# A generative language model decodes contextual constraints on codon choice for mRNA design

Marjan Faizi[1], Helen Sakharova[2], and Liana F. Lareau[1,2,3,4,*]

[1] *California Institute for Quantitative Biosciences, University of California, Berkeley, CA 94720, USA*
[2] *Center for Computational Biology, University of California, Berkeley, CA 94720, USA*
[3] *Department of Bioengineering, University of California, Berkeley, CA 94720, USA*
[4] *Chan Zuckerberg Biohub, San Francisco, CA 94158, USA*
[*] Correspondence to: lareau@berkeley.edu

June 6, 2025

**Abstract**

The genetic code allows multiple synonymous codons to encode the same amino acid, creating a vast sequence space for protein-coding regions. Codon choice can impact mRNA function and protein output, a consideration newly relevant with advances in mRNA technology. Genomes preferentially use some codons, but simple optimization methods that select preferred codons miss complex contextual patterns. We present Trias, an encoder-decoder language model trained on millions of eukaryotic coding sequences. Trias learns codon usage rules directly from sequence data, integrating local and global dependencies to generate species-specific codon sequences that align with biological constraints. Without explicit training on protein expression, Trias generates sequences and scores that correlate strongly with experimental measurements of mRNA stability, ribosome load, and protein output. The model outperforms commercial codon optimization tools in generating sequences resembling high-expression codon sequence variants. By modeling codon usage in context, Trias offers a data-driven framework for synthetic mRNA design and for understanding the molecular and evolutionary principles behind codon choice.

## 1 Introduction

Synthetic messenger RNA technology has emerged as a transformative tool for biotechnology, opening new frontiers in synthetic biology, genome engineering, and medicine. mRNAs encoding proteins of interest can be delivered into cells for translation by the cell's own machinery. A key challenge is to design mRNA sequences that drive high levels of heterologous protein expression. Natural mRNAs are shaped by overlapping constraints: coding sequences must ensure accurate protein synthesis, while the coding region and untranslated regions (UTRs) also control aspects of mRNA fate like translation initiation rate, structural stability, and half-life. All of these aspects of post-transcriptional regulation are subject to optimization for synthetic mRNA design (Sample *et al.*, 2019; Lewis *et al.*, 2024; Leppek *et al.*, 2022; Zhang *et al.*, 2023).

Design of the coding sequence presents a particular challenge. While the desired protein sequence may be predetermined, there are many ways to encode the same protein sequence in mRNA. Most amino acids can be specified by two or more synonymous codons, so codon choice becomes a combinatorially complex problem. For example, the SARS-CoV-2 spike protein has 1,273 amino acids and can be encoded by more than $10^{632}$ different codon sequences (Kim *et al.*, 2024). Within this vast range of possibilities, codon choice is not random; organisms favor certain codons, and these preferences vary by species. Yet, the underlying rules guiding optimal codon selection remain poorly understood (Wu and Bazzini, 2023).

Codon choice has subtle but powerful impacts on the outcomes of translation. Common codons are decoded faster and more accurately due to more abundant tRNAs, while rare codons slow translation (Liu, 2020). Slow translation can destabilize mRNAs through codon optimality mediated decay (Presnyak *et al.*, 2015; Bazzini *et al.*, 2016; Narula *et al.*, 2019). It can also reduce the number of proteins made per mRNA; codon choice alone can create an eight-fold range in protein output (Tunney *et al.*, 2018; Barrington *et al.*, 2023; Lyons *et al.*, 2024). However, fast translation is

not always ideal. Multi-domain proteins can require precisely timed translation for proper folding and function (Kim *et al.*, 2015; Yu *et al.*, 2015; Kimchi-Sarfaty *et al.*, 2007; Kirchner *et al.*, 2017). Decoding the rules for codon choice thus has important applications for biotechnology and therapeutics (Lin *et al.*, 2023a).

These observations underscore the need for sophisticated approaches to sequence design that consider biological context. The classic approach to codon optimization considers species-specific codon frequencies as well as mRNA structure and GC content. Most commercial tools rely on manually defined features and cannot fully capture the dependencies between codons. In parallel to these feature-based methods, some deep learning approaches learn directly from experimental data on ribosome distribution (Tunney *et al.*, 2018; Jain *et al.*, 2023). These models represent a major step forward by learning codon patterns from functional measurements, but they are based on hard to acquire data and may not capture complex sequence dependencies.

Recent advances in machine learning, particularly large language models (LLMs), have introduced powerful data-driven approaches that capture complex information from biological sequences. These models learn the underlying distributions and patterns from millions of naturally occurring sequences (Simon *et al.*, 2024). LLMs can serve as foundation models that generalize across a wide range of biological tasks, either in a zero-shot setting (without fine-tuning or task-specific examples) or after fine-tuning. For example, the Evolutionary Scale Model (ESM), trained on millions of protein sequences, was fine-tuned to predict protein structure elements and mutation effects (Rives *et al.*, 2021; Lin *et al.*, 2023b). Similarly, genomic foundation models have been trained on millions of DNA sequences and fine-tuned for tasks ranging from variant effect prediction to the generation of novel CRISPR-Cas systems (Benegas *et al.*, 2023; Karollus *et al.*, 2024; Nguyen *et al.*, 2024a,b; Brixi *et al.*, 2025).

Building on the success of these large language models, several codon language models have recently been developed and trained on large datasets of mRNA coding sequences (Li *et al.*, 2024; Constant *et al.*, 2023; Outeiral and Deane, 2024; Sidi *et al.*, 2024; Fallahpour *et al.*, 2025; Naghipourfar *et al.*, 2024). These data-driven models have the potential to uncover complex codon usage rules that traditional methods may not capture, providing an improved tool for codon optimization and a framework for studying synonymous codon choice. However, their use for synthetic mRNA design raises a fundamental question, as these models may not align perfectly with the optimization tasks to which they are applied. Natural mRNA sequences are optimized by evolution for their functional need, not necessarily for maximal expression. Understanding what these models learn about natural codon usage and how this knowledge translates to synthetic applications is critical for their successful use in sequence design.

While codon language models show promise, most existing models are based on encoder-only architectures that do not directly generate sequences. For example, CodonBERT (Li *et al.*, 2024) and CO-BERTa (Constant *et al.*, 2023) are fine-tuned to estimate sequence-level properties such as mRNA half-life or protein expression to support the design of synthetic sequences. These encoder-only models are well suited for classification and regression tasks, while encoder-decoder and decoder-only models are autoregressive architectures designed for sequence generation. Encoder-decoder models first map the input sequence into a latent representation, which is then decoded token by token into the target output (Cho *et al.*, 2014). Decoder-only models use the input as a prompt and generate tokens autoregressively without an intermediate representation. Some codon models have explored generative capabilities. Constant *et al.* (2023) introduced an encoder-decoder model to reverse-translate protein sequences into codons for *E. coli*, and Sidi *et al.* (2024) trained an encoder-decoder model on short fragments from two bacterial and two yeast genomes, limiting the model's ability to capture long-range dependencies across full-length coding sequences. Naghipourfar *et al.* (2024) presented EnCodon (encoder-only) and DeCodon (decoder-only) for prediction of synonymous variant effects and mRNA properties, but neither model was designed to generate codon sequences for a given protein.

Here, we present Trias, a generative language model based on an encoder-decoder architecture, trained on the full-length coding regions of 10 million vertebrate mRNA sequences for the task of reverse-translating protein sequences into codon sequences. Trias is distinct in focusing on eukaryotic sequences, capturing their distinct post-transcriptional sequence constraints to enable sequence design for expression in human cells. We show that Trias captures codon usage patterns that align with biological constraints by learning to generate coding sequences one codon at a time. The model favors frequent codons while preserving rare codons in functionally relevant contexts. With no training on experimental data, Trias predicted scores for synonymous variants of GFP that correlated strongly with measurements of their protein expression, ribosome load, and mRNA half-life in human cells.

Together, these findings highlight the potential of Trias as a data-driven method for codon optimization in recombinant protein production and mRNA-based therapeutics, as well as a tool for studying how codon usage is shaped by sequence context. To facilitate its broad application, we have released Trias on Hugging Face, where it enables reverse translation of protein sequences up to 2,046 amino acids in length into codon sequences for over 600 vertebrate species.

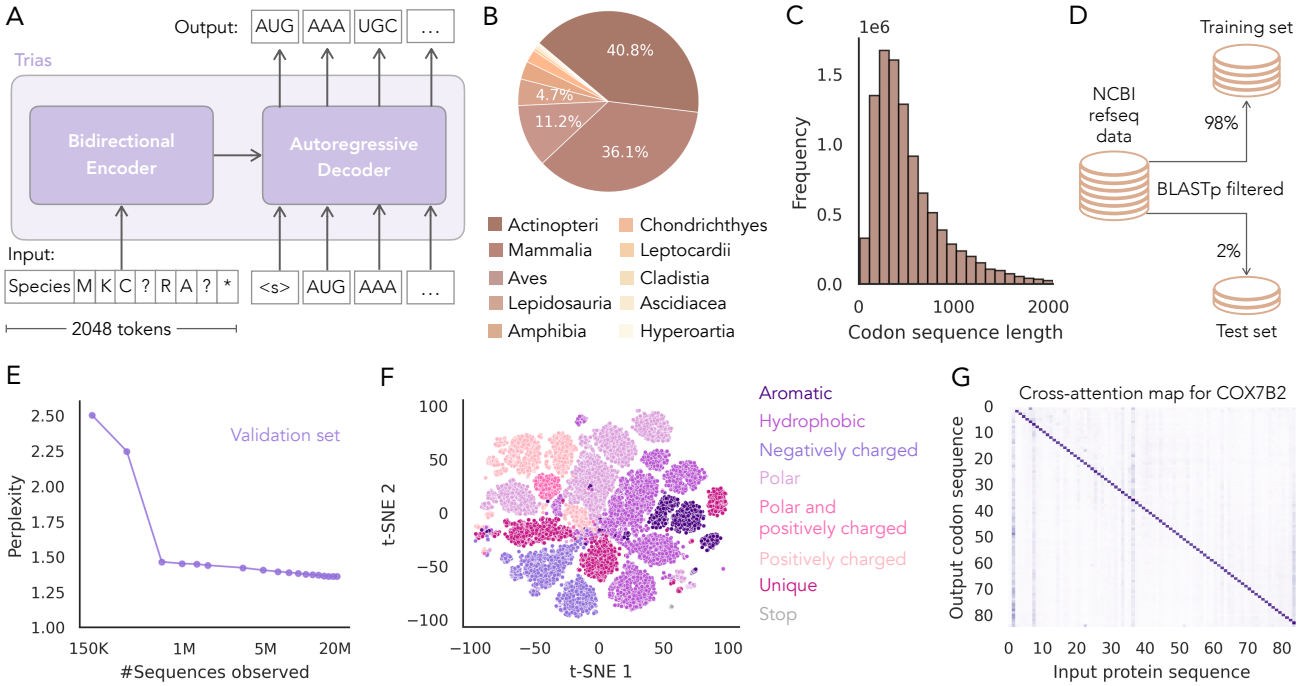


**Figure 1: Model overview, dataset composition and training performance.** (A) Trias is an encoder-decoder model that reverse translates protein sequences into codon sequences while incorporating species-specific codon preferences. During training, 15% of input tokens are randomly masked (indicated as '?'). (B) The model was trained on 10M codon sequences from 640 vertebrate species. (C) Distribution of codon sequence lengths in the dataset. (D) An initial test set was chosen at random and sequences with protein similarity to the initial test set were moved from the training to test set. To monitor training, we additionally randomly sampled 1% of the training data as a validation set. (E) Perplexity during training, showing a decrease from 2.5 to below 1.5 on the validation set. (F) Projection of the final encoder embedding layer for 20 randomly selected human protein sequences, with each dot representing an amino acid. (G) Cross-attention map for example gene, showing how each codon attends to the input amino acid sequence and the species tag. Each row corresponds to a codon, and each column to an amino acid. The example shown represents a single attention head, averaged across all layers.

## 2 Results

### 2.1 Trias learns codon preferences and amino acid properties through generative training

Trias is a sequence-to-sequence model designed to reverse translate protein sequences into codon sequences (Figure 1A). It is based on the BART architecture with bidirectional encoder and autoregressive decoder modules, commonly used in natural language processing tasks such as machine translation (Lewis *et al.*, 2019). The bidirectional encoder module of Trias processes the entire input protein sequence at once, capturing long-range dependencies between amino acids. It produces a latent representation of the sequence that is passed to the decoder. The autoregressive decoder then generates the corresponding codon sequence token by token, conditioning each predicted codon on the previously generated codons and the encoder's representation. The model is trained by providing the encoder module with a corrupted version of each protein sequence with 15% of amino acid positions masked and then calculating the reconstruction loss of the generated codon sequence against the true codon sequence. To allow the model to learn species-specific preferences for codon usage, we prepend species tags to the protein input sequence.

Trias has 47 million parameters. It was trained on 10 million sequences from 640 vertebrate species sourced from the NCBI RefSeq database (Figure 1B). Sequences can be up to 2,046 tokens long (Figure 1C), exceeding the input limits of previous models. To ensure robust evaluation and prevent information leakage between the training and test sets, we removed homologous sequences based on protein-level similarity (Figure 1D); highly similar sequences between sets could allow the model to simply leverage sequence similarity rather than learning generalizable codon preferences.

We evaluated training convergence by tracking perplexity on a validation dataset held out from the training set (Figure 1E). Perplexity quantifies the model's uncertainty in predicting codons, representing the average number of codon choices the model considers per amino acid. Lower values indicate more confident predictions with reduced

uncertainty and fewer equally probable alternatives. During training, perplexity dropped from 2.5 to below 1.5 after 600K examples, indicating that the model quickly learned synonymous codon preferences. We also confirmed that, although Trias was not explicitly provided with the codon table, it quickly learned to select valid synonymous codons for each amino acid (Supplementary Figure S1 and S2).

We confirmed that the encoder module learned meaningful latent representations of amino acids by visualizing the t-SNE projection of the final encoder hidden layer for 20 randomly selected human protein sequences (Figure 1F, Supplementary Figure S3). Amino acids clustered by biochemical properties, similar to ESM-1b (Rives *et al.*, 2021), with distinct groups for hydrophobic, polar, and charged residues. Next, to illustrate how Trias conditions its codon predictions on protein context, we visualized cross-attention weights for an example gene, COX7B2 (Figure 1G, Supplementary Figure S4). Cross-attention indicates how much each predicted codon attends to different positions in the input protein sequence. The strong diagonal signal indicates that codon choice is primarily influenced by the identity of its target amino acid. The vertical band at the start highlights the species tag, which influences codon selection across the sequence. Off-diagonal signals show that the model also attends to distant residues, suggesting that Trias incorporates both local and long-range context during codon generation.

### 2.2 Trias captures functional constraints on codon choice for sequence generation

To evaluate the generative performance of Trias, we generated codon sequences for each protein in the human test set ($\sim$6,000 sequences, including multiple splice isoforms of some genes) using a greedy search decoding strategy that selects the most probable codon at each step based on the previously generated sequence. We then compared these generated sequences to their wild-type counterparts by analyzing codon usage (measured as scaled relative synonymous codon usage, sRSCU, where a maximum value of 1 indicates that a sequence always uses the most common codons), GC content, and mRNA structure (measured as minimum free energy, MFE) (Figure 2A).

Comparison of wild-type and Trias-generated sequences for each gene revealed a distinctive pattern (Figure 2A). The naturally occurring sequences had a unimodal distribution of sRSCU, but the generated sequences fell into two populations. The bulk of the sequences, spanning the full range of natural sRSCU values, were generated to have more optimal codons than their natural counterparts, but a distinct, smaller population of sequences were generated with low sRSCU that more closely matched, or even under-estimated, the low natural sRSCU of these sequences. GC content followed a similar bimodal pattern, and with a stronger overall correlation, as did MFE. Sequences with high sRSCU values, where the model overestimated codon usage, exhibited higher GC content and stability.

As GC content, mRNA structure, and codon choice are intertwined, with an intrinsic link between G-C basepairs and mRNA structure stability, we probed whether Trias learned fundamental preferences related to structure and mRNA stability. Although the overall correlation was weak, codons with higher sRSCU values tend to have higher GC content in humans (Supplementary Figure S5). The higher correlation between the GC content of wild-type and Trias-generated sequences compared to the correlation of sRSCU suggests that GC preferences might drive Trias' codon preferences.

To establish baseline expectations, we compared Trias to two models: a random model that samples uniformly from the set of synonymous codons at each position and a frequency-based baseline model that always selects the most common codon for each amino acid. On average, sequences generated by Trias were 50% similar to their wild-type counterparts (Figure 2B), based on codon identity, compared to 44% for the frequency-based baseline model and 37% for the random model. These baseline models did not show the bimodal distribution of sRSCU, GC content, and MFE observed in Trias outputs (Supplementary Figure S6). These results indicate that Trias learns sequence-specific codon preferences beyond random selection and simple frequency-based patterns.

We then explored possible explanations for the bimodal outcomes. Examples of the two classes are shown in Figure 2C. We selected four genes with naturally low sRSCU (0.81–0.83): two where Trias preserved rare codon usage (TUT4 and ZNF701), and two where it predicted more frequent codons than in the wild-type (RPL17 and MPP3). We plotted %MinMax profiles, showing relative codon usage fluctuations along a sequence, for the real and generated sequences, and assessed codon-level similarity and Euclidean distance between their codon usage profiles (Figure 2C,D). Generated sequences for TUT4 and ZNF701 closely matched the profiles of their wild-type sequences, while RPL17 and MPP3 deviated more strongly.

As ribosomal proteins such as RPL17 tend to be highly expressed, and expression has been linked to codon usage in some model organisms, we investigated whether Trias implicitly learns which genes are highly expressed and assigns them higher sRSCU values. We used gene expression values (measured as mRNA abundance from RNA-seq) from GTEx, taking the median expression across all tissues for 810 genes in our human test set that matched GTEx entries.

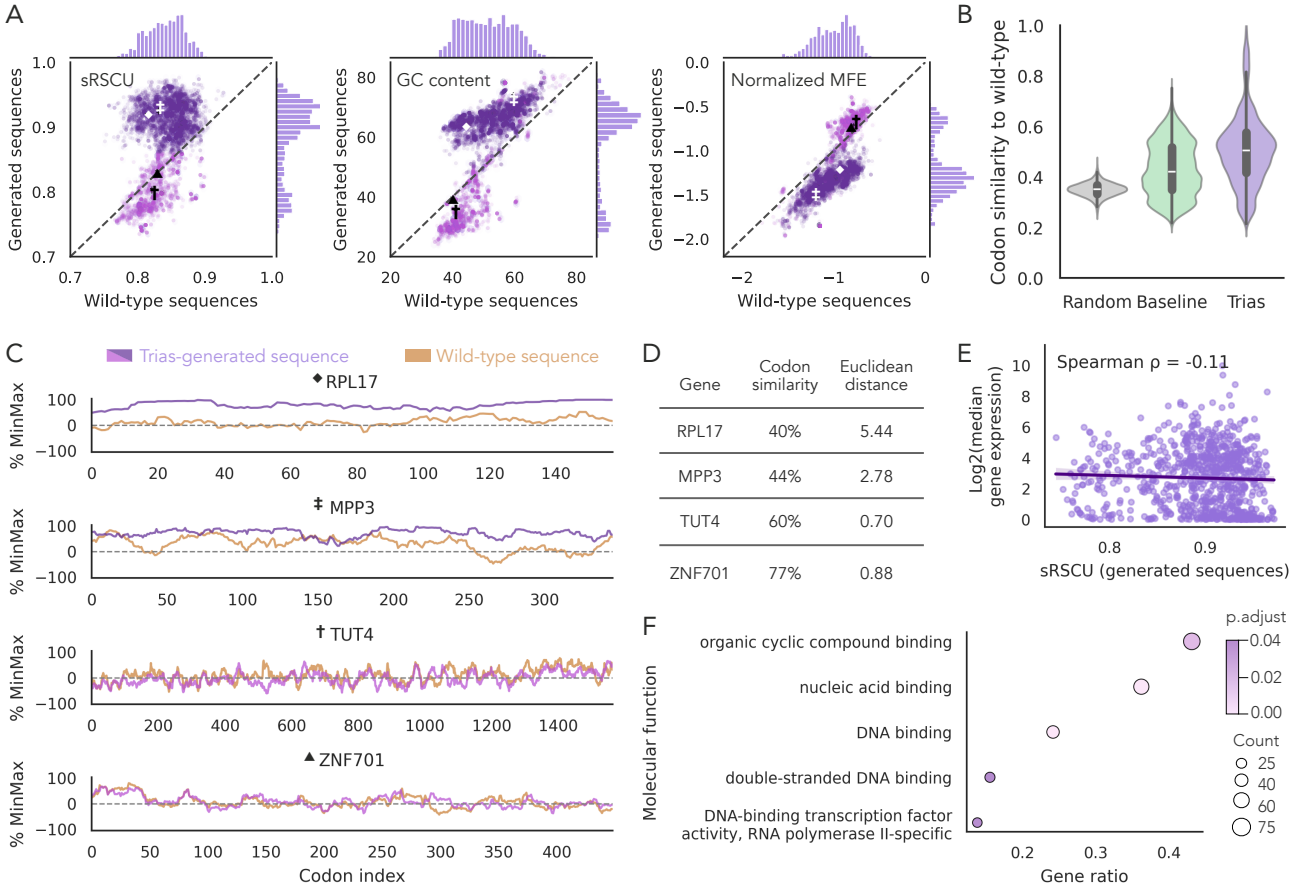


**Figure 2: Sequence-level analysis of Trias-generated human codon sequences.** (A) Comparison of Trias-generated sequences to their wild-type counterparts across scaled relative synonymous codon usage (sRSCU), GC content, and minimum free energy (MFE, normalized by sequence length, computed using the ViennaRNA package (Lorenz *et al.*, 2011)) in the human test set (∼6,000 sequences, including isoforms). Sequences were colored purple or pink according to high or low generated sRSCU values to facilitate comparison across metrics. (B) Codon similarity between Trias-generated sequences and wild-type sequences, benchmarked against a baseline model selecting the most frequent codon at each position and a model that randomly selects synonymous codons for every amino acid. (C) %MinMax profiles illustrate codon usage fluctuations across four genes with naturally low sRSCU (0.81–0.83), marked with symbols in panel A. TUT4 and ZNF701 maintain low sRSCU similar to wild-type. RPL17 and MPP3 have a higher predicted sRSCU (∼0.92). (D) Codon similarity and Euclidean distance between profiles (normalized by sequence length) quantify deviations from the wild-type codon usage profiles. (E) Log2-transformed median gene expression across all tissues in the GTEx dataset plotted against predicted sRSCU values for 810 human test set genes found in GTEx. (F) Gene ontology (GO) enrichment analysis of genes with predicted low sRSCU ($\leq$0.86). Shown are significantly enriched GO terms (adjusted p-value $< 0.05$).

Interestingly, we found no significant correlation between predicted or natural sRSCU and gene expression (Figure 2E, Supplementary Figure S7). These results suggest that highly expressed human genes do not necessarily favor frequent codons, in keeping with evolutionary analyses (Bénitière *et al.*, 2025).

Next, to investigate whether the bimodal outcomes reflect functional constraints, we performed a gene ontology (GO) analysis on the subset of genes with predicted low sRSCU ($\leq$0.86) (Figure 2F). The GO analysis revealed a significant enrichment for transcription factor genes in these predicted low-sRSCU genes, suggesting that rare codon usage in these genes may be under selection to maintain functional roles.

### 2.3 Sequence context influences rare codon predictions

We next identified contexts in which Trias prefers rare codons. Defining rare codons as those with sRSCU $< 0.8$ (Figure 3A), we classified all 4 million codon predictions from the human test set sequences into four categories based

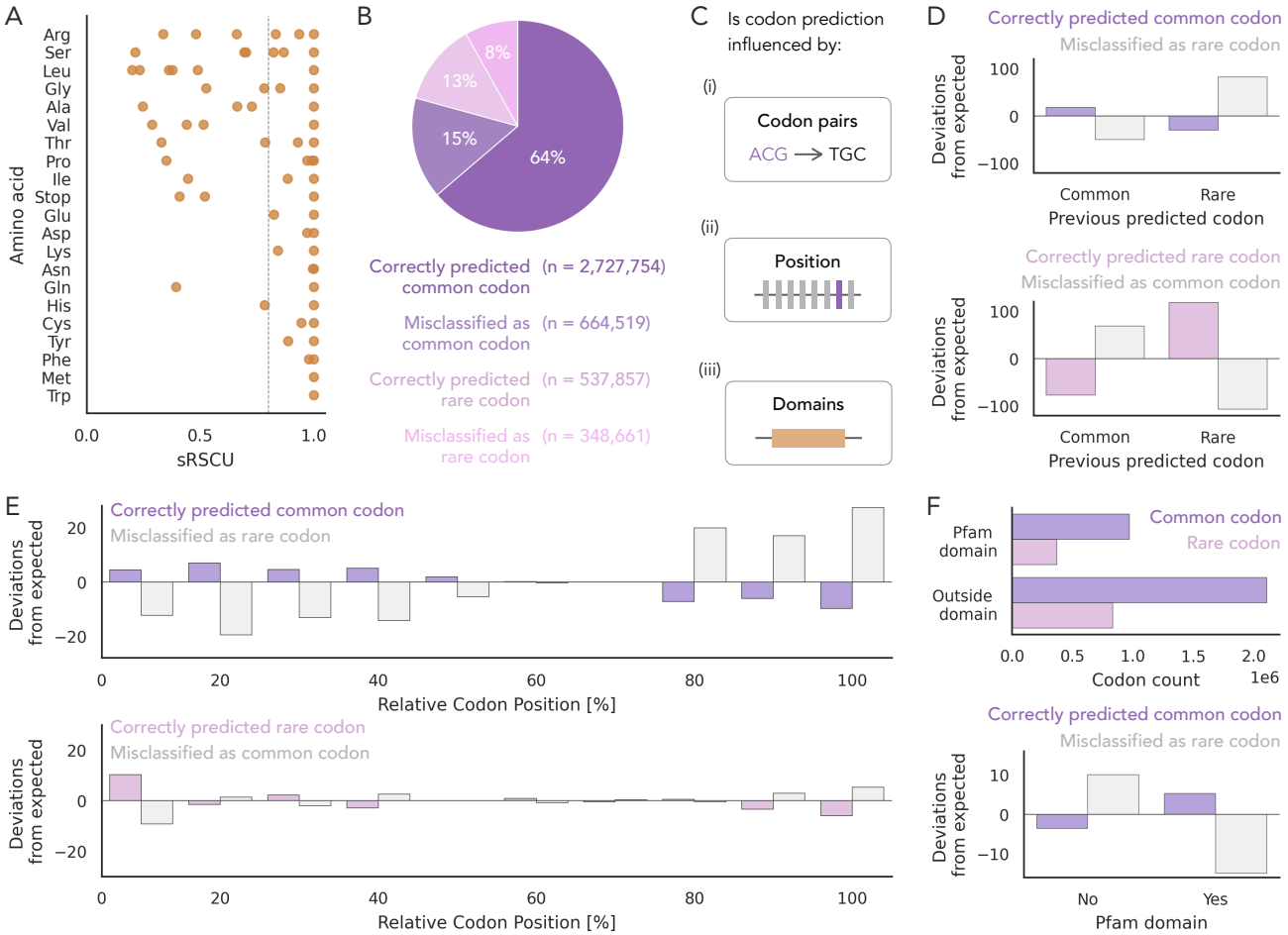


**Figure 3: Contextual influences on codon prediction.** (A) Rare codons were defined using an sRSCU threshold of <0.8. (B) Predictions for all 4,351,017 codons from ∼6,000 human sequences in the test set were categorized into four classes based on whether the predicted codon was common or rare and whether it matched the wild-type codon. (C) To test whether codon predictions are influenced by previous codon identity, sequence position, or domain context, we performed separate chi-squared tests for true common codons (correctly predicted common and misclassified as rare) and true rare codons (correctly predicted rare and misclassified as common). For each analysis, observed prediction frequencies were compared to expected frequencies under the assumption of independence between prediction outcome and the feature. (D–F) Standardized residuals are plotted to show deviations from expected frequencies. (D) Prediction outcomes were significantly associated with the identity of the previous codon (p-value < 1e-5). The strongest deviations were observed for rare codons, which were more likely to follow rare codons than expected. (E) Prediction outcomes varied significantly with codon position (p-value < 1e-5). The largest deviations were observed at the 5′ end (first 40%) and 3′ end (last 30%) for true common codons, and within the first 10% for true rare codons. (F) Over 11,000 Pfam domains were identified using InterProScan across the test set proteins. Codon predictions were significantly affected for true common codons (p-value < 1e-5), but not for true rare codons.

on their wild-type identity and model prediction: (i) correctly predicted common codon, (ii) misclassified as rare codon (common codons incorrectly predicted as rare), (iii) correctly predicted rare codon, and (iv) misclassified as common codon (rare codons incorrectly predicted as common) (Figure 3B). We then tested whether predictions were affected by the preceding codon, codon position, and protein domain localization (Figure 3C).

We found that codon predictions were influenced by the identity of the previously predicted codon (chi-squared test, p-value<1e-5). Rare codons were more likely to follow other rare codons than expected (Figure 3D, Supplementary Figure S8). A similar trend was observed for common codons, where common codon predictions more often followed other common codons, although the effect was less pronounced.

We also found that the codon position, measured across the relative length of each coding sequence, significantly influences the model's predictions (chi squared test, p < 1e-5; Figure 3E). Common codons misclassified as rare were

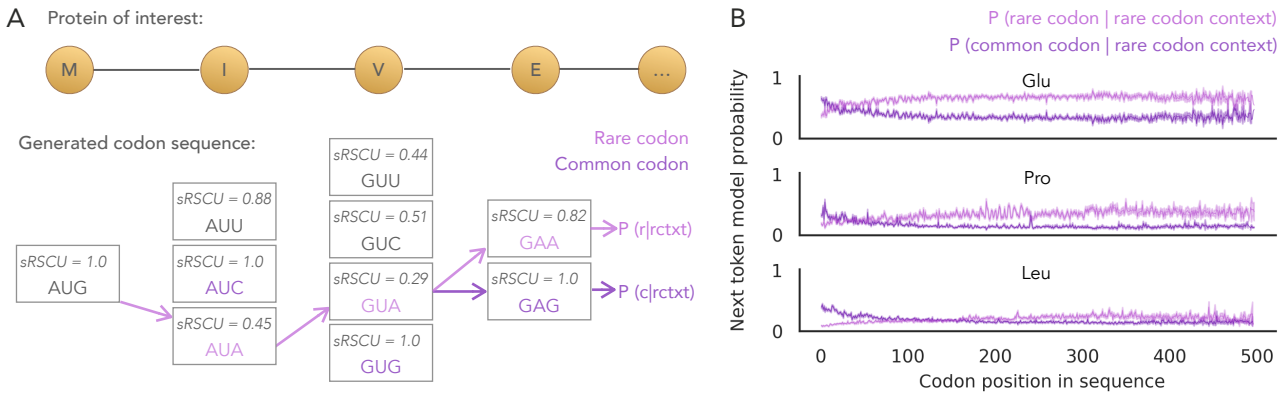


**Figure 4: Effect of codon context on next token probabilities in Trias.** (A) Next token probability is calculated for selecting the rarest (lowest sRSCU) or most common (highest sRSCU) codon, given that all preceding codons are rare. These are denoted as $P(r \mid \text{rctxt})$ and $P(c \mid \text{rctxt})$, with rctxt: rare codon context; r: rare codon; c: common codon. (B) Mean probabilities and confidence intervals across ∼2,500 human sequences in the test set (sequence length ≤500 codons) are shown for three amino acids: glutamate (Glu), proline (Pro), and leucine (Leu).

less frequent near the 5′ end (first 40%) and more frequent toward the 3′ end (last 30%) of sequences. Correctly predicted rare codons were enriched within the first 10% at the 5′ end. This suggests that Trias tends to overestimate rare codon usage toward the end of coding regions, but more accurately predicts rare codons at the beginning.

Finally, we investigated whether codon predictions were affected by location within or outside of protein domains. Around 30% of rare codons and 30% of common codons were located within annotated Pfam domains (Figure 3F). Within domains, common codons were more often predicted as such, and misclassification of common codons as rare was strongly reduced. Outside of domains, common codons were more frequently misclassified as rare. (No significant effect was found for rare codons.) These results suggest that codon usage within domains is more constrained, and Trias more closely aligns with wild-type codon choice for common codons in these regions.

### 2.4 Rare codon context biases generative models toward continued rare codon use

As shown in Figure 2A, Trias-generated sequences generally favor high-frequency codons even for genes whose natural sequence uses rare codons. However, a distinct subset of sequences are generated with rare codons, resulting in codon usage patterns that more closely resemble their wild-type counterparts. This may reflect the tendency of rare codons to cluster in natural sequences (Clarke and Clark, 2024) and is consistent with our earlier observation (Figure 3D) that rare codons are more likely when preceded by other rare codons.

To better understand how rare codon context influences predictions, we designed an *in silico* experiment on the human test set sequences (Figure 4A). At each position, we constrained all preceding codons to be rare (lowest sRSCU per amino acid) and evaluated the probability of selecting the next codon as rare versus frequent (highest sRSCU). We hypothesized that Trias would adapt to a history of seeing rare codons by assigning higher probabilities to subsequent rare codons throughout the sequence, even when globally such choices would be suboptimal. Consistent with our hypothesis, the preference for rare codons increased with the length of the preceding rare-codon context (Figure 4B). This trend was observed across all amino acids, with smaller probability differences for amino acids that have more synonymous codons (Supplementary Figure S9). We conclude that the model learned a form of internal consistency: although it is likely to predict preferred codons in most situations, it can detect genes that have stretches of rare codons and continue generating additional rare codons in these sequences.

### 2.5 Trias generates codon sequences that align with experimental data

To evaluate the connection between Trias predictions and protein output, we benchmarked its predictions against experimental data. A recent study from Bicknell *et al.* (2024) measured the output of 30 different encodings of GFP, differing only in their codon composition and spanning a range of sRSCU and MFE values, that were transfected into HEK293 cells as N1-methylpseudouridine-modified GFP mRNAs (Supplementary Figure S10). We scored each sequence with Trias by computing the negative log-likelihood of the model's probabilities, normalized by sequence

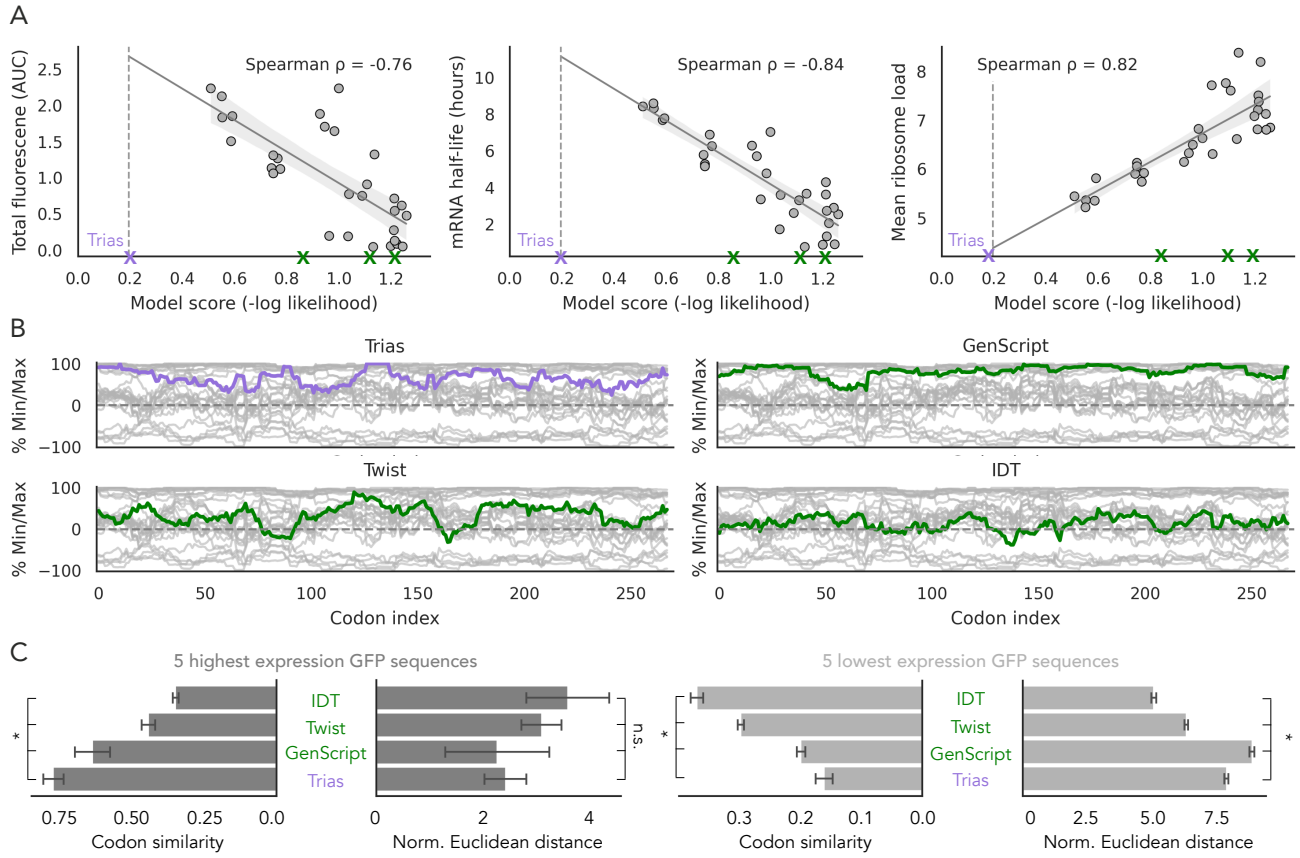


**Figure 5: Trias-predicted codon sequences correlate with experimental data and differ from commercial codon optimization tools.** (A) Correlation between Trias-predicted sequence likelihood and experimentally measured properties of 30 GFP codon variants expressed in HEK293 cells (data from Bicknell *et al.* (2024)). The purple cross marks the Trias-generated sequence obtained via greedy decoding (score = 0.19), with the regression line extrapolated up to this point. AUC: area under the curve quantifies total GFP fluorescence over time. (B) %MinMax codon usage profiles of all 30 GFP variants (gray) compared to sequences generated by Trias and three commercial codon optimization tools (GenScript, Twist, IDT). Trias likelihood scores for sequences generated by these tools were 0.86, 1.13, and 1.22 respectively (green markers in A). (C) Codon similarity and normalized Euclidean distances comparing Trias and commercial tool-generated sequences to the five highest- and lowest-expression GFP variants. Asterisks (*) denote statistical significance (p-value<0.05); error bars show standard deviation.

length, and tested correlations with experimental measurements (Figure 5A). Trias scores correlated strongly with the measured GFP protein output ($\rho = -0.76$), mRNA half-life ($\rho = -0.84$), and mean ribosome load ($\rho = 0.82$). Despite being trained only on natural sequences without any experimental data, Trias learns codon patterns that align with experimentally measured properties. In contrast, sRSCU alone is insufficient as a predictor (Supplementary Figure S10B), highlighting Trias' ability to capture additional sequence context beyond codon frequency.

In addition to scoring the experimental GFP variants, we generated a new human GFP codon sequence using a greedy decoding approach. Its likelihood score of 0.20 exceeded the scores of all experimentally tested sequences. Given the strong correlation between model scores and experimental protein expression, we extrapolated the regression line in Figure 5A to estimate a potential expression level of around 2.67 for this Trias-optimized sequence. We also compared this optimal Trias-generated GFP sequence to sequences designed by commercially available codon optimization tools from GenScript[1], Twist [2] and IDT[3]. Trias ranked these sequences significantly lower, with likelihood scores of 0.86, 1.13 and 1.22, respectively (Figure 5A). We found that the Trias and GenScript sequences predominantly followed the upper range of the possible %MinMax profiles, indicating a preference for more frequently used codons (Figure 5B).

[1]https://www.genscript.com/gensmart-free-gene-codon-optimization.html
[2]https://www.twistbioscience.com/resources/digital-tools/codon-optimization-tool
[3]https://www.idtdna.com/pages/tools/codon-optimization-tool

While GenScript consistently chose frequent codons, Trias used a wider range of codon choices, perhaps reflecting more nuanced rules.

We then compared the sequences produced by Trias and commercial codon optimization tools to the highest- and lowest-expressing GFP variants (Figure 5B,C). The Trias-generated sequence showed high codon similarity to the five highest-expressing variants (76%) and low similarly to the lowest-expressing variants (16%). The sequences generated by GenScript, Twist and IDT were significantly less similar to the high expression variants. Similarly, the %MinMax profiles of Trias and GenScript sequences more closely resembled those of the highest-expressing variants, while Twist and IDT matched more closely with mid-sRSCU GFP variants (Supplementary Figure S11). Taken together, these comparisons with empirical results suggest that Trias-generated sequences more effectively capture codon usage patterns linked to high protein expression compared to commercially optimized sequences.

## 3 Discussion

Trias is a species-aware encoder-decoder model trained on 10 million vertebrate coding sequences to reverse translate protein sequences into codon sequences. Trias learns codon usage directly from natural sequences, capturing biologically meaningful patterns beyond simple frequency-based approaches. This data-driven strategy presents a promising alternative for designing codon-optimized mRNA sequences.

Our analysis reveals that Trias favors common codons and generates sequences with high mRNA stability. It also retains rare codons in biologically relevant positions, indicating that it captures functional constraints on codon choice. Gene-specific differences in rare codon usage suggest Trias learns rules beyond general frequency trends.

An implicit assumption in many applications of nucleotide foundation models for sequence design is that higher-likelihood sequences will correspond to higher protein expression. While this may hold in some cases, natural mRNAs are not necessarily optimized for high expression. Despite this conceptual limitation and without explicit training on experimental protein expression data, Trias' predicted scores for GFP codon variants show strong correlations with protein output, mRNA half-life, and ribosome load. Thus, our results validate the approach of learning from natural optimization in designing synthetic mRNAs for high expression.

The exceptions to this pattern are important to consider. We find that Trias' autoregressive decoding introduces a form of local internal consistency, where rare codon choices are more likely to be followed by additional rare codons. This behavior can lead to generation of sequences with codon usage that may echo nature but does not correspond to high expression. These findings have broader implications for the use of generative language models in predictive settings. Off-the-shelf models trained on natural sequences may carry biases that affect downstream applications, not only for zero-shot prediction of protein expression, but also for tasks such as variant effect prediction, where sequence likelihoods are often used as proxies for functional impact or pathogenicity. For Trias, fine-tuning on protein expression data could help reduce these biases and further enhance its ability to design codon sequences optimized for high protein expression.

Foundation models of nucleic acid sequence have become promising tools for synthetic mRNA sequence design, with broad applications in recombinant protein production and mRNA-based therapeutics. Unlike commercial codon optimization tools that rely on predefined rules, Trias leverages a data-driven approach that generalizes across species and functional contexts. Our benchmarking demonstrated that Trias-generated sequences more closely resemble high-expression GFP variants than those from commercial tools. In contrast to other deep learning models, Trias focuses on vertebrate sequences to gain specific insight into the rules most likely to apply to human mRNA design; these rules can differ in key ways from prokaryotic translation constraints. Despite training on a smaller, phylogenetically narrower dataset, our model achieves strong performance, supporting the value of models tailored to specific biological scenarios. Trias is distinct in integrating generative sequence design into its architecture, limiting our ability to compare it directly to published codon language models that focus on predictive tasks rather than sequence generation. Nonetheless, our benchmarking against experimental protein expression data demonstrates that Trias effectively captures biologically relevant codon usage patterns that translate to improved protein output.

Ultimately, a complete model for synthetic mRNA design must incorporate regulatory features beyond the codon sequence itself. Untranslated regions (UTRs) play key roles in mRNA stability and translation efficiency (Mayr, 2017; Leppek *et al.*, 2018), and the UTR sequence can affect the impact of codon choice (Lyons *et al.*, 2024). Understanding these interactions could further refine the model's ability to design synthetic mRNA sequences; recent work has shown that incorporating UTRs can improve predictions of mRNA properties and protein expression (Li *et al.*, 2025). A model of synthetic mRNA must also consider the impact of chemical modifications such as with N1-methylpseudouridine that are necessary to reduce innate immune reactions to therapeutic mRNAs (Metkar *et al.*, 2024). These modifications,

not found in natural mRNAs, may affect mRNA structure and translation efficiency. Therefore, fine-tuning Trias on chemically modified sequences will be important to account for these effects and adapt codon choices accordingly. Future work to expand the scope of our model can fully unlock the potential of generative models for mRNA design in synthetic biology and therapeutic applications.

## 4 Methods

### 4.1 Datasets

#### *4.1.1 Training and test dataset*

We downloaded all vertebrate sequences from the NCBI RefSeq database[4] on August 5th, 2024. To ensure data quality, we applied the following filtering criteria: sequences labeled as "pseudo" were excluded, and only records containing the feature tags "CDS" and "translation" were kept. We verified that each sequence was complete, initiated with the start codon "ATG" and terminated with one of the three stop codons ("TAA" "TGA" or "TAG"). Additionally, the translated protein sequences were required to start with methionine "M" and end with a stop symbol "*".

After processing the downloaded records, we filter the mRNA sequences according to the length of the coding region. This step resulted in approximately 26 million sequences. To reduce redundancy, we clustered sequences with 90% or greater nucleotide similarity using MMSeqs2 and kept only one representative sequence per cluster. This step reduced the data set to around 10.2 million sequences.

To evaluate model performance, we created a train-test split from our curated dataset of 10.2 million coding sequences. For the test set, we randomly selected 300 human genes from each expression category (high, medium, and low) based on GTEx expression levels, ensuring a diverse range of expression profiles. These 900 genes were translated into protein sequences and used to query the rest of the dataset using BLAST. Any sequences in the remaining dataset with significant protein similarity to these test genes were removed and added to the test set to avoid overlap. This resulted in a test set of approximately 200,000 sequences, including the original 900 human genes. In total, the test set contains approximately 6,000 human coding sequences, including multiple splice isoforms of some genes. The final training set consisted of all 10 million remaining sequences from the curated dataset after removing those with significant protein similarity to the test genes. This approach ensured that the test set, used for performance evaluation, did not include sequences with close homologues in the training set. To monitor training progress, we randomly sampled 1% of the training data as a validation set, with sampling balanced across species.

#### *4.1.2 Gene expression data from GTEx*

The GTEx dataset provides gene expression data across 54 human tissues from approximately 1,000 individuals, measured via bulk RNA sequencing (RNA-seq). For this study, gene expression data were obtained from the GTEx Portal[5] on December 11th 2024, from the GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct.gz file, which contains median gene-level TPM values per tissue. Genes with no expression (TPM = 0) in any tissue were excluded. We then computed the median TPM across all tissues and retained only genes that matched those in our human test set.

### 4.2 Model architecture and training

Trias is based on the HuggingFace implementation of the BART encoder-decoder architecture (Lewis *et al.*, 2019). The input sequence length is set to 2,048 tokens, with the first token reserved for a species tag to ensure species-specific optimization. The vocabulary consists of 64 codon tokens, 20 amino acid tokens, 640 species-specific tokens, and 4 special tokens (</s>, <unk>, <pad>, <mask>).

The model uses the BartForConditionalGeneration class, with 15% of amino acid tokens in the protein sequence masked during training. Flash attention is applied to improve memory efficiency and computational speed. Optimization is performed using the Adam optimizer with a cosine learning rate schedule and warm-up steps. The learning rate starts at 1e-4 and a weight decay of 0.01 is applied. Dropout is set to 0.1. The model is trained for 650,000 steps, corresponding to approximately two epochs.

[4] https://ftp.ncbi.nlm.nih.gov/refseq/release/
[5] https://www.gtexportal.org/home/

Training is carried out on an NVIDIA A100 GPU with 80 GB of memory. The batch size is set to 8, and gradient accumulation is used with an accumulation factor of 4, resulting in an effective batch size of 32. The architecture consists of 6 encoder layers and 6 decoder layers, each with 8 attention heads. The hidden size of the feedforward networks is 2,048 for both the encoder and decoder. This configuration results in a model with approximately 47M trainable parameters.

### 4.3 Data analysis and performance metrics

#### *4.3.1 Gene ontology enrichment analysis*

To investigate functional constraints on codon usage, we performed a gene ontology (GO) enrichment analysis on genes with predicted low and high sRSCU values. Prior to analysis, we removed isoforms from the human test set, resulting in a non-redundant set of 879 genes (down from 6,165 sequences, including isoforms). For each gene, we aggregated predicted and natural sRSCU values by taking the mean across isoforms. We classified genes into two groups based on their predicted sRSCU: genes with low sRSCU ($\leq$0.86) and high sRSCU (>0.86). We then performed GO analysis to identify functional terms overrepresented in these subsets compared to the full set of 879 human test genes. GO term enrichment was assessed using the gprofiler Python package, with the background set defined as all 879 genes in the test set. We applied Bonferroni correction for multiple hypothesis testing and set a significance threshold of p-value $< 0.05$. GO terms related to molecular function, biological process, and cellular component categories were considered in the analysis.

#### *4.3.2 Codon usage calculation*

To quantify codon usage bias, we computed the scaled relative synonymous codon usage (sRSCU) for every codon in each species based on species-specific codon frequency in the training set. sRSCU normalizes codon frequency within synonymous codon categories, allowing direct comparisons across different amino acids. It measures how frequently a given codon is used relative to the expected uniform distribution of synonymous codons for an amino acid. The relative synonymous codon usage (RSCU) for a given codon $c$ is calculated following Sharp *et al.* (1986):

$$\mathrm{RSCU}_{i,j} = \frac{c_{i,j}}{\frac{1}{n_i}\sum_{j=1}^{n_i} c_{i,j}}, \tag{1}$$

where $c_{i,j}$ is the observed frequency of the $jth$ codon for the $ith$ amino acid in the training set for a specific species, and $n_i$ is the number of synonymous codons for the $ith$ amino acid. To derive sRSCU, we scale RSCU values by dividing by the highest RSCU value for each amino acid within a species:

$$\mathrm{sRSCU}_{i,j} = \frac{\mathrm{RSCU}_{i,j}}{\max(\mathrm{RSCU}_{i,1}, \mathrm{RSCU}_{i,2}, ..., \mathrm{RSCU}_{i,n_i})} \tag{2}$$

This ensures that the most frequently used synonymous codon receives a score of 1, while other synonymous codons are scaled accordingly. To calculate the sRSCU for a gene, we computed the average sRSCU value of all codons within that gene.

#### *4.3.3 Expected value calculation for chi-squared tests*

We used chi-squared tests to assess whether prediction outcomes were associated with specific sequence features (previous codon, codon position, and domain localization). Codon predictions from the test set were categorized into four classes (correctly predicted common, misclassified as rare, correctly predicted rare, or misclassified as common). Observed frequencies for all four prediction categories across each sequence feature are reported in Supplementary Tables S1–S4. For the chi-squared tests, we constructed separate contingency tables for true common codons (correctly predicted common and misclassified as rare) and true rare codons (correctly predicted rare and misclassified as common). Expected counts were calculated under the null hypothesis that codon prediction categories are independent of the tested feature. Expected values $E_{ij}$ were computed as follows:

$$E_{ij} = \frac{R_i \cdot C_j}{N}, \tag{3}$$

where $R_i$ is the total count for feature level $i$, $C_j$ is the total count for prediction category $j$, and $N$ is the total number of codon predictions. To identify which feature levels contributed most to significant associations, we also computed standardized residuals.

#### *4.3.4 Pfam domain annotation*

We used InterProScan to identify protein domains in our test set. It scans protein sequences against a comprehensive collection of databases to annotate known domains, families, and functional sites. InterProScan was run locally on all ∼6,000 human protein sequences in the test set. From the resulting annotations, we extracted Pfam domain matches and kept only those with e-values <1e-5 to ensure high-confidence predictions. The e-value indicates the number of matches expected by chance, with lower values reflecting higher significance. Codons were labeled as within a domain if they overlapped a detected Pfam domain, and as outside a domain otherwise.

#### *4.3.5 Codon usage profiles*

To analyze codon usage variability along sequences, we computed %MinMax profiles, which quantify deviations in codon usage relative to expected frequency distributions. This approach is adapted from Rodriguez *et al.* (2018) and provides a measure of codon bias by comparing observed codon frequencies within a sliding window (size = 18) to the expected frequency range. Positive %MinMax values indicate a preference for frequently used codons, while negative values reflect increased usage of rare codons.

To compare codon usage patterns between sequences, we used normalized Euclidean distance between their %MinMax profiles. The Euclidean distance was computed across the profile and normalized by sequence length. This approach provides a quantitative measure of similarity in codon usage fluctuations along the sequence.

#### *4.3.6 Sequence likelihood scoring*

To assess the likelihood of a codon sequence $S$, we define a scoring function $L(S)$ based on the negative log-likelihood assigned by the model. Trias decodes codon sequences autoregressively, generating codons one at a time, where each codon is predicted based on the previously generated codons and the input protein sequence. The likelihood of a sequence $S = (c_1, c_2, ..., c_N)$ is given by:

$$L(S) = -\frac{1}{N}\sum_{i=1}^{N} \log P(c_i|c_1, c_2, ..., c_{i-1}), \tag{4}$$

where $P(c_i|c_1, c_2, ..., c_{i-1})$ is the model's predicted probability for codon $c_i$, given the preceding sequence $(c_1, c_2, ..., c_{i-1})$. The log-likelihood is normalized by sequence length $N$. A lower $L(S)$ value indicates a sequence that is more likely under the model, meaning the model assigns high probabilities to codon choices at each position. A score of 0 corresponds to a sequence where the model is fully confident in every prediction ($P(c_i) = 1$), while higher scores reflect more uncertain predictions. Trias uses greedy decoding to generate codon sequences, a decoding strategy in which at each step the model selects the most probable codon. Greedy decoding ensures high-likelihood outputs.

## Code availability

The code used to train Trias is available at https://github.com/lareaulab/Trias, and model weights are hosted on https://huggingface.co/lareaulab/Trias.

## Acknowledgments

This work was supported by grants from the National Institute of General Medical Sciences R01GM132104, the Rose Hills Foundation, the Shirl and Kay Curci Foundation, and the Chan Zuckerberg Biohub (to L.F.L.). H.S. was supported by the Department of Defense through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. We thank Kinsey Long for insightful input into the results.

## References


Barrington, C. L., Galindo, G., Koch, A. L., Horton, E. R., Morrison, E. J., Tisa, S., Stasevich, T. J., and Rissland, O. S. (2023). Synonymous codon usage regulates translation initiation. *Cell Reports*, **42**(12).

Bazzini, A. A., Viso, F. D., Moreno-Mateos, M. A., Johnstone, T. G., Vejnar, C. E., Qin, Y., Yao, J., Khokha, M. K., and Giraldez, A. J. (2016). Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition. *The EMBO Journal*, **35**(19), 2087–2103.

Benegas, G., Batra, S. S., and Song, Y. S. (2023). Dna language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, **120**(44), e2311219120.

Bénitière, F., Lefébure, T., and Duret, L. (2025). Variation in the fitness impact of translationally optimal codons among animals. *Genome Research*, **35**(3), 446–458.

Bicknell, A. A., Reid, D. W., Licata, M. C., Jones, A. K., Cheng, Y. M., Li, M., Hsiao, C. J., Pepin, C. S., Metkar, M., Levdansky, Y., Fritz, B. R., Andrianova, E. A., Jain, R., Valkov, E., Köhrer, C., and Moore, M. J. (2024). Attenuating ribosome load improves protein output from mRNA by limiting translation-dependent mRNA decay. *Cell Reports*, **43**, 114098.

Brixi, G., Durrant, M. G., Ku, J., Poli, M., Brockman, G., Chang, D., Gonzalez, G. A., King, S. H., Li, D. B., Merchant, A. T., Naghipourfar, M., Nguyen, E., Ricci-Tam, C., Romero, D. W., Sun, G., Taghibakshi, A., Vorontsov, A., Yang, B., Deng, M., Gorton, L., Nguyen, N., Wang, N. K., Adams, E., Baccus, S. A., Dillmann, S., Ermon, S., Guo, D., Ilango, R., Janik, K., Lu, A. X., Mehta, R., Mofrad, M. R., Ng, M. Y., Pannu, J., Ré, C., Schmok, J. C., John, J. S., Sullivan, J., Zhu, K., Zynda, G., Balsam, D., Collison, P., Costa, A. B., Hernandez-Boussard, T., Ho, E., Liu, M.-Y., McGrath, T., Powell, K., Burke, D. P., Goodarzi, H., Hsu, P. D., and Hie, B. L. (2025). Genome modeling and design across all domains of life with Evo 2. *bioRxiv*.

Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *arXiv*.

Clarke, T. F. and Clark, P. L. (2024). Rare Codons Cluster. *PloS One*, **3**(10), e3412.

Constant, D. A., Gutierrez, J. M., Sastry, A. V., Viazzo, R., Smith, N. R., Hossain, J., Spencer, D. A., Carter, H., Ventura, A. B., Louie, M. T. M., Kohnert, C., Consbruck, R., Bennett, J., Crawford, K. A., Sutton, J. M., Morrison, A., Steiger, A. K., Jackson, K. A., Stanton, J. T., Abdulhaqq, S., Hannum, G., Meier, J., Weinstock, M., and Gander, M. (2023). Deep learning-based codon optimization with large-scale synonymous variant datasets enables generalized tunable protein expression. *bioRxiv*.

Fallahpour, A., Gureghian, V., Filion, G. J., Lindner, A. B., and Pandi, A. (2025). CodonTransformer: a multispecies codon optimizer using context-aware neural networks. *Nature Communications*, **16**(3205).

Jain, R., Jain, A., Mauro, E., LeShane, K., and Densmore, D. (2023). ICOR: improving codon optimization with recurrent neural networks. *BMC Bioinformatics*, **24**(132).

Karollus, A., Hingerl, J., Gankin, D., Grosshauser, M., Klemon, K., and Gagneur, J. (2024). Species-aware dna language models capture regulatory elements and their evolution. *Genome Biology*, **25**(1), 83.

Kim, S. J., Yoon, J. S., Shishido, H., Yang, Z., Rooney, L. A., Barral, J. M., and Skach, W. R. (2015). Translational tuning optimizes nascent protein folding in cells. *Science*, **348**(6233), 444–448.

Kim, Y.-A., Mousavi, K., Yazdi, A., Zwierzyna, M., Cardinali, M., Fox, D., Peel, T., Coller, J., Aggarwal, K., and Maruggi, G. (2024). Computational design of mRNA vaccines. *Vaccine*, **42**(7), 1831–1840.

Kimchi-Sarfaty, C., Oh, J. M., Kim, I.-W., Sauna, Z. E., Calcagno, A. M., Ambudkar, S. V., and Gottesman, M. M. (2007). A "silent" polymorphism in the mdr 1 gene changes substrate specificity. *Science*, **315**(5811), 525–528.

Kirchner, S., Cai, Z., Rauscher, R., Kastelic, N., Anding, M., Czech, A., Kleizen, B., Ostedgaard, L. S., Braakman, I., Sheppard, D. N., and Ignatova, Z. (2017). Alteration of protein function by a silent polymorphism linked to tRNA abundance. *PLoS Biology*, **15**(5), e2000779.

Leppek, K., Das, R., and Barna, M. (2018). Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nature Reviews Molecular Cell Biology*, **19**, 158–174.

Leppek, K., Byeon, G. W., Kladwang, W., Wayment-Steele, H. K., Kerr, C. H., Xu, A. F., Kim, D. S., Topkar, V. V., Choe, C., Rothschild, D., Tiu, G. C., Wellington-Oguri, R., Fujii, K., Sharma, E., Watkins, A. M., Nicol, J. J., Romano, J., Tunguz, B., Diaz, F., Cai, H., Guo, P., Wu, J., Meng, F., Shi, S., Eterna Participants, Dormitzer, P. R., Solorzano, A., Barna, M., and Das, R. (2022). Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics. *Nature Communications*, **13**(1), 1536.

Lewis, C. J., Xie, L., Bhandarkar, S., Jin, D., Abdallah, K. S., Draycott, A. S., Chen, Y., Thoreen, C. C., and Gilbert, W. V. (2024). Quantitative profiling of human translation initiation reveals regulatory elements that potently affect endogenous and therapeutically modified mRNAs. *bioRxiv*, page 2024.02.28.582532.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv*.

Li, S., Moayedpour, S., Li, R., Bailey, M., Riahi, S., Kogler-Anele, L., Miladi, M., Miner, J., Zheng, D., Wang, J., Balsubramani, A., Tran, K., Zacharia, M., Wu, M., Gu, X., Clinton, R., Asquith, C., Skaleski, J., Boeglin, L., Chivukula, S., Dias, A., Montoya, F. U., Agarwal, V., Bar-Joseph, Z., and Jager, S. (2024). CodonBERT large language model for mRNA vaccines. *Genome Research*, **34**(7), 1027–1035.

Li, S., Noroozizadeh, S., Moayedpour, S., Kogler-Anele, L., Xue, Z., Zheng, D., Montoya, F. U., Agarwal, V., Bar-Joseph, Z., and Jager, S. (2025). mRNA-LM: full-length integrated SLM for mRNA analysis. *Nucleic Acids Research*, **53**(3), gkaf044.

Lin, B. C., Kaissarian, N. M., and Kimchi-Sarfaty, C. (2023a). Implementing computational methods in tandem with synonymous gene recoding for therapeutic development. *Trends in Pharmacological Sciences*, **44**(2), 73–84.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. (2023b). Evolutionary-scale prediction of atomic-level proteinstructure with a language model. *Science*, **379**, 1123–1130.

Liu, Y. (2020). A code within the genetic code: codon usage regulates co-translational protein folding. *Cell Communication and Signaling*, **18**(1), 145.

Lorenz, R., Bernhart, S. H., zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, **6**(26).

Lyons, E. F., Devanneaux, L. C., Muller, R. Y., Freitas, A. V., Meacham, Z. A., McSharry, M. V., Trinh, V. N., Rogers, A. J., Ingolia, N. T., and Lareau, L. F. (2024). Translation elongation as a rate limiting step of protein production. *bioRxiv*, pages 2023–11.

Mayr, C. (2017). Regulation by 3'-Untranslated Regions. *Annual Review of Genetics*, **51**, 171–194.

Metkar, M., Pepin, C. S., and Moore, M. J. (2024). Tailor made: the art of therapeutic mRNA design. *Nature Reviews Drug Discovery*, **23**, 77–83.

Naghipourfar, M., Chen, S., Howard, M. K., Macdonald, C. B., Saberi, A., Hagen, T., Mofrad, M. R. K., Coyote-Maestas, W., and Goodarzi, H. (2024). A Suite of Foundation Models Captures the Contextual Interplay Between Codons. *bioRxiv*.

Narula, T., Ellis, J., Taliaferro, J. M., and Rissland, O. S. (2019). Coding regions affect mRNA stability in human cells. *RNA*, **25**(12), 1751–1764.

Nguyen, E., , Poli, M., Faizi, M., Thomas, A. W., Sykes, C. B., Wornow, M., Patel, A., Rabideau, C., Massaroli, S., Bengio, Y., Ermon, S., Baccus, S. A., and Re, C. (2024a). HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. *Advances in Neural Information Processing Systems*, **36**.

Nguyen, E., Poli, M., Durrant, M. G., Kang, B., Katrekar, D., Li, D. B., Bartie, L. J., Thomas, A. W., King, S. H., Brixi, G., Sullivan, J., Ng, M. Y., Lewis, A., Lou, A., Ermon, S., Baccus, S. A., Hernandez-Boussard, T., Ré, C., Hsu, P. D., and Hie, B. L. (2024b). Sequence modeling and design from molecular to genome scale with Evo. *Science*, **386**(746).

Outeiral, C. and Deane, C. M. (2024). Codon language embeddings provide strong signals for protein engineering. *Nature Machine Intelligence*, **6**, 170–179.

Presnyak, V., Alhusaini, N., Chen, Y.-H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K. E., Graveley, B. R., and Coller, J. (2015). Codon optimality is a major determinant of mRNA stability. *Cell*, **160**(6), 1111–1124.

Rives, A., Meiera, J., Sercu, T., Goyala, S., Linb, Z., Liua, J., Guoc, D., Otta, M., Zitnicka, C. L., Mad, J., and Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, **118**(15), e2016239118.

Rodriguez, A., Wright, G., Emrich, S., and Clark, P. L. (2018). %MinMax: A versatile tool for calculating and comparing synonymous codon usage and its impact on protein folding. *Protein Science*, **27**(1), 356–362.

Sample, P. J., Wang, B., Reid, D. W., Presnyak, V., McFadyen, I. J., Morris, D. R., and Seelig, G. (2019). Human 5' UTR design and variant effect prediction from a massively parallel translation assay. *Nature Biotechnology*, **37**(7), 803–809.

Sharp, P. M., Tuohy, T. M., and Mosurski, K. R. (1986). Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research*, **14**(13), 5125–43.

Sidi, T., Bahiri-Elitzur, S., Tuller, T., and Kolodny, R. (2024). Predicting gene sequences with AI to study codon usage patterns. *PNAS*, **122**(1), e2410003121.

Simon, E., Swanson, K., and Zou, J. (2024). Language models for biological research: a primer. *Nature Methods*, **21**, 1422–1429.

Tunney, R., McGlincy, N. J., Graham, M. E., Naddaf, N., Pachter, L., and Lareau, L. F. (2018). Accurate design of translational output by a neural network model of ribosome distribution. *Nat Struct Mol Biol*, **25**(7), 577–582.

Wu, Q. and Bazzini, A. A. (2023). Translation and mRNA Stability Control. *Annual Review of Biochemistry*, **92**, 227–45.

Yu, C.-H., Dang, Y., Zhou, Z., Wu, C., Zhao, F., Sachs, M. S., and Liu, Y. (2015). Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. *Molecular Cell*, **59**(5), 744–754.

Zhang, H., Zhang, L., Lin, A., Xu, C., Li, Z., Liu, K., Liu, B., Ma, X., Zhao, F., Jiang, H., Chen, C., Shen, H., Li, H., Mathews, D. H., Zhang, Y., and Huang, L. (2023). Algorithm for optimized mRNA design improves stability and immunogenicity. *Nature*, **621**(7978), 396–403.

## SUPPLEMENT

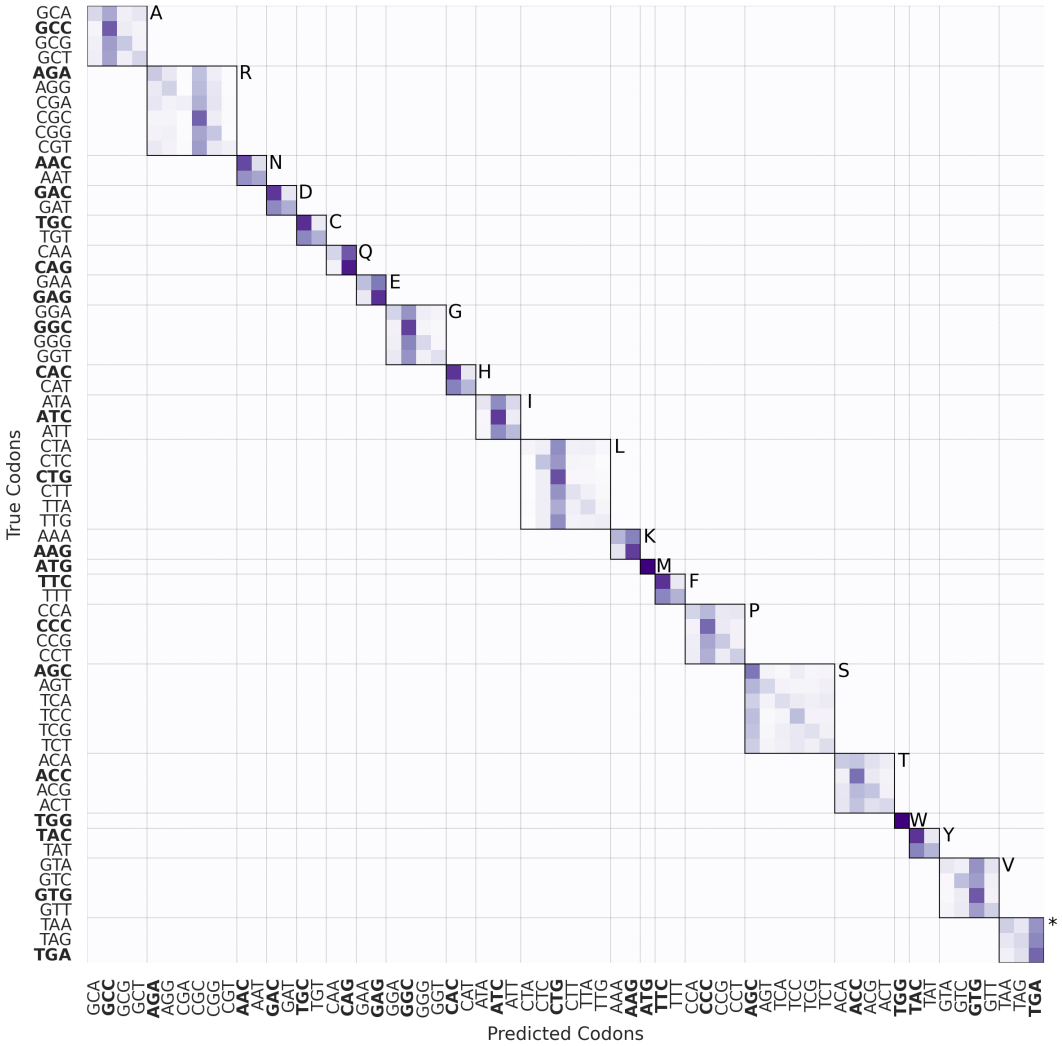


**Figure S1: Percentage of Trias-predicted codons matching wild-type.** Results are based on 4,351,017 codons from ∼6,000 human test set sequences. Bold codons indicate the most frequent codon per amino acid, with synonymous codons grouped in boxes. The one-letter amino acid code is displayed.

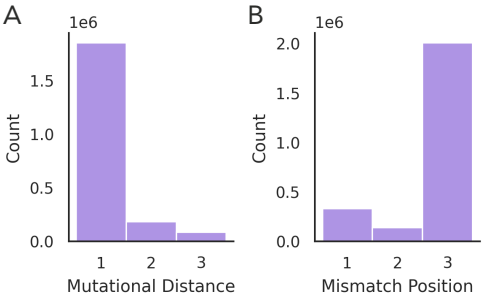


**Figure S2: Codon mismatches and their positions.** Among 4,351,017 codons in the human test set, approximately half differ from the wild-type codon, while only 0.002% result in an incorrect amino acid. (A) The majority of mismatches involve a single nucleotide substitution. (B) These substitutions predominantly occur at the third codon position.

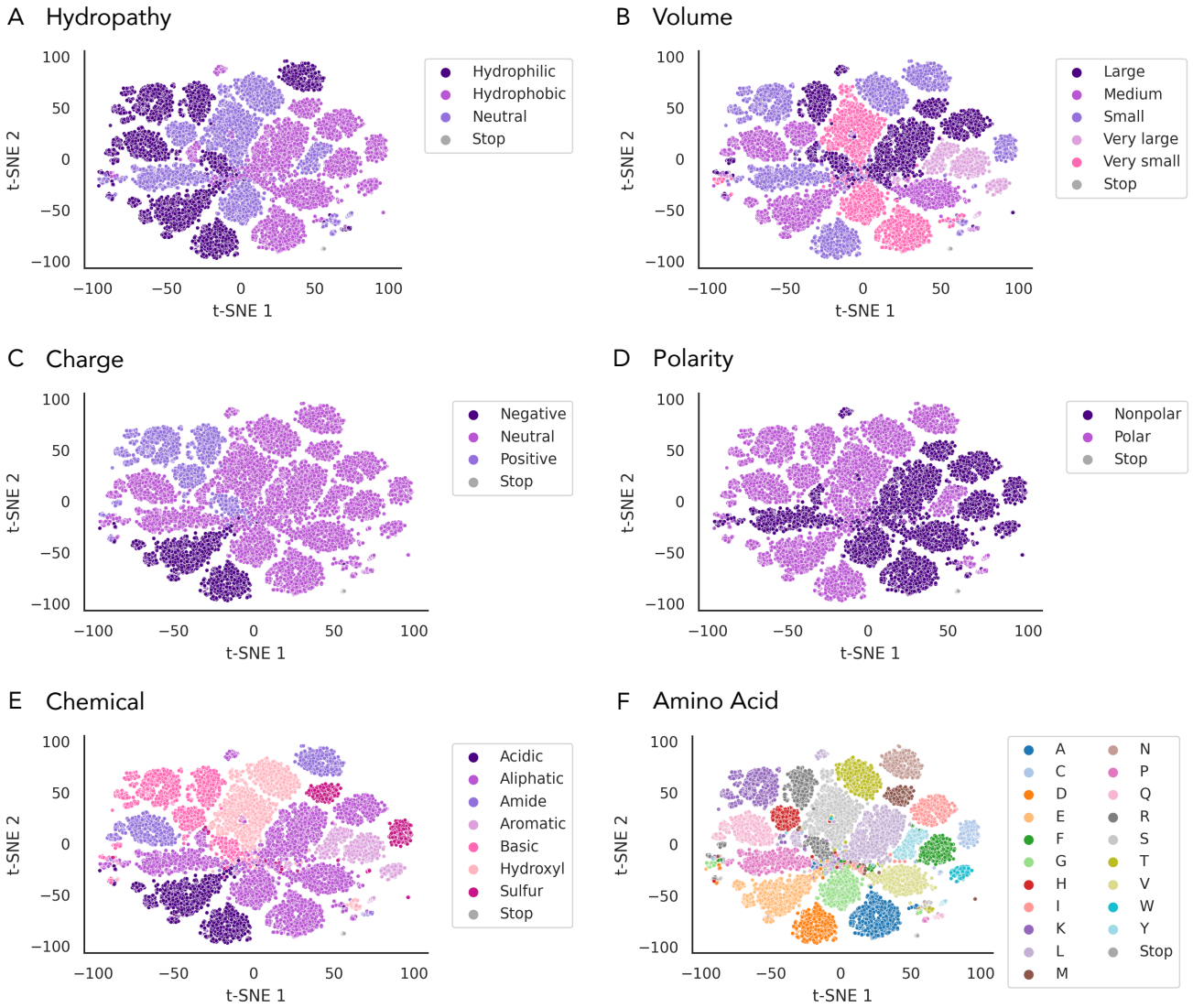


**Figure S3: Projection of final encoder embeddings with different labels.** Each dot represents an amino acid in 20 randomly selected protein sequences from the human test set. The same projection is shown with different labels based on classifications from this reference table[6].

[6]https://www.imgt.org/IMGTeducation/Aide-memoire/_UK/aminoacids/IMGTclasses.html

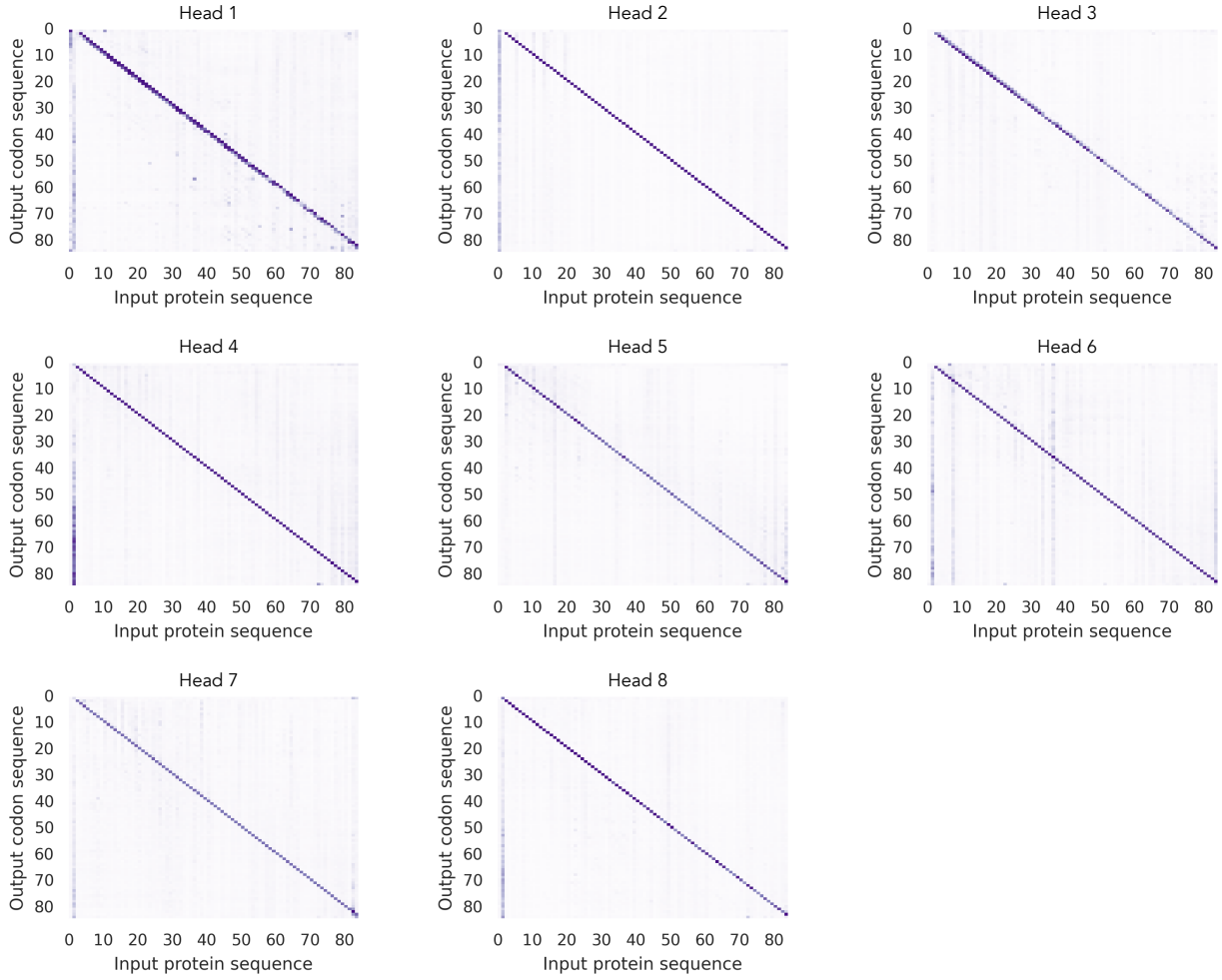


**Figure S4: Cross-attention maps for all decoder heads averaged for COX7B2.** Each panel shows attention from output codons (y-axis) to input amino acids (x-axis) for one decoder head. Attention weights are averaged across all decoder layers.

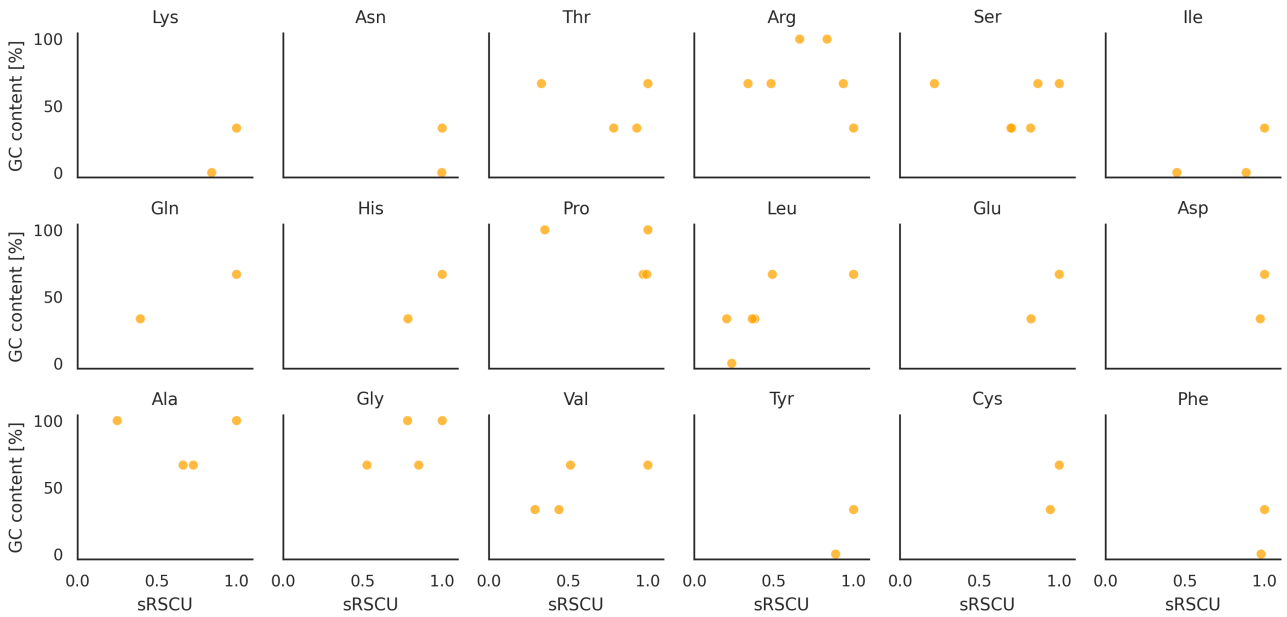


**Figure S5: Relationship between codon GC content and sRSCU.** For each amino acid, codons were plotted according to their GC content and sRSCU value in humans.

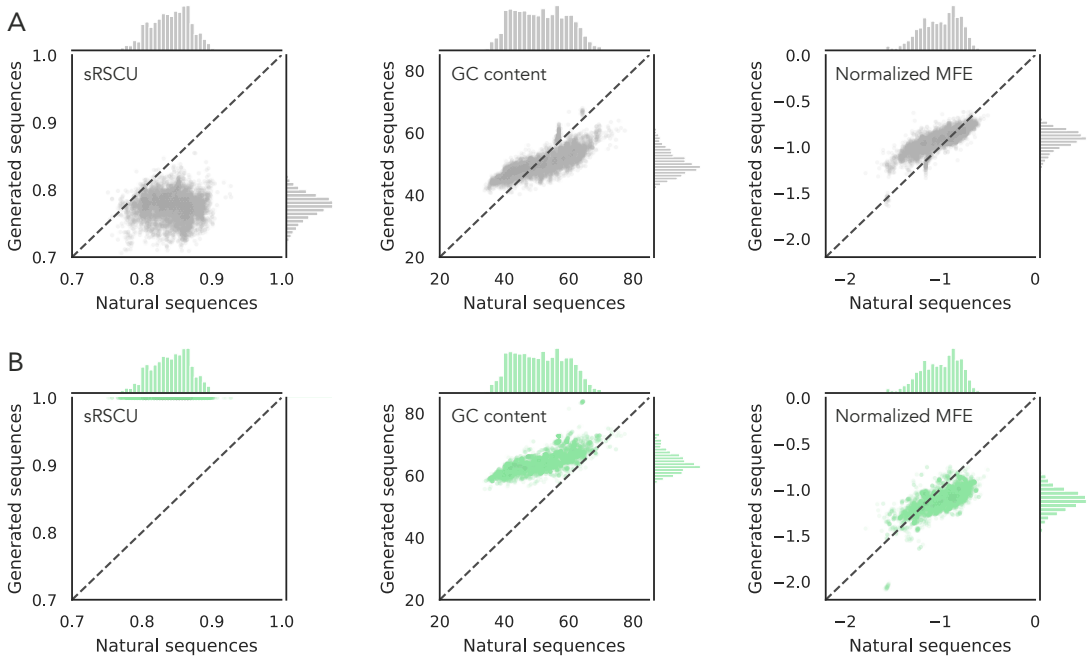


**Figure S6:** Comparison of generated sequences from a random model (A) and a frequency-based model (B) to their wild-type counterparts across scaled relative synonymous codon usage (sRSCU), GC content, and minimum free energy (MFE, normalized by sequence length) in the human test set (∼6,000 sequences, including isoforms). The random model selects a random synonymous codon for every amino acid site and the frequency-based model selects for every site the most frequent codon.

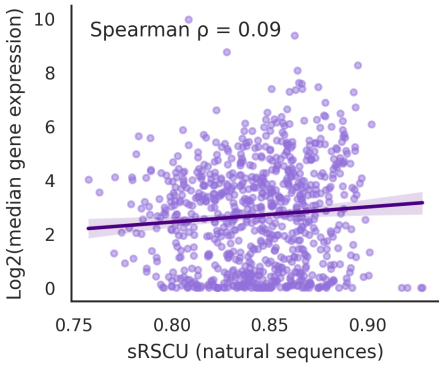


**Figure S7:** Log2-transformed median gene expression across all tissues in the GTEx dataset plotted against wild-type sRSCU values for 810 human genes in our test set that matched GTEx entries.

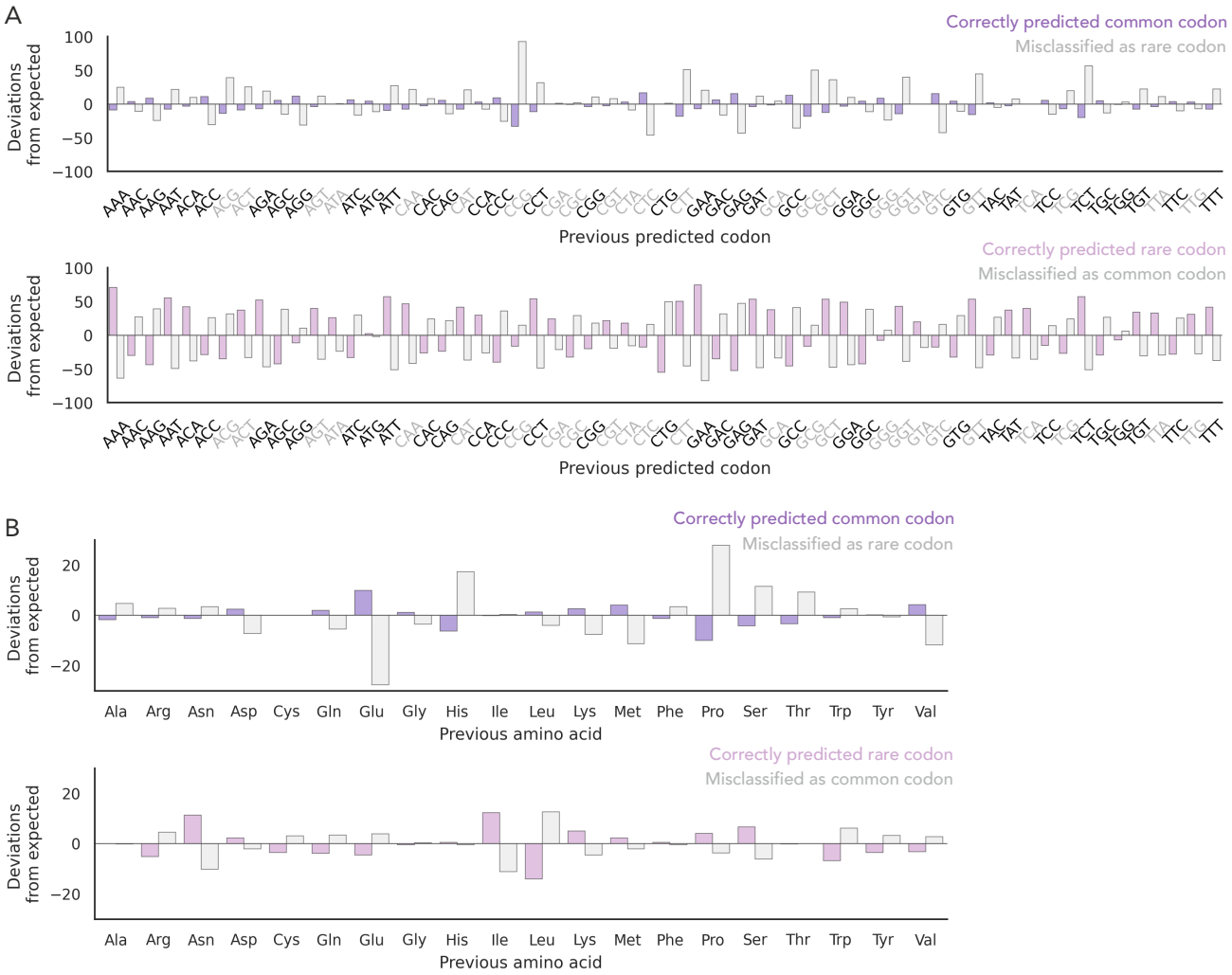


**Figure S8: Influence of previous codon or amino acid on codon prediction.** (A) Standardized residuals are plotted to show deviations from expected frequencies of next codon prediction classes based on the identity of the previously predicted codon. Rare codons are shown in gray, and common codons in black. (B) Same analysis as in (A), but grouped by the identity of the preceding amino acid.

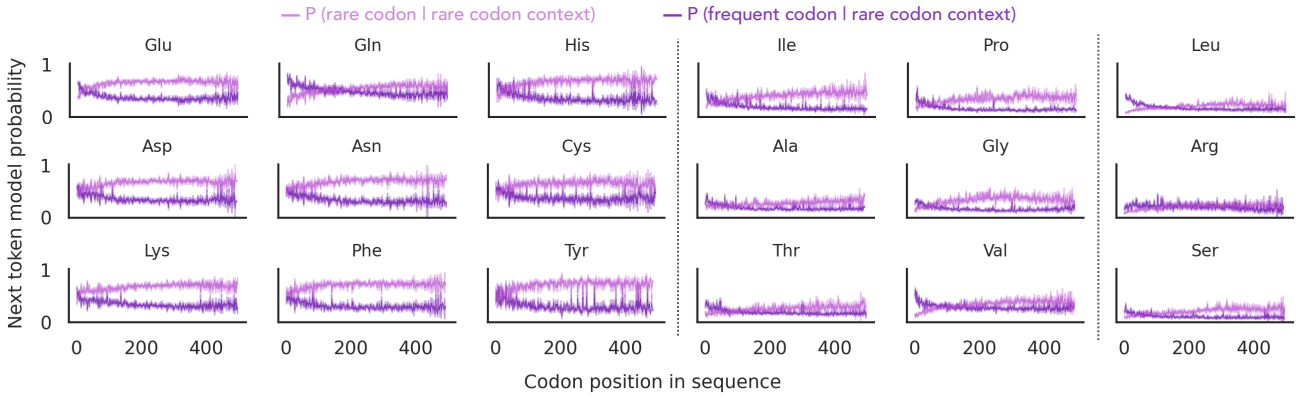


**Figure S9: Rare codon context biases next token probabilities across amino acids.** Model probability of selecting the next codon as frequent (highest sRSCU) or rare (lowest sRSCU), given that all preceding codons are rare. Amino acids are grouped by the number of synonymous codons (2, 3–4, and 6), with vertical dashed lines marking group boundaries. Mean and confidence interval are shown for ∼2,500 human test set sequences (sequence length ≤500 codons).

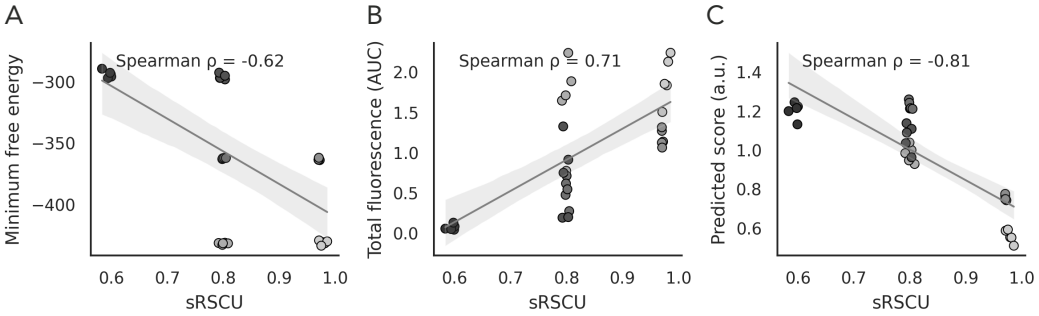


**Figure S10: GFP codon variants from Bicknell *et al.* (2024).** (A) Minimum free energy (MFE) and codon optimality (sRSCU) values for 30 GFP variants. Variants were selected based on a grid of three distinct sRSCU and three different MFE values, with five variants chosen per region. Different shades of gray indicate the selection regions based on MFE and sRSCU. (B) The same 30 variants, with sRSCU plotted against total GFP fluorescence, quantified as the area under the curve (AUC) from fluorescence measurements in HEK293 cells. (C) sRSCU values plotted against scores predicted by Trias for each of the 30 GFP variants, based on the negative sequence likelihood.

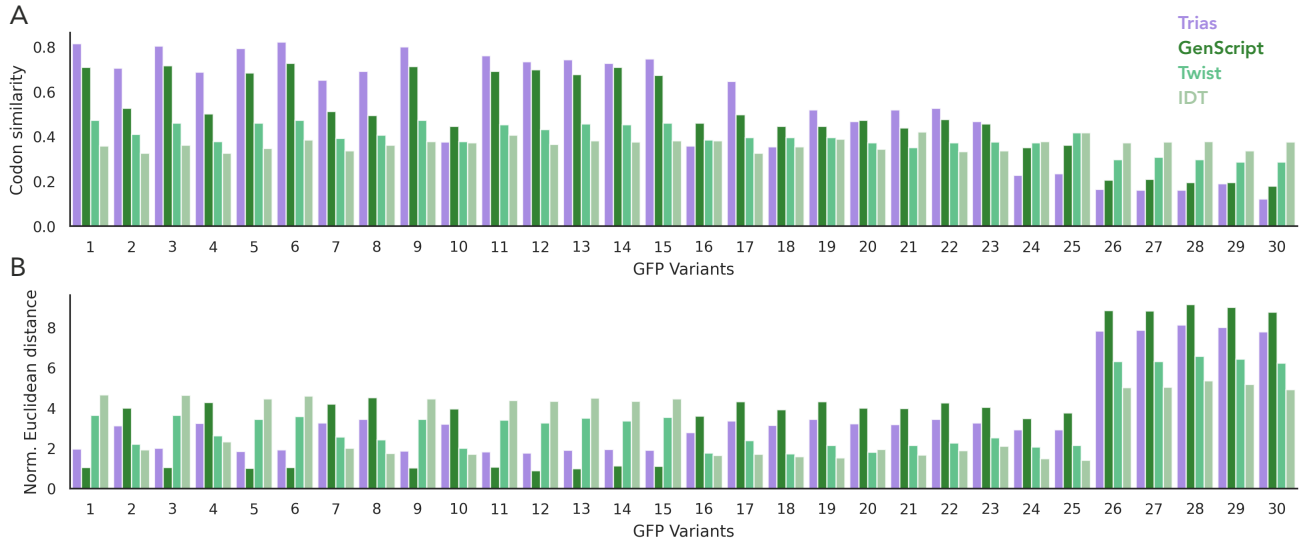


**Figure S11: Comparison of Trias and commercial tool sequences to experimentally validated GFP Variants.** Codon similarity and normalized Euclidean distances between sequences generated by Trias or commercial tools (GenScript, Twist, IDT) and all 30 GFP variants, sorted by experimentally measured protein expression (variant 1 = highest expression, variant 30 = lowest).

| Prev. codon category | Correctly predicted common codon | Misclassified as rare codon | Correctly predicted rare codon | Misclassified as common codon |
|---|---|---|---|---|
| Common | 1,061,767 | 260,437 | 393,287 | 541,552 |
| Rare | 260,443 | 87,260 | 143,166 | 121,303 |

**Table S1: Observed prediction counts by codon pair context.** Counts of prediction categories for the second codon in each codon pair, grouped by whether the preceding predicted codon was rare or common.

| Amino acid | Correctly predicted common codon | Misclassified as rare codon | Correctly predicted rare codon | Misclassified as common codon |
|---|---|---|---|---|
| Gln | 145,099 | 13,760 | 19,167 | 50,557 |
| His | 48,976 | 9,996 | 19,905 | 29,988 |
| Ile | 131,475 | 7,205 | 7,136 | 32,271 |
| Ala | 74,091 | 29,103 | 97,426 | 85,737 |
| Gly | 135,556 | 24,055 | 35,274 | 78,994 |
| Pro | 219,287 | 38,118 | 8,734 | 18,139 |
| Val | 76,757 | 28,831 | 69,577 | 69,098 |
| Thr | 102,188 | 42,316 | 40,215 | 46,668 |
| Arg | 96,149 | 69,136 | 50,250 | 27,534 |
| Ser | 184,841 | 50,159 | 62,205 | 86,616 |
| Leu | 107,792 | 35,009 | 126,564 | 137,253 |

**Table S2: Observed prediction counts by amino acid.** Counts of prediction categories grouped by the identity of the encoded amino acid. Only amino acids with both rare and common codons are included.

| Bin | Correctly predicted common codon | Misclassified as rare codon | Correctly predicted rare codon | Misclassified as common codon |
|---|---|---|---|---|
| 1 | 269,652 | 31,902 | 57,671 | 65,883 |
| 2 | 273,384 | 30,882 | 55,089 | 68,889 |
| 3 | 273,852 | 32,251 | 54,818 | 66,539 |
| 4 | 276,504 | 32,329 | 52,767 | 66,672 |
| 5 | 274,536 | 33,939 | 53,790 | 66,453 |
| 6 | 274,820 | 35,057 | 52,564 | 64,523 |
| 7 | 273,940 | 34,997 | 52,998 | 65,705 |
| 8 | 270,057 | 38,732 | 53,507 | 65,796 |
| 9 | 270,812 | 38,227 | 52,281 | 66,292 |
| 10 | 270,197 | 40,345 | 52,372 | 67,767 |

**Table S3: Observed prediction counts by relative sequence position.** Counts of prediction categories across ten bins representing relative coding sequence positions (each bin corresponds to 10% of the total sequence length).

| | Correctly predicted common codon | Misclassified as rare codon | Correctly predicted rare codon | Misclassified as common codon |
|---|---|---|---|---|
| Pfam domain | 1,863,628 | 243,789 | 371,947 | 459,143 |
| Other | 864,126 | 104,872 | 165,910 | 205,376 |

**Table S4: Observed prediction counts by domain localization.** Counts of prediction categories grouped by whether the codon is located within or outside an annotated Pfam domain.