretation of statistical significance. Moreover, EnCodon currently models sequence features without explicitly incorporating cell-type-specific tRNA abundance, RNA modifications, or secondary structure dynamics, factors known to modulate translation *in vivo*. Integrating such context-dependent data and validating predictions through perturbation assays, ribosome profiling, and synthetic mRNA design experiments will be essential for translating EnCodon's insights into mechanistic and therapeutic applications.

Acknowledgements

We thank Brian Plosky, Chiara Ricci-Tam, Michelle Gill, Maria Korshunova, Lina Brilliantova, and Yingfei Wang for their valuable comments and insights during the preparation of this manuscript. WJG's contribution was as a paid consultant and was not part of his Stanford University duties or responsibilities.

Code Availability

All code for model training, evaluation, and analysis is available at the https://github.com/NVIDIA-Digital-Bio/CodonFM GitHub repository. Pretrained CodonFM model checkpoints can be accessed on Hugging Face (https://huggingface.co/collections/nvidia/clara-biology) and NVIDIA NGC (https://catalog.ngc.nvidia.com/orgs/nvidia/teams/clara/models/nv_codonfm_encodon).

References

1. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.* **12**, 32–42 (2011).
2. Grantham, R., Gautier, C., Gouy, M., Mercier, R. & Pavé, A. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.* **8**, r49–r62 (1980).
3. Ikemura, T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.* **146**, 1–21 (1981).
4. Sharp, P. M. & Li, W. H. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* **24**, 28–38 (1986).
5. Reis, M. dos, Savva, R. & Wernisch, L. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* **32**, 5036–5044 (2004).
6. Tuller, T., Waldman, Y. Y., Kupiec, M. & Ruppin, E. Translation efficiency is determined by both codon bias and folding energy. *Proc. Natl. Acad. Sci.* **107**, 3645–3650 (2010).
7. Hershberg, R. & Petrov, D. A. Selection on codon bias. *Annu. Rev. Genet.* **42**, 287–299 (2008).
8. Rudolph, K. L. M. *et al.* Codon-Driven Translational Efficiency Is Stable across Diverse Mammalian Cell States. *PLOS Genet.* **12**, e1006024 (2016).
9. Dittmar, K. A., Goodenbour, J. M. & Pan, T. Tissue-Specific Differences in Human Transfer RNA Expression. *PLOS Genet.* **2**, e221 (2006).

10. Quax, T. E. F., Claassens, N. J., Söll, D. & van der Oost, J. Codon Bias as a Means to Fine-Tune Gene Expression. *Mol. Cell* **59**, 149–161 (2015).
11. Buhr, F. *et al.* Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. *Mol. Cell* **61**, 341–351 (2016).
12. Gamble, C. E., Brule, C. E., Dean, K. M., Fields, S. & Grayhack, E. J. Adjacent Codons Act in Concert to Modulate Translation Efficiency in Yeast. *Cell* **166**, 679–690 (2016).
13. Hanson, G. & Coller, J. Codon optimality, bias and usage in translation and mRNA decay. *Nat. Rev. Mol. Cell Biol.* **19**, 20–30 (2018).
14. Brule, C. E. & Grayhack, E. J. Synonymous Codons: Choose Wisely for Expression. *Trends Genet. TIG* **33**, 283–297 (2017).
15. Sauna, Z. E. & Kimchi-Sarfaty, C. Understanding the contribution of synonymous mutations to human disease. *Nat. Rev. Genet.* **12**, 683–691 (2011).
16. Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T. & Lehner, B. Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. *Cell* **156**, 1324–1335 (2014).
17. Sharma, Y. *et al.* A pan-cancer analysis of synonymous mutations. *Nat. Commun.* **10**, 2569 (2019).
18. Pardi, N., Hogan, M. J., Porter, F. W. & Weissman, D. mRNA vaccines — a new era in vaccinology. *Nat. Rev. Drug Discov.* **17**, 261–279 (2018).
19. Teijaro, J. R. & Farber, D. L. COVID-19 vaccines: modes of immune activation and future challenges. *Nat. Rev. Immunol.* **21**, 195–197 (2021).

20. Karikó, K. *et al.* Incorporation of Pseudouridine Into mRNA Yields Superior Nonimmunogenic Vector With Increased Translational Capacity and Biological Stability. *Mol. Ther. J. Am. Soc. Gene Ther.* **16**, 1833–1840 (2008).
21. Gustafsson, C., Govindarajan, S. & Minshull, J. Codon bias and heterologous protein expression. *Trends Biotechnol.* **22**, 346–353 (2004).
22. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in Escherichia coli. *Science* **324**, 255–258 (2009).
23. Puigbò, P., Guzmán, E., Romeu, A. & Garcia-Vallvé, S. OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. *Nucleic Acids Res.* **35**, W126–W131 (2007).
24. Paremskaia, A. I. *et al.* Codon-optimization in gene therapy: promises, prospects and challenges. *Front. Bioeng. Biotechnol.* **12**, (2024).
25. Mortimer, S. A., Kidwell, M. A. & Doudna, J. A. Insights into RNA structure and function from genome-wide studies. *Nat. Rev. Genet.* **15**, 469–479 (2014).
26. Mauger, D. M. *et al.* mRNA structure regulates protein expression through changes in functional half-life. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 24075–24083 (2019).
27. Mustoe, A. M. *et al.* Pervasive Regulatory Functions of mRNA Structure Revealed by High-Resolution SHAPE Probing. *Cell* **173**, 181-195.e18 (2018).
28. de Smit, M. H. & van Duin, J. Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 7668–7672 (1990).
29. Yu, C.-H. *et al.* Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. *Mol. Cell* **59**, 744–754 (2015).

30. Chamary, J. V. & Hurst, L. D. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol.* **6**, R75 (2005).
31. Hia, F. *et al.* Codon bias confers stability to human mRNAs. *EMBO Rep.* **20**, e48220 (2019).
32. Presnyak, V. *et al.* Codon Optimality Is a Major Determinant of mRNA Stability. *Cell* **160**, 1111–1124 (2015).
33. O'Leary, N. A. *et al.* Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Sci. Data* **11**, 732 (2024).
34. Vaswani, A. *et al.* Attention is All you Need. in *Advances in Neural Information Processing Systems* (eds Guyon, I. et al.) vol. 30 (Curran Associates, Inc., 2017).
35. Kaplan, J. *et al.* Scaling Laws for Neural Language Models. Preprint at https://doi.org/10.48550/arXiv.2001.08361 (2020).
36. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
37. Zhou, X. *et al.* Integrating de novo and inherited variants in 42,607 autism cases identifies mutations in new moderate-risk genes. *Nat. Genet.* **54**, 1305–1319 (2022).
38. Zheng, D. *et al.* Predicting the translation efficiency of messenger RNA in mammalian cells. Preprint at https://doi.org/10.1101/2024.08.11.607362 (2024).
39. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
40. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. in *Proceedings of the 2019*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers) (eds Burstein, J., Doran, C. & Solorio, T.) 4171–4186 (Association for Computational Linguistics, 2019). doi:10.18653/V1/N19-1423.

41. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. Preprint at https://doi.org/10.48550/arXiv.1301.3781 (2013).
42. Clark, K., Luong, M.-T., Le, Q. V. & Manning, C. D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020* (OpenReview.net, 2020).
43. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
44. Chen, J. *et al.* Interpretable RNA Foundation Model from Unannotated Data for Highly Accurate RNA Structure and Function Predictions. Preprint at https://doi.org/10.48550/arXiv.2204.00300 (2022).
45. Dalla-Torre, H. *et al.* Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat. Methods* **22**, 287–297 (2025).
46. Outeiral, C. & Deane, C. M. Codon language embeddings provide strong signals for use in protein engineering. *Nat. Mach. Intell.* **6**, 170–179 (2024).

47. Zou, S. *et al.* A Large-Scale Foundation Model for RNA Function and Structure Prediction. 2024.11.28.625345 Preprint at https://doi.org/10.1101/2024.11.28.625345 (2024).
48. Li, S. *et al.* CodonBERT large language model for mRNA vaccines. *Genome Res.* **34**, 1027–1035 (2024).
49. Penić, R. J., Vlašić, T., Huber, R. G., Wan, Y. & Šikić, M. RiNALMo: general-purpose RNA language models can generalize well on structure prediction tasks. *Nat. Commun.* **16**, 5671 (2025).
50. Akiyama, M. & Sakakibara, Y. Informative RNA base embedding for RNA structural alignment and clustering by deep representation learning. *NAR Genomics Bioinforma.* **4**, lqac012 (2022).
51. Wang, N. *et al.* Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning. *Nat. Mach. Intell.* **6**, 548–557 (2024).
52. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
53. Zhang, H., Xu, M. S., Fan, X., Chung, W. K. & Shen, Y. Predicting functional effect of missense variants using graph attention neural networks. *Nat. Mach. Intell.* **4**, 1017–1028 (2022).
54. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).
55. Nieuwkoop, T. *et al.* Revealing determinants of translation efficiency via whole-gene codon randomization and machine learning. *Nucleic Acids Res.* **51**, 2363–2376 (2023).