Predicting Translation Efficiency from Synonymous Coding Sequences Using Codon-Level Embeddings

Rushmila Islam
Rushmila.Islam@gmail.com

December 15, 2025

1 Introduction

Accurately predicting translation efficiency is essential for understanding gene expression dynamics and improving synthetic gene design [1]. Two mRNAs can encode the same protein yet differ by an order of magnitude in protein output due to regulatory features that influence translation initiation, elongation, and termination [2, 3]. Translation efficiency determines protein abundance, cellular behaviour, optimisation strategies in biotechnology, and the functional consequences of synonymous and non-synonymous mutations, with particular importance for mRNA-based therapeutics where precise control of protein output is critical. Translation occurs in three major stages, each influenced by distinct sequence features. **Translation initiation** is frequently rate-limiting and is largely governed by the $5'$ untranslated region ($5'$UTR), including ribosome-binding motifs (e.g., Shine–Dalgarno or Kozak sequences), local RNA secondary structure near the start codon, and upstream open reading frames. Inefficient initiation limits translation regardless of downstream codon quality. In contrast, **translation elongation** depends on features within the coding sequence (CDS), such as synonymous codon usage, tRNA abundance, codon-pair preferences, and mRNA structure along the transcript. These factors determine ribosome velocity, pausing, co-translational folding, and the rate at which functional protein is produced [4]. The first 10–20 codons are particularly influential in modulating ribosome flow. **Termination** generally contributes less variability but can be affected by stop-codon identity and downstream structure.

Understanding how synonymous codon substitutions modulate elongation remains a longstanding challenge in molecular biology, evolutionary genomics, and biotechnology. Although synonymous changes do not alter the amino-acid sequence, they can produce large differences in protein expression by altering translation kinetics, mRNA stability, and folding dynamics. Traditional measures such as the Codon Adaptation Index (CAI) and tRNA Adaptation Index (tAI) capture global codon usage trends but fail to represent higher-order dependencies, positional effects, or structural context within the CDS. Recent advances in natural language processing have therefore inspired data-driven approaches in which codons are treated as tokens and coding sequences as sentences. Models such as *Codon2Vec* [5], a shallow embedding method based on Word2Vec principles, and *CodonFM* [6], a transformer-based foundation model trained on billions of codons, offer new opportunities to learn contextual representations that better reflect translation-relevant biology. A critical consideration, however, is that translation efficiency is shaped by both initiation and elongation. Initiation effects are strongly influenced by the $5'$UTR, as demonstrated by large-scale studies showing that UTR accessibility and local mRNA structure can dominate expression outcomes when UTRs are varied [7, 8].

In this work, we use the *mRFP* synonymous-codon dataset, in which the promoter, ribosome binding site, and $5'$UTR are held constant across all constructs, while only the CDS varies through synonymous substitutions. This controlled design isolates *elongation-driven* effects and minimises initiation variability, allowing observed expression differences to be attributed primarily to codon-dependent elongation dynamics and CDS-internal features. Within this fixed-initiation framework, we evaluate whether transformer-based codon language models capture codon-dependent translation efficiency more effectively than shallow embedding approaches. Using the *mRFP* expression dataset, we compare **CodonFM**, a contextual codon foundation model, with **Codon2Vec**, a local co-occurrence embedding model. By analysing how well embeddings from each method predict experimentally measured fluorescence values, we assess their ability to represent elongation-relevant sequence features and determine whether large pre-trained codon models provide measurable advantages for understanding gene expression and informing mRNA-based therapeutic design.

2 Related Works

Synonymous codon usage has long been recognised as a major determinant of translation efficiency and protein abundance. Plotkin and Kudla [9] synthesised extensive molecular and evolutionary evidence showing that codon bias arises from tRNA availability, mutation pressure, context-dependent selection, and mRNA structural constraints. Their work established that synonymous substitutions can influence ribosome speed, co-translational folding, and overall expression output, motivating the development of models capable of capturing these multi-layered dependencies. A substantial body of experimental work has since quantified how synonymous variation within the coding sequence modulates expression. Goodman *et al.* [3] generated one of the most comprehensive synonymous variant libraries of the *mRFP* gene, demonstrating that N-terminal codon identity and local RNA structure strongly influence early elongation and protein production. Bol *et al.* [10] extended these findings genome-wide, showing that codon usage affects not only elongation rate but also steady-state mRNA abundance, indicating coupled regulation between translation and RNA stability. Together, these studies highlight that codon-mediated elongation effects are substantial but mechanistically complex, depending on codon context, position, and mRNA structure.

At the same time, numerous high-impact studies have shown that translation initiationregulated primarily by the 5′UTRis often the dominant source of expression variability when UTR sequences differ. Cambray *et al.* [7] systematically analysed more than 244,000 synthetic 5′UTRs and demonstrated that variation in ribosome binding site strength, spacing, and local RNA accessibility can overshadow elongation effects by orders of magnitude. Similarly, Goodman, Church, and Kosuri [8] showed that N-terminal codon bias is shaped jointly by mRNA structure in the 5′UTR and the first few codons, reinforcing the tight coupling between initiation and early elongation. These findings establish that translation efficiency depends on both initiation and elongation, but also underscore the importance of experimental designs that isolate individual regulatory layers. The *mRFP* dataset used in this study provides such a controlled setting: because all constructs share an identical promoter, ribosome binding site, and 5′UTR [3, 8], observed variation in fluorescence can be attributed specifically to synonymous differences within the coding sequence. Unlike UTR-focused studies, our work therefore isolates codon-driven elongation effects under fixed initiation conditions.

Parallel to these experimental advances, computational approaches have been developed to model the relationship between codon composition and gene expression. Wint *et al.* [5] introduced Codon2Vec, which applies Word2Vec-style embeddings to learn local co-occurrence patterns among codons. While effective for capturing short-range dependencies, Codon2Vec is limited by its shallow architecture and fixed context window. In contrast, Ji *et al.* [11] demonstrated with DNABERT that transformer-based masked language models can learn rich, context-aware representations of genomic sequences that generalise across prediction tasks. Building on this paradigm, NVIDIAs CodonFM [6, 12] introduced large-scale codon-resolution transformer models trained on billions of codons, enabling contextual embeddings that capture long-range dependencies and translation-relevant semantics. Our study builds on these developments by conducting a focused benchmark comparison between CodonFM and Codon2Vec on the synonymous *mRFP* dataset. By evaluating their ability to predict experimentally measured fluorescencea proxy for translation efficiencywe assess how effectively shallow versus transformer-based embeddings capture elongation-driven regulatory signals, positioning our work as complementary to initiation-focused studies while providing a systematic evaluation of codon-level representation learning under a controlled initiation framework.

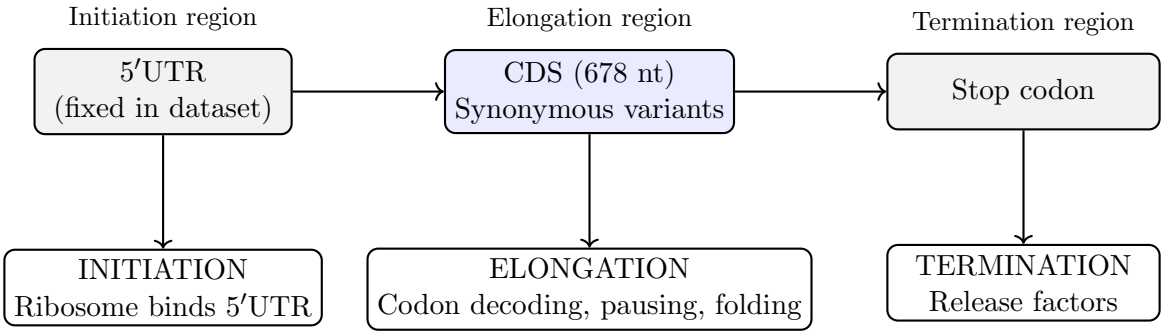

Figure 1: Translation stages in the mRFP constructs. The 5′UTR and promoter are fixed across all variants, isolating codon-dependent elongation effects arising from the CDS.

3 Methodology

3.1 Dataset Preprocessing and Embedding Generation

We use a synonymous-codon expression dataset based on monomeric red fluorescent protein (*mRFP*), originally introduced by Goodman *et al.* [3] and subsequently adopted in studies of codon-dependent translation efficiency [10]. The dataset comprises 1,459 distinct coding-sequence variants, all encoding an identical amino acid sequence but differing through synonymous codon substitutions. Importantly, all constructs share the same promoter, ribosome binding site, and 5′ untranslated region (5′UTR), such that variation in measured expression primarily reflects elongation-related effects arising from differences within the coding sequence. For each variant, the dataset provides the full coding sequence and a corresponding quantitative expression measurement. Specifically, the available information includes:

- **Coding sequence (CDS):** a fixed-length open reading frame of 678 nucleotides (226 codons) encoding *mRFP*
- **Expression measurement:** fluorescence intensity, which serves as a continuous proxy for translation efficiency and functional protein output
- **Predefined data partitioning:** a split into training (1,021 sequences), validation (219 sequences), and test (219 sequences) subsets, which is preserved throughout all experiments to enable consistent model comparison

Preprocessing. For the Codon2Vec pipeline, coding sequences were standardised by converting all characters to uppercase, filtering for valid nucleotide symbols, and mapping RNA bases to their DNA equivalents (U→T). Sequences were then truncated to ensure lengths divisible by three and segmented into non-overlapping codon triplets for downstream embedding generation. For the CodonFM pipeline, coding sequences were provided directly to the Encodon tokenizer without additional sequence-level modification. Expression measurements were converted to floating-point values but otherwise retained in their original scale. In both pipelines, the predefined training, validation, and test splits supplied with the dataset were preserved to ensure consistent and comparable evaluation across models.

Embedding Generation. We evaluate two codon-representation approaches that differ fundamentally in architectural complexity, context modelling, and representational capacity. Codon2Vec models coding sequences as ordered sequences of codon tokens, analogous to words in natural

language. A skip-gram Word2Vec model is trained on the full set of synonymous *mRFP* coding sequences to learn distributed representations of codons based on their local co-occurrence patterns. The model is trained with an embedding dimensionality of 100, a context window of five codons, negative sampling with ten negative examples, and ten training epochs. This procedure yields vector representations for the 60 sense codons observed in the dataset. For downstream analysis, codon-level vectors within each coding sequence are aggregated by mean pooling to produce a fixed-length 100-dimensional sequence embedding. CodonFM employs a transformer-based architecture trained at codon resolution on large-scale coding sequence corpora. We use a pretrained CodonFM Encodon model in inference mode to generate contextual sequence embeddings. Each full coding sequence is processed by the model to produce a single 2048-dimensional embedding that captures codon identity, positional information, and long-range contextual dependencies. Embeddings are generated for all 1,459 *mRFP* variants, resulting in a feature matrix of size 1459×2048 that is aligned with predefined dataset splits.

3.2 Predictive Modelling

To evaluate how effectively the learned embeddings capture information relevant to translation efficiency, we formulate fluorescence prediction as a supervised regression task. Separate regression models are trained on the Codon2Vec and CodonFM sequence embeddings using identical data splits to ensure a fair comparison. We evaluate a range of regression approaches, including linear regression, random forest regression, gradient-boosted decision trees (XGBoost), and shallow feed-forward neural networks. Model hyperparameters are selected using the predefined validation set provided with the dataset, while final performance is reported on the held-out test set. This strategy ensures consistent evaluation across embedding types. Model performance is assessed using the coefficient of determination (R^2), root mean squared error (RMSE), and Spearman rank correlation, which together quantify absolute prediction accuracy and rank-order agreement between predicted and observed expression levels.

4 Result Analysis

4.1 Experimental Setup

All experiments were conducted in Google Colab using NVIDIA T4 GPUs. Training the Codon2Vec embedding model required less than one minute on the full *mRFP* corpus, while generating CodonFM embeddings for all 1,459 sequences required approximately five minutes using the pretrained transformer. Downstream regression model training was computationally lightweight, requiring less than five seconds for Random Forest models and under twenty seconds for multilayer perceptrons.

4.2 Predictive Performance and Interpretation

We compare Codon2Vec and CodonFM at both statistical and biological levels. Quantitatively, model performance metrics measure how effectively each embedding captures variation in experimentally measured fluorescence across synonymous *mRFP* variants. Because initiation-related features are held constant in the dataset, differences in predictive accuracy can be attributed primarily to codon-dependent elongation effects encoded within the embeddings.

Across all evaluated models, CodonFM embeddings substantially outperform Codon2Vec. As shown in Table 1, the Random Forest model trained on CodonFM embeddings achieves ap-

Embedding Model	Test R^2	RMSE	Spearman ρ
Codon2Vec + Random Forest	~ 0.20	higher	~ 0.55
CodonFM + Random Forest	~ 0.44	lower	~ 0.73

Table 1: Predictive performance on the held-out test set using Random Forest regression. CodonFM embeddings consistently outperform Codon2Vec across all evaluation metrics.

proximately double the explained variance on the held-out test set, alongside lower prediction error and higher rank-order correlation. This performance gap indicates that transformer-based contextual embeddings capture elongation-relevant regulatory signals that are not accessible through shallow, local co-occurrence models.

4.3 Embedding Structure and Biological Plausibility

Beyond predictive accuracy, we examined the structure of the learned embedding spaces to assess biological coherence. For Codon2Vec, dimensionality reduction of codon-level embeddings reveals clustering driven primarily by local co-occurrence patterns, with limited separation among functionally distinct synonymous codons. In contrast, CodonFM embeddings yield tighter predicted-versus-observed expression relationships, reduced overfitting, and smoother residual distributions across training, validation, and test splits. Feature importance analyses from tree-based models further suggest that CodonFM distributes predictive signal across multiple embedding dimensions, consistent with encoding long-range context, positional effects, and higher-order dependencies within the coding sequence. Collectively, these observations support the conclusion that transformer-based embeddings provide a more biologically informative representation of elongation-related regulatory features.

5 Conclusion and Future Works

This study evaluated whether codon-level sequence embeddings can capture elongation-driven determinants of translation efficiency from synonymous coding sequences. Using a controlled *mRFP* dataset in which the promoter, ribosome binding site, and 5′UTR are fixed, we isolated codon-dependent effects within the coding sequence and compared a shallow embedding approach (Codon2Vec) with a transformer-based foundation model (CodonFM). Across all regression models and evaluation metrics, CodonFM embeddings consistently achieved substantially higher predictive accuracy, indicating that contextual, long-range modelling of coding sequences provides a more informative representation of elongation-related regulatory signals than local co-occurrence embeddings.

The scope of this work is limited to elongation under fixed initiation conditions. Future work will extend this framework to datasets with systematic variation in 5′UTR sequences, enabling joint modelling of translation initiation and elongation. Additional directions include integrating complementary experimental measurements such as ribosome profiling or mRNA stability data, and exploring task-specific fine-tuning of codon-level transformer models. Together, these extensions would support the development of more comprehensive and mechanistically grounded models of translation efficiency, with direct relevance to synthetic biology and mRNA-based therapeutic design.

References

[1] Hila Gingold and Yitzhak Pilpel. Determinants of translation efficiency and accuracy. *Molecular Systems Biology*, 7:481, 2011.

[2] Grzegorz Kudla, Andrew W. Murray, David Tollervey, and Joshua B. Plotkin. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science*, 324(5924):255–258, 2009.

[3] Daniel B. Goodman, George M. Church, and Sriram Kosuri. Causes and effects of n-terminal codon bias in bacterial genes. *Science*, 342(6157):475–479, 2013.

[4] Andrei A. Komar. A pause for thought: Synonymous codons and co-translational folding. *Cell*, 97(2):143–146, 2009.

[5] Rayson Wint, Tien-Fu F. Ng, and Daniel Cheung. Codon2vec: Distributed representation of codons for gene expression prediction. *Bioinformatics*, 37(22):4082–4090, 2021.

[6] NVIDIA Digital Biology Team. Codonfm: Foundation models for coding sequences. https://github.com/NVIDIA-Digital-Bio/CodonFM, 2024. GitHub repository.

[7] Guillaume Cambray, Siyu Guimaraes, Ariel Y. Mutalik, Vikram Srikumar, Taner Arkin, Christopher A. Endy, and Adam P. Arkin. Quantifying the sequence determinants of translation efficiency in bacteria. *Nature*, 551:547–552, 2017.

[8] Daniel B. Goodman, George M. Church, and Sriram Kosuri. N-terminal codon bias is shaped by mrna structure and translation initiation. *Nature*, 512(7514):441–445, 2015.

[9] Joshua B. Plotkin and Grzegorz Kudla. Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics*, 12(1):32–42, 2011.

[10] Gregory Boël, Rob De Smet, Anastasia Kostova, Alessandra Mangelinck, Eric O. Bollen, Veronica Beerten, Jelle Steen, J. De Baets, K. Callewaert, Wim Versées, Joost Schymkowitz, and Frederic Rousseau. Codon influence on protein expression in *E. coli* correlates with mrna levels. *Nature*, 529:358–363, 2016.

[11] Yuan Ji, Yuzhen Zhao, Qipeng Guo, Tianjiao Liu, Zhiqiang Hu, and Jianyang Zeng. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

[12] Kyle Gion and Lindsey Schell. Introducing the codonfm open model for rna design and analysis. https://developer.nvidia.com/blog/introducing-the-codonfm-open-model-for-rna-design-and-analysis/, October 2025. NVIDIA Developer Blog.