Deep Generative Optimization of mRNA Codon Sequences for Enhanced Protein Production and Therapeutic Efficacy

Yupeng Li[1,#], Fan Wang[1,#], Jiaqi Yang[1], Zirong Han[2,3], Linfeng Chen[1], Wenbing Jiang[1], Hao Zhou[1], Tong Li[1], Zehua Tang[1], Jianxiang Deng[1], Xin He[1], Gaofeng Zha[4], Jiekai Hu[5], Yong Hu[5], Linping Wu[6], Changyou Zhan[7], Caijun Sun[2,3,8,9], Yao He[1,*], Zhi Xie[1,*]

[1]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, 510060, China.

[2]School of Public Health (Shenzhen), Sun Yat-sen University, Shenzhen, 518107, China.

[3]Shenzhen Key Laboratory of Pathogenic Microbes and Biosafety, Shenzhen Campus of Sun Yat-sen University, Shenzhen, 518107, China.

[4]Scientific Research Center, The Seventh Affiliated Hospital. Sun Yat-sen University, Shenzhen, 518107, China.

[5]ENO Bio mRNA Innovation Institute, Rhegen Biotechnology Co., Ltd. Shenzhen, China

[6]Center for Chemical Biology and Drug Discovery, China-New Zealand Joint Laboratory of Biomedicine and Health, Guangdong Provincial Key Laboratory of Biocomputing, Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou, 510530, China.

[7]Department of Pharmacology, School of Basic Medical Sciences, Fudan University, Shanghai, 200032, China.

[8]Key Laboratory of Tropical Disease Control (Sun Yat-sen University), Ministry of Education; Guangzhou, 514400, China.

[9]State Key Laboratory of Anti-Infective Drug Discovery and Development, School of Pharmaceutical Sciences, Sun Yat-sen University, Guangzhou, 510006, China.

[#]Equally contributed.

[*]Corresponding author. Email: xiezhi@gmail.com (Z.X.), scheyao@hotmail.com (Y. He)

ABSTRACT

Messenger RNA (mRNA) therapeutics show immense promise, but their efficacy is limited by suboptimal protein expression. Here, we present RiboCode, a deep learning framework that generates mRNA codon sequences for enhanced protein production. RiboCode introduces several advances, including direct learning from large-scale ribosome profiling data, context-aware mRNA optimization and generative exploration of a large sequence space. *In silico* analysis demonstrate RiboCode's robust predictive accuracy for unseen genes and cellular environments. *In vitro* experiments show substantial improvements in protein expression, with up to a 72-fold increase, significantly outperforming past methods. In addition, RiboCode achieves cell-type specific expression and demonstrates robust performance across different mRNA formats, including $m^1\Psi$-modified and circular mRNAs, an important feature for mRNA therapeutics. *In vivo* mouse studies show that optimized influenza hemagglutinin mRNAs induce ten times stronger neutralizing antibody responses against influenza virus compared to the unoptimized sequence. In an optic nerve crush model, optimized nerve growth factor mRNAs achieve equivalent neuroprotection of retinal ganglion cells at one-fifth the dose of the unoptimized sequence. Collectively, RiboCode represents a paradigm shift from rule-based to data-driven, context-sensitive approach for mRNA therapeutic applications, enabling the development of more potent and dose-efficient treatments.

INTRODUCTION

Messenger RNA (mRNA) therapy has emerged as a promising approach for treating diseases. This innovative therapeutic strategy harnesses the cell's protein synthesis machinery to produce desired proteins encoded by the delivered mRNA[1–3], leading to the application of mRNA therapies in various fields, such as vaccine development and protein replacement therapy[4]. The successful development and deployment of mRNA vaccines during the COVID-19 pandemic have further highlighted the transformative potential of this technology[5].

Despite the remarkable progress in mRNA vaccines, achieving efficient and consistent protein translation from delivered mRNA molecules remains a key challenge, particularly critical for protein replacement therapy where sustained, precise, and often higher levels of protein expression are required in specific cellular contexts. However, the biological instability of mRNA and the complex regulatory mechanisms governing mRNA translation in cells can lead to suboptimal protein expression[6–8]. Therefore, improving the expression of mRNA is a key challenge for enhancing the therapeutic efficacy and reducing the required dose of mRNA-based treatments.

An amino acid can be encoded by multiple synonymous codons, ranging from one to six codons per amino acid. Codon optimization is a strategy to improve protein expression by changing the synonymous codon of an mRNA molecule while maintaining the encoded amino acid sequence. The choice of synonymous codons can largely impact the efficiency of mRNA translation and the stability of the mRNA molecule[6,7]. For example, it has been shown that optimal codon usage can enhance ribosome engagement and increase translation elongation rates, ultimately leading to higher protein production[8]. Additionally, codon choice can influence mRNA stability, which is important as mRNAs are prone to degradation. The minimum free energy (MFE) of an mRNA, a computational indicator of its secondary structure stability, is often used to assess this aspect. A lower (more negative) MFE indicates a more stable secondary structure, as more energy would be required to disrupt the base pairing and

unfold the RNA[6]. Therefore, codon optimization is a critical step in the design of mRNA-based therapies to achieve maximal protein production, leading to better therapeutic efficacy.

Computational tools have been developed for codon optimization, most for DNAs, employing various strategies to select optimal codons. Past methods rely on codon usage bias derived from highly expressed genes in a given species, such as codon adaptation index (CAI)[9]. These methods aim to mimic the codon usage patterns of efficiently translated endogenous mRNAs. More recently, LinearDesign has been developed for mRNA optimization, by increasing CAI and reducing MFE, to jointly optimize translation and mRNA stability[6]. LinearDesign uses a linear programming approach to explore a wider space of sequence variants compared to previous methods and showed superior performance over the previous codon optimization methods.

Despite the development of the previous methods, several limitations hinder their effectiveness in consistently improving the protein expression of mRNA molecules. Firstly, the existing methods primarily rely on predefined sequence features, such as CAI, to guide codon selection. However, these metrics often fail to correlate with the experimentally measured protein expression levels[10,11], indicating that they do not accurately capture the complex factors governing mRNA translation. Secondly, the existing methods do not adequately account for the activity of translational regulators that influence mRNA translation[12]. This lack of context-aware optimization may reduce the effectiveness of the optimized mRNA sequences in specific cellular environments. Furthermore, the existing methods explore a limited space of codon sequences due to computational constraints and the reliance on predefined rules. This restricted search space may prevent the discovery of novel and highly optimized sequences that could potentially yield significant improvements in protein expression.

Deep learning has achieved remarkable success in tasks such as image recognition, natural language processing, and protein structure prediction, where it has outperformed conventional algorithms by learning complex patterns and relationships

104 from vast amounts of data[13,14]. In the context of mRNA codon optimization, a deep
105 learning approach may enable the model to capture the complex interplay between
106 codon usage and cellular context, without relying on predefined rules. Moreover, deep
107 learning models can explore a vast sequence space and discover novel patterns that may
108 not be apparent to human experts or accessible through traditional optimization
109 methods[15]. This ability has been exemplified in the field of protein engineering, where
110 deep learning has been used to design novel protein sequences with improved stability,
111 binding affinity, and catalytic activity[16–18]. For the codon optimization problem, if we
112 assume an average of three synonymous codons per amino acid, the number of possible
113 codon sequences for a 500 amino acid protein would 3^{500}, which is more than 10^{238}.
114 This number is much larger than the estimated number of atoms ($\sim 10^{80}$) in the
115 observable universe[19]. A deep learning-based codon optimization method could
116 potentially generate novel and highly optimized mRNA sequences by exploring the
117 immense sequence space.

118 Massive parallel reporter assays (MPRA) are commonly used to study the effects
119 of regulatory sequences on gene expression[20]. However, it is not suitable for optimizing
120 coding sequences due to the short sequence limitation, which is generally less than 300
121 base pairs, for high-throughput DNA synthesis. Additionally, MPRA experiments often
122 rely on artificial reporter constructs and may not fully recapitulate the complex
123 regulatory landscape of endogenous mRNA molecules. Ribosome profiling (Ribo-seq)
124 is a powerful experimental technique that provides a snapshot of actively translating
125 ribosomes on mRNA molecules, offering a direct and quantitative measurement of
126 translation at a transcriptome-wide scale[21,22]. By training a deep learning model on
127 Ribo-seq data from diverse codon sequences in diverse cell types and conditions, we
128 may capture the complex relationship between codon usage and translation which
129 occurs in the natural cellular context. However, such attempts have been lacking so far.

130 In this study, we present RiboCode, a deep learning model for mRNA codon
131 optimization that enhances protein expression by directly learning complex relationship

of mRNA codon sequences to their translation level from large-scale Ribo-seq data. Our prediction model demonstrated robust performance, while analysis of RiboCode's optimization strategies revealed a complex interplay between sequence characteristics and translation. *In vitro* experiments showed up to a 72-fold increase in protein expression, significantly outperforming past methods. RiboCode also achieved cell-type specific expression, and maintained robust performance across unmodified, $m^1\Psi$-modified, and circular mRNA formats. *In vivo*, optimized influenza virus hemagglutinin (HA) mRNA induced approximately ten times stronger neutralizing antibody responses in mice, while optimized nerve growth factor (NGF) mRNA achieved equivalent neuroprotection of retinal ganglion cells at one-fifth the dose in an optic nerve crush mouse model. This data-driven approach to codon optimization advances our understanding of mRNA translation and facilitates the development of more effective mRNA therapeutics.

RESULTS

RiboCode is a Deep Learning Framework for mRNA Codon Optimization

RiboCode is a deep learning-based framework for optimizing mRNA codon sequences to enhance protein production. It integrates three key components: a translation prediction model, an MFE prediction model, and a codon optimizer that explores and optimizes codon choices guided by the prediction models (Figure 1a).

The translation prediction model estimates the translation amount of a given codon sequence by learning the translational expression of diverse mRNA sequences from Ribo-seq experiments (Figures 1b and S1). In contrast to previous tools that rely on optimizing predefined features such as CAI, our deep learning model automatically extracts relevant features by training on 320 paired Ribo-seq and RNA sequencing (RNA-seq) datasets from 24 different human tissues and cell lines, encompassing translation measurements of over 10,000 mRNAs per dataset[23,24]. In addition, the model incorporates not only codon sequences but also mRNA abundances and cellular context

that is presented by gene expression profiles from RNA-seq, via a multi-head attention mechanism. This integrative approach enables the prediction of mRNA translation by jointly considering these important factors influencing translation.

To address mRNA stability, we developed an MFE prediction model. Current MFE prediction tools, such as RNAfold[25] and Linearfold[26], use dynamic programming, which is non-differentiable and incompatible with our codon optimizer described below. Our MFE model employs a deep neural network architecture and undergoes an iterative optimization process, to simultaneously improve its predictive capability and optimize sequences for lower MFE values (Figure S2).

The codon optimizer of RiboCode begins with the original codon sequence of a given protein (Figure 1c). The prediction models then predict a fitness score for this sequence. Using a gradient ascent optimization approach based on activation maximization (AM)[27], the optimizer adjusts the codon distribution to maximize the fitness score. A synonymous codon regularizer ensures that only synonymous codons encoding the same amino acids as the original sequence are considered, preserving the protein's amino acid sequence. Through iterative cycles of sequence generation, prediction, and optimization, the system produces codon sequences with improved properties. RiboCode can optimizes mRNA translation, stability or both, by interfacing with both the translation and MFE models, which uses a parameter w of 0 to optimize translation only, w of 1 to optimize MFE only and, a value between 0 to 1 to jointly optimize both.

By combining data-driven predictions with high-throughput sequence generation, RiboCode overcomes limitations of conventional heuristic approaches. It enables the exploration of a vast mRNA codon space, potentially uncovering optimized sequences.

Evaluation of Translation Prediction Model

We first evaluated the RiboCode's performance and generalizability using three cross-validation datasets: "new gene", "new environment", and "new gene in new

environment", which represented unseen genes, unseen cell types and unseen genes in unseen cell types during training (Figure S3). The model achieved a coefficient of determination (R^2) of 0.81, 0.89, and 0.81 for the three datasets, respectively (Figures 2a), indicating its robustness and ability to generalize.

To understand the relative importance of the three model inputs, we performed ablation analysis, revealing that mRNA abundances were the most important contributor to the prediction of translation (Figure 2b, Table S2), in agreement with an early study of yeast translation found that the most predictive variable for translation was the mRNA expression of the gene[22]. The incorporation of codon sequences lifted the R^2 by 0.15, and further inclusion of cellular environment improved the R^2 by 0.06. The ablation analysis demonstrated that all the inputs contributed to predicting mRNA translation.

We next investigated whether our model captured complex sequence features beyond common translation-related metrics. While our model learned relevant sequence features directly from the raw codon sequences, we tried to include common translation-related sequence metrics including CAI, MFE, and codon frequencies as additional model inputs and found these metrics did not improve prediction accuracy (Table S1). This suggested that the model could capture the sequence patterns that were predictive of translation, beyond these sequence metrics.

We explored alternative approaches to incorporating cellular context information. We directly incorporated the meta information of Ribo-seq datasets into the model, including cell types and experimental conditions and found it did not improve the performance (Table S1). This indicated that the gene expression profiles used in the model were an effective proxy to capture the relevant cellular environment influencing mRNA translation.

To validate the model's ability to capture cellular environment information, we leveraged a proteomics dataset from GTEx[28], which measured the relative protein expression of genes across 32 human tissues. We predicted the mRNA translation levels

of 9,582 genes in these tissues. Notably, the predicted translation levels of 6,805 out of 9,582 genes across these tissues (71%) were significantly correlated with their measured protein levels (p-value < 0.05 after False Discovery Rate (FDR) adjustment, Pearson's correlation), with a median correlation value of 0.56 (Figure 2c).

To exclude random correlations, we conducted a control experiment to calculate the correlations with shuffled cellular environment profiles for 1,000 times. With randomized data, only 0.02% of genes showed significant correlations, compared to 71% with the correct environment. The median correlation coefficients with shuffled environments were near zero and significantly lower than those with the correct environment ($p < 0.0005$, one-sample t-test, Figure 2d). These results demonstrated that our model effectively captured cellular context information.

Finally, we investigated the positional importance of coding sequences in translation prediction. We analyzed the importance of each nucleotide position for the model's prediction. The results showed that the coding sequences close to the translation start site (TSS) were more important (Figure 2e). This is consistent with a general knowledge that codons near TSS had a greater impact on protein synthesis, by influencing translation initiation[8].

Overall, the data-driven approach of RiboCode enabled robust predictive capabilities with biological relevance by learning important sequence patterns directly from the Ribo-seq data.

RiboCode's Optimization Strategies for Enhanced mRNA Translation

Having established the efficacy of our translation prediction model, we next explored how this model could be leveraged to generate sequences with enhanced translation potential. We first generated codon sequences of Gaussia luciferase (Gluc) (Figure S4). T-distributed stochastic neighbor embedding (t-SNE) indicated that the model established an association between the sequence space and translation levels. The red area in the upper right showed that a wide space of high translation sequences was explored (Figure 3a). We next explored how RiboCode optimized translation and

stability independently or jointly. A widely used Gluc sequence (MF882921.1) was used as a reference for comparison, which had a predicted translation level of 5.9 and an MFE value of -216 (Figure 3b). By optimizing the sequence for translation (w=0), the predicted translation level increased to around 25. On the other hand, codon sequences optimized for MFE (w=1) reduced the MFE from -150 kcal/mol to around -350 kcal/mol, with a similar translation level to the reference. With joint optimization ($0<w<1$), RiboCode explored a wider sequence space, achieving both enhanced translation and reduced MFE.

To understand RiboCode's optimization strategy, we analyzed codon usage patterns between generated sequences with enhanced and reduced translation, as well as between high- and low-translated endogenous sequences. We found that codons preferentially used in highly translated endogenous sequences were also favored in RiboCode-generated sequences with enhanced translation. Notably, the differences in codon usage between RiboCode's enhanced and reduced translation sequences were more pronounced than those observed between high- and low-translated endogenous sequences (Figures 3c). To assess the generalizability of these findings, we extended our analysis to multiple genes across various cell types. Consistently, we observed the same pattern of biased codon usage in all the cases (Figure S5). This suggests that RiboCode not only mimics but amplifies the codon usage patterns of efficiently translated endogenous mRNAs, potentially leading to even greater improvements in translation.

We next examined how RiboCode utilized sequence features during generation and optimization. Analysis of sequence features across different mRNAs revealed complex and variable relationships with translation (Figures 3d and S6). Notably, highly translated mRNAs generally showed an increase in uridine content (U%), which may reduce secondary structure formation and facilitate smoother ribosome movement during translation[7]. Additionally, these mRNAs mostly exhibited a decrease in Effective Number of Codons (ENC), suggesting a selection against rare or inefficient codon pairs

to enhance translation[29]. Variations in CAI, Codon Pair Bias (CPB), GC content (GC%) and MFE across different mRNAs suggested that while these features could influence translation, their impact might be more mRNA- or context-dependent.

In short, these findings highlight RiboCode's ability to capture complex sequence-translation relationships, offering a sophisticated approach to mRNA optimization that goes beyond traditional codon optimization metrics.

Experimental Validation Demonstrates RiboCode's Versatility and Efficacy

While our *in silico* analyses demonstrated the potential of RiboCode to optimize codon sequences, we next sought to validate these findings experimentally. We first validated RiboCode's ability to optimize codon sequences for enhanced protein expression. For Gluc, protein expression levels of all RiboCode-optimized sequences showed a dramatic enhancement in protein production compared to the reference, with up to a 72-fold increase (RD2) (Figure 4a, Table S3), which also significantly outperformed the LinearDesign-optimized sequences (p-value=0.019, one-sided Mann-Whitney U test). The predicted translational levels showed a strong positive correlation with the experimentally measured protein levels, although the p-value was slightly above 0.05 (Correlation coefficient=0.71, p-value=0.077, Pearson's correlation, Figures 4b and S7). In contrast, the CAI showed negative correlation with the experimental measurements (Correlation coefficient=-0.15), indicating that CAI is not a reliable predictor of protein expression levels in this context and may even lead to counterproductive optimization strategies. We additionally optimized another commonly used reporter gene, firefly luciferase (Fluc). The experimental validation showed a 17-fold increase in protein expression compared to the WT Fluc (Figure S8).

We next experimentally evaluated RiboCode's ability to design mRNAs with cell type specificity. We designed Gluc mRNA variants optimized for preferential expression in HEK293T cells and compared expression in HEK293T cells to both A549 and ARPE19 cells. For HEK293T vs A549, RiboCode predicted expression ratios ranging from 1.65 to 1.80, which closely matched experimental ratios of 1.35 to 1.73

(Figure 4c, Table S4). In the HEK293T vs ARPE19 comparison, while predicted ratios were around 1.41 for all variants, experimental results showed higher ratios between 2.57 and 3.13 (Figure 4d, Table S5). All designed mRNAs consistently demonstrated preferential expression in HEK293T cells across both comparisons. These results showed RiboCode's capacity to capture cellular environment and design mRNAs with enhanced expression in desired cell types.

Modified mRNAs, such as those with 1-methylpseudouridine ($m^1\Psi$) modifications, and circular RNAs are used in mRNA therapy instead of unmodified mRNAs due to their improved stability and reduced immunogenicity[3,7,30]. We therefore assessed the effectiveness of RiboCode in enhancing translation in these alternative mRNA forms. Among the four codon variants, $m^1\Psi$-modified RD2 and RD4 showed higher protein expression levels compared to the reference, with up to a 4.6-fold higher expression at 48 hours post-transfection (Figure 4e, Table S3). Moreover, all four RiboCode-generated codon variants in the circular form outperformed the reference (Figure 4f, Table S3). These results demonstrate that RiboCode optimization enhances protein production in both $m^1\Psi$-modified and circular mRNAs, illustrating its reliability and versatility.

These experimental validations demonstrate RiboCode's ability to significantly enhance protein expression, optimize for specific cell types, and improve translation across various mRNA forms, highlighting its potential as a powerful tool for mRNA therapeutic development.

RiboCode Enhances Immunogenicity of mRNA-based Influenza Vaccines

Having established the robustness of our optimization approach, we next aimed to demonstrate its practical application in the development of mRNA-based vaccines. Influenza A viruses are responsible for causing respiratory infections, leading to annual epidemics that result in millions of human infections worldwide[31]. HA, a glycoprotein found on the surface of influenza A viruses, plays a crucial role in the viral infection process and is the primary target for the development of influenza vaccines. Although

most of the vaccines were developed using inactivated influenza viruses, mRNA-based influenza vaccines are currently actively developed[32].

To enhance the expression of HA and potentially improve the efficacy of HA-based vaccines, we optimized the HA coding sequence. Two RiboCode-optimized HA sequences showed enhanced *in vitro* protein expression compared to the WT (Figure 5a). Particularly, RD1 showed substantial improvement compared to the WT and LinearDesign-optimized sequences. In addition, RD1 exhibited considerably higher expression levels compared to the WT sequence in both $m^1\Psi$-modified and circular mRNA forms (Figures 5b and 5c). These results again highlight the robustness and versatility of the RiboCode-optimized sequence.

We further assessed the *in vivo* immunogenicity induced by the optimized sequence, for both the prime and boost responses, where split virus influenza vaccine (SV) was served as the positive control (Figure 5d). The RD1 sequence induced significantly stronger neutralizing antibody responses, measured by the micro-neutralization (MN) titers, compared to the WT sequence and SV. For the prime response, RD1 elicited significantly higher MN titers compared to WT, with approximately 4.4-fold increase (Figure 5e, mean MN titers: RD1=2,560, WT=580; p-value=0.008, one-sided Wilcoxon test). The difference was more pronounced for the boost response, with RD1 inducing a 9.6-fold increase in MN titers compared to WT (Figure 5e, mean MN titers: RD1=83,200, WT=8,640; p-value=0.002, one-sided Wilcoxon test). These results demonstrated that the RiboCode-optimized sequence significantly enhanced both the initial and boosted immune responses. This dramatic improvement in immunogenicity underscores RiboCode's potential to enable more effective vaccines with lower doses.

Enhanced Protein Expression and Therapeutic Efficacy with optimized NGF mRNA

Having demonstrated the efficacy of RiboCode in optimizing mRNA for vaccine development, we next explored its potential in protein replacement therapy. We focused

on NGF as a promising candidate for treating glaucoma, which is a leading cause of irreversible blindness[33] and causes death of retinal ganglion cells (RGCs). Our recent study demonstrated that mRNA-based NGF therapy provided robust neuroprotection for RGCs in an optic nerve crush (ONC) mouse model[34].

To improve the neuroprotection efficacy, we optimized the codon sequences of human NGF mRNA. The protein expression levels of three RiboCode-designed sequences were more than 2-fold higher compared to that of the WT while the LinearDesign sequences did not show improvement (Figure 6a). We further assessed the best performed sequence (RD3) in both $m^1\Psi$-modified mRNA and circular mRNA forms. Notably, with $m^1\Psi$-modification, RD3 achieved 8.4- and 9.8-fold higher protein levels compared to the WT at 24h and 48h, respectively (Figure 6b). With mRNA circulation, RD3 also achieved a more than 2-fold higher expression than the WT at both 24h and 48h (Figure 6c). These results again demonstrated the robustness of RiboCode-optimized sequences across different mRNA forms.

Based on its superior performance in initial *in vitro* tests, we selected RD3 for further *in vivo* studies. To evaluate the *in vivo* expression of optimized NGF mRNA, we intravitreally administered both the RD3 and WT sequences. Each mRNA was m1Ψ-modified and encapsulated within LNP and administered at two doses: 100 ng/μl and 500 ng/μl. The RD3 sequence demonstrated significantly higher NGF protein expression compared to the WT sequence at both doses. Remarkably, RD3 at 100 ng/μl achieved even slightly higher expression than WT at 500 ng/μl (Figure 6d).

We then investigated the therapeutic potential of optimized NGF mRNA using an ONC mouse model, which mimics RGC injury and resulted in significant RGC loss (Figures 6e-g). Treatment with NGF mRNA showed clear neuroprotective effects, preserving more RGCs after injury. Notably, mice treated with 100 ng/μl RD3 showed significantly higher RGC counts than those treated with the same dose of WT mRNA (Figures 6h and 6i, p-value=0.0002, one-sided Wilcoxon test). Moreover, these counts were comparable to those in mice treated with 500 ng/μl WT mRNA.

To sum, the optimized sequence exhibited superior protein expression both *in vitro* and *in vivo*, while maintaining therapeutic efficacy at one-fifth the dose of the unoptimized sequence. These results demonstrated the effectiveness of RiboCode in optimizing NGF mRNA for the treatment of RGC injury.

DISCUSSION

In this study, we present RiboCode, a novel deep learning-based framework for mRNA codon optimization. The generative optimization framework, guided by the deep learning prediction model, enables the efficient exploration of the immense space of possible codon sequences. This allows RiboCode to discover novel, highly optimized sequences that may not be accessible to traditional optimization methods. RiboCode-optimized sequences demonstrate superior performance in various mRNA formats, including unmodified, $m^1\Psi$ modified, and circular mRNAs, highlighting its broad applicability in the rapidly evolving field of mRNA therapeutics. *In vitro* and *in vivo* experiments using the optimized sequences of therapeutically relevant proteins show substantial enhancements in protein expression compared to the unoptimized sequences. These improvements further translate into increased therapeutic efficacy, as demonstrated by significantly enhanced immune responses to an optimized influenza vaccine and markedly improved RGC protection in mice with optic nerve injury.

The superior performance of RiboCode can be attributed to several factor. Firstly, RiboCode's deep learning model learns directly from diverse nature sequences with translation measurements, enabling it to capture complex patterns of codon sequences for mRNAs with high translation level. In contrast, previous approaches, such as LinearDesign, relies on limited sequence features, which may not fully capture the intricacies of mRNA translation. Second, our model considered the cellular contexts of mRNA translation whereas the previously codon optimization tools such as CAI-based methods or LinearDesign did not. Third, RiboCode's generative optimization framework allows it to explore a large sequence space and to discover novel, highly optimized sequences that may not be accessible to the traditional approaches.

The findings of our study have important implications for the field of mRNA therapeutics. Firstly, RiboCode can generate and evaluate a vast number of novel codon combinations. This capability allows RiboCode to optimize mRNA sequences beyond the limitations of evolutionary constraints, potentially uncovering more efficient codon usage patterns not found in natural transcripts. By transcending natural sequence limitations, RiboCode represents a significant advancement in mRNA optimization, potentially leading to levels of protein expression and therapeutic efficacy that surpass what can be achieved with naturally evolved sequences. Second, by substantially increasing protein production, the optimized sequences can improve the potency and reduce the required dose of mRNA-based treatments, potentially mitigating side effects and enhancing patient outcomes. This is particularly relevant for applications such as protein replacement therapies, where achieving high levels of protein expression is crucial for therapeutic success. Third, RiboCode's robustness and versatility across different mRNA formats, including modified and circular mRNAs, expand the range of therapeutic applications for which it can be employed. As the field of mRNA therapeutics continues to evolve and new mRNA formats are developed to enhance stability, reduce immunogenicity, and improve delivery[30,35], RiboCode's ability to optimize sequences for these diverse formats will be invaluable.

While our study demonstrates the significant potential of RiboCode in optimizing mRNA codon sequences for enhanced protein expression and therapeutic efficacy, there are several limitations to address and future directions to explore. Firstly, we focused exclusively on optimizing the codon sequences of mRNA molecules to enhance protein expression in this study. While this approach yielded significant improvements in protein production and therapeutic efficacy, our future work will expand upon this foundation by jointly optimizing both the untranslated regions and codon sequences. By addressing both coding and non-coding regions, we will develop a more comprehensive optimization strategy that maximizes the potential of mRNA-based therapeutics across a wider range of applications and cellular contexts. Second,

although we attempted to incorporate mRNA structure-related features such as MFE into our model, we found no significant improvement in prediction accuracy. However, this does not conclusively rule out the importance of mRNA structure in translation. Future directions of RiboCode could benefit from more sophisticated integration of structural information, such as local secondary structures or ribosome pause sites. Finally, future studies could expand the validation to more advanced preclinical models, optimizing a broader range of genes and proteins, incorporating additional optimization objectives, and elucidating the underlying mechanisms.

In conclusion, RiboCode represents a paradigm shift from rule-based to data-driven mRNA optimization, potentially uncovering entirely new principles of efficient translation that were previously inaccessible. RiboCode will provide a versatile tool for researchers to maximize the potential of mRNA-based therapeutics, paving the way for more effective treatments in various medical applications.

DATA AVAILABILITY

The translation model and optimization framework of RiboCode are available on GitHub (https://github.com/wangfanfff/RiboCode).

ACKNOWLEDGMENTS

We are grateful to the researchers who shared Ribo-seq & RNA-seq data and to the authors of RPFdb, whose contributions made this research possible. We are also immensely grateful to the Precision Medicine Center of Sun Yat-sen University for its long-term support.

AUTHORS' CONTRIBUTIONS

Z.X. conceived, designed and supervised the project. YP.L. collected and preprocessed the datasets, evaluated the model performance and analyzed the data. Y. He, F.W. and ZH.T. developed the translation deep model. H.Z. developed the MFE deep model. F.W. and H.Z. developed the AM framework. JX.D. tested the model. JQ.Y. and LF.C. prepared mRNAs and conducted *in vitro* experiments. WB.J. conducted *in vivo* NGF

experiments. ZR.H. and CJ.S. conducted the *in vivo* HA experiments. GF.Z encapsulated mRNA with LNP. X.H., H.J., Y. Hu, LP.W., and CY.Z. contributed to project discussion. YP.L., Z.X., Y. He and F.W. wrote the manuscript. All authors read and approved the manuscript.

REFERENCES

1. Baden, L. R. *et al.* Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *N Engl J Med* **384**, 403–416 (2021).

2. Gebre, M. S. *et al.* Optimization of non-coding regions for a non-modified mRNA COVID-19 vaccine. *Nature* **601**, 410–414 (2022).

3. Pardi, N., Hogan, M. J., Porter, F. W. & Weissman, D. mRNA vaccines — a new era in vaccinology. *Nat Rev Drug Discov* **17**, 261–279 (2018).

4. Qin, S. *et al.* mRNA-based therapeutics: powerful and versatile tools to combat diseases. *Signal Transduct Target Ther* **7**, 166 (2022).

5. Fang, E. *et al.* Advances in COVID-19 mRNA vaccine development. *Sig Transduct Target Ther* **7**, 94 (2022).

6. Zhang, H. *et al.* Algorithm for optimized mRNA design improves stability and immunogenicity. *Nature* **621**, 396–403 (2023).

7. Leppek, K. *et al.* Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics. *Nat Commun* **13**, 1536 (2022).

8. Hanson, G. & Coller, J. Codon optimality, bias and usage in translation and mRNA decay. *Nat Rev Mol Cell Biol* **19**, 20–30 (2018).

9. Sharp, P. M. & Li, W. H. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**, 1281–1295 (1987).

10. Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E. M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* **25**, 117–124 (2007).

11. Li, G.-W., Burkhardt, D., Gross, C. & Weissman, J. S. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* **157**, 624–635 (2014).

12. Waldman, Y. Y., Tuller, T., Shlomi, T., Sharan, R. & Ruppin, E. Translation efficiency in humans: tissue specificity, global optimization and differences between developmental stages. *Nucleic Acids Research* **38**, 2964–2974 (2010).

13. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. **596**, 583–589 (2021).

14. Cao, C. *et al.* Deep Learning and Its Applications in Biomedicine. *Genomics, Proteomics & Bioinformatics* **16**, 17–32 (2018).

15. Zrimec, J. *et al.* Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nat Commun* **11**, 6141 (2020).

16. Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).

17. Sumida, K. H. *et al.* Improving Protein Expression, Stability, and Function with ProteinMPNN. *J. Am. Chem. Soc.* **146**, 2054–2061 (2024).

18. Bennett, N. R. *et al.* Improving de novo protein binder design with deep learning. *Nat Commun* **14**, 2625 (2023).

19. observable universe. https://en.wikipedia.org/wiki/Observable_universe.

20. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**, 271–277 (2012).

21. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).

22. Weinberg, D. E. *et al.* Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. *Cell Rep* **14**, 1787–1799 (2016).

23. Xie, S.-Q. *et al.* RPFdb: a database for genome wide information of translated mRNA generated from ribosome profiling. *Nucleic Acids Res* **44**, D254-258 (2016).

24. Wang, H. *et al.* RPFdb v2.0: an updated database for genome-wide information of translated mRNA generated from ribosome profiling. *Nucleic Acids Research* **47**, D230–D234 (2019).

25. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**, 26 (2011).

26. Huang, L. *et al.* LinearFold: linear-time approximate RNA folding by 5'-to-3' dynamic programming and beam search. *Bioinformatics* **35**, i295–i304 (2019).

27. Linder, J. & Seelig, G. Fast activation maximization for molecular sequence design. *BMC Bioinformatics* **22**, 510 (2021).

28. Jiang, L. *et al.* A Quantitative Proteome Map of the Human Body. *Cell* **183**, 269-283.e19 (2020).

29. Wright, F. The 'effective number of codons' used in a gene. *Gene* **87**, 23–29 (1990).

30. Chen, R. *et al.* Engineering circular RNA for enhanced protein production. *Nat Biotechnol* **41**, 262–272 (2023).

31. Molinari, N.-A. M. *et al.* The annual impact of seasonal influenza in the US: measuring disease burden and costs. *Vaccine* **25**, 5086–5096 (2007).

32. Myers, M. L. *et al.* Commercial influenza vaccines vary in HA-complex structure and in induction of cross-reactive HA antibodies. *Nat Commun* **14**, 1763 (2023).

33. Lambiase, A. *et al.* Experimental and clinical evidence of neuroprotection by nerve growth factor eye drops: Implications for glaucoma. *Proc Natl Acad Sci U S A* **106**, 13469–13474 (2009).

34. Jiang, W. *et al.* Circular RNA-based therapy provides sustained and robust neuroprotection for retinal ganglion cells. *Molecular Therapy - Nucleic Acids* **35**, 102258 (2024).

35. Chaudhary, N., Weissman, D. & Whitehead, K. A. mRNA vaccines for infectious diseases: principles, delivery and clinical translation. *Nat Rev Drug Discov* **20**, 817–838 (2021).

36. Joshi NA & Fass JN. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files. (2011).

37. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).

38. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

39. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

40. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

41. Jia, L. & Qian, S.-B. Therapeutic mRNA Engineering from Head to Tail. *Acc Chem Res* **54**, 4272–4282 (2021).

42. Ho, J. J. D. *et al.* A network of RNA-binding proteins controls translation efficiency to activate anaerobic metabolism. *Nat Commun* **11**, 2677 (2020).

43. Luan, Y. *et al.* Deficiency of ribosomal proteins reshapes the transcriptional and translational landscape in human cells. *Nucleic Acids Research* **50**, 6601–6617 (2022).

44. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535–550 (2016).

45. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).

46. Zrimec, J., Buric, F., Kokina, M., Garcia, V. & Zelezniak, A. Learning the Regulatory Code of Gene Expression. *Front. Mol. Biosci.* **8**, 673363 (2021).

47. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–999 (2016).

48. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. in *International conference on machine learning* (2015).

49. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The journal of machine learning research* **15**, 1929–1958 (2014).

50. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).

51. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. Preprint at http://arxiv.org/abs/1711.05101 (2019).

52. Nair, V. & Hinton, G. E. Rectified Linear Units Improve Restricted Boltzmann Machines. in *Proceedings of the 27th international conference on machine learning (ICML-10)* (2010).

53. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (2016).

54. Miyato, T., Dai, A. M. & Goodfellow, I. Adversarial Training Methods for Semi-Supervised Text Classification. Preprint at http://arxiv.org/abs/1605.07725 (2021).

55. Maas, A. L., Hannun, A. Y. & Ng, A. Y. Rectifier nonlinearities improve neural network acoustic models. in *Proc. icml* (2013).

56. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks. in *International conference on machine learning* 3319–3328 (2017).

57. Bach, S. *et al.* On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE* **10**, e0130140 (2015).

58. Shrikumar, A., Greenside, P. & Kundaje, A. Learning Important Features Through Propagating Activation Differences. in *International conference on machine learning* 3145–3153 (2017).

59. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

60. Wang, Y. *et al.* L226Q Mutation on Influenza H7N9 Virus Hemagglutinin Increases Receptor-Binding Avidity and Leads to Biased Antigenicity Evaluation. *J Virol* **94**, e00667-20 (2020).

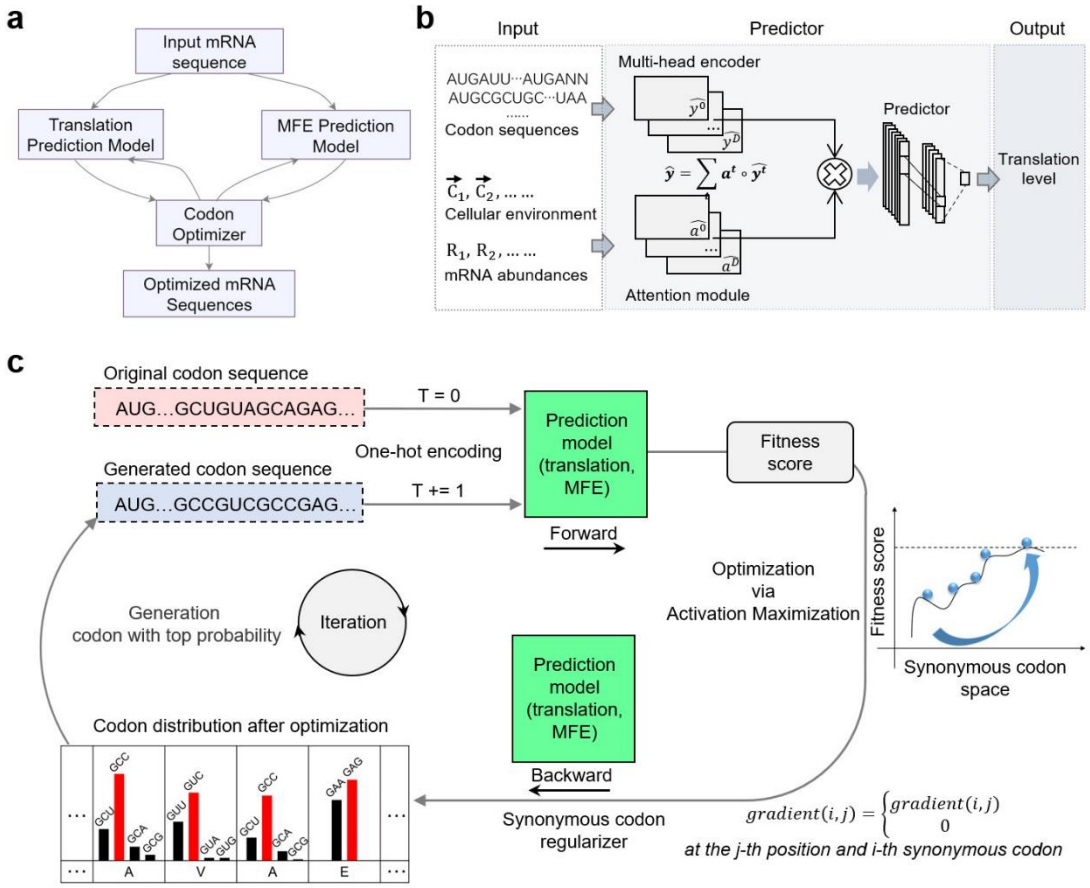


Figure 1. Predictive and Generative Optimization of RiboCode.

a. RiboCode contains three main components, a codon optimizer, a translation prediction model and an MFE prediction model.

b. The framework of the prediction model for translation. The input includes codon sequences in one-hot encoding, corresponding mRNA abundances and, the cellular environment which is presented by vectors of gene expression profiles from RNA-seq. The model integrates these inputs using a multi-head attention mechanism. From the fused representations, a convolutional neural network extracts features and outputs the predicted translation level of mRNA.

c. Iterative optimization of codon sequences. RiboCode predicts fitness of an original sequence (T=0), then uses activation maximization to generate optimized synonymous variants (T+=1). A synonymous regularizer maintains amino acid sequence. This process iterates until peak fitness is achieved.

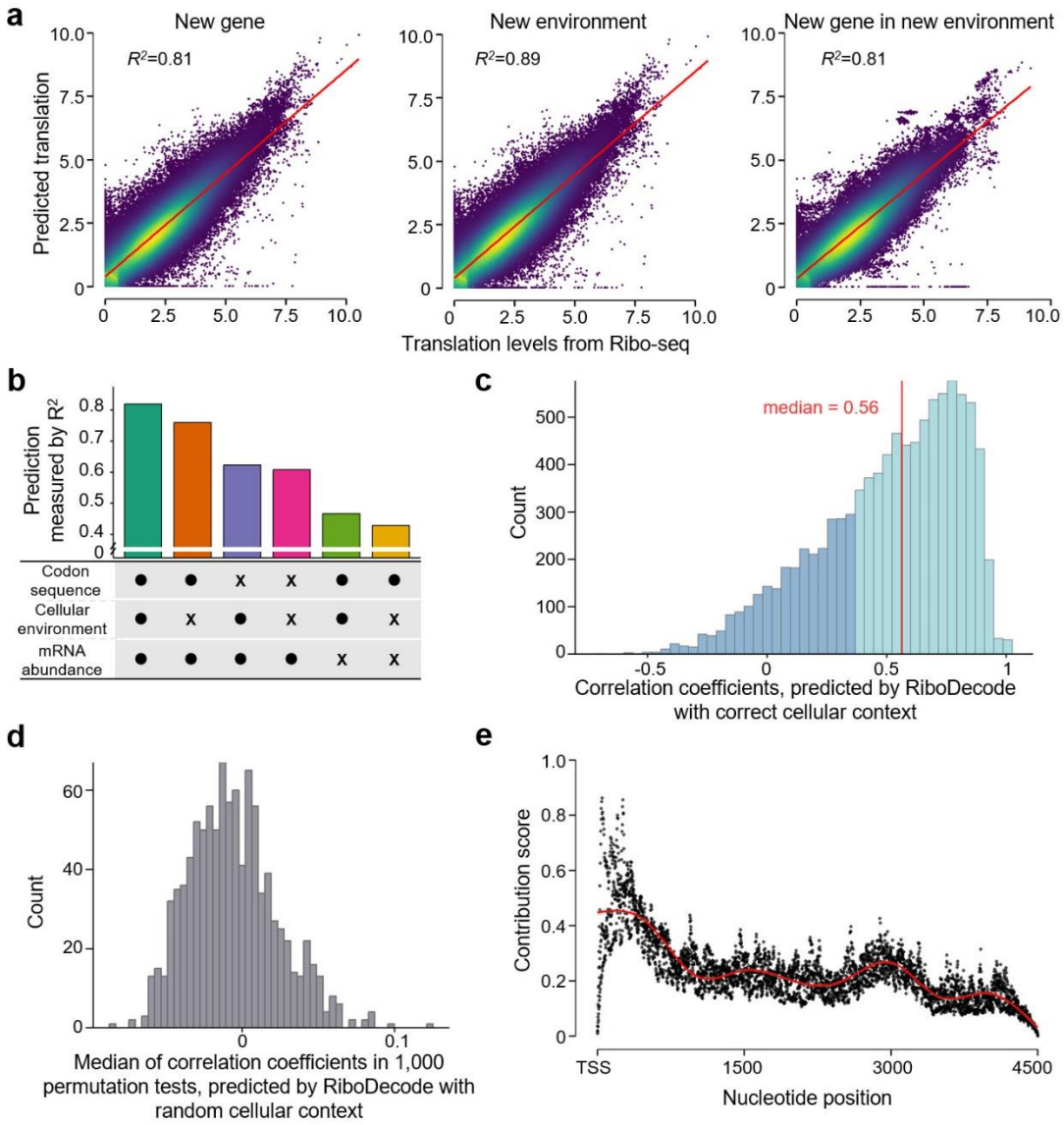


Figure 2. Evaluation of the Translation Prediction Model.

a. Experimentally measured translation levels by Ribo-seq versus predicted translation levels in the three validation datasets. Red lines denote the linear fit.

b. Ablation analysis shows the contributions of the three inputs to the prediction model. The table below shows the ablation status of the inputs, with dots and crosses representing the presence and absence of corresponding elements, respectively.

c. Comparison of protein measurements across 32 human tissues by mass spectrum and predicted translation levels from RiboCode. We calculated the Pearson's correlation coefficient for all 9,582 genes, of which 6,805 (71%) showed significant correlations (p-value < 0.05 after FDR adjustment for multiple testing). Red vertical line indicates the median of correlation coefficient. Light blue represents a significant correlation. Darker blue represents insignificance.

d. Median of Pearson's correlation coefficients between protein level and predicted translation level with randomized environment input for 1,000 simulations. Only 0.02% of results were significant (p-value < 0.05 after FDR adjustment for multiple-testing).

e. The importance of each nucleotide position for the translation prediction. The x-axis represents the nucleotide position from the TSS (translation starting site). Integrated Gradients attribution method was used to obtain the importance score for each nucleotide position (black dots). The red line denotes the local polynomial regression fit.

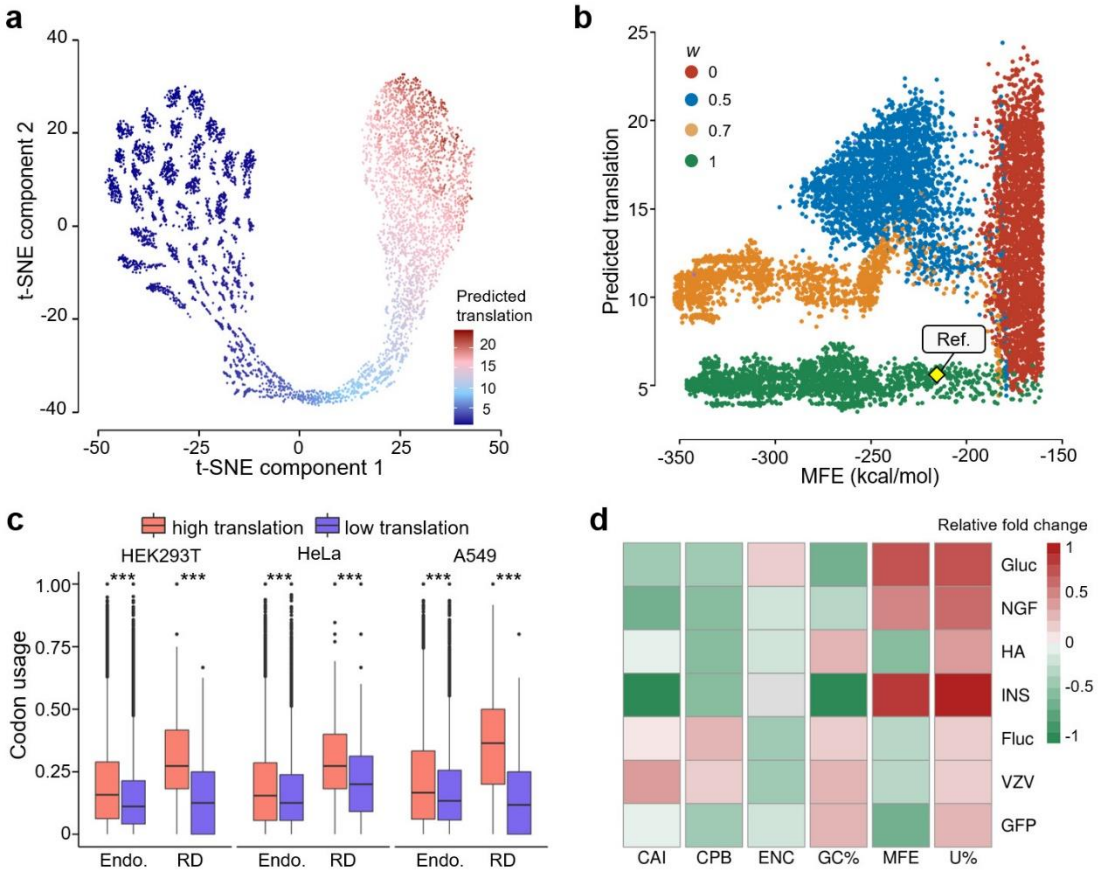


Figure 3. Strategies of Enhanced Translation in Generated Sequences.

a. Generation of Gluc codon sequences with low translation level (the upper left area) and high translation level (the upper right area) (w=0, Figure S4). T-SNE of codon sequences is shown. Each dot represents one sequence, and the color represents the predicted translation level.

b. Generation and optimization of Gluc codon sequences using different w of 0, 0.5, 0.7, and 1. Each dot represents one sequence, positioned in its predicted translation level (y-axis) and MFE (x-axis). The position of the reference sequence (MF882921.1) is shown.

c. Codons that appeared more frequently in highly translated endogenous sequences were also used more often in highly translated Gluc sequences generated by RiboCode. "RD": RiboCode-generated sequences, "Endo.": endogenous genes. (***: p<0.001, t-test).

d. Changes of sequence features of optimized sequences compared to the unoptimized, for different mRNAs. For each column (feature), cells in red represent that the values were higher in optimized sequences than unoptimized ones while cells in green are opposite. The difference of ENC for INS shows no significance (the cell in gray). Abbreviations: GFP: green fluorescent protein, Fluc: firefly luciferase, INS: insulin, VZV: varicella zoster virus glycoprotein E, and HA: influenza A hemagglutinin.

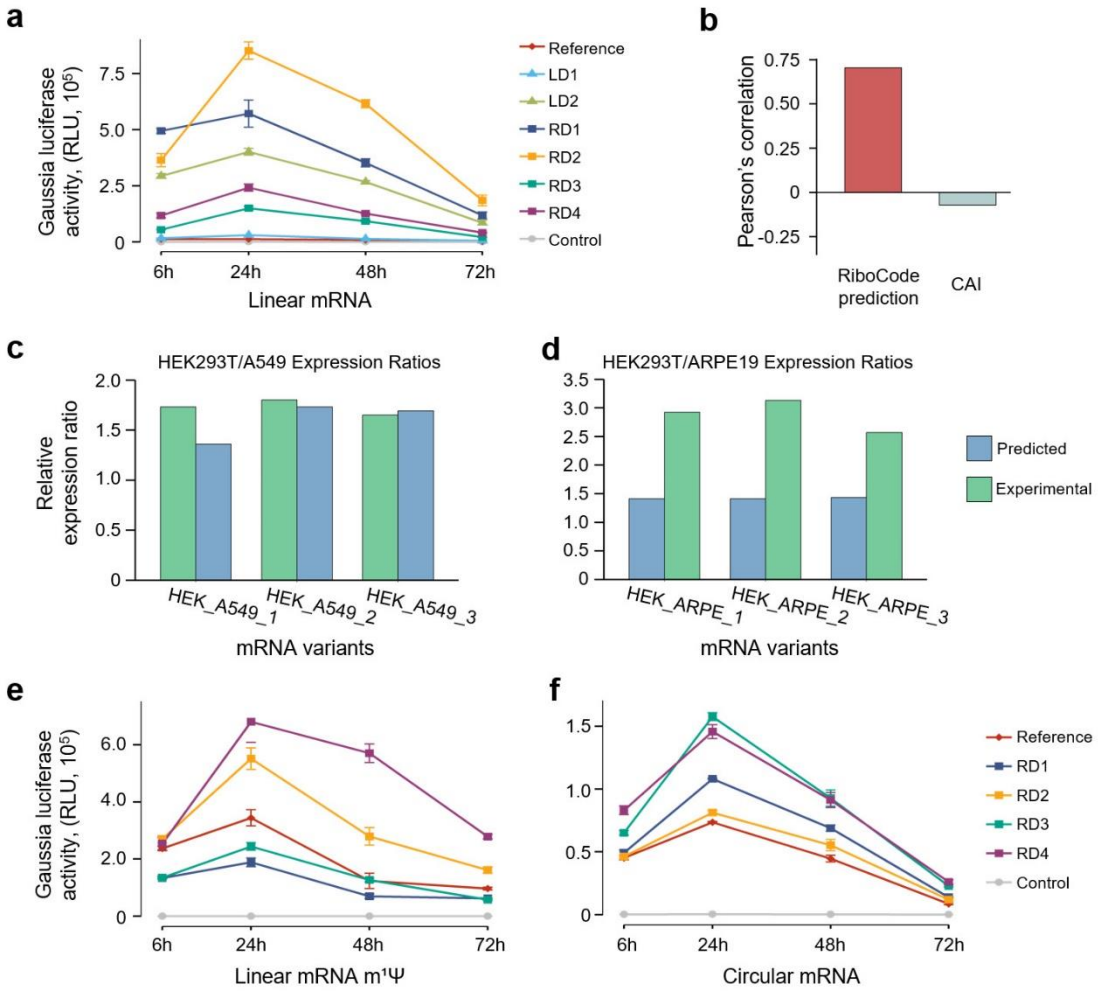


Figure 4. Robustness of Optimization across Unmodified, Modified, and Circular mRNAs.

a. Protein expression of Gluc was measured by fluorescence intensity. RD sequences were designed by RiboCode. LD sequences were designed by LinearDesign. "RLU": relative light units.

b. Correlation of experimentally measured protein expression values of Gluc at 24 hours versus predicted values of RiboCode and CAI (see Figure S7 for the details).

c. Relative protein expression levels of mRNA variants in HEK293T vs. ARPE19 cells. The ratio of larger than 1 indicates high protein expression level in HEK293T than that in ARPE19.

d. Relative protein expression levels of mRNA variants in HEK293T vs. A549 cells. The ratio of larger than 1 indicates high protein expression level in HEK293T than that in A549.

e, f. Protein expression of generated Gluc codon variants in (e) the linear mRNA form with $m^1\Psi$ modification and (f) the circular mRNA form.

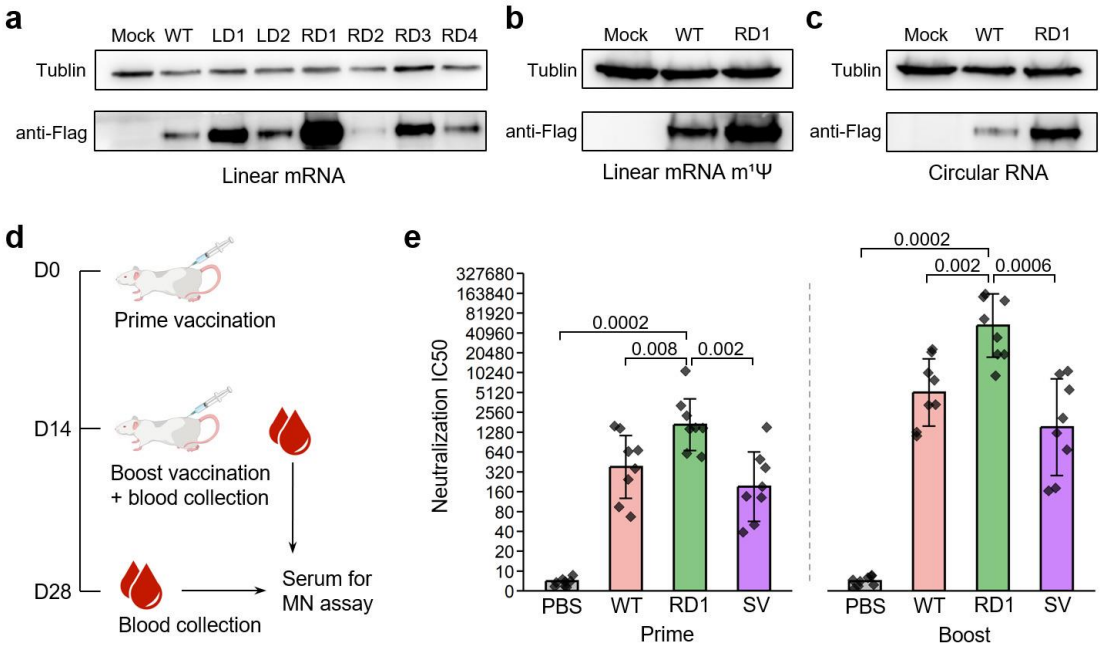


Figure 5. More Effective mRNA-based Influenza Vaccines Through Codon Optimization

a. Western blot analysis shows protein expression of the HA variants in HEK293T cells 24 hours after transfection. RD sequences were designed by RiboCode. LD sequences were designed by LinearDesign.

b, c. Protein expression RD1 in (b) the linear mRNA form with $m^1\Psi$ modification and (c) the circular mRNA form.

d. HA mRNA immunization and analysis: BALB/c mice were intramuscularly inoculated with two doses (10μg mRNA for each dose) with an interval of two weeks. The mouse serum was collected at 14 days and 28 days for MN assay.

e. Levels of neutralizing antibodies against influenza viruses after prime and boost vaccination. IC50, half-maximal inhibitory concentration. PBS and split virus influenza vaccine (SV) were used as the negative and positive controls, respectively. One-sided Wilcox test was used to calculate p values shown in the figure.

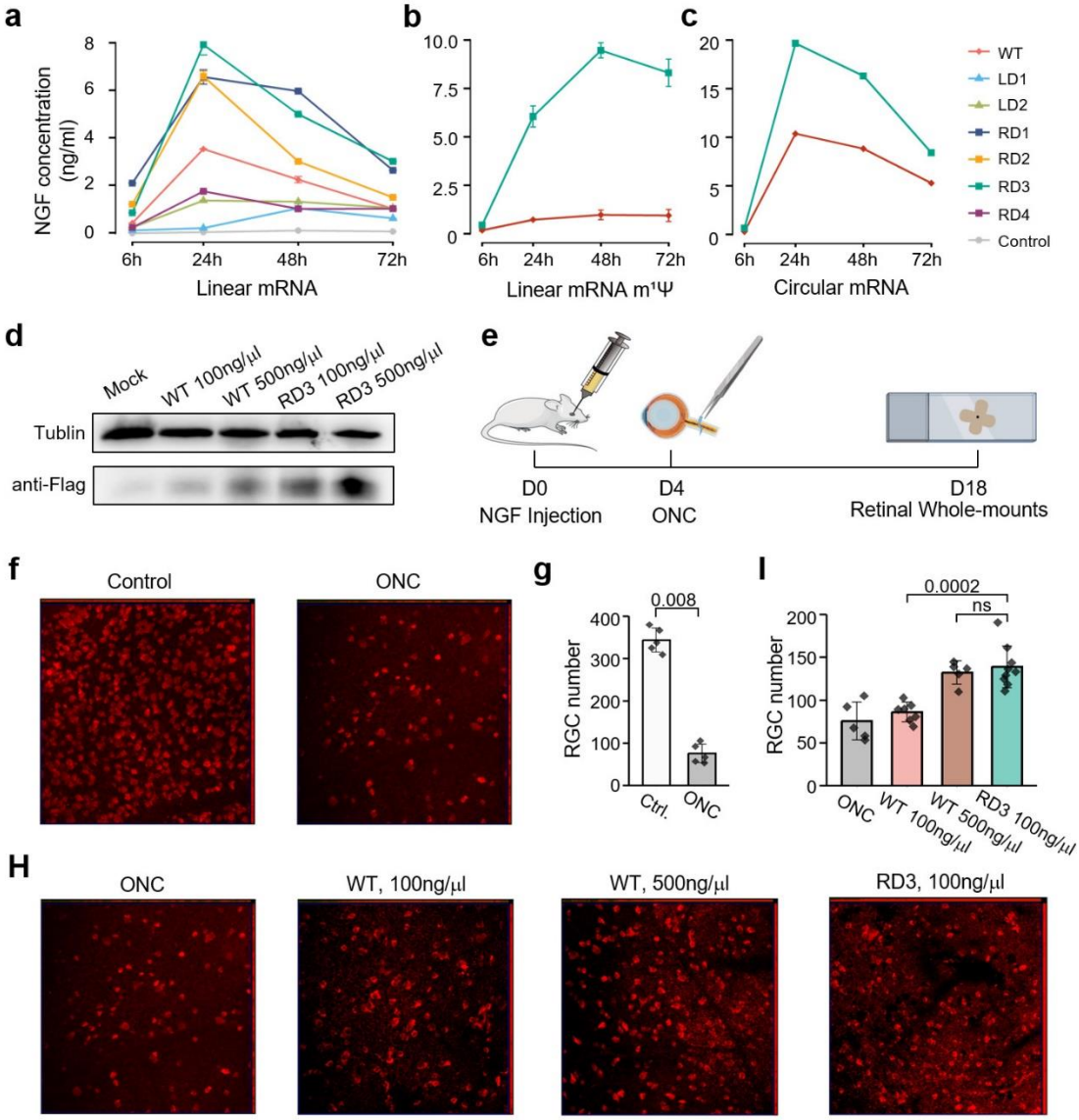


Figure 6. Enhanced Protein Expression and Therapeutic Efficacy with Optimized NGF mRNA.

a. NGF protein expression in HEK293T cells, measured by ELISA. RD sequences were designed by RiboCode. LD sequences were designed by LinearDesign.

b, c. Protein expression levels of RD3 in m$^1\Psi$-modified (b) and circular (c) mRNA formats.

d. *In vivo* protein expression of RD3. The m$^1\Psi$-modified mRNAs were injected into mouse retinas by intravitreal injection, and protein levels were measured by western blot 48 hours post-injection.

e. Timeline of NGF mRNA therapy in ONC model: At Day 0 (D0), the m$^1\Psi$-modified NGF mRNAs were injected into mouse retinas by intravitreal injection. At Day 4 (D4), the optic nerve was subjected to a physical crush injury. At Day 18 (D18), RGC numbers were quantified using immunofluorescence staining.

f, g. RGC numbers measured by immunofluorescence staining in the ONC mouse were significantly reduced compared to that of the control (one-sided Wilcox test).

h. RGC numbers in the ONC mouse retina after injection of NGF $m^1\Psi$-modified mRNA with 100 and 500 ng/μl dosages.

i. The RGC number in mice treated with 100 ng/μl RD3 was similar to those treated with 500 ng/μl WT and significantly higher than those treated with 100 ng/μl WT. One-sided Wilcoxon test was used to calculate p-values shown in the figure; ns: $p>0.05$.

Supplementary Materials

Deep Generative Optimization of mRNA Codon Sequences for Enhanced Protein Production and Therapeutic Efficacy

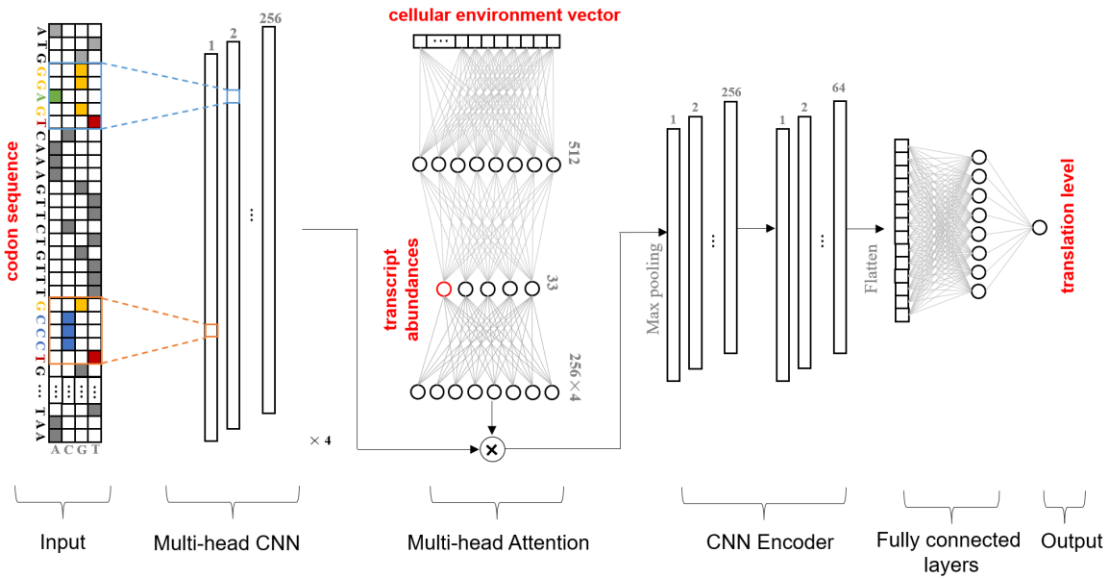
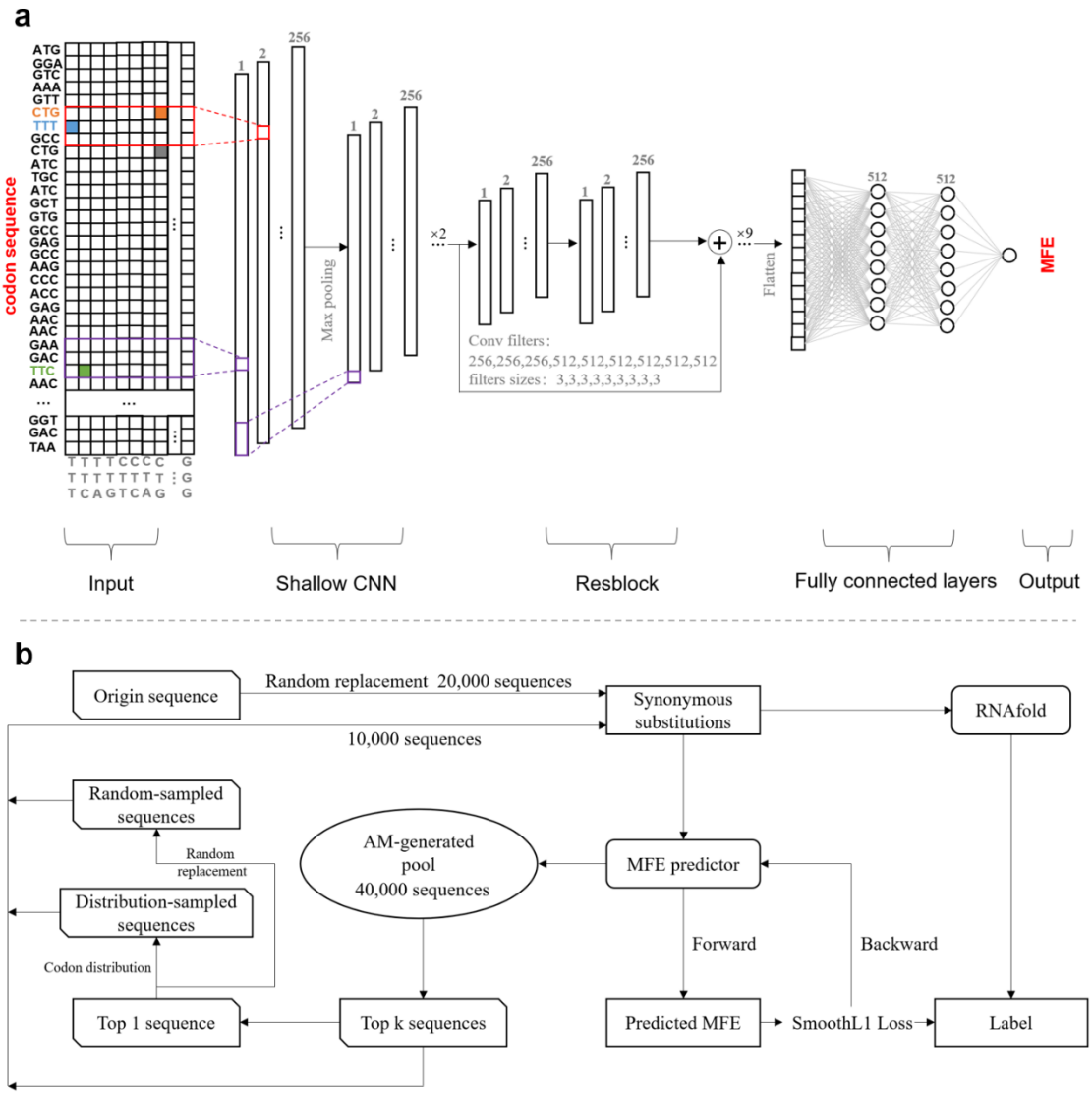


Figure S1. Schematic diagram of the translation model structure. Multi-head CNN contains four heads of CNN. Multi-head Attention is constructed by cell environment vector and transcript abundances through the fully connected layers. CNN Encoder consists of two max pooling layers and one convolutional layer.



729

730 **Figure S2. MFE Model Architecture and Optimization Process**

731 **a.** MFE Model Architecture: a shallow CNN with two convolutional layers, nine Resblocks, and
732 three fully connected layers.

733 **b.** The MFE optimization contains four steps, including initial sampling, initial training, sequence
734 generation and model retraining.

735

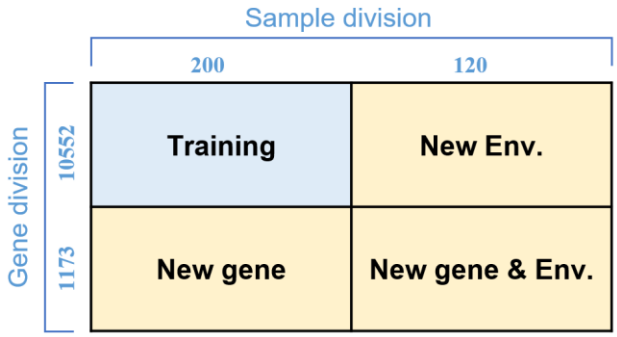


Figure S3. Training and internal validation sets of the translation model. Schematic representation of how the dataset was divided into training (90% of genes) and test (10% of genes) sets. The 320 total datasets were split into 200 for training and 120 for validation, creating three validation sets: "new gene", "new cellular environment", and "new gene in new cellular environment".

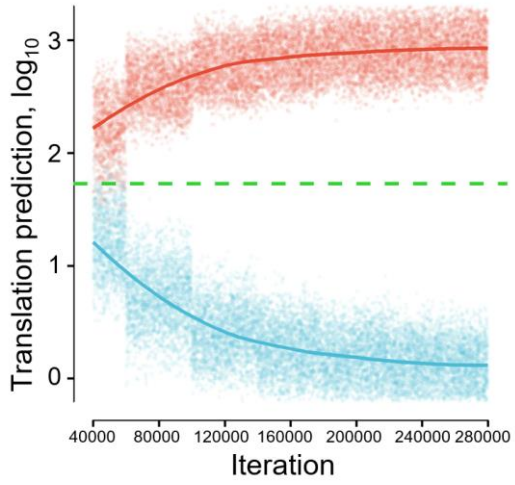


Figure S4. Generated Gluc codon variants. Plot showing the predicted translation levels of Gluc codon variants generated using the translation model (w=0). The x-axis represents generation iterations, with red and blue lines indicating enhanced and reduced predicted translation, respectively. The green dashed line shows the predicted translation level of the reference Gluc sequence.

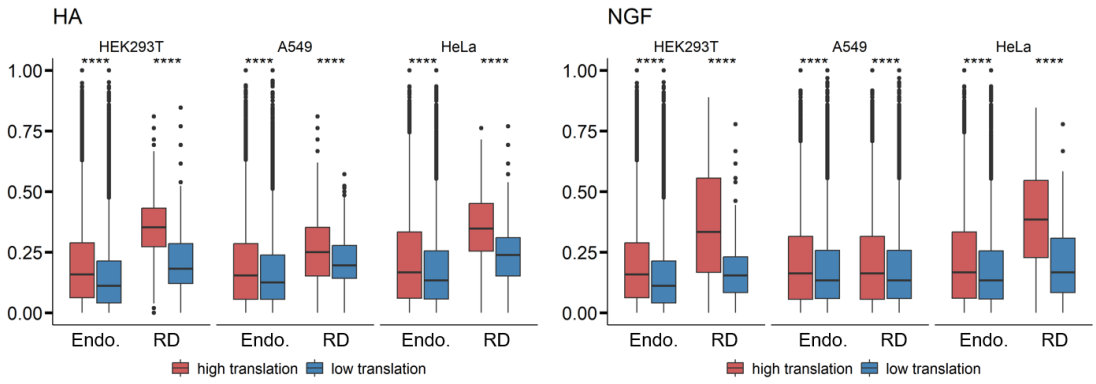
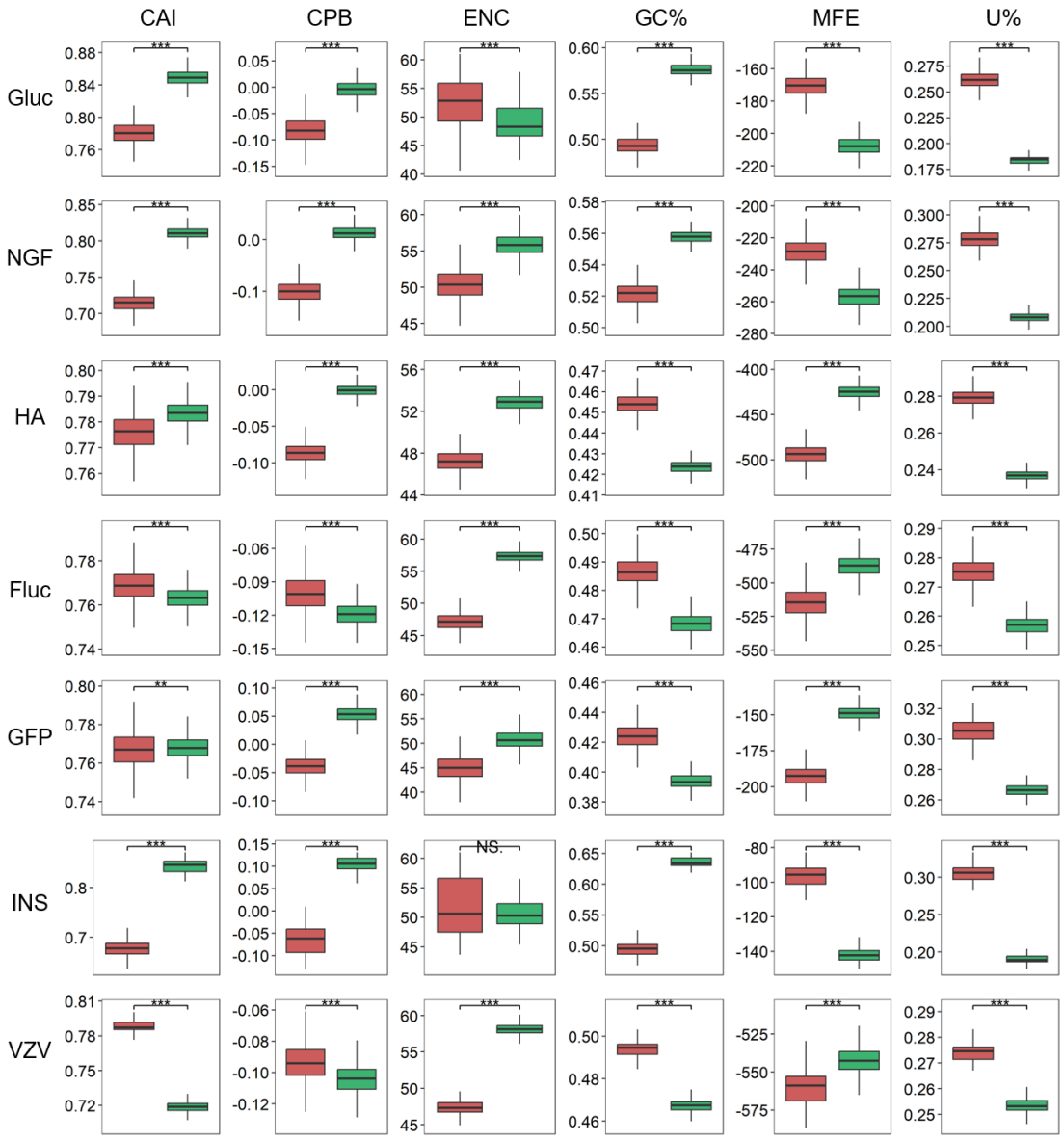


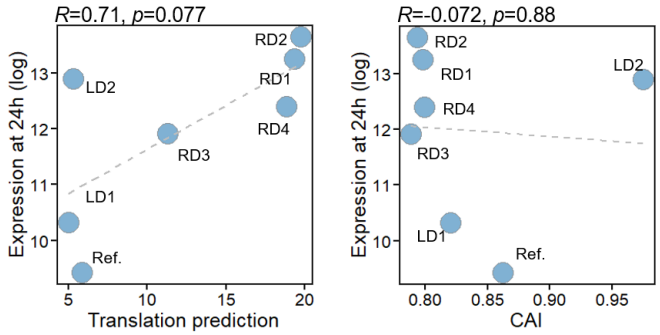
Figure S5. Codon Usage Similarity Between Endogenous and RiboCode-Generated Sequences. Comparison of codon usage patterns in endogenous high-translation sequences and low-translation sequences, as well as the RiboCode-designed sequences with enhanced and reduced translation for HA, and NGF in different cellular environments (HEK293T, A549, and HeLa). "Endo.": endogenous. "RD": RiboCode-designed. (t-test, ****: $p<0.0001$).



758

759 **Figure S6. Features of generated codon sequences for different genes.** The graphs show various
760 sequence features for optimized (red) and unoptimized (green) codon sequences for different genes
761 (w=0). (t-test, ***: p<0.001). The full names of mRNAs are noted in the main text.

762



763

764 **Figure S7. Correlation of experimental expression with predicted measures.** The graphs show
765 correlations between experimental protein expression (in log) and translation predicted by
766 RiboCode, and CAI. Pearson's correlation coefficients are provided.

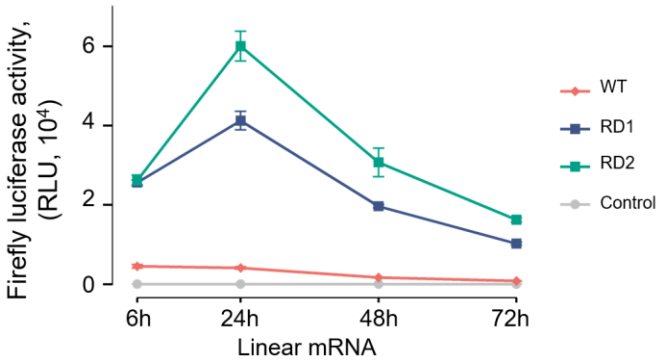


Figure S8. Fluc protein expression in transfected HEK293T cells. Protein expression was determined by fluorescence intensity. RD1 and RD2 were generated by RiboCode (w=0 and 0.5, respectively). Compared to the wild-type (WT), RD2 expression increased by 14.5-fold at 24h, and 17.9-fold at 48h. "RLU": relative light unit.

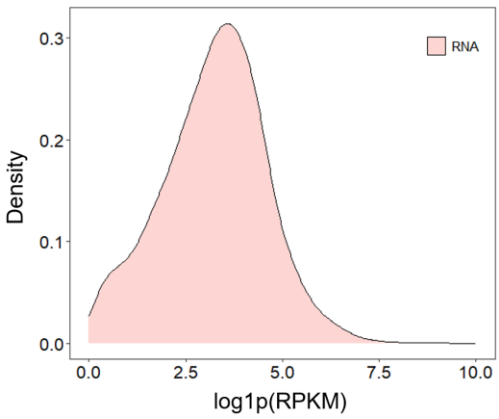


Figure S9. Distribution of mRNA expression across all samples. The graph shows the distribution of mRNA expression for 11,725 genes across all samples. Expression counts were transformed to log(y×RPKM+1), where y was set to 5 to maximize the correlation between mRNA abundance and translation level. RPKM stands for reads per kilobase of transcript per million reads mapped. The default mRNA count for codon optimization was set to 4.5 based on the median value of this distribution.

Input	Training set (R^2)	Test set (R^2)		
		New gene	New Env.	New gene & Env.
basic	**0.86021**	**0.81329**	**0.88719**	**0.80838**
+ codon frequency	0.86108	0.800528	0.883919	0.790282
+ MFE	0.86086	0.809827	0.889674	0.800357
+ CAI	0.86127	0.802417	0.880232	0.791394
+ cell type	0.8609	0.808396	0.890336	0.799348
+ treatment	0.86174	0.803956	0.885125	0.796533

Table S1. Evaluation of translation prediction model on internal test sets. This table compares the performance (R^2) of the basic translation model with models incorporating additional inputs on various test sets. The basic input includes codon sequence, mRNA abundance, and cellular environment represented by gene expression profiles from RNA-seq. "Env." stands for cellular environment.

Input	Training set (R^2)	Test set (R^2)
CDS + mRNA abundances + cellular environment	0.85924	0.819247
mRNA abundances + cellular environment	0.7331	0.62351
CDS + mRNA abundances	0.80825	0.760404
CDS + cellular environment	0.58303	0.467216
CDS	0.56135	0.429161
mRNA abundances	0.70318	0.608484

Table S2. Ablation analysis of translation model inputs. This table shows the results of an ablation analysis investigating the contribution of the three main inputs (codon sequences, mRNA abundances, cellular environment) to the translation model's performance. “CDS”: codon sequence.

Sequence ID	Predicted translation level	Linear mRNA		Linear mRNA ($m^1\Psi$)		Circular mRNA		MFE	CAI
		24h	48h	24h	48h	24h	48h		
Reference	5.88	12369.33	8546.333	344260.5	123505	73567.67	44620	-216.4	0.808
RD1	19.34	571477.7	352015.7	188385.8	69464.5	108143	68785.67	-195.7	0.713
RD2	19.75	851809	615058.3	551472.8	279046	81295	55305	-195.3	0.703
RD3	11.28	149905	92193.33	244131.3	126622.7	157691	92775.67	-342.4	0.748
RD4	18.85	241644	125983.3	680751.3	570459.5	145763.7	91233.33	-258.4	0.736
LD1	4.99	30501.67	13699.67					-346.2	0.766
LD2	5.29	400987.7	267094.3					-302.5	0.952

Table S3. Predicted translation levels and experimental protein expression of Gluc mRNA variants. This table presents data for various Gluc mRNA constructs, including the reference sequence, RiboCode-generated variants (RD1-RD4 with different *w* values), and LinearDesign-generated variants (LD1-LD2). For each construct, the table shows predicted translation levels, experimental protein expression in different mRNA formats (linear, $m^1\Psi$-modified, and circular) at 24h and 48h post-transfection, as well as MFE and CAI values. Protein expression was measured by fluorescence intensity in HEK293T cells.

Sequence ID	Protein expression level (24h)		Relative expression level		Expression ratio of HEK239T/A549
	In HEK293T	In A549	In HEK293T	In A549	
Reference	24527.5	28271.25	1	1	1
HEK_A549_1	38522.25	32671	1.57	1.16	1.35
HEK_A549_2	29584	19712.75	1.21	0.70	1.73
HEK_A549_3	23215.25	21544	1.29	0.76	1.70

Table S4. Experimental protein expression of Gluc mRNA variants with cellular differential expression. This table presents experimental protein expression for various Gluc mRNA construct, including the reference sequence and RiboCode-designed variants. The designed variants were predicted to express more in HEK293T than in A549 (HEK_A549_1 to 3, w=0). For each construct, the table shows original protein expression level, expression level relative to the reference, and expression ratio of HEK239T/A549. Protein expression was measured by fluorescence intensity.

Sequence ID	Protein expression level (24h)		Relative expression level		Expression ratio of HEK239T/ARPE19
	In HEK293T	In ARPE19	In HEK293T	In AREP19	
Reference	24527.5	106308.5	1	1	1
HEK_ARPE_1	23215.25	34516	0.95	0.32	2.97
HEK_ARPE_2	17301	23928.5	0.71	0.23	3.09
HEK_ARPE_3	31671.25	53428.25	1.29	0.50	2.58

Table S5. Experimental protein expression of Gluc mRNA variants with cellular differential expression. This table presents experimental protein expression for various Gluc mRNA construct, including the reference sequence and RiboCode-designed variants. The designed variants were predicted to express more in HEK293T than in ARPE19 (HEK_ARPE_1 to 3, w=0). For each construct, the table shows original protein expression level, expression level relative to the reference, and expression ratio of HEK239T/ARPE19. Protein expression was measured by fluorescence intensity.