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

Rushmila Islam

1. Introduction

Accurately predicting translation efficiency is central to understanding gene expression and optimising synthetic gene design. Translation efficiency determines how effectively an mRNA is converted into functional protein and directly impacts protein abundance, cellular behaviour, and mRNA-based therapeutic performance. Notably, mRNAs encoding identical amino acid sequences can differ by orders of magnitude in protein output due to regulatory features acting at different stages of translation. Translation occurs in three main stages: initiation, elongation, and termination. Initiation is often rate-limiting and is primarily governed by features within the 5' untranslated region (5'UTR), including ribosome-binding motifs and local RNA structure. In contrast, elongation is controlled by features within the coding sequence, such as synonymous codon usage, tRNA abundance, codon-pair preferences, and mRNA structure, which together influence ribosome dynamics and co-translational protein folding. The first 10–20 codons are particularly important in establishing elongation dynamics. Although synonymous codon substitutions do not alter protein sequence, they can substantially affect translation kinetics, mRNA stability, and folding pathways. Traditional codon-usage metrics such as CAI and tAI provide coarse summaries of codon bias but fail to capture positional and contextual dependencies. Recent advances in natural language processing have therefore inspired data-driven approaches that treat coding sequences as a language, enabling embedding models to learn contextual codon representations. Shallow methods such as Codon2Vec (Wint, 2021) capture local co-occurrence patterns whereas transformer-based models such as CodonFM (NVIDIA Digital Biology, 2024) capture long-range dependencies. A key challenge in evaluating such models is disentangling elongation effects from initiation effects driven by 5′UTR variation. We address this using a controlled dataset in which the promoter, ribosome binding site, and 5'UTR are fixed, and only the coding sequence varies through synonymous substitutions, allowing focused analysis of elongation-driven translation efficiency.

2. Related Works

Synonymous codon usage has long been recognised as a major determinant of translation efficiency and protein abundance. Early molecular and evolutionary studies established that codon bias arises from a combination of tRNA availability, mutation pressure, selection for translational accuracy, and constraints imposed by mRNA structure (Kudla, 2011). These findings overturned the assumption that synonymous mutations are functionally neutral. Large-scale experimental studies have directly quantified the impact of synonymous variation on protein expression. Comprehensive libraries of synonymous variants demonstrated that codon identity, particularly near the N-terminus, strongly influences translation efficiency by modulating early elongation dynamics and local RNA structure (Daniel B. Goodman, 2013). Genome-wide analyses further revealed that codon usage affects not only ribosome elongation rates but also steady-state mRNA levels, highlighting coupled regulation between translation and RNA stability (Gregory Boël, 2016). At the same time, numerous high-impact studies have shown that translation initiation is often the dominant source of expression variability when 5'UTR sequences differ. Systematic analyses of large synthetic UTR libraries demonstrated that ribosome binding site strength, spacing, and RNA accessibility near the start codon can overshadow elongation effects by orders of magnitude (Guillaume Cambray, 2017). These results underscore the importance of isolating regulatory layers when studying translation efficiency. The mRFP dataset used in this work provides exactly such a controlled setting. Because all constructs share an identical promoter, ribosome binding site, and 5'UTR, observed variation in protein expression can be attributed specifically to synonymous differences within the coding sequence (Daniel B. Goodman, 2015). This design enables focused investigation of codon-driven elongation effects. Computational approaches to modelling codon-dependent expression have evolved alongside experimental advances. Codon2Vec introduced distributed representations of codons based on local co-occurrence, enabling machine-learning models to predict gene expression from codon composition (Rayson Wint, 2021). More recently, transformer-based models such as DNABERT (Yuan Ji, 2021) and CodonFM (Kyle Gion, 2025) have demonstrated the ability to learn rich, context-aware representations of biological sequences. CodonFM operates at codon resolution and is trained on billions of codons, making it a promising foundation model for translation-related tasks.

3. Methodology

We use a synonymous-codon expression dataset based on monomeric red fluorescent protein (mRFP), originally introduced in (Daniel B. Goodman, 2013) and subsequently adopted in studies of codon-dependent translation efficiency (Gregory Boël, 2016). 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. 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

3.1 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.

3.2 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 1459 x 2048 that is aligned with dataset splits.

3.3 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 $\mathcal{R}^2$, root mean squared error (RMSE), and Spearman rank correlation, which together quantify prediction accuracy and rank-order agreement between predicted and observed expression levels.

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

Embedding Model	**Test $\mathcal{R}^2$**	**RMSE**	**Spearman ρ**
Codon2Vec + Random Forest	~0.20	Higher	~0.55
CodonFM + Random Forest	~0.44	Lower	~0.73

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 perceptron.

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 approximately 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 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.

Reference

Daniel B. Goodman, G. M. (2013). Causes and effects of N-terminal codon bias in bacterial genes. *Science, 342*(6157), 475–479.

Daniel B. Goodman, G. M. (2015). N-terminal codon bias is shaped by mRNA structure and translation initiation. *Nature, 512*(7514), 441–445.

Essential Cell Biology. (2019). In H. H. Alberts, *Essential Cell Biology* (p. 269).

Gregory Boël, R. D. (2016). Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature, 529*, 358–363.

Guillaume Cambray, S. G. (2017). Quantifying the sequence determinants of translation efficiency in bacteria. *Nature, 551*, 547–552.

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

Kyle Gion, L. S. (2025, Oct 28). *NVIDIA*. Retrieved from https://developer.nvidia.com/blog/introducing-the-codonfm-open-model-for-rna-design-and-analysis/

NVIDIA Digital Biology. (2024). *CodonFM: Foundation Models for Coding Sequences*. Retrieved December 12, 2025, from https://github.com/NVIDIA-Digital-Bio/CodonFM

Rayson Wint, T.-F. F. (2021). *Codon2Vec v1.0*. Retrieved December 12, 2025, from https://github.com/rhondene/Codon2Vec

Rayson Wint, T.-F. F. (2021). Codon2Vec: Distributed representation of codons for gene expression prediction. *Bioinformatics, 37*(22), 4082–4090.

Wint, R. (2021). *U.S. Department of Energy Office of Scientific and Technical Information*. Retrieved December 2025, from https://doi.org/10.11578/dc.20211123.7

Yuan Ji, Y. Z. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics, 37*(15), 2112–2120.