nature portfolio

Peer Review File

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

Corresponding Author: Professor Zhi Xie

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The study introduces RiboCode, a deep learning framework that optimizes mRNA codon sequences to enhance protein production. This framework uses ribosome profiling data to generate codon-optimized mRNA sequences, resulting in increased protein expression across various experimental systems. This approach represents a distinction from traditional codon optimization methods, which largely depend on empirical rules or heuristic strategies. By integrating multiple layers of modeling, the authors successfully demonstrate the ability to design mRNA sequences that outperform their native counterparts in protein production.

One of the most compelling aspects of this work is its versatility, as evidenced by its application to therapeutic proteins and its robust performance in generating optimized designs across diverse mRNA formats, including m1Ψ-modified and circular mRNAs. These findings underscore the potential of machine learning in addressing the challenges of mRNA design, particularly in the context of mRNA-based therapies.

A key strength of this study lies in the integration of three components: a translation prediction model, a minimum free energy (MFE) prediction model for evaluating mRNA secondary structure stability, and a codon optimizer that fine-tunes sequences based on the predictions from the former two models. This combined approach allows the framework to tackle both translational efficiency and structural considerations, offering a potentially comprehensive solution to mRNA optimization. Overall, this work provides an innovative contribution to the field by demonstrating how deep learning can overcome limitations of traditional methods. The authors effectively illustrate the promise of machine learning in enhancing protein production and advancing the development of mRNA-based therapeutics. However, the manuscript requires revisions to address key concerns regarding the disconnect between translation efficiency and protein production, the role of MFE in the optimization process, and more detailed data analysis and documentation. Once these concerns are addressed, the work could significantly contribute to the field of mRNA optimization and synthetic biology.

Major concerns:
1. The paper does not sufficiently address the disconnect between so-called translation efficiency (measured by ribosome profiling data) and protein production, which complicates the interpretation of the results. Protein levels were measured at multiple time points (6, 24, 48, and 72 hours), but the temporal dynamics of RNA stability and protein degradation were not accounted for. Without this consideration, the authors' claims that their optimization model improves protein production may overestimate the role of mRNA codon optimization alone.

2. While the authors provide strong evidence that their framework can generate codon-optimized mRNA sequences for enhanced protein production, some claims, particularly regarding the use of MFE in the optimization process, are not fully substantiated. The framework's reliance on MFE as a factor in codon optimization is mentioned but not sufficiently tested or discussed in the results. This leaves uncertainty about the role of MFE in the optimization process.

3. The authors train their own differentiable MFE model, but its performance is not evaluated in the manuscript. The authors should evaluate their model and compare it to existing differentiable MFE models such as that presented in the work "Differentiable partition function calculation for RNA" or the preprint "Scalable Differentiable Folding for mRNA Design" to

demonstrate novelty. Furthermore, the novel “integrative training” of the MFE model is outlined without any demonstration of an increase in performance relative to a standard training approach, which should be rectified.

Minor concerns:
4. The paper would benefit from better descriptions of cleanup of the paired Ribo-seq and RNA sequencing datasets from different human tissues and cell lines, as this would align it with current high-standards in scientific publishing. One such preprint which can be used as an example is Predicting the translation efficiency of messenger RNA in mammalian cells | bioRxiv

5. The correlation between predicted and actual protein levels for Gluc ($p = 0.077$ in Figure 4b) is marginal and raises questions about the strength of the prediction model. The manuscript could acknowledge this limitation, which may be due to the small number of tested constructs, and emphasize the need for further validation through much larger data sets.

6. The study shows higher expression ratios in HEK293T cells versus ARPE19 (Figure 4d) but does not discuss potential underlying reasons for this discrepancy (e.g., differential ribosome abundance or mRNA stability).

7. It is remarked that “We recommend selecting one or more candidates from each of the three optimization rounds (w=0, w=0.5, w=0.7) for experimental validation”. However, the authors do not specify the value of w for the sequences presented in their own experimental validation, for example those shown in Figures 4c-f. This should be included to justify the recommendation.

8. It is not clear what the constants alpha and beta in the loss function of the codon optimizer are; they should be described explicitly.

9. It was unclear whether the computer models presented had enough data for training, especially the model predicting translation efficiency. Following best practices in machine learning literature, the authors should describe the number of parameters in the model and the number of data used for supervised training to let the reader assess whether the model was ‘overtrained’ or ‘undertrained’ (based on data/parameter ratio). There should also be a clear description of the number of layers, dimensions of kernels, etc. for each of the models, including not just the prediction models but also the design models.

(Remarks on code availability)
I have briefly inspected the README.md at the Github repository but not followed the link to download from Google drive due to concerns about this link preserving anonymity. The authors and journal editorial team needs to make the code available to us via a .zip or fileshare link hosted at the journal website.

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
In this study, Li et al. introduce RiboCode, a context-sensitive deep learning framework for optimizing mRNA codon sequences to improve protein production. Unlike previous approaches that rely on predefined sequence metrics, RiboCode learns the impact of cellular environment and mRNA abundance on protein translation using paired RNA-seq and ribosome profiling data, allowing it to predict and optimize translation levels in specific cell lines. The authors validate RiboCode’s performance through in vitro and in vivo experiments, demonstrating improvements in protein expression for m1Ψ-modified and circular mRNAs. These optimizations translate to enhanced immune responses in a mouse model of influenza vaccination and improved neuroprotection in an optic nerve injury model.

I do not doubt the technical quality of this work. However, it seems that the current manuscript has not fully explored the potential of this context-sensitive approach in practical applications. I have provided the following comments, which I hope will help improve this work:

- I am not sure whether the context-aware approach specifically contributes to the efficiency of designed mRNA in Figure 5/6, or if the model still relies on a general rule learnt from ribosome profiling data. For example, how do the authors set the cellular environment for these designed sequences? It would be fine to use a median or default cellular state, but this should be clarified. Additionally, I suggest the authors discuss practical considerations for applying their tool in vaccine or mRNA therapy optimization, especially the choice of cellular context and mRNA level.

- To compare the performance of RD and LD (LinearDesign), the authors selected four w values for RD and two lambda values for LD. Could they provide rationale for these choices? My concerns are as follows: (1) the number of designs for the two methods are different, and clearly RD explores a larger parameter space. (2) for LinearDesign, lambda ranges from 0 (structure-only optimization) to infinity (translation-only optimization), and a larger lambda value would be more comparable to the w=0 case for RD. (3) Table S3 shows that the MFE changes from -350 to -300 as lambda increases from 0 to 4, while the optimal RD1 design has an MFE around -200. It seems that the authors should further increase lambda for LD to achieve a similar MFE.
In short, I would suggest the authors ensure a fair comparison to fully demonstrate the advantages of RD over LD.

- It would be helpful to show that the optimized codon sequences designed by RD vary across cellular environments. For example, the authors could generate a t-SNE plot of the optimized Gluc codon sequences with w=0 in different cell lines (HEK293T, HeLa, and A549) and visualize the actual difference.

- The performance of the MFE prediction model is not benchmarked in the current manuscript. I recommend that the authors provide a comparison between the predicted MFE values and those calculated using RNAfold to demonstrate the accuracy and reliability of their prediction model.

- Figure S1 needs clarification: the term “multi-head” is usually associated with attention mechanisms, and it is unclear how this concept is applied to CNN in the context; the “multi-head attention” part in Figure S1 doesn’t seem to function like a transformer layer and it appears to be a simple multiplication of outputs from two network branches; the “CNN Encoder” term is also misleading - does this imply there’s a decoder as well? I would suggest the authors revise Figure S1, Figure 1b, and the methods section to better explain the structure of their neural network.

- The software package should provide users with the flexibility to define the cellular environment, i.e., allow the optimization of codon sequences based on user-specified gene expression profile.

Minor comments:
- Figure 4 c/d, figure titles and captions do not match.
- Translation levels of Gluc sequences in Figure S4 seem much higher than those in Figure 3b. Are these plots accurate?
- The Y axis label is missing for Figure S5.
- Table S5, "ARPE19" is misspelled as "AREP19".
- The name "RiboCode" has already been used in a prior work (https://github.com/xryanglab/RiboCode). I recommend the authors consider renaming their software to avoid any confusion.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
Peer Review Report for NCOMMS-24-67515-T:

The authors introduce RiboCode, a novel deep learning-based framework designed to optimize mRNA codon sequences to enhance protein production and therapeutic efficacy. Unlike traditional pre-defined metric-based codon optimization, they claim their deep learning-based method better learned complex relationships governing translation and explored a bigger sequence space. RiboCode leveraged large-scale ribosome profiling (Ribo-seq), cellular environment (e.g., RNA-seq), and RNA secondary structure prediction datasets to better capture the interplay between codon usage, cellular context, translation efficiency, and stability.
RiboCode’s sequence optimizer framework combined a trained translation prediction model, an MFE active learning model, and a codon optimizer that employed activation maximization to explore and optimize mRNA sequences for high translation and/or stability. The authors validated its effectiveness through in vitro experiments and in vivo studies, demonstrating increased protein expression, and improvements in vaccine immunogenicity and therapeutic efficacy. Additionally, they highlighted its compatibility with chemical modifications and mRNA formats, including m1Ψ-modified and circular mRNAs which are highly relevant to the field of mRNA therapeutics.
This study underscored the advantages of using deep learning approaches to mRNA codon optimization.
While the manuscript presents a novel and promising approach, the points raised below highlight significant gaps that need to be addressed to ensure clarity, rigor, and reliability. Overall, we recommend major revisions to improve the robustness of the findings and alignment with the claims made. With these revisions, the manuscript could make a valuable contribution to the field of mRNA therapeutics.

Major comments:

Major Comment #1: Additional in silico validation is needed to support the claims

1. Sequence Space Exploration: The manuscript emphasizes the exploration of sequence space as a key strength of RiboCode, but no direct comparison with LinearDesign or other existing tools (Ribotree (PMID: 34520542), CDSFold (PMID:

26589279)) are provided to demonstrate this claim. Including such comparisons is essential to validate the broader sequence exploration ability of RiboCode.

2. MFE Model Learning: The manuscript lacks evidence for how well the MFE model learns to predict MFE values. Results on unseen mRNAs (separated by families) and their associated metrics would provide confidence in the generalizability of the MFE model. A recent article from Szikszai highlights this issue (PMID: 35748706).

3. "New Genes" Definition: The term "new genes" in the study may be misleading if the sequences come from the same gene families or are not entirely novel. A clarification or additional analysis is necessary to validate this claim.

4. Energy Landscape Visualization: Since a local optimizer is employed, the manuscript should analyze the energy landscape of different sequences. This would help demonstrate whether the optimizer avoids local minima and, if not, why this is acceptable in the given context.

5. CAI Threshold (Fig 3c): The threshold of 10% for CAI calculation is arbitrary. Since the differences between “Endo” and “RD” are small but statistically significant, a smaller threshold (e.g., 1%) may make the result statistically not significant. The authors should explain the rationale or explore other thresholds.

6. Model Extrapolation (Figs 2 & 3): Figure 2 shows a maximum predicted translation of 10, while Figure 3 reports values up to 25. This discrepancy suggests potential extrapolation by the model, which should be addressed and clarified.

7. Ablation Studies and Error Bars: Error bars are missing in the ablation studies and other analyses (e.g., different dataset splits). Including error bars and biological replicates would enhance the reliability of the findings.

8. Cellular Context Justification: Cellular context contributes only marginally (R2=0.06), yet it is heavily justified as a critical component of the model. This justification should be toned down or revised to align with the results.

9. Key References Missing:

1. The authors have overlooked important contributions to the field, including Mauger et al., 2019 (PMID: 31712433) and Ribotree (PMID: 34520542). These references are foundational in mRNA translation and optimization research and should be cited with appropriate acknowledgment. The relevance of these methods (along with LinearDesign) for therapeutic mRNA design were reviewed in Metkar et al. (PMID: 38030688).
2. Other codon optimization algorithms that utilize deep learning Lines 115-117. LLM bases approaches like CodonBERT (PMID: 38951026), and ChatNT (https://doi.org/10.1101/2024.04.30.591835) have recently come out. These methods need to be acknowledged in the Introduction. While these methods do not leverage RiboSeq, the authors can point to this to indicate how their deep learning approach covers this aspect.
3. Other machine learning models for ribo-seq data: Line 129: There have been deep-learning approaches exploring Ribo-Seq data like RiboShape (PMID: 27307616), Riboformer (PMID: 38443396) and more recently RiboTIE (PMID: 38585907), RiboNN (https://doi.org/10.1101/2024.08.11.607362). The statement should be changed to acknowledge other methods so far. The authors should contrast their model and findings with a recent study from Can Cenik lab and Sanofi (RiboNN). The model uses a similar approach and also the UTR sequences in input with a much bigger compendium of datasets. While that study is not yet peer-reviewed, the authors can make their manuscript better by providing some insight on the differences between the other study and theirs to better inform the readers.

Major Comment #2: The definition of ‘translation level’ is not clear and the relationship with mRNA abundance.

1. Definition of translation: The authors do not provide a clear definition of “translation level” in either the result or methods section. Is ‘translation level’ the RPKM from Ribo-Seq or is it ‘translation efficiency’ defined for each transcript, RPKM from Ribo-Seq divided by RPKM from paired RNA-Seq? authors need to explicitly specify the definition for better interpretability of the results. Dependence on mRNA Abundance
The definition of "translation level," based on RPKM from both Ribo-Seq and paired RNA-Seq, inherently incorporates mRNA abundance as both an input and a component of the target variable. While this approach is valid in deep learning due to the ability to model non-linear relationships, it raises the risk of overemphasizing mRNA abundance at the expense of other features, particularly given that mRNA abundance already accounts for the highest R2R2 among individual features. The ablation analysis indicates that the model performs slightly better than mRNA abundance alone, with codon sequence and cellular environment contributing incrementally to R2 (0.76 vs. 0.6). However, the authors should explicitly acknowledge in the Results or Discussion that the high contribution of mRNA abundance may stem from its dual role as both an input and a component of the target variable. Additionally, it would be valuable for the authors to clarify whether the model has been rigorously tested to ensure it captures meaningful relationships beyond mapping mRNA abundance to translation levels.

2. Relative Contributions of Transcription vs. Translation: The R2 of 0.6 for transcription dominance suggests that the model primarily learns transcriptional rather than translational regulation. The authors should address how this impacts the utility of the method for optimizing mRNA therapeutics, where translational rules may play a more critical role.
Further, we recommend that the authors discuss their results in the context of the findings by Schwanhäusser et al. (PMID: 21593866), which demonstrated that post-transcriptional regulation at the level of translation played an equal or greater role in determining protein expression compared to transcriptional regulation. The current study, in contrast, emphasizes the

predictive power of transcription-level measures (mRNA abundance). The authors should explore how this discrepancy impacts the interpretation of their results and the relative contributions of transcription and translation to protein expression. A deeper discussion of these points would significantly enhance the manuscript's clarity and broader relevance.

3. Effect of mRNA abundance input for codon optimization: mRNA abundance is an input parameter for the translation prediction model and for every step of the codon optimizer, a constant mRNA abundance was set to 4.5 (line 647). Have the authors tested the performance of the codon optimizer by varying this input?

4. Missing 5'UTR Context: Translation initiation is a limiting step, yet the model appears to lack information about the 5'UTR. If this is not incorporated, we think, its potential impact on translation prediction needs more discussion

5. Importance Plot and Padding: The per-position importance plot may be biased due to 3' padding, as low importance at the 3' end could be an artifact. Additional analyses are needed to account for this.

Major Comment #3: Incorporating MFE Optimization in the Discussion and Results

Some of the RD sequences designed for experimental validation incorporate an MFE optimization component (w>0), which is not thoroughly discussed in the manuscript. According to the methods (line 688), the weights for RD1, RD2, RD3, and RD4 are w=0,0,0.7,0.5, respectively, indicating that RD3 and RD4 were optimized for both translation level and lower MFE. This design choice warrants further clarification and discussion.

Specific Recommendations:
1. Explicit Labels for Weights (w) in Figures: Figures that compare RD sequences, particularly Figures 4, 5, and 6, should include explicit labels indicating the translation level and MFE weights (w) used for each sequence. This would improve the clarity and interpretability of the data.

2. Consistent Nomenclature Across RD Sequences: The order and naming of RD sequences across Figures 4, 5, and 6 are inconsistent. For example, the order of RD sequences based on w for HA differs from that of Gluc and NGF. Standardizing the nomenclature based on the weights would make data interpretation more intuitive and consistent.

3. Discussion of Results Based on Different w Values: The manuscript should explicitly discuss the varying performance of RD1 and RD2 (designed for translation only) compared to RD3 and RD4 (designed for both translation and stability). This is particularly relevant for circular RNAs, where RD3 and RD4 outperform RD1 and RD2 in protein expression. The authors should explore and discuss how optimizing for MFE contributes to higher protein expression in circular RNA contexts.

4. Contrasting MFE and Translation-Only Models: While the Discussion notes that MFE did not have a general impact on translation, it fails to emphasize that MFE optimization significantly contributed to the performance of top-performing sequences in experimental validation. The authors should address this contrast and discuss how MFE-optimized sequences outperformed translation-only models in specific contexts.

Major Comment #4: Experimental Validation: mRNA variants, cellular context, and different mRNA formats.

1. Luciferase Feature Analysis (Fig 3d): The two luciferase proteins exhibit opposite CAI importance. This variability requires further explanation, especially in relation to their sequence features.
2. Cellular Context Claims (Fig 4): The difference between the predicted and experimental ratios for the HEK293 vs other cellular contexts is stark. This undermines the claim that the model learns cellular context. The authors should refine this claim or redesign sequences to better reflect cellular differences. Second, a barplot with error bars would be a better representation for this data.

3. Performance on Modified Chemistries: With m1Ψ-modified sequences, the performance drops to a 4.6-fold improvement from 72-fold. This suggests that the model does not effectively capture rules for modified chemistries and should be discussed.

4. Circular mRNA in Fig 6: Circular mRNAs perform better than linear in Figure 6 but not in previous figures. This discrepancy should be explained.

Major Comment #5: Missing/additional controls

1. Reference Sequence CAI: The reference Gluc sequence has a CAI of 0.8, Mauger et al., 2019 showed to have very low expression. This may undermine the claim of a 72-fold improvement. A higher CAI (~1.0) control should be included.

2. Fluc Performance (Fig S8): The 17-fold improvement for Fluc (vs. 72-fold for Gluc) suggests the model's translation rules

are not well-generalized. Sequence feature comparisons could be performed between the RD and reference sequences for these 2 related proteins to explain this discrepancy. Missing CAI control here as well.

3. Dose Response (Fig 5): Unlike Figure 6, Figure 5 lacks a dose-response experiment. Including this would strengthen the conclusions.

4. Missing Controls (Fig. 5 and 6): No codon-optimized (e.g., high CAI) sequences were used as reference. As Kudla et al., (PMID: 16700628) have demonstrated that WT sequences with low GC3% usually do not express well.

Minor Comments

Minor Comment #1: Line 74-76. MFE is discussed only in relation to mRNA stability. Lower MFE/higher mRNA structure can also be a hurdle for translation elongation and needs to be acknowledged here.

Minor Comment #2: Lines 91-95. The authors can be more specific and provide some examples of any trans-factors that affect translation that may not be captured by traditional methods.

Minor Comment #3: Predicted translation values below zero in Figure S4 suggest a model error. The authors should ensure that predictions are constrained to valid ranges.

Minor Comment #4: The r=0.7 correlation in Figure S7 is derived from two distinct groups of points, making it a poor representation of the data. Alternative visualizations or metrics should be considered.

Minor Comment #5: Figure 5a requires bar plots quantifying the Western blot signals for a more accurate comparison.

Minor Comment #6: Lines 1050-1051, 1065-1066, 1081-1082. The w values for the RD sequences need to be shown either in the figure legends or at least in the figure captions.

Minor Comment #7: Fix grammatical errors –
1. Line 386: …to several factors.
2. Line 392: …mRNA translation whereas the previously codon optimization…

(Remarks on code availability)

Reviewer #5

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)
The code is accessible as a .whl file. The code can run the model with given parameters. I did encounter an issue with installation that I state below.

The instructions for installation should be more clear in README file. The run.py has a 'python' command. It is increasingly common for modern Linux distributions and macOS to have only python3 installed and not python. The following error was encountered which needed some troubleshooting.

```
Traceback (most recent call last):
File "~/.local/bin/ribo-code", line 8, in <module>
sys.exit(main())
File "~/.local/lib/python3.8/site-packages/RiboCode/run.py", line 24, in main
result = subprocess.run(command)
File "/usr/lib/python3.8/subprocess.py", line 493, in run
with Popen(*popenargs, **kwargs) as process:
File "/usr/lib/python3.8/subprocess.py", line 858, in __init__
self._execute_child(args, executable, preexec_fn, close_fds,
File "/usr/lib/python3.8/subprocess.py", line 1704, in _execute_child
raise child_exception_type(errno_num, err_msg, err_filename)
FileNotFoundError: [Errno 2] No such file or directory: 'python'
```

I had to specify sudo ln -s /usr/bin/python3 /usr/bin/python or use alias python='python3' in my ~/.bashrc to avoid this error.

The authors should either replace python with python3 in their run.py or provide more explicit instructions.

Reviewer #6

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
While the authors have addressed some of our previous concerns, we still have reservations regarding the training and validation of their models:

1. One concern remains the imbalance between the amount of training data and the number of parameters in both the translation and MFE models (our prior point 9). The data/parameter ratios, 0.12 for the translation model and 0.02 for the MFE model, are far below the generally accepted minimum for robust training (much greater than 1), raising doubts about the generalizability of the models and the validity of the chosen hyperparameters. We recommend the authors explicitly address this issue of undertraining in the main text.

2. We remain unconvinced by the authors' explanation for the α and β parameters, specifically, their assigned values of 100, used to scale the translation and MFE losses (our prior point 8). They cite ‘empirical experience’ but do not provide specific results that justify constraining the translation loss to 0-10 and the MFE loss to 0-100. Moreover, the revised manuscript states that “incorporating MFE optimization played a complementary role in modulating mRNA secondary structure stability”, but based on the reported loss values, the MFE seems to be weighted more highly. The authors should rectify this disparity.

3. Finally, the authors' response to our previous comment regarding the evaluation of their model against RNAfold (Figure S13) remains inadequate (our previous point 3). The sequence diversity used in Figure S13 is still insufficient, being limited to sequences synonymous with only three seed sequences. Without knowing the seed sequence(s) used for training the MFE model, it is difficult to assess its generalizability based on these tests. In addition, the gradient of the line of best fit is significantly different between the three plots (0.12, 0.16, 0.03). By separating the data into three plots, the R-value is artificially raised. The authors could rectify these issues by testing their model on a wider diversity of sequences and providing the R-value for the predictions across all sequences.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors have addressed most of my previous concerns and have provided an improved manuscript. I have a few additional comments that I hope will further strengthen the work.

1) I appreciate the authors' effort to revise Figure S1. Tables S9 and S10 would benefit from clearer notation for model parameters. Several names are ambiguous or contain minor errors—for example, “max-pool” should be “max-pooling.” It is also difficult to understand parameter names such as “Out” or “encoder”. For guidance on documenting neural network

layers and parameters, the authors might refer to RiboTIE (doi.org/10.1038/s41467-025-56543-0; Supplementary Algorithm 1).

2) The new comparison between RD and LD in Figure S11 is helpful. To aid interpretation, please consider adding design details for LD3 and LD4 to Table S3 (e.g., CAI values and predicted translation level). A minor point: the authors mentioned that RD1 and RD2 in Figure S11 are identical to those in Fig. 4a, yet the values seem different. In Figure 4a, RD1 is ~5 at the 6 h time point, whereas in Figure S11 it is ~2.5. Could the authors double-check these data?

3) Figure S13 shows a strong correlation between the MFE model and RNAfold predictions, but the absolute scales are quite different. For INS sequences, RNAfold MFE values range from –250 to –220, whereas the MFE model spans only –207.25 to –206.25 (30-fold difference). Could the authors comment on the source of this discrepancy? This might not affect the sequence optimization part since the weight parameter w could compensate for this effect. But additional clarification of the MFE model's training and benchmarking protocol would be helpful.

some minor errors:
1) Line 917: "Gaussian luciferase" should be "Gaussia luciferase".
2) Table S5: both columns for 24h protein expression are labeled "In HEK293T". Column label "In AREP19" should be "In ARPE19".

(Remarks on code availability)
I successfully installed the software package. To make RiboDecode more broadly useful, it would help to allow optimization of user-defined coding sequences, rather than limiting inputs to the three predefined examples (gluc, NGF, H1N1).

Reviewer #4

(Remarks to the Author)
Please see the attached document.

(Remarks on code availability)

Reviewer #5

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #6

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have addressed most of our comments. On seeing this revision, however, it was surprising to read that the authors' neural MFE model is specific to particular sequences. There already exist models which can predict MFE for any sequence without needing to be re-trained for each individual mRNA, like the Krueger/Ward model. The manuscript mentions that this prior model isn't available, but it seems that it has been out for a while: https://github.com/rkruegs123/jax-rnafold. The authors should test the model in their pipeline, to evaluate if their MFE model provides any benefit compared to it.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors have addressed my previous concerns.

A few minor points:
(1) It is unusual to use loge(RPKM) in the axis label in Fig. 2a. If the base of the logarithm is e, it is more conventional to use ln() instead.

(2) The manuscript mentions that the MFE model needs to be trained for new CDS inputs. It might be helpful if the authors could provide additional guidance in the GitHub repository on how the MFE model is trained for a new sequence. For example, how many training samples are typically generated? This information would facilitate broader adoption of the method.

(Remarks on code availability)

Reviewer #6

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Version 3:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have satisfactorily addressed our comments. As a minor suggestion to help readers, the authors should include the time to train their model in the time taken to perform the optimization (since this must be done for every new sequence, unlike JAX-RNAFold).

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

(Remarks on code availability)

Reviewer #6

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The study introduces RiboCode, a deep learning framework that optimizes mRNA codon sequences to enhance protein production. This framework uses ribosome profiling data to generate codon-optimized mRNA sequences, resulting in increased protein expression across various experimental systems. This approach represents a distinction from traditional codon optimization methods, which largely depend on empirical rules or heuristic strategies. By integrating multiple layers of modeling, the authors successfully demonstrate the ability to design mRNA sequences that outperform their native counterparts in protein production.

One of the most compelling aspects of this work is its versatility, as evidenced by its application to therapeutic proteins and its robust performance in generating optimized designs across diverse mRNA formats, including m1Ψ-modified and circular mRNAs. These findings underscore the potential of machine learning in addressing the challenges of mRNA design, particularly in the context of mRNA-based therapies.

A key strength of this study lies in the integration of three components: a translation prediction model, a minimum free energy (MFE) prediction model for evaluating mRNA secondary structure stability, and a codon optimizer that fine-tunes sequences based on the predictions from the former two models. This combined approach allows the framework to tackle both translational efficiency and structural considerations, offering a potentially comprehensive solution to mRNA optimization.
Overall, this work provides an innovative contribution to the field by demonstrating how deep learning can overcome limitations of traditional methods. The authors effectively illustrate the promise of machine learning in enhancing protein production and advancing the development of mRNA-based therapeutics. However, the manuscript requires revisions to address key concerns regarding the disconnect between translation efficiency and protein production, the role of MFE in the optimization process, and more detailed data analysis and documentation. Once these concerns are addressed, the work could significantly contribute to the field of mRNA optimization and synthetic biology.

Major concerns:
1. The paper does not sufficiently address the disconnect between so-called translation efficiency (measured by ribosome profiling data) and protein production, which complicates the interpretation of the results. Protein levels were measured at multiple time points (6, 24, 48, and 72 hours), but the temporal dynamics of RNA stability and protein degradation were not accounted for. Without this consideration, the authors' claims that their optimization model improves protein production may overestimate the role of mRNA codon optimization alone.

We thank the reviewer for raising this question. We agree that our model increases protein production by optimizing translation efficiency and does not account for other factors such as protein degradation. To reflect this and avoid the confusion, we replaced "protein production" by "mRNA translation" in the manuscript, as follows:

1. Title: Deep Generative Optimization of mRNA Codon Sequences for Enhanced mRNA Translation and Therapeutic Efficacy.
2. In Abstract: Here, we present RiboDecode, a deep learning framework that generates mRNA codon sequences for enhanced mRNA translation.
3. In Introduction: In this study, we present RiboDecode, a deep learning model for mRNA codon optimization that enhances mRNA translation by directly learning complex relationship…
4. In Results: RiboDecode is a deep learning-based framework for optimizing mRNA codon sequences.
5. In Discussion: While our study demonstrates the significant potential of RiboDecode in optimizing mRNA codon sequences for enhanced mRNA translation and therapeutic efficacy, there are several limitations…
6. In Discussion, we deleted: ~~While this approach yielded significant improvements in protein production and therapeutic efficacy,~~ our future work…

2. While the authors provide strong evidence that their framework can generate codon-optimized mRNA sequences for enhanced protein production, some claims, particularly regarding the use of MFE in the optimization process, are not fully substantiated. The framework's reliance on MFE as a factor in codon optimization is mentioned but not sufficiently tested or discussed in the results. This leaves uncertainty about the role of MFE in the optimization process.

We appreciate the reviewer's insightful comment regarding the role of MFE in our optimization framework. To clarify, the incorporation of the MFE prediction model is intended to address mRNA stability by reducing the free energy associated with secondary structure formation. However, we observed that while optimizing codons for enhanced mRNA translation, the MFEs of some mRNA increased and some decreased (Figure 3D). This variability underscores that MFE is not a direct predictor of translation efficiency but serves as a complementary parameter that can help mitigate unfavorable secondary structures when balanced with translation optimization.

For the four mRNAs we experimentally tested, we found that incorporating MFE optimization for Fluc, HA, and NGF led to optimal performance, indicating that a joint optimization strategy (with w between 0 and 1) can be beneficial. However, we also acknowledge that the best w value for each mRNA is not straightforward to determine and likely varies depending on the specific sequence and experimental context. In our revised manuscript, we expanded the discussion:

Second, our results indicate that while our primary goal was to enhance translation through codon optimization, incorporating MFE optimization played a complementary role in modulating mRNA secondary structure stability. However, we observed the best w value appears to vary depending on the specific mRNA and its context, suggesting that a one-size-fits-all approach may not be appropriate. Consequently, we recommend that experimental designs include the testing of multiple w values to identify the optimal balance for each mRNA target. Further research into the determinants of the optimal w value will be critical to refine this strategy and could lead to more systematic approaches for integrating MFE into mRNA design.

3. The authors train their own differentiable MFE model, but its performance is not evaluated in the manuscript. The authors should evaluate their model and compare it to existing differentiable MFE models such as that presented in the work "Differentiable partition function calculation for RNA" or the preprint "Scalable Differentiable Folding for mRNA Design" to demonstrate novelty.

Thanks, and per the suggestions, we evaluated our MFE model in two aspects in the revision: 1. We validated the MFE values of mRNAs between our model and RNAfold. 2. We compared our model to LinearDesign and the paper suggested by the reviewer.

Comparison to RNAfold:
We found that our MFE value highly agreed with the one from RNAfold, showing the reliability of our MFE model. We showed the result in the Methods section of the revision.
We evaluated the MFE values of mRNAs between our model and RNAfold. We found that our MFE value highly agreed with the one from RNAfold (Figure S13), showing the reliability of our MFE model.

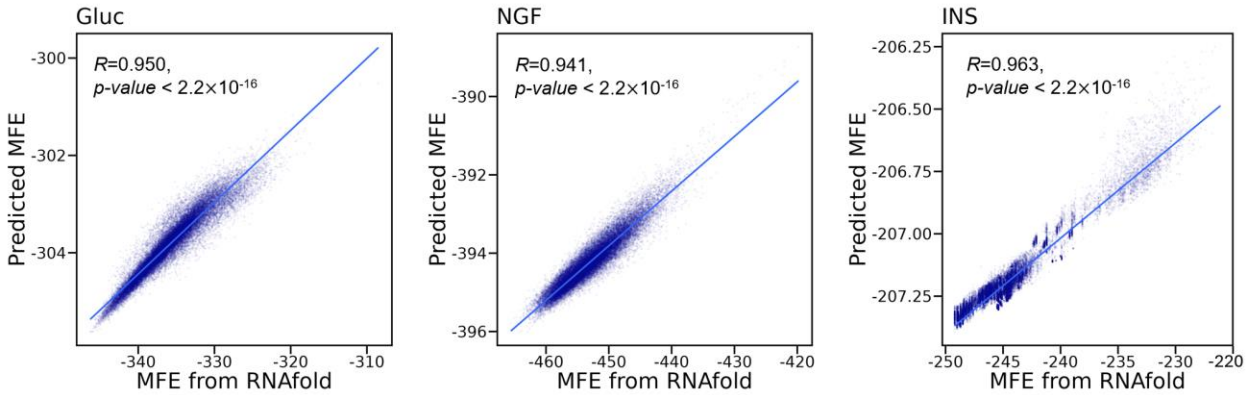


Figure S13. Evaluation of the Predictive Performance of the MFE Model to RNAfold. Pearson's correlation coefficients are provided. Blue lines denote the linear fit. All values are reported in kcal/mol.

Comparison with LinearDesign and the model in "Scalable Differentiable Folding for mRNA Design (Krueger et al. 2024)":

The paper “Differentiable partition function calculation for RNA” presented a prediction model and the preprint “Scalable Differentiable Folding for mRNA Design” presented the MFE optimization model, where we referred it as Krueger’s model. We tested the lowest of MFE values that LinearDesign, Krueger’s model and our MFE model can optimize. As Krueger’s model is not publicly available, we optimized the same mRNA sequences in the Krueger’s paper and compared the results in their paper. We found that all the three models significantly decreased MFE values of given mRNAs while Krueger’s model and LinearDesign can achieve lower MFE values. As mentioned in the paper, LinearDesign is not a differentiable model, so we can’t use it in our framework. Once the Krueger’s model is publicly available, we will test and incorporate it into our framework.

We have incorporated the following content into the Methods of the revision.
A recent study presented a differentiable MFE optimization model (Krueger’s model)[68]. We compared the lowest of MFE values that LinearDesign, Krueger’s model and our MFE model could optimize. As Krueger’s model is not publicly available yet, we optimized the same mRNA sequences in the Krueger’s paper and compared the results presented in the paper (Table S12). We found that all the three models significantly decreased MFE values of given mRNAs while Krueger’s model and LinearDesign can achieve lower MFE values. As LinearDesign is not a differentiable model, we cannot use it in our framework. Once the Krueger’s model is publicly available in the future, we will test and incorporate it into our framework.

mRNA	Unoptimized	LinearDesign	Krueger’s model	RiboDecode
MEV	-88.20	-114.84	-114.92	-111.17
Mini-GFP	-150.20	-207.65	-208.59	-202.39
Nanoluciferase	-332.70	-452.34	-452.38	-410.399
spike RBD	-278.20	-411.55	-412.59	-378.5
eGFP + degron	-358.40	-546.92	-547.71	-503.799

Table S12. Optimized MFEs of five benchmark mRNA sequences by LinearDesign, Krueger et al. 2024, and RiboDecode. The results of LinearDesign and Krueger et al. 2024 are derived from the study titled "Scalable Differentiable Folding for mRNA Design". All values are reported in kcal/mol.

Furthermore, the novel “integrative training” of the MFE model is outlined without any demonstration of an increase in performance relative to a standard training approach, which should be rectified.

Regarding the integrative training of MFE model, we included the use of *w* and potential benefits of MFE in the revision. Please see the texts in Question 2.

Minor concerns:

4. The paper would benefit from better descriptions of cleanup of the paired Ribo-seq and RNA sequencing datasets from different human tissues and cell lines, as this would align it with current high-standards in scientific publishing. One such preprint which can be used as an example is Predicting the translation efficiency of messenger RNA in mammalian cells | bioRxiv

We appreciate the reviewer's suggestion. To enhance the reader's understanding, we have included the details about the Ribo-seq data processing:
The following steps were implemented for processing the Ribo-seq data in accordance with RPFdb: First, to prevent adapter interference in downstream analyses, the 3′ adapter sequences were manually extracted for each dataset from the original publications or the corresponding MultiQC[3] outputs. Adapter sequences, if present at the ends of sequencing reads, were subsequently removed using Cutadapt (version 1.16)[4]. Next, to minimize rRNA and tRNA contamination, sequences corresponding to rRNA and tRNA were retrieved for each species from ENSEMBL[5] and UCSC[6] databases and removed post-mapping using Bowtie2[7]. Finally, to ensure the retention of high-quality ribosome-protected footprints, which exhibit a characteristic read-length distribution reflecting the size of a translating ribosome on the RNA, only footprints within the 25–34 nucleotide length range were retained after contaminant removal and alignment.

5. The correlation between predicted and actual protein levels for Gluc (p = 0.077 in Figure 4b) is marginal and raises questions about the strength of the prediction model. The manuscript could acknowledge this limitation, which may be due to the small number of tested constructs, and emphasize the need for further validation through much larger data sets.

Thanks for the suggestion. We have acknowledged the limitation in the revision: The predicted translational levels showed a positive correlation with the experimentally measured protein levels (Correlation coefficient=0.71, p-value=0.077, Pearson's correlation, Figure S10). However, the p value is marginal, potentially due to the small number of constructs tested. Further validation with larger datasets is needed.

6. The study shows higher expression ratios in HEK293T cells versus ARPE19 (Figure 4d) but does not discuss potential underlying reasons for this discrepancy (e.g., differential ribosome abundance or mRNA stability).

We appreciate your comment regarding Figure 4d. The different expression levels of HEK293T versus ARPE19 demonstrate RiboDecode's capability to optimize codon sequences considering cellular context, such as ribosome proteins, RNA binding proteins and tRNAs, which influence both mRNA translation and stability. However, the observed discrepancy between predicted and experimental expression ratios, particularly in the HEK293T vs. ARPE19 comparison, suggests that while the model captures context-dependent effects, further refinements may be needed to fully represent the complexities of cellular environments.

We added contents in Results regarding the discrepancy of our experimental and predicted ratios: These results confirm RiboDecode's capability for context-aware design, while the discrepancy in the ARPE19 comparison indicates potential for refining the model to better capture quantitative differences across diverse cellular environments.

7. It is remarked that "We recommend selecting one or more candidates from each of the three

optimization rounds (w=0, w=0.5, w=0.7) for experimental validation". However, the authors do not specify the value of w for the sequences presented in their own experimental validation, for example those shown in Figures 4c-f. This should be included to justify the recommendation.

In our previous manuscript, the *w* parameters of the experimental sequences had been mentioned in "Parameters used for Sequence Optimization"：For Gluc, HA, and NGF sequence optimization, four sequences were designed by RiboDecode (RD1, RD2, RD3, and RD4: w=0, 0, 0.7 and 0.5, respectively) …

Furthermore, we have incorporated these *w* parameter values into the figures:

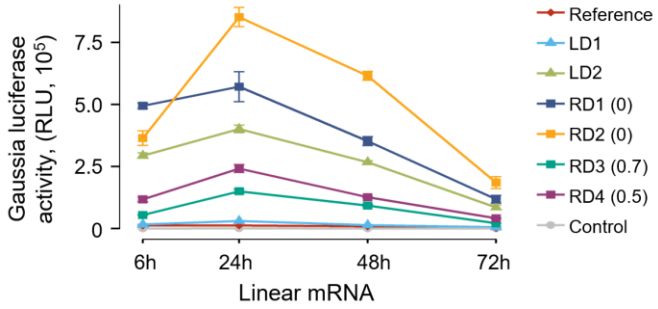


Figure 4a. Protein expression of Gluc was measured by fluorescence intensity. RD sequences were designed by RiboDecode, with w parameter indicated in parentheses…

In addition, we included the use of *w* and potential benefits of MFE in the revision. Please see the texts in Question 2.

8. It is not clear what the constants alpha and beta in the loss function of the codon optimizer are; they should be described explicitly.

Sorry we overlooked this point, and we appreciate the comment.

We have added the following content: The introduction of α and β is intended to balance the loss in the translation optimization and MFE optimization processes. We set the values of both α and β to 100 based on empirical experience, to prevent the translation model's loss to vary within a range of 0-10 while the MFE model's loss fluctuates within a range of 0-100. (Table S11).

Type	Parameter name	Values	Value range
Global	α	100	fixed
	β	100	fixed
	parameters	274.30 K	fixed

Table S11. Hyperparameters for Codon Optimizer.

9. It was unclear whether the computer models presented had enough data for training, especially the model predicting translation efficiency. Following best practices in machine learning literature, the authors should describe the number of parameters in the model and the number of data used for supervised training to let the reader assess whether the model was 'overtrained' or 'undertrained' (based on data/parameter ratio). There should also be a clear description of the number of layers, dimensions of kernels, etc. for each of the models, including not just the prediction models but also the design models.

Thanks for the suggestion. We have added more details of the model in the revised manuscript (Table S9, S10, S11).

Reviewer #1 (Remarks on code availability):

I have briefly inspected the README.md at the Github repository but not followed the link to download from Google drive due to concerns about this link preserving anonymity. The authors and journal editorial team needs to make the code available to us via a .zip or fileshare link hosted at the journal website.

Per the suggestion, we uploaded our software to figshare (https://figshare.com/articles/software/RiboDecode_zip/28916288?file=54138881).

Reviewer #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3 (Remarks to the Author):

In this study, Li et al. introduce RiboCode, a context-sensitive deep learning framework for optimizing mRNA codon sequences to improve protein production. Unlike previous approaches that rely on predefined sequence metrics, RiboCode learns the impact of cellular environment and mRNA abundance on protein translation using paired RNA-seq and ribosome profiling data, allowing it to predict and optimize translation levels in specific cell lines. The authors validate RiboCode's performance through in vitro and in vivo experiments, demonstrating improvements in protein expression for m1Ψ-modified and circular mRNAs. These optimizations translate to enhanced immune responses in a mouse model of influenza vaccination and improved neuroprotection in an optic nerve injury model.

I do not doubt the technical quality of this work. However, it seems that the current manuscript has not fully explored the potential of this context-sensitive approach in practical applications. I have provided the following comments, which I hope will help improve this work:

- I am not sure whether the context-aware approach specifically contributes to the efficiency of designed mRNA in Figure 5/6, or if the model still relies on a general rule learnt from ribosome profiling data. For example, how do the authors set the cellular environment for these designed sequences? It would be fine to use a median or default cellular state, but this should be clarified. Additionally, I suggest the authors discuss practical considerations for applying their tool in vaccine or mRNA therapy optimization, especially the choice of cellular context and mRNA level.

We appreciate your valuable comments and would like to address them as follows:

1. We apologize for not including the cellular condition details in Figures 5 and 6. These details have now been added to the 'Parameters used for sequence optimization' section in the Methods: To preliminarily evaluate mRNA optimization *in vitro*, we first performed codon optimization for HA and NGF for the HEK293T cellular environment and measured their protein expression in HEK293T cells (Figures 5a-c and 6a-c).

2. To assist users in selecting the appropriate cellular environment for sequence optimization, we provided the following contents in Methods: When optimizing mRNA sequences using RiboDecode, we recommend using the RNA expression profile of the target or the most similar tissue or cell type through either experimental sequencing or publicly available datasets. To facilitate implementation, we have

incorporated a user-defined environment input parameter into the software package, accompanied by comprehensive documentation and step-by-step instructions.

- To compare the performance of RD and LD (LinearDesign), the authors selected four w values for RD and two lambda values for LD. Could they provide rationale for these choices? My concerns are as follows: (1) the number of designs for the two methods are different, and clearly RD explores a larger parameter space. (2) for LinearDesign, lambda ranges from 0 (structure-only optimization) to infinity (translation-only optimization), and a larger lambda value would be more comparable to the w=0 case for RD. (3) Table S3 shows that the MFE changes from -350 to -300 as lambda increases from 0 to 4, while the optimal RD1 design has an MFE around -200. It seems that the authors should further increase lambda for LD to achieve a similar MFE.
In short, I would suggest the authors ensure a fair comparison to fully demonstrate the advantages of RD over LD.

We appreciate your valuable input on the RD and LD sequence comparison.

We previously selected the λ parameter for LinearDesign guided by the results presented in their article (PMID: 37130545). In the case of the SARS-CoV-2 spike protein mRNA, the structure-only optimization (λ= 0) yielded the best results. Increasing the λ value serves the purpose of enhancing the CAI. We noted that LinearDesign effectively optimized the mRNA of the VZV gE protein when λ was 4. Additionally, we found that at λ = 4, the CAI of Gluc nearly reached its maximum value (0.97). Considering these findings, we therefore used the λ values to 0 and 4 for our analysis. Per the suggestion, we tested additional LinearDesign-optimized sequences. We increased the λ parameter in LinearDesign and observed that the MFE reached its upper limit at λ = 100. Consequently, we synthesized additional LinearDesign sequences at λ = 16 (-266.8 kcal/mol) and λ = 100 (-246.70 kcal/mol). We have incorporated the following texts into the manuscript:
We noticed that the two RiboDecode-designed mRNAs (RD1 and RD2) had the best performance which also had the highest MFE values of around -200. In contrast, the LinearDesign sequences had MFE values of -350 and -300. To rule out the increased protein production was associated to higher MFE value, we tested additional LinearDesign sequences with similar MFEs (-266.8 and -246.70 kcal/mol, respectively). Compared to four LinearDesign sequences, RiboDecode-designed sequences outperformed (Figure S11).

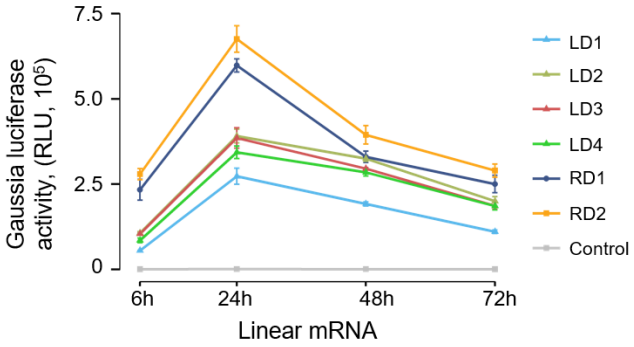


Figure S11. Protein expression of Gluc was measured by fluorescence intensity. RD1and RD2 were the same shown in Figure 4a, which had MFE values of -195.7 and -195.3 kcal/mol, respectively. LD1, LD2, LD3, and LD4 were designed using LinearDesign, with the MFEs of -346.2, -302.5, -266.8, and -246.7 kcal/mol, respectively. “RLU”: relative light units.

- It would be helpful to show that the optimized codon sequences designed by RD vary across cellular environments. For example, the authors could generate a t-SNE plot of the optimized Gluc codon sequences with w=0 in different cell lines (HEK293T, HeLa, and A549) and visualize the actual difference.

Per the suggestion, we optimized the Gluc coding sequence with our model in three cell lines: HEK293T, HeLa, and A549 (w=0). Then, we extracted high-dimensional codon embeddings using CodonBERT (PMID: 38951026) and generated t-SNE plots. The results indicate that the sequences generated under the three distinct cellular environments exhibit different positional distributions in the t-SNE space (Figure SX), suggesting that these sequences may reflect cellular environments.
We have included these in the manuscript: Furthermore, designing Gluc codons for different cell lines showed that the generated codons had distinct sequence patterns for different cellular contexts, reflecting the differences in cellular environment (Figure S7).

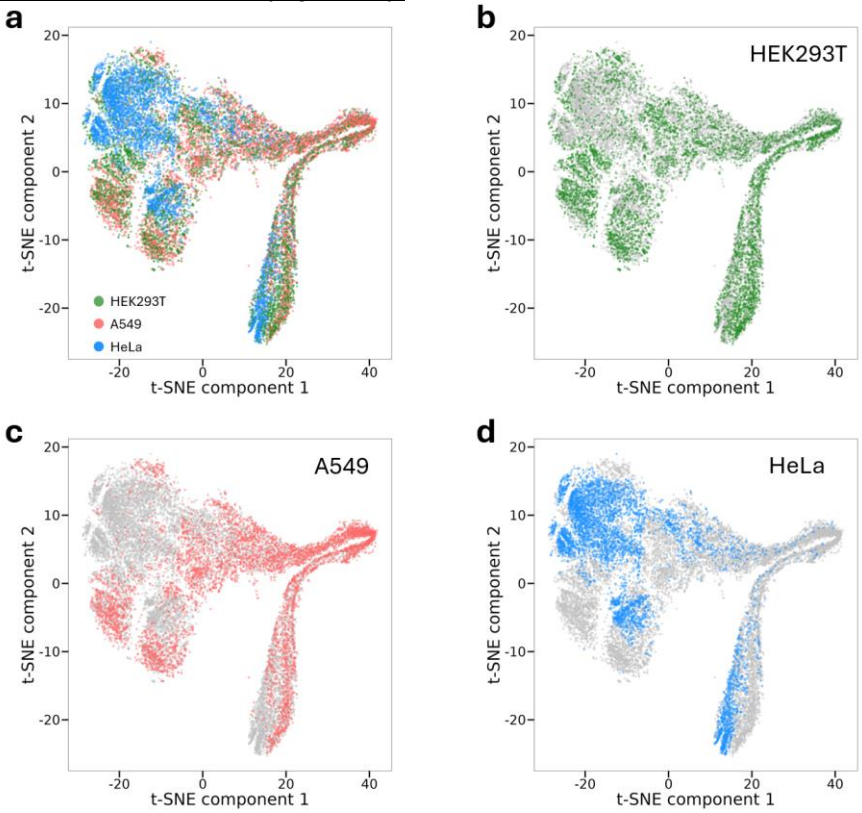


Figure S7. Gluc codon sequences were optimized by RiboDecode in different cellular contexts (w=0). The sequence embeddings were obtained using CodonBERT and visualized in a t-SNE plot. Each dot represents a single sequence, with the color indicating the cell type.

- The performance of the MFE prediction model is not benchmarked in the current manuscript. I recommend that the authors provide a comparison between the predicted MFE values and those calculated using RNAfold to demonstrate the accuracy and reliability of their prediction model.

Thank you for bringing this issue. Per the suggestion, we tested and found that our the MFE value highly agreed with the one from RNAfold, showing the reliability of our MFE model. We showed the result in the Methods section of the revision. We evaluated the MFE values of mRNAs between our model and RNAfold. We found that our MFE value highly agreed with the one from RNAfold (Figure S13), showing the reliability of our MFE model.

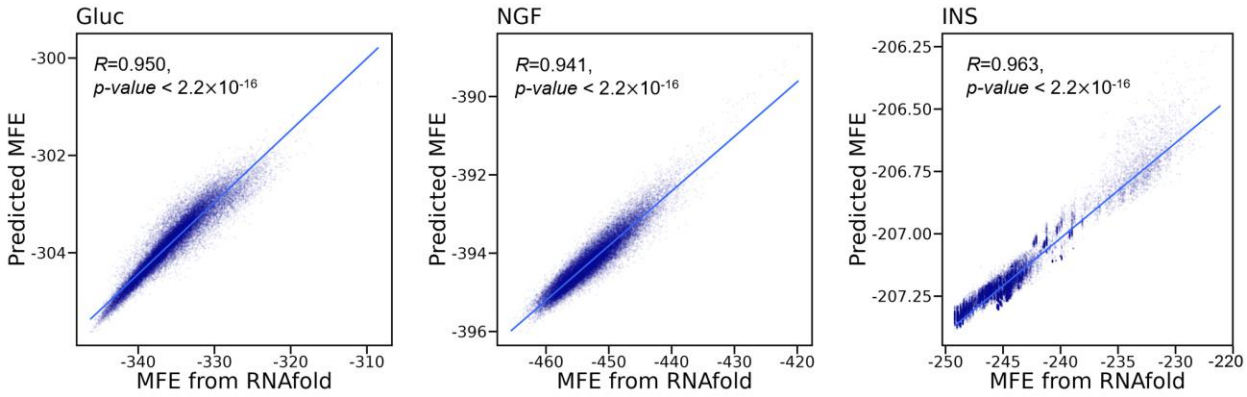


Figure S13. Evaluation of the Predictive Performance of the MFE Model to RNAfold. Pearson's correlation coefficients are provided. Blue lines denote the linear fit. All values are reported in kcal/mol.

- Figure S1 needs clarification: the term "multi-head" is usually associated with attention mechanisms, and it is unclear how this concept is applied to CNN in the context; the "multi-head attention" part in Figure S1 doesn't seem to function like a transformer layer and it appears to be a simple multiplication of outputs from two network branches; the "CNN Encoder" term is also misleading - does this imply there's a decoder as well? I would suggest the authors revise Figure S1, Figure 1b, and the methods section to better explain the structure of their neural network.

Thank you for pointing out these issues. We realized that terms like "multi-head" and "CNN Encoder" could indeed cause confusion. Therefore, we modified the texts in the revision and improved Figure S1 to make the model structure clearer.

We revised the content in 'Translation Model Architecture' section "In parallel, convolutional neural networks (CNNs) processes the one-hot encoded codon sequence, generating embeddings with four distinct feature sets."

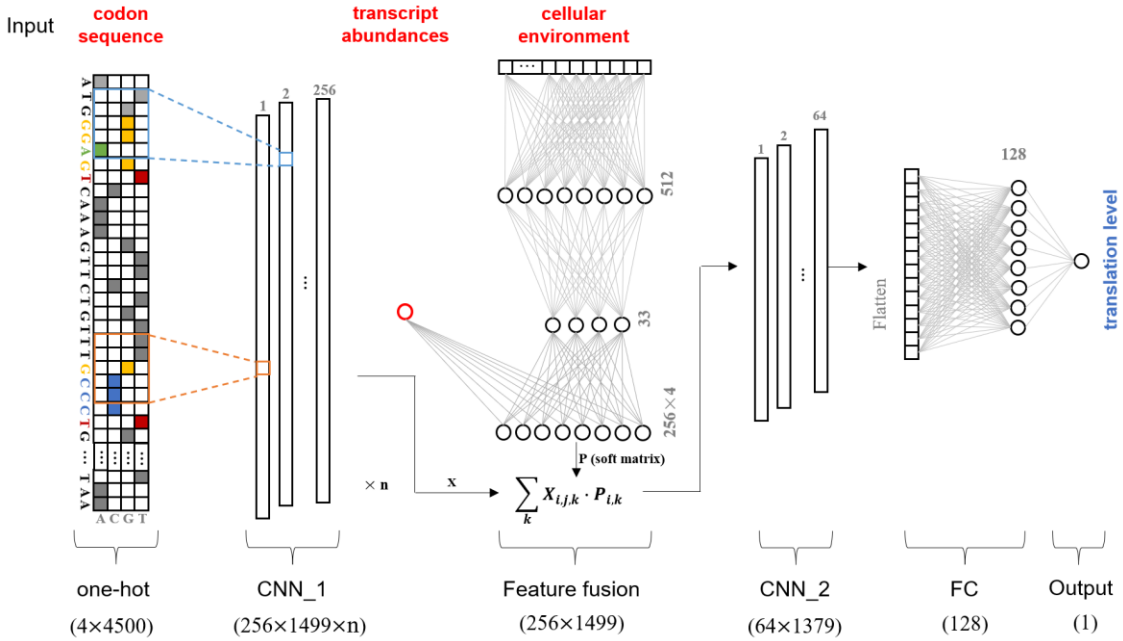


Figure S1. Schematic diagram of the translation model structure. The model accepts one-hot encoded nucleotide sequences as input, processed by convolutional neural networks (CNNs). A feature fusion

mechanism is dynamically established through fully connected layers (FC) that integrate codon sequences, cellular environment features, and transcript abundance data (see Methods for implementation details). The dimensional shapes of each structural component are annotated in parentheses below the diagram. The number of CNN_1 was set to 4 for our model.

- The software package should provide users with the flexibility to define the cellular environment, i.e., allow the optimization of codon sequences based on user-specified gene expression profile.

We have incorporated a user-defined environment input parameter into the software package, accompanied by corresponding instructions in README:
"To use a custom cellular environment for prediction, run the following command: *pred-translation --cds ATGGACGGGTAG --csv env_file.csv*
Here, 'csv' specifies the custom cellular environment file. We provide a standard template file, env_file.csv, with the following format: The first column contains human gene IDs, which must remain unchanged. The second column lists the corresponding mRNA RPKM values, provided by the user. Missing values can be replaced with 0.
To use a custom cellular environment for optimization, run the following command: *ribo-code --cds gluc --env custom --mfe_weight 0 --optim_epoch 10 --csv env_file.csv*
The parameters after 'env' can be defined manually to save the environment naming method of the generated sequence. 'csv' specifies the custom cellular environment file, same format as the csv file of the TranslationModel."

Minor comments:
- Figure 4 c/d, figure titles and captions do not match.

Thank you for pointing this out. We are sorry we mislabeled the captions in our previous figure. Per the suggestion from another reviewer, we have adjusted the presentation of the results by combing data from the three experiments for each cell line with error bars and displaying the two cell lines separately. This provides a more intuitive visualization.

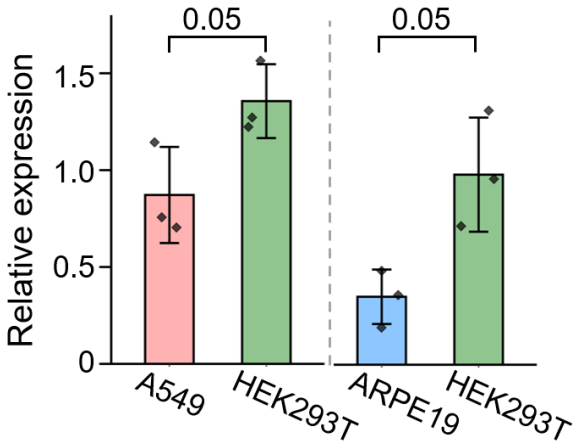
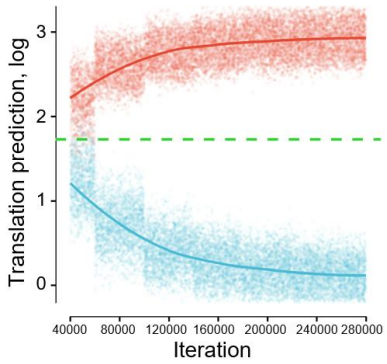


Figure 4b. Relative expression of experimentally measured protein expression values of mRNA variants designed by RiboDecode at 24 hours in different cells. The error bars denote standard variation. One-sided Wilcoxon test was used to calculate p-values shown in the figure.

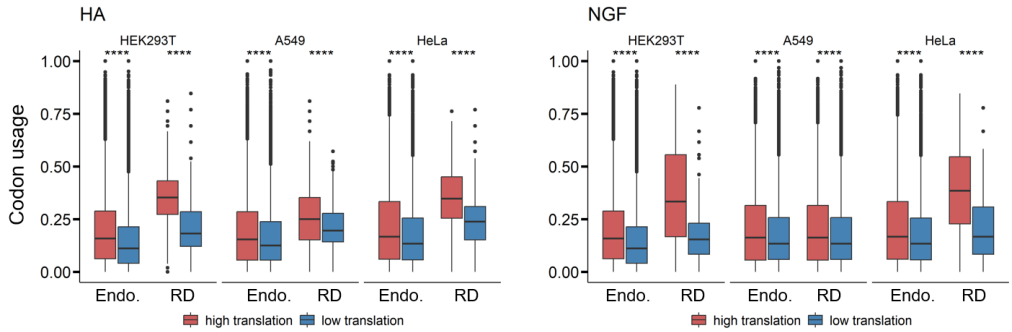
- Translation levels of Gluc sequences in Figure S4 seem much higher than those in Figure 3b. Are these plots accurate?

Thank you for point this out. We are sorry for the error. Upon careful verification, we confirm that the y-axis labeling in Figure S4 (now it is Figure S5) is incorrect. The translation prediction levels for the Gluc sequence were subjected to a natural logarithm (log) transformation, rather than a base-10 logarithm (log10) transformation. We added the content to the figure legend: The y-axis represents predicted translation level transformed to log(RPKM×5+1) (where the log denotes natural logarithm).



- The Y axis label is missing for Figure S5.

Thank you. We have added it.



- Table S5, "ARPE19" is misspelled as "AREP19".

Thank you. We have modified it.

- The name " RiboCode " has already been used in a prior work (https://github.com/xryanglab/RiboCode). I recommend the authors consider renaming their software to avoid any confusion.

Thanks for the suggestion. We changed the name to “RiboDecode”, which had not been used before.

Reviewer #4 (Remarks to the Author):

Peer Review Report for NCOMMS-24-67515-T:

The authors introduce RiboCode, a novel deep learning-based framework designed to optimize mRNA codon sequences to enhance protein production and therapeutic efficacy. Unlike traditional pre-defined metric-based codon optimization, they claim their deep learning-based method better learned complex relationships governing translation and explored a bigger sequence space. RiboCode leveraged large-scale ribosome profiling (Ribo-seq), cellular environment (e.g., RNA-seq), and RNA secondary structure prediction datasets to better capture the interplay between codon usage, cellular context, translation efficiency, and stability.
RiboCode’s sequence optimizer framework combined a trained translation prediction model, an MFE active learning model, and a codon optimizer that employed activation maximization to explore and

optimize mRNA sequences for high translation and/or stability. The authors validated its effectiveness through in vitro experiments and in vivo studies, demonstrating increased protein expression, and improvements in vaccine immunogenicity and therapeutic efficacy. Additionally, they highlighted its compatibility with chemical modifications and mRNA formats, including m1Ψ-modified and circular mRNAs which are highly relevant to the field of mRNA therapeutics.
This study underscored the advantages of using deep learning approaches to mRNA codon optimization. While the manuscript presents a novel and promising approach, the points raised below highlight significant gaps that need to be addressed to ensure clarity, rigor, and reliability. Overall, we recommend major revisions to improve the robustness of the findings and alignment with the claims made. With these revisions, the manuscript could make a valuable contribution to the field of mRNA therapeutics.

Major comments:

Major Comment #1: Additional in silico validation is needed to support the claims

1. Sequence Space Exploration: The manuscript emphasizes the exploration of sequence space as a key strength of RiboCode, but no direct comparison with LinearDesign or other existing tools (Ribotree (PMID: 34520542), CDSFold (PMID: 26589279)) are provided to demonstrate this claim. Including such comparisons is essential to validate the broader sequence exploration ability of RiboCode.

Thank you for your suggestion.
We added the following content to a new section in Methods, titled 'Comparative analysis of sequence space across design Methods': We generated 1,000 full-length Gluc codon sequences using Ribotree, LinearDesign, CDS-fold, and RiboDecode, respectively. High-dimensional sequence embeddings were extracted using CodonBERT, followed by t-SNE dimensionality reduction for visualization.

In addition, we added the following content to 'RiboDecode's Optimization Strategies for Enhanced mRNA Translation' section in Results: Moreover, RiboDecode-generated sequences spanned a broader embedding space compared to those produced by Ribotree, CDSfold, and LinearDesign (Figure S6, see Methods), suggesting enhanced sequence diversity.

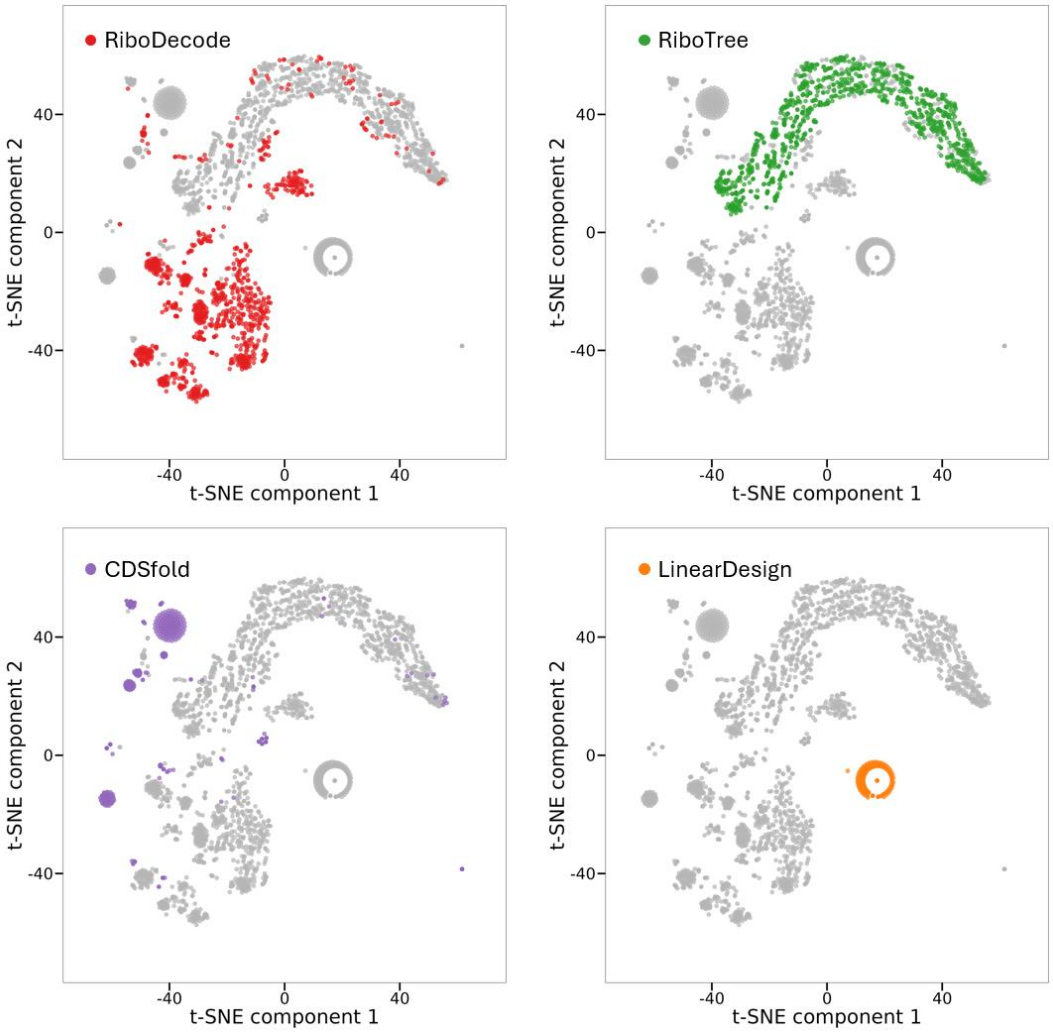


Figure S6. t-SNE distribution of Gluc codon sequences generated by RiboDecode, RiboTree, CDSfold, and LinearDesign (see Methods). The results show that the sequence space of RiboDecode is larger than other methods.

2. MFE Model Learning: The manuscript lacks evidence for how well the MFE model learns to predict MFE values. Results on unseen mRNAs (separated by families) and their associated metrics would provide confidence in the generalizability of the MFE model. A recent article from Szikszai highlights this issue (PMID: 35748706).

Thanks for the suggestion. We benchmarked our model to RNAfold and found that our the MFE value highly agreed with the ones from RNAfold, showing the reliability of our MFE model. We showed the result in the Methods section of the revision. We evaluated the MFE values of mRNAs between our model and RNAfold. We found that our MFE value highly agreed with the one from RNAfold (Figure S13), showing the reliability of our MFE model.

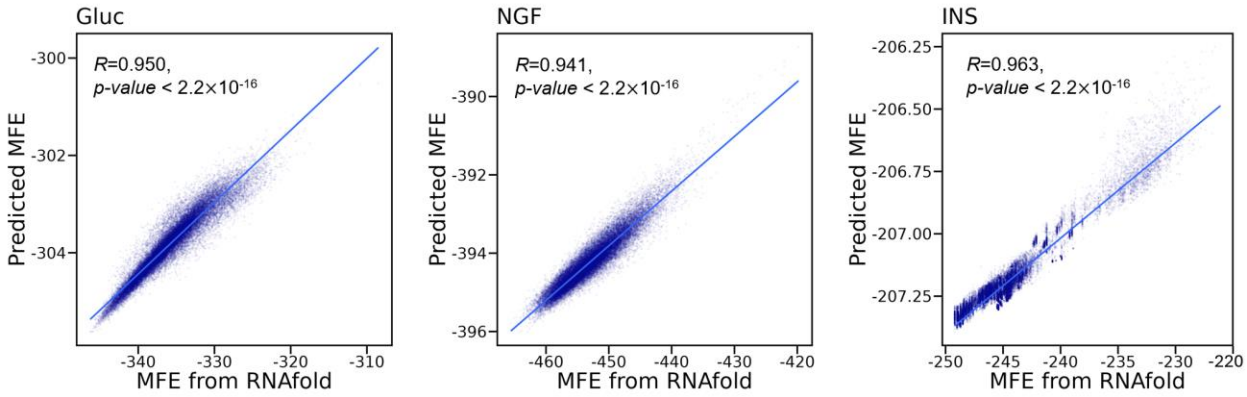


Figure S13. Evaluation of the Predictive Performance of the MFE Model to RNAfold. Pearson's correlation coefficients are provided. Blue lines denote the linear fit. All values are reported in kcal/mol.

For the second question, our model cannot predict MFEs for unseen mRNAs. For an unseen mRNA, our model must first train the sequences with MFEs labelled by RNAfold. Our requirement for the MFE model is that it be differentiable which will be compatible with our optimizer. By the time we started our study, there was no differentiable MFE optimization model. We therefore developed our own differentiable MFE model. As another reviewer suggested, we found that the recently developed a differentiable general MFE model (Paper name: “Scalable Differentiable Folding for mRNA Design”) also could efficiently optimize MFE. Once this model is publicly available (it is in biorxiv right now), we will comprehensively evaluate it and incorporate it into our integrative framework. We have discussed this limitation in Discussion in the revision.
Third, our MFE model cannot predict MFEs for unseen mRNAs. For an unseen mRNA, our model must first train the sequences with MFEs labelled by RNAfold. A general MFE model should be developed and used in future.

3. "New Genes" Definition: The term "new genes" in the study may be misleading if the sequences come from the same gene families or are not entirely novel. A clarification or additional analysis is necessary to validate this claim.
We are sorry for the misleading word. We meant genes which were not seen during training. We have renamed it as "unseen genes" in our revised manuscript.

4. Energy Landscape Visualization: Since a local optimizer is employed, the manuscript should analyze the energy landscape of different sequences. This would help demonstrate whether the optimizer avoids local minima and, if not, why this is acceptable in the given context.

Thank you for the suggestions. We agree that our model uses a gradient ascent optimization based on activation maximization (AM), which is a local optimizer. We provided several lines of evidence that our method effectively avoided local optima or it was acceptable in our case.

1. We present the translation optimization trajectory (Figure S4) and the sequence space of translation level & MFE (Figure 3a, 3b) for the Gluc codon sequences generated by RiboDecode. The results demonstrate that RiboDecode achieves highly optimized sequences, with the predicted translation level increasing from 5.9 to 25 and the MFE decreasing from -216 to -350 kcal/mol—approaching the lowest MFE achieved by LinearDesign, which is more like a global optimizer (-346.2 kcal/mol).

2. Per the suggestion, we now include the fitness score trajectory during optimization (w=0.5; Figure S3), which shows a consistent improvement in fitness score. Based on these results, we believe our method effectively avoids poor local optima and the optima it finds are sufficient and acceptable.

We added the fitness score trajectory to 'RiboDecode is a Deep Learning Framework for mRNA Codon Optimization' section in Results: Using a gradient ascent optimization approach based on activation maximization (AM), the optimizer adjusts the codon distribution to maximize the fitness score (Figure S3).

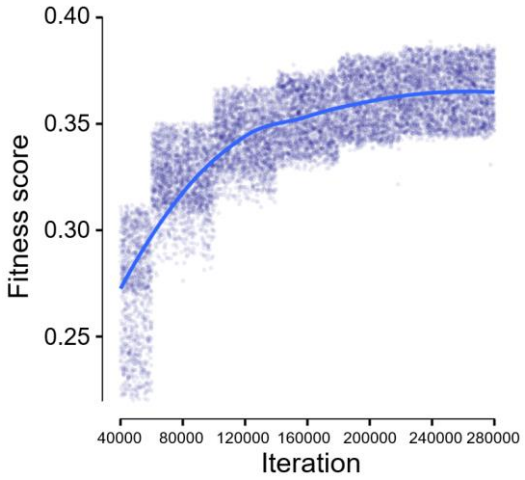
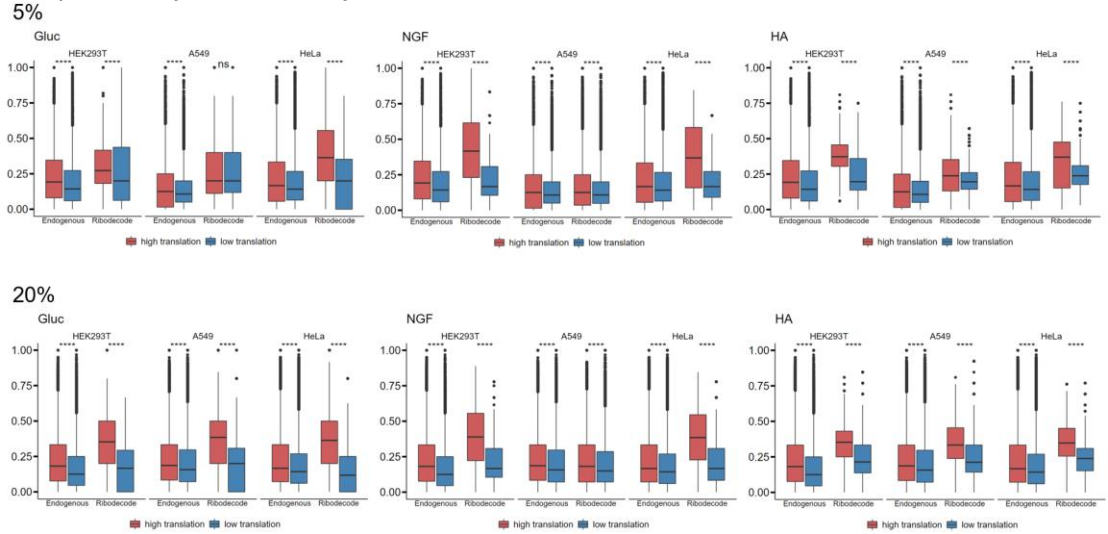


Figure S3. Fitness score trajectory during the iterative optimization of Gluc codon sequences (*w*=0.5). The fitness score is defined as the inverse of the loss function in our codon frame work (see Methods).

5. CAI Threshold (Fig 3c): The threshold of 10% for CAI calculation is arbitrary. Since the differences between "Endo" and "RD" are small but statistically significant, a smaller threshold (e.g., 1%) may make the result statistically not significant. The authors should explain the rationale or explore other thresholds.

1. We conducted the same analysis at thresholds of 5% and 20%. The results were consistent with those obtained at the 10% threshold, except Gluc in A549 cells at the 5% threshold. We have added the following content to 'Codon usage analysis' section in Methods: We further evaluated the impact of varying expression thresholds (top/bottom 5%, 10%, and 20%), with consistent results observed across these cutoff values, except Gluc in A549 cells at the 5% threshold.
2. A threshold of 1% would yield an inadequate sample size of approximately 20 genes, rendering the subsequent analysis statistically unreliable.



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

6. Model Extrapolation (Figs 2 & 3): Figure 2 shows a maximum predicted translation of 10, while Figure 3

reports values up to 25. This discrepancy suggests potential extrapolation by the model, which should be addressed and clarified.

The expression values in the original manuscript were log converted. We appreciate your attention to this and have accordingly revised the labels in figure 2a: …The translation levels from Ribo-seq were log-transformed (see Methods).

7. Ablation Studies and Error Bars: Error bars are missing in the ablation studies and other analyses (e.g., different dataset splits). Including error bars and biological replicates would enhance the reliability of the findings.

Thank you for pointing this out. We presented all the results for the cross-validation (Figure S4) and ablation analysis (Figure 2b):

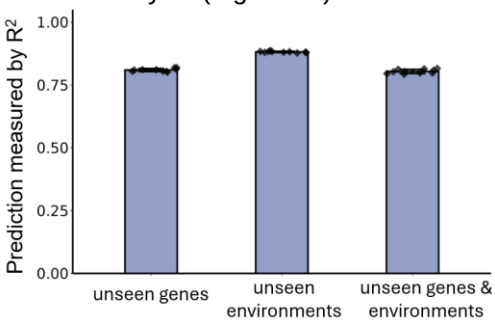


Figure S4b. 10-fold cross-validation on the test sets, where each dot on the bar represents a test. The error bars denote standard variation.

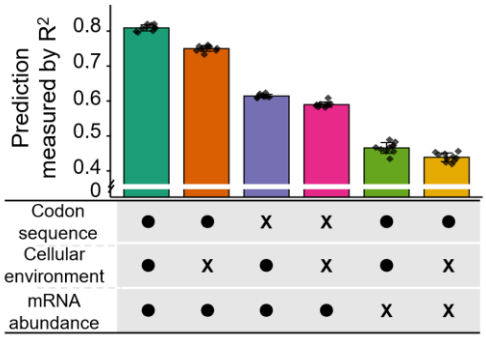


Figure 2b. …10-fold cross-validation for the ablation analyzes, where each dot on the bar represents a test. The error bars denote standard variation.

8. Cellular Context Justification: Cellular context contributes only marginally (R2=0.06), yet it is heavily justified as a critical component of the model. This justification should be toned down or revised to align with the results.

We agree with your view that the contribution of the cellular context in the model is marginal. We revised the phrases in the manuscript and toned down the justification.

1. Removed the underlined portion in the Abstract: "RiboDecode introduces several advances, including ... ~~context-aware mRNA optimization~~ ...
2. Revised the sentence in the Abstract from "RiboDecode achieves mRNA design in different cellular context" to RiboDecode enables mRNA design with consideration of cellular context.
3. Revised the sentence in the Introduction from "RiboDecode also achieved mRNA design in different cellular context" to RiboDecode also considered cellular context.

4. Added the following content to the 'Experimental Validation Demonstrates RiboDecode's Versatility and Efficacy' section in Results: … These results confirm RiboDecode's capability for context-aware design, while the discrepancy in the ARPE19 comparison indicates potential for refining the model to better capture quantitative differences across diverse cellular environments.

9. Key References Missing:
1. The authors have overlooked important contributions to the field, including Mauger et al., 2019 (PMID: 31712433) and Ribotree (PMID: 34520542). These references are foundational in mRNA translation and optimization research and should be cited with appropriate acknowledgment. The relevance of these methods (along with LinearDesign) for therapeutic mRNA design were reviewed in Metkar et al. (PMID: 38030688).
2. Other codon optimization algorithms that utilize deep learning Lines 115-117. LLM bases approaches like CodonBERT (PMID: 38951026), and ChatNT (https://doi.org/10.1101/2024.04.30.591835) have recently come out. These methods need to be acknowledged in the Introduction. While these methods do not leverage RiboSeq, the authors can point to this to indicate how their deep learning approach covers this aspect.
3. Other machine learning models for ribo-seq data: Line 129: There have been deep-learning approaches exploring Ribo-Seq data like RiboShape (PMID: 27307616), Riboformer (PMID: 38443396) and more recently RiboTIE (PMID: 38585907), RiboNN (https://doi.org/10.1101/2024.08.11.607362). The statement should be changed to acknowledge other methods so far. The authors should contrast their model and findings with a recent study from Can Cenik lab and Sanofi (RiboNN). The model uses a similar approach and also the UTR sequences in input with a much bigger compendium of datasets. While that study is not yet peer-reviewed, the authors can make their manuscript better by providing some insight on the differences between the other study and theirs to better inform the readers.

Thank you for your suggestions. We have included the following texts and these references in Introduction:

1. Mauger et al., 2019 (PMID: 31712433) and Ribotree: Previous studies have demonstrated that mRNA structure critically influences its stability *in vivo*[8] and *in solution*[9], as well as translation[17].

2. CodonBERT and ChatNT: Recent advances in codon optimization research have seen the emergence of deep learning-based algorithms, particularly large language models (LLMs) trained on cross-species nucleotide sequences. These models have been implemented for predictive modeling of mRNA translation efficiency and degradation kinetics[10,11]. Nevertheless, there persists an urgent demand for developing a rigorously validated optimization framework to improve mRNA-encoded protein expression specifically tailored for therapeutic applications.

3. RiboShape (PMID: 27307616), Riboformer (PMID: 38443396), RiboTIE (PMID: 38585907), and RiboNN：Recent studies have leveraged Ribo-seq to develop translation-focused deep learning models[12–14]. For instance, RiboNN predicts mRNA translation efficiency in mammalian cells by integrating mRNA sequences with ribosome profiling data, revealing translation-stability regulatory mechanisms[15]. However, the field critically requires a rational design framework that translates data-derived translational signatures into codon optimization strategies, enabling high-throughput exploration of sequence space and de novo generation of optimized mRNA constructs for therapeutic development.

Major Comment #2: The definition of 'translation level' is not clear and the relationship with mRNA abundance.

1. Definition of translation: The authors do not provide a clear definition of "translation level" in either the result or methods section. Is 'translation level' the RPKM from Ribo-Seq or is it 'translation efficiency' defined for each transcript, RPKM from Ribo-Seq divided by RPKM from paired RNA-Seq? authors need to explicitly specify the definition for better interpretability of the results.

Thanks for the comments and suggestions.

In our manuscript, the term "translation level" refers to "RPKM derived from Ribo-seq". We have added the following texts to the manuscript:

In Introduction: Ribosome profiling sequencing (Ribo-seq) is a powerful experimental technique that provides a snapshot of actively translating ribosomes on mRNA molecules, where the translation level of an mRNA can be derived from the reads per kilobase per million (RPKM) of Ribo-seq.

The first appearance in Results: T-distributed stochastic neighbor embedding (t-SNE) indicated that the model established an association between the sequence space and translation levels (similar to Ribo-seq-derived RPKM values).

Dependence on mRNA Abundance
The definition of "translation level," based on RPKM from both Ribo-Seq and paired RNA-Seq, inherently incorporates mRNA abundance as both an input and a component of the target variable. While this approach is valid in deep learning due to the ability to model non-linear relationships, it raises the risk of overemphasizing mRNA abundance at the expense of other features, particularly given that mRNA abundance already accounts for the highest R2 among individual features. The ablation analysis indicates that the model performs slightly better than mRNA abundance alone, with codon sequence and cellular environment contributing incrementally to R2 (0.76 vs. 0.6). However, the authors should explicitly acknowledge in the Results or Discussion that the high contribution of mRNA abundance may stem from its dual role as both an input and a component of the target variable. Additionally, it would be valuable for the authors to clarify whether the model has been rigorously tested to ensure it captures meaningful relationships beyond mapping mRNA abundance to translation levels.

2. Relative Contributions of Transcription vs. Translation: The R2 of 0.6 for transcription dominance suggests that the model primarily learns transcriptional rather than translational regulation. The authors should address how this impacts the utility of the method for optimizing mRNA therapeutics, where translational rules may play a more critical role.

Thanks for the comments. As these two questions are related, we would like to respond them together. We acknowledge that mRNA abundance serves as both an input feature and is intrinsically linked to the target variable (Ribo-seq RPKM). This relationship likely contributes to its high predictive importance observed in the ablation analysis (Figure 2b).

However, several lines of evidence indicate the model captures meaningful biological relationships beyond simply mapping abundance levels, rigorously testing its capabilities:

1. Incremental Contributions: The model's predictive performance (R2) significantly improves with the addition of codon sequence and cellular context information (increasing R^2 by 0.15 and 0.06 respectively over abundance alone), demonstrating it learns from these distinct inputs.
2. Learning Complex Features: Incorporating established metrics like CAI did not improve the model's predictive accuracy. This suggests RiboDecode identifies more complex, inherent sequence features governing translation, rather than relying solely on simpler, predefined rules or just mapping abundance.
3. Experimental Validation: Crucially, the model's ability to generate sequences yielding substantial, experimentally validated increases in protein expression *in vitro* and enhanced therapeutic effects *in vivo* provides strong evidence that it learns and optimizes functional, sequence-based translational signals. This functional validation confirms it captures meaningful relationships beyond simply reflecting transcript abundance.
4. Context Specificity: The model also demonstrated the ability to design sequences for preferential expression in specific cellular contexts, requiring it to learn context-dependent translational nuances.

In summary, while the dual role of mRNA abundance influences its predictive weight, the incremental improvements from other features, the capture of complex sequence patterns, and, most importantly, the robust experimental validation of optimized outputs confirm that RiboDecode effectively learns and utilizes sequence-intrinsic translational information beyond merely mapping transcript levels.

We have added texts to Discussion in the revision acknowledging this potential influence.
While mRNA abundance serves as both an input feature and intrinsically linked to the Ribo-seq RPKM target variable, potentially explaining its high predictive weight in ablation analysis, the model still demonstrated significant improvements in prediction accuracy by integrating codon sequences and cellular context. Importantly, RiboDecode's in vitro and in vivo validation, where optimized sequences substantially enhanced protein expression and therapeutic efficacy, confirms its ability to optimize biologically meaningful translational signals rather than merely reflecting transcript abundance.

Further, we recommend that the authors discuss their results in the context of the findings by Schwanhäusser et al. (PMID: 21593866), which demonstrated that post-transcriptional regulation at the level of translation played an equal or greater role in determining protein expression compared to transcriptional regulation. The current study, in contrast, emphasizes the predictive power of transcription-level measures (mRNA abundance). The authors should explore how this discrepancy impacts the interpretation of their results and the relative contributions of transcription and translation to protein expression. A deeper discussion of these points would significantly enhance the manuscript's clarity and broader relevance.

Thank you for this suggestion regarding Schwanhäusser et al.. We agree their work highlighted the major role of translation-level control in determining protein expression.

Our finding that mRNA abundance is the *strongest predictor* in our model reflects its prerequisite role as the template molecule, rather than diminishing the importance of translation efficiency itself. RiboDecode *does* explicitly consider translational control through its codon sequence and cellular environment inputs, trained on Ribo-seq data reflecting active translation. It does not, however, model post-translational regulation, as its predictions are based on ribosome occupancy prior to subsequent protein modifications or degradation.

Our experimental results, showing significantly enhanced protein expression and *in vivo* efficacy through optimizing these mRNA features, strongly validate the power of targeting the translational elements our model considers. Thus, our findings complement the Schwanhäusser study by demonstrating an effective method to leverage sequence-based translational control for enhancing protein production. We have added discussion points reflecting this context to the manuscript.

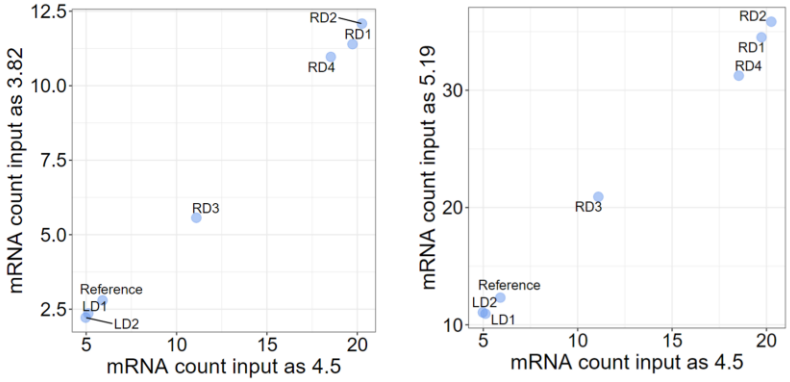
These findings also align with the broader framework of gene expression regulation: while mRNA abundance is a prerequisite for protein synthesis (consistent with its predictive dominance), our approach directly targets translational control, a critical layer highlighted by Schwanhäusser et al.43, by leveraging Ribo-seq-derived codon usage and cellular context to maximize translational efficiency. Although the model excludes post-translational events (as Ribo-seq captures ribosome activity prior to protein maturation), the achieved gains in protein output and in vivo efficacy robustly validate translational optimization as a key determinant of functional protein levels, complementing established regulatory hierarchies.

3. Effect of mRNA abundance input for codon optimization: mRNA abundance is an input parameter for the translation prediction model and for every step of the codon optimizer, a constant mRNA abundance was set to 4.5 (line 647). Have the authors tested the performance of the codon optimizer by varying this input?

Thank you for this suggestion. We evaluated the effect of varying the mRNA abundance input on the translation prediction model, which informs the codon optimizer. As described in the Methods section, the input mRNA abundance is based on a log transformation ($y=\ln(RPKM\times5+1)$). Our default input value of $y=4.5$ corresponds to an RPKM of approximately 17.8.

Following the reviewer's suggestion, we tested the model using alternative abundance inputs corresponding to half (RPKM=8.9, $y\approx3.82$) and double (RPKM=35.6, $y\approx5.19$) the default RPKM value. We found that the predicted translation levels for sequences, when calculated using these different abundance inputs ($y=3.82$, 4.5, or 5.19), were highly correlated with each other. This indicates that while

the absolute predicted translation score shifts with the abundance input, the relative scores between different sequences are largely preserved within this tested range. Since the codon optimizer primarily relies on these relative differences to guide sequence generation towards higher predicted translation, our results suggest that the optimizer's performance and the resulting optimized sequences are robust to reasonable variations in the input mRNA abundance value.



Comparation of translation prediction with different mRNA abundance (Pearson's correlation)

4. Missing 5'UTR Context: Translation initiation is a limiting step, yet the model appears to lack information about the 5'UTR. If this is not incorporated, we think, its potential impact on translation prediction needs more discussion

Thanks for the suggestion, we added this in Discussion: " Firstly, we focused exclusively on optimizing the codon sequences while not explicitly modeling the 5'UTR. Recognizing the critical role of the 5'UTR in regulating translation initiation, our future work aims to expand RiboDecode to jointly optimize both UTRs and the codon sequence."

5. Importance Plot and Padding: The per-position importance plot may be biased due to 3' padding, as low importance at the 3' end could be an artifact. Additional analyses are needed to account for this.

For the results in Figure 2c, we calculated the importance scores based on the non-padding regions. We added the following content to the 'Nucleotide position contribution for translation prediction' in Methods: For the final analytical outcomes and visualization, the importance score assigned to each nucleotide position was calculated as the mean of absolute values derived from non-padding regions.

Major Comment #3: Incorporating MFE Optimization in the Discussion and Results

Some of the RD sequences designed for experimental validation incorporate an MFE optimization component (w>0), which is not thoroughly discussed in the manuscript. According to the methods (line 688), the weights for RD1, RD2, RD3, and RD4 are w=0,0,0.7,0.5, respectively, indicating that RD3 and RD4 were optimized for both translation level and lower MFE. This design choice warrants further clarification and discussion.

Specific Recommendations:
1. Explicit Labels for Weights (w) in Figures: Figures that compare RD sequences, particularly Figures 4, 5, and 6, should include explicit labels indicating the translation level and MFE weights (w) used for each sequence. This would improve the clarity and interpretability of the data.

Thank you for your suggestion. We have included the w parameters to the sequences in Figures 4, 5, and 6. For example:

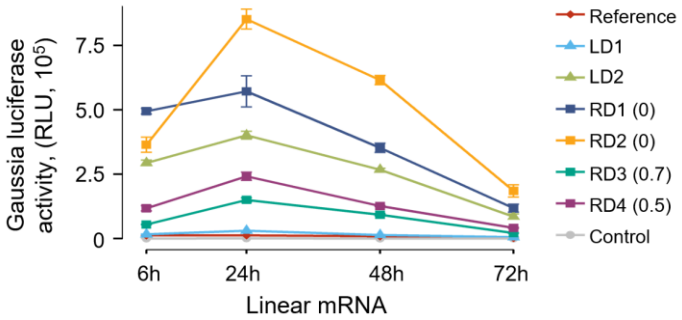


Figure 4a. Protein expression of Gluc was measured by fluorescence intensity. RD sequences were designed by RiboDecode, with w parameter indicated in parentheses…

In addition, we included the translation level and MFE weights (w) of Gluc, Fluc, HA and NGF in Table S3, S6, S7 and S8.

2. Consistent Nomenclature Across RD Sequences: The order and naming of RD sequences across Figures 4, 5, and 6 are inconsistent. For example, the order of RD sequences based on w for HA differs from that of Gluc and NGF. Standardizing the nomenclature based on the weights would make data interpretation more intuitive and consistent.

We have renamed the HA sequences to make their name and w consistent with that of Gluc and NGF. For example:

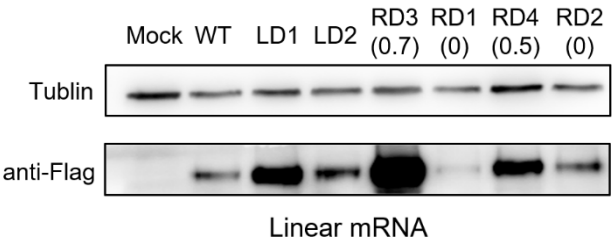


Figure 5a. Western blot analysis shows protein expression of the HA variants in HEK293T cells 24 hours after transfection. RD sequences were designed by RiboDecode, with w parameter indicated in parentheses. LD sequences were designed by LinearDesign.

3. Discussion of Results Based on Different w Values: The manuscript should explicitly discuss the varying performance of RD1 and RD2 (designed for translation only) compared to RD3 and RD4 (designed for both translation and stability). This is particularly relevant for circular RNAs, where RD3 and RD4 outperform RD1 and RD2 in protein expression. The authors should explore and discuss how optimizing for MFE contributes to higher protein expression in circular RNA contexts.

The observation that optimized codon sequences with lower MFE exhibit higher protein expression in circular RNA format aligns with a recent study (doi: https://doi.org/10.1101/2023.07.09.548293). The authors developed an optimization strategy for circRNAs and found that circRNAs with lower MFE demonstrated enhanced protein expression *in vivo*. The results suggested that this was likely because codon sequences with lower MFE facilitate the proper folding of the internal ribosome entry site (IRES), thereby enabling more efficient ribosome recruitment and translation. However, as we only experimentally tested a few numbers of circular RNAs in our study, we cannot validate the generalization of this statement. We discussed our recommended strategy of choosing *w* in Discussion. Please see the answer in the next question for details.

4. Contrasting MFE and Translation-Only Models: While the Discussion notes that MFE did not have a general impact on translation, it fails to emphasize that MFE optimization significantly contributed to the performance of top-performing sequences in experimental validation. The authors should address this

contrast and discuss how MFE-optimized sequences outperformed translation-only models in specific contexts.

Thanks for the suggestion, we added this in Discussion:
Second, our results indicate that while our primary goal was to enhance translation through codon optimization, incorporating MFE optimization played a complementary role in modulating mRNA secondary structure stability. However, we observed the best w value appears to vary depending on the specific mRNA and its context, suggesting that a one-size-fits-all approach may not be appropriate. Consequently, we recommend that experimental designs include the testing of multiple w values to identify the optimal balance for each mRNA target. Further research into the determinants of the optimal w value will be critical to refine this strategy and could lead to more systematic approaches for integrating MFE into mRNA design.

Major Comment #4: Experimental Validation: mRNA variants, cellular context, and different mRNA formats.

1. Luciferase Feature Analysis (Fig 3d): The two luciferase proteins exhibit opposite CAI importance. This variability requires further explanation, especially in relation to their sequence features.

The observed difference in CAI trends between Gluc and Fluc underscores this context-dependence. Potential contributing factors could include differences in their specific amino acid compositions, overall sequence lengths, or resulting structural elements, which interact differently with the cellular translation machinery and influence which codons are optimal in each specific case. Furthermore, we found that CAI, as well as CPB, GC% and MFE, influenced translation but be more mRNA specific. It may be because of sequence or structure feature changes.
In the “RiboDecode’s Optimization Strategies for Enhanced mRNA Translation” section of Results, we discussed this: 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, due to complex sequence or structure feature changes.

2. Cellular Context Claims (Fig 4): The difference between the predicted and experimental ratios for the HEK293 vs other cellular contexts is stark. This undermines the claim that the model learns cellular context. The authors should refine this claim or redesign sequences to better reflect cellular differences. Second, a barplot with error bars would be a better representation for this data.

Thanks for the suggestions. Regarding the discrepancy between the experimental and predicted ratios. We revised the texts in the revision: We next evaluated RiboDecode’s ability to design for cellular context by optimizing Gluc mRNA for preferential expression in HEK293T cells over A549 and ARPE19 cells. The designed variants successfully exhibited the intended higher expression in HEK293T compared to both cell lines (Figure 4b, Table S4). Predicted expression ratios favoring HEK293T closely matched experimental results in the comparison against A549 (~1.7-fold predicted vs. ~1.55-fold experimental). Preferential expression was also achieved against ARPE19 (Figure 4b, Table S5), although the experimental fold-change (~2.8-fold) doubled the prediction (~1.4-fold). These results confirm RiboDecode's capability for context-aware design, while the discrepancy in the ARPE19 comparison indicates potential for refining the model to better capture quantitative differences across diverse cellular environments.

In addition, we have toned down the cellular context claims in multiple places in the manuscript (in response to the Major comments #1, Question 8).

Also, per the suggestion, we included error bars in the plots. In addition, we combined the three experiments in each cell line and showed two cell lines separately, which gives a more intuitive visualization. We provided the predicted values and discussion on the texts as above.

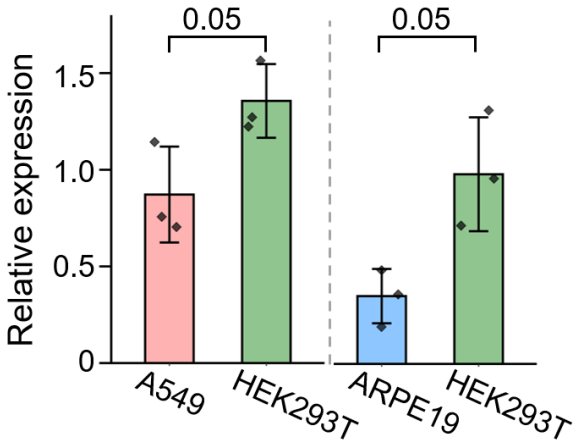


Figure 4b. Relative expression of experimentally measured protein expression values of mRNA variants designed by RiboDecode at 24 hours in different cells. The error bars denote standard variation. One-sided Wilcoxon test was used to calculate p-values shown in the figure.

3. Performance on Modified Chemistries: With m1Ψ-modified sequences, the performance drops to a 4.6-fold improvement from 72-fold. This suggests that the model does not effectively capture rules for modified chemistries and should be discussed.

Thank you for this observation regarding the performance with m1Ψ-modified sequences.

We acknowledge that the fold-improvement for the optimized Gluc sequence compared to its reference was lower when using m1Ψ modification (up to 4.6-fold) compared to the unmodified linear version (up to 72-fold). There are two primary factors contributing to this difference in *relative* improvement:

1. Higher Baseline Expression of the m1Ψ Reference: As shown in Figure 4a and 4e, the m1Ψ modification itself significantly boosted the baseline protein expression of the Gluc reference sequence compared to the unmodified version. Since the starting expression level for the m1Ψ reference was already considerably higher, the subsequent fold-improvement achieved through RiboCode optimization appears smaller in relative terms.
2. Model Training Data: The RiboCode model was trained on large-scale Ribo-seq datasets derived from endogenous, *unmodified* mRNAs. Therefore, the sequence rules it learned primarily reflect the translation dynamics associated with standard nucleotides and do not explicitly capture rules potentially unique to m1Ψ incorporation, as well as circular RNAs.

We agree that the specific interplay with modified chemistries warrants further investigation, as mentioned in the discussion.

We acknowledged this limitation in Discussion: Finally, the current model was trained exclusively on Ribo-seq data from endogenous, unmodified mRNAs. Although our results showed significant expression enhancements for optimized sequences in both m1Ψ-modified and circular forms compared to their respective controls, the relative fold-improvement compared to the unmodified format varied between constructs. Future iterations could potentially incorporate data from modified and circular transcripts to further refine optimization rules.

4. Circular mRNA in Fig 6: Circular mRNAs perform better than linear in Figure 6 but not in previous figures. This discrepancy should be explained.

Thank you for this comment regarding the performance of circular mRNAs. Our data consistently show optimization improves expression *within* the circular format for Gluc, HA, and NGF compared to controls. However, the figures do not provide a systematic comparison of *relative* performance between linear and circular formats across these different genes, making it difficult to confirm the suggested discrepancy. Any variations in how circularization impacts expression relative to linear formats could potentially stem from

gene-specific factors, such as differences in inherent mRNA stability, translation efficiency in the circular form, or interactions with the cellular machinery. Currently, our data do not support definitive conclusions about discrepancies in the relative performance of circular vs. linear formats across genes. This aspect is acknowledged in the Discussion (please see above question for this discussion).

Major Comment #5: Missing/additional controls

1. Reference Sequence CAI: The reference Gluc sequence has a CAI of 0.8, Mauger et al., 2019 showed to have very low expression. This may undermine the claim of a 72-fold improvement. A higher CAI (~1.0) control should be included.

Our original reference Gluc had a CAI of 0.8. We found that the Gluc-LD2, designed by LinearDesign, exhibits a high CAI (0.97) but demonstrated lower expression levels compared to our designed sequences, RD1 and RD2. This indicates that a high CAI was not necessarily correlated with high expression. However, as the reviewer suggested, to avoid potential overinterpretation, we did not specify 72-fold change in expression in the revised manuscript, from “72-fold increase” to “significant increase”.

2. Fluc Performance (Fig S8): The 17-fold improvement for Fluc (vs. 72-fold for Gluc) suggests the model's translation rules are not well-generalized. Sequence feature comparisons could be performed between the RD and reference sequences for these 2 related proteins to explain this discrepancy. Missing CAI control here as well.

Thank you for the comments regarding the Firefly Luciferase (Fluc) performance and comparison to Gaussia Luciferase (Gluc).

1. Differing Fold-Improvement & Generalization: We acknowledge the difference in maximum fold-improvement observed between Fluc and Gluc. We noted the magnitude of fold-improvement can be highly dependent on the specific wild-type sequence's baseline expression and intrinsic properties. Given that RiboDecode performed well across multiple distinct proteins (Gluc, Fluc, HA, NGF), we believe the learned rules show reasonable generalization, although the achievable fold-increase varies between targets. However, As previously addressed, we have recognized that the 72-fold augmentation in Gluc expression may be attributable to the diminished expression of the reference sequence, and as such, we will refrain from highlighting this number further.

2. Sequence Feature Comparison: Figure 3d in the manuscript provides a high-level comparison of feature changes upon optimization for several proteins, including Fluc. As discussed in the manuscript, the optimal balance of features like CAI, MFE, U%, etc., appears highly gene-specific. The contrasting results between Gluc and Fluc underscore this context-dependency, suggesting a simple feature comparison might not fully capture the complex interplay determining translation efficiency for each specific sequence.

3. CAI Control & Additional Fluc Data: We appreciate the comment regarding controls and have performed the suggested additional experiments comparing WT Fluc with sequences optimized by both RiboDecode (RD1-RD4) and LinearDesign (LD1, LD2), with results shown in the figure you provided.
    - All optimized sequences significantly increased Fluc activity compared to WT.
    - The LinearDesign sequence LD2 (high CAI = 0.952) yielded the highest expression, achieving approximately a 41-fold increase over WT at 24h. LD1 (moderate CAI = 0.766) showed around a 7-fold improvement.
    - RiboDecode sequences (with lower CAIs ~0.71-0.73) achieved substantial improvements ranging from 6-fold to 16-fold over WT (e.g., RD4 achieved ~16-fold).
    - These results indicate that for Fluc, high CAI correlates well with high expression. However, RiboDecode, optimizing for complex features learned from Ribo-seq data rather than primarily maximizing CAI, still yielded significant (up to 16-fold) improvements. This contrasts

sharply with our Gluc results, where high CAI correlated negatively with expression and RiboDecode significantly outperformed LinearDesign.

Conclusion: The additional experiments confirm RiboDecode effectiveness for Fluc, while the differing optimal strategies (high CAI best for Fluc, RiboDecode complex optimization best for Gluc) strongly underscore the gene-specific nature of mRNA optimization. This highlights the value of context-dependent, data-driven approaches like RiboDecode that can capture complex sequence features beyond single metrics like CAI.

We have updated the manuscript reflecting the new experiments in Results:

We additionally optimized another commonly used reporter gene, firefly luciferase (Fluc). Seven sequences, including four RiboDecode-optimized (RD1-RD4), two LinearDesign-optimized (LD1, LD2) and a WT, were transfected into HEK293T cells, and activity was measured over 72 hours (Figure S12, Table S6). All optimized sequences significantly outperformed the WT. The LD1 (CAI of 0.766) and LD2 (CAI=0.952) yielded the increased expression with about 7-and 41-fold changes, respectively. RiboDecode sequences (CAI of ~0.71-0.73) also achieved substantial improvements with 6- to 16-fold over the WT. The superior performance of the LD2 sequence with a high CAI value here—contrasting with results for Gluc—underscores that the optimal codon strategy is gene-specific. This highlights the need for context-dependent approaches like RiboDecode that capture complex features beyond single metrics like CAI.

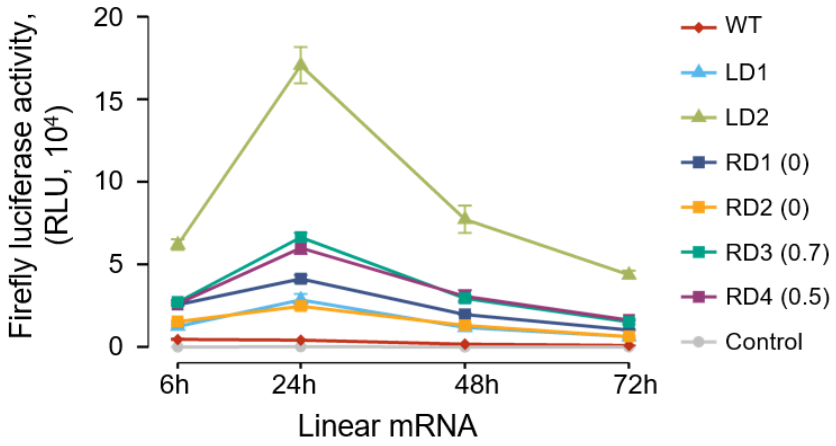


Figure S12. Fluc protein expression in transfected HEK293T cells. Protein expression was determined by fluorescence intensity. RD1, RD2, RD3, and RD4 were designed using RiboDecode, with the *w* parameters indicated in brackets. The detailed information of the sequences is provided in Table S6. "RLU": relative light unit.

Gene	Sequence ID	Parameter	MFE	CAI
Fluc	Reference	NA	-216.4	0.808
Fluc	RD1	w = 0	-195.7	0.713
Fluc	RD2	w = 0	-195.3	0.703
Fluc	RD3	w = 0.7	-342.4	0.748
Fluc	RD4	w = 0.5	-258.4	0.736
Fluc	LD1	λ=0	-346.2	0.766
Fluc	LD2	λ=4	-302.5	0.952

Table S6. This table presents sequence features for various Fluc mRNA constructs, including the wild-type sequence, RiboDecode-generated variants (RD1-RD4 with different w values), and LinearDesign-generated variants (LD1-LD2).

3. Dose Response (Fig 5): Unlike Figure 6, Figure 5 lacks a dose-response experiment. Including this would strengthen the conclusions.

Thanks for the comments. You are correct that Figure 5 does not include a dose-response experiment for the HA vaccine, whereas Figure 6 presents data for NGF at two different doses.

The rationale for this difference lies in the distinct goals and biological contexts of each application:

1. NGF Therapeutic (Figure 6): The goal was *local protein replacement therapy* in the retina. NGF exerts its therapeutic effect directly within the retina where it is expressed following intravitreal injection. Therefore, demonstrating that optimized mRNA could achieve a comparable therapeutic effect (neuroprotection) with lower *local* protein levels (resulting from a lower dose) was essential to show enhanced potency and dose-sparing potential in the target tissue.
2. HA Vaccine (Figure 5): The goal of the vaccine, administered intramuscularly, is to generate a *systemic immune response*. While the HA antigen is expressed locally (primarily in muscle and draining lymph nodes) to initiate this response, the desired outcome (protective immunity) is systemic and measured via antibody titers in the blood. High local concentration of the antigen is not the therapeutic endpoint itself. For this proof-of-concept study, demonstrating that RiboDecode optimization significantly boosted the systemic immune response (~10-fold higher neutralizing antibodies) at a single, relevant dose (10 µg) was sufficient to validate the optimization strategy's effectiveness for enhancing immunogenicity. This 10µg dose is a relevant dose level used within the range reported for preclinical influenza mRNA vaccine studies in mice (PMID: 30135514, 32164372). We have added these references into the manuscript and changed the text in Methods: BALB/c mice received two intramuscular doses of 10 µg mRNA each, with the dose determined by previous studies[73,74], administered on day 0 (prime) and day 14 (boost).

Therefore, the experiments were designed differently to best address the primary question for each application (dose-sparing potential for NGF therapeutic vs. enhanced immunogenicity for HA vaccine) within the scope of this initial validation study.

4. Missing Controls (Fig. 5 and 6): No codon-optimized (e.g., high CAI) sequences were used as reference. As Kudla et al., (PMID: 16700628) have demonstrated that WT sequences with low GC3% usually do not express well.

Thank you for this comment. We would like to clarify the sequences included as controls in Figures 5 and 6. In Figure 5, the HA-LD2 sequence had a CAI of 0.96 and GC3% of 88%. In Figure 6, the NGF-LD2 sequence had a CAI of 0.95 and GC3% of 82%. These sequences can effectively serve as the requested 'codon-optimized. These LD2 sequences, along with our RiboDecode sequences, consistently outperformed the wild-type (WT) sequences, which likely aligns with the expectation from Kudla et al. that WT sequences with potentially lower GC3% might exhibit lower expression.

Minor Comments

Minor Comment #1: Line 74-76. MFE is discussed only in relation to mRNA stability. Lower MFE/higher mRNA structure can also be a hurdle for translation elongation and needs to be acknowledged here.

Thank you for your reminder. We emphasized that MFE also affects translation in Discussion: …incorporating MFE optimization played a complementary role in modulating mRNA secondary structure stability and translation.

Minor Comment #2: Lines 91-95. The authors can be more specific and provide some examples of any trans-factors that affect translation that may not be captured by traditional methods.

Translation factors (for example, eIF4F and eIF2), RNA-binding proteins, tRNAs, and ribosomal RNAs could regulate mRNA translation (PMID: 34341537). However, predefined sequence features cannot capture the expression profiles of these factors in cells or their interactions with mRNA sequences.
We added the following content to the Introduction: Secondly, the existing methods do not adequately account for the activity of translational regulators that influence mRNA translation, such as translation factors and RNA-binding proteins.

Minor Comment #3: Predicted translation values below zero in Figure S4 suggest a model error. The authors should ensure that predictions are constrained to valid ranges.

The y-axis in Figure S4 represents the predicted values after log transformation, which is why negative values are present.

Minor Comment #4: The r=0.7 correlation in Figure S7 is derived from two distinct groups of points, making it a poor representation of the data. Alternative visualizations or metrics should be considered.

We acknowledge the reviewer's point that the data points in the scatter plot comparing RiboDecode's translation prediction versus experimental expression (Figure S7, left panel) appear somewhat clustered. We also agree that Pearson's correlation coefficient (R), while standard, can be influenced by such structures.

We acknowledged the shortcomings of p-value in Results: The predicted translational levels showed a positive correlation with the experimentally measured protein levels (Correlation coefficient=0.71, p-value=0.077, Pearson's correlation, Figures 4b and S7). However, the p value is marginal, potentially due to the small number of constructs tested.
In addition, a purpose of Figure S7 was to directly compare the predictive performance of RiboDecode against a traditional metric, CAI, using the same experimental data. The figure visually demonstrates that RiboDecode's predictions exhibit a positive trend with experimental expression (R=0.71, p=0.077), whereas CAI shows essentially no correlation (R=-0.072, p=0.88) for these sequences. This comparison highlights the improved predictive capability of our data-driven approach over the CAI metric in this context.

Minor Comment #5: Figure 5a requires bar plots quantifying the Western blot signals for a more accurate comparison.

Thanks for the suggestion. We have added a bar plot and text in the manuscript: Three out of four RiboDecode-optimized HA sequences and two LinearDesign-optimized sequences showed higher in vitro protein expression compared to the WT (Figures 5a, Table S7). Particularly, RD3 showed approximate 5-fold increase compared to the WT and LinearDesign-optimized sequences (Figure S15).

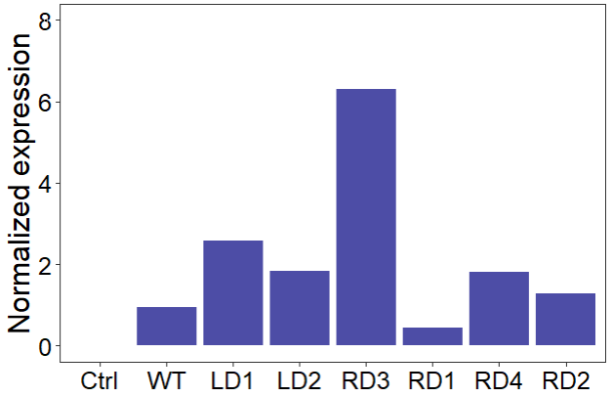


Figure S15. The gel bands of the Western blot in Figure 5a were quantified using GelAnalyzer. The expression values were normalized by Tublin.

Minor Comment #6: Lines 1050-1051, 1065-1066, 1081-1082. The w values for the RD sequences need to be shown either in the figure legends or at least in the figure captions.

Thank you for your suggestion. We have assigned the w parameters to the sequences in the Figures.

Minor Comment #7: Fix grammatical errors –

1. Line 386: …to several factors.
2. Line 392: …mRNA translation whereas the previously codon optimization…

Thank you for pointing out these issues and we have corrected them.

Reviewer #5 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #5 (Remarks on code availability):

The code is accessible as a .whl file. The code can run the model with given parameters. I did encounter an issue with installation that I state below.

The instructions for installation should be more clear in README file. The run.py has a 'python' command. It is increasingly common for modern Linux distributions and macOS to have only python3 installed and not python. The following error was encountered which needed some troubleshooting.

Traceback (most recent call last):
File "~/.local/bin/ribo-code", line 8, in <module>
sys.exit(main())
File "~/.local/lib/python3.8/site-packages/ RiboCode/run.py", line 24, in main
result = subprocess.run(command)
File "/usr/lib/python3.8/subprocess.py", line 493, in run
*with Popen(*popenargs, **kwargs) as process:*
File "/usr/lib/python3.8/subprocess.py", line 858, in __init__
self._execute_child(args, executable, preexec_fn, close_fds,
File "/usr/lib/python3.8/subprocess.py", line 1704, in _execute_child
raise child_exception_type(errno_num, err_msg, err_filename)
FileNotFoundError: [Errno 2] No such file or directory: 'python'

I had to specify sudo ln -s /usr/bin/python3 /usr/bin/python or use alias python='python3' in my ~/.bashrc to avoid this error.

The authors should either replace python with python3 in their run.py or provide more explicit instructions.

Thank you for your feedback. We have replaced ‘python’ with ‘python3’ and uploaded the package.

Reviewer #6 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):
While the authors have addressed some of our previous concerns, we still have reservations regarding the training and validation of their models:

We appreciate the reviewer's insightful comments. To better address the concerns regarding the MFE model, we have chosen to first respond to the third question.

3. Finally, the authors' response to our previous comment regarding the evaluation of their model against RNAfold (Figure S13) remains inadequate (our previous point 3). The sequence diversity used in Figure S13 is still insufficient, being limited to sequences synonymous with only three seed sequences. Without knowing the seed sequence(s) used for training the MFE model, it is difficult to assess its generalizability based on these tests. In addition, the gradient of the line of best fit is significantly different between the three plots (0.12, 0.16, 0.03). By separating the data into three plots, the R-value is artificially raised. The authors could rectify these issues by testing their model on a wider diversity of sequences and providing the R-value for the predictions across all sequences.

We thank the reviewer for this comment and apologize for not explaining our MFE model with sufficient clarity. Our MFE model is not a general-purpose predictor; it is trained independently for each individual gene being optimized. Therefore, the results presented in Figure S13 show three separate MFE prediction models, each trained on a different mRNA. For this reason, displaying the results independently is the appropriate approach, and combining them would be statistically invalid.

We have addressed this limitation in the revised Discussion section, stating: "Third, our MFE model cannot predict MFEs for unseen mRNAs. For an unseen mRNA, the model must first train the sequences with MFEs labelled by RNAfold. A general MFE model should be developed and used in future."

1. One concern remains the imbalance between the amount of training data and the number of parameters in both the translation and MFE models (our prior point 9). The data/parameter ratios, 0.12 for the translation model and 0.02 for the MFE model, are far below the generally accepted minimum for robust training (much greater than 1), raising doubts about the generalizability of the models and the validity of the chosen hyperparameters. We recommend the authors explicitly address this issue of undertraining in the main text.

We thank the reviewer for raising this important point about the data-to-parameter ratios and the potential for overfitting. We have now explicitly addressed this concern in the revised manuscript by adding detailed explanations and new analyses for both models.

For the translation model, we have clarified how the model architecture inflates the parameter count and how our dropout strategy effectively mitigates the risk of overfitting. We have added the following explanation to the Methods section:
"The translation model's architecture, which uses a fixed 4,500 bp input, results in a high nominal parameter count. This is an artifact of flattening the zero-padded input sequences required for the majority of genes, which have a median length of only 1,200 nt. To mitigate the resulting risk of overfitting, we implemented a strong dropout regularization strategy (rate of 0.9) before the final layers. This high dropout rate forces the model to learn robust features rather than fitting to noise in the padded regions. A comparative study (Figure S15) confirms this strategy is essential: models with no or low dropout overfit immediately, whereas the 0.9 dropout rate eliminates overfitting and ensures stable, improving performance on all validation sets."

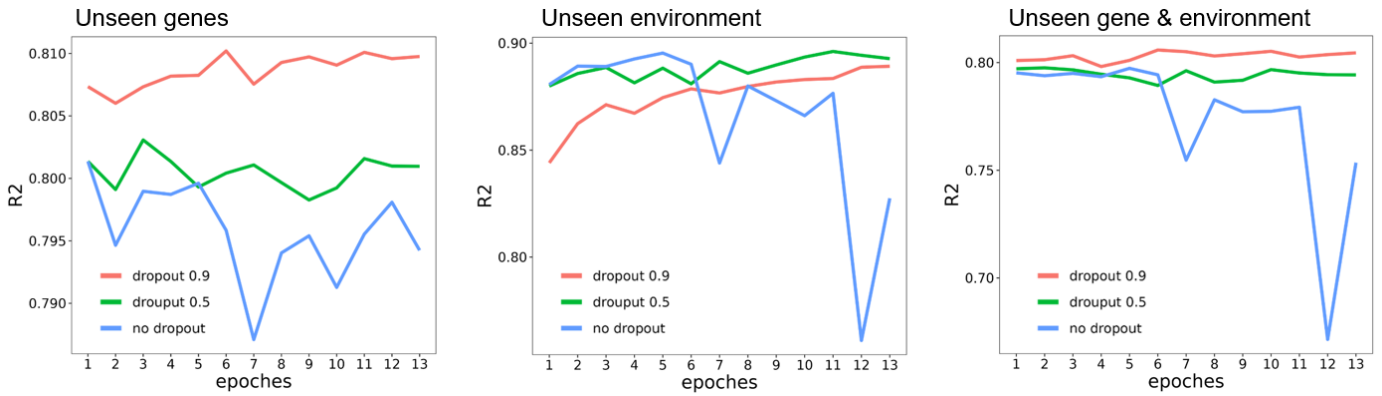


Figure S15. Evaluation of translation prediction models trained with varying dropout rates. R^2 values across three test sets are shown.

For the MFE model, we have added an analysis justifying our use of a compact and computationally efficient training set. We added the following text to the manuscript:
"To determine the optimal training set size, we compared our model trained on a compact set of 210,000 sequences against a model trained on an extended dataset of 11.5 million sequences (achieving a data-to-parameter ratio > 1). Both models achieved comparable optimization performance and yielded MFE predictions highly correlated with values from RNAfold. Given the similar performance, we selected the 210K training set for our framework, as it reduces computational time by approximately 75% without sacrificing model reliability or accuracy."

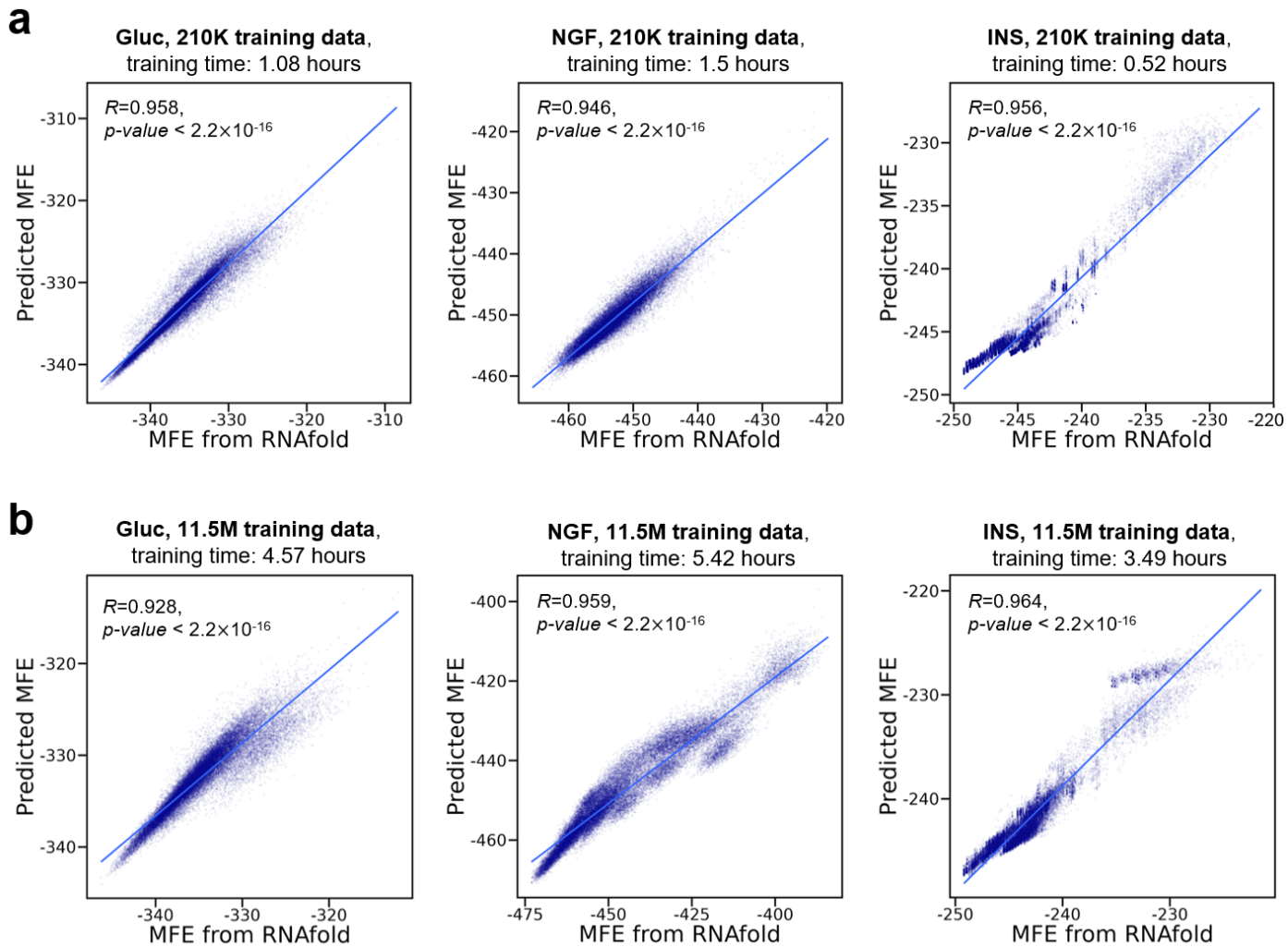


Figure S13. Evaluation of the Predictive Performance of the MFE Model to RNAfold. a. MFE models for different

mRNAs trained by 210K samples. b. MFE models for different mRNAs trained by 11.5M samples.

2. We remain unconvinced by the authors' explanation for the α and β parameters, specifically, their assigned values of 100, used to scale the translation and MFE losses (our prior point 8). They cite 'empirical experience' but do not provide specific results that justify constraining the translation loss to 0-10 and the MFE loss to 0-100. Moreover, the revised manuscript states that "incorporating MFE optimization played a complementary role in modulating mRNA secondary structure stability", but based on the reported loss values, the MFE seems to be weighted more highly. The authors should rectify this disparity.

We appreciate the reviewer's attention to this point. Regarding the scaling parameters α and β for the translation and MFE losses:
The output range (translation levels) of the translation model (S□) typically is between 0 and ~100, while MFE values generally range between -100 and -1000. To ensure numerical stability and comparability between the two loss terms during optimization, we introduced α and β to normalize them. For translation, we define $L_t = \sqrt{(r - S_t)^2}/\alpha$. If the target value r is 100 and the predicted S□ is 25, then with α = 100, we obtain L□ = 0.75 — bringing the loss into a 0–1 scale. For MFE, we define L□ = -β / S□. If S□ = –350 and β = 100, then L□ = 0.29 — again within a similar range. Without the α and β, the loss L□ and L□ would be about 75 and 0.003 (-1/-350), which are not in the same order of magnitude.
This normalization does not imply that MFE is weighted more heavily — the actual influence of each component is governed by the weight parameter (*w*) in: $\mathrm{Loss} = \mathrm{L_m} * \mathrm{weight} + \mathrm{L_t} * (1 - \mathrm{weight})$.

To enhance clarity, we have revised the manuscript text to explicitly state: "The output range of the translation model (S□) typically lies between 0 and ~100 (e.g., 0 to 25 for Gluc), while RNAfold-derived MFE values generally range between -100 and -1000 (e.g., -350 to -150 for Gluc). To ensure comparable magnitudes and numerical stability during optimization, we introduced scaling factors α and β to normalize both loss components. For most codon sequences exhibiting translation prediction values < 100 and MFE values > -1000 kcal/mol, α = β = 100 provide appropriate normalization. However, we recommend parameter adjustments under extreme value conditions: α should be increased to 1000 when translation predictions exceed 100, while β should be elevated to 1000 when MFE values fall below -1000 kcal/mol.

Additionally, the statement that "…incorporating MFE optimization played a complementary role..." might be misinterpreted as implying a subordinate importance of MFE optimization. To avoid ambiguity, we have revised the text to emphasize the synergistic coordination: "...incorporating MFE optimization synergistically modulating mRNA secondary structure stability and translation efficiency".

Reviewer #3 (Remarks to the Author):
The authors have addressed most of my previous concerns and have provided an improved manuscript. I have a few additional comments that I hope will further strengthen the work.

1. I appreciate the authors' effort to revise Figure S1. Tables S9 and S10 would benefit from clearer notation for model parameters. Several names are ambiguous or contain minor errors—for example, "max-pool" should be "max-pooling." It is also difficult to understand parameter names such as "Out" or "encoder". For guidance on documenting neural network layers and parameters, the authors might refer to RiboTIE (doi.org/10.1038/s41467-025-56543-0; Supplementary Algorithm 1).

We appreciate your suggestion. We have accordingly updated the parameter nomenclature in Tables S9-S10 and supplemented detailed descriptions in the revised manuscript:

Category	Parameter Name	Description	Values	Value Range
Training Configuration	num_epochs	Number of training epochs	20	fixed
	learning_rate	Initial learning rate	0.001	fixed
	lr_scheduler_milestones	Epochs when learning rate will drop	[10]	fixed
	lr_scheduler_gamma	Learning rate decay factor at each milestone	0.4	fixed
	weight_decay	L2 regularization coefficient	0.01	fixed
	bias_initialization	Custom bias parameter initialization	(0, 1)	random variable
	batch_size	Training batch size	100	fixed
Data & parameter	training_data	Number of training data	2.11M	fixed
	model_parameters	Total trainable parameters	17.27M	fixed
CNN Architecture	conv1d_kernel_size	1D convolution kernel size	[5, 30]	fixed
	conv1d_filters	Number of 1D convolution filters	[64, 256]	fixed
	conv1d_dilation	Dilation rate for 1D convolution	[1, 4]	fixed
	conv1d_stride	Stride for 1D convolution	[1, 3]	fixed
Pooling & Regularization	maxpooling_kernel_size	Max-pooling kernel size	3	fixed
	maxpooling_stride	Max-pooling stride	1	fixed
	dropout_rate	Dropout probability	[0.5, 0.9]	fixed
Fully Connected	output_layers	Number of final output layers	2	fixed
	dense_layer_units	Fully connected layer dimensions	[32, 128, 512]	fixed
	dropout_rate	Dropout probability	0.025	fixed

Table S9. Hyperparameters for the translation model. It includes global parameters and specific parameters for the Convolutional Neural Network and Fully Connected layers. The table specifies whether each parameter is fixed or variable within a given range.

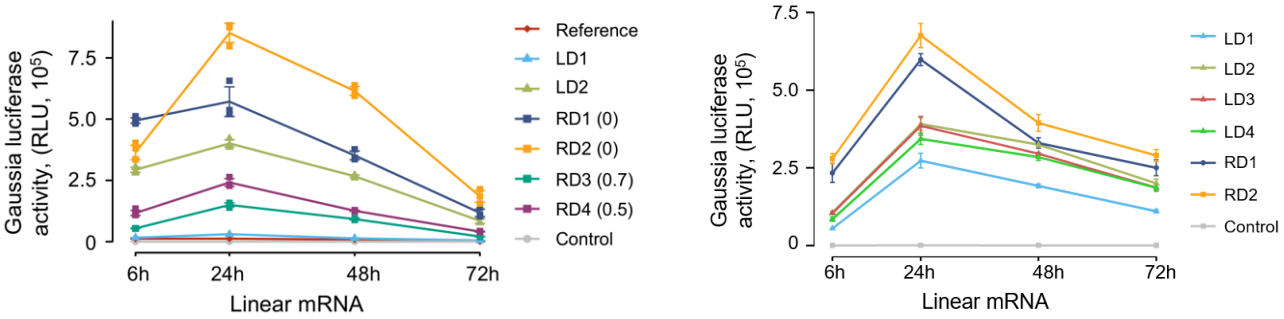
Category	Parameter Name	Description	Values	Value Range

Training Configuration	num_epochs	Number of training epochs	20	fixed
	learning_rate	Initial learning rate	0.001	fixed
	lr_scheduler_milestones	Epochs when learning rate will drop	[1]	fixed
	lr_scheduler_gamma	Learning rate decay factor at each milestone	0.1	fixed
	weight_decay	L2 regularization coefficient	5e-4	fixed
	epsilon_greedy	Epsilon parameter for exploration	(0.1, 1)	random variable
	batch_size	Training batch size	128	fixed
Data & parameter	training_data	Number of training data	210,000	fixed
	model_parameters	Total trainable parameters	10.88M	fixed
CNN Architecture	conv1d_kernel_size	1D convolution kernel size	[1, 3]	fixed
	conv1d_filters	Number of 1D convolution filters	[128, 256]	fixed
	conv1d_dilation	Dilation rate for 1D convolution	1	fixed
	conv1d_stride	Stride for 1D convolution	1	fixed
Pooling & Regularization	maxpooling_kernel_size	Max-pooling kernel size	3	fixed
	maxpooling_stride	Max-pooling stride	2	fixed
	dropout_rate	Dropout probability	0.5	fixed
	leaky_relu_slope	Negative slope for LeakyReLU	0.2	fixed
ResNet Components	residual_blocks	Number of BasicBlock residual units	8	fixed
	adaptive_pooling_output	Adaptive average pooling output size	1	fixed
Fully Connected	output_layers	Number of final output layers	2	fixed
	dense_layer_units	Fully connected layer dimensions	[256, 512]	fixed

Table S10. Hyperparameters for MFE model. Like Table S9, it includes global parameters and specific parameters for the CNN and Fully Connected layers. The table indicates whether each parameter is fixed or variable within a specified range, providing insights into the model's architecture and training process.

2. The new comparison between RD and LD in Figure S11 is helpful. To aid interpretation, please consider

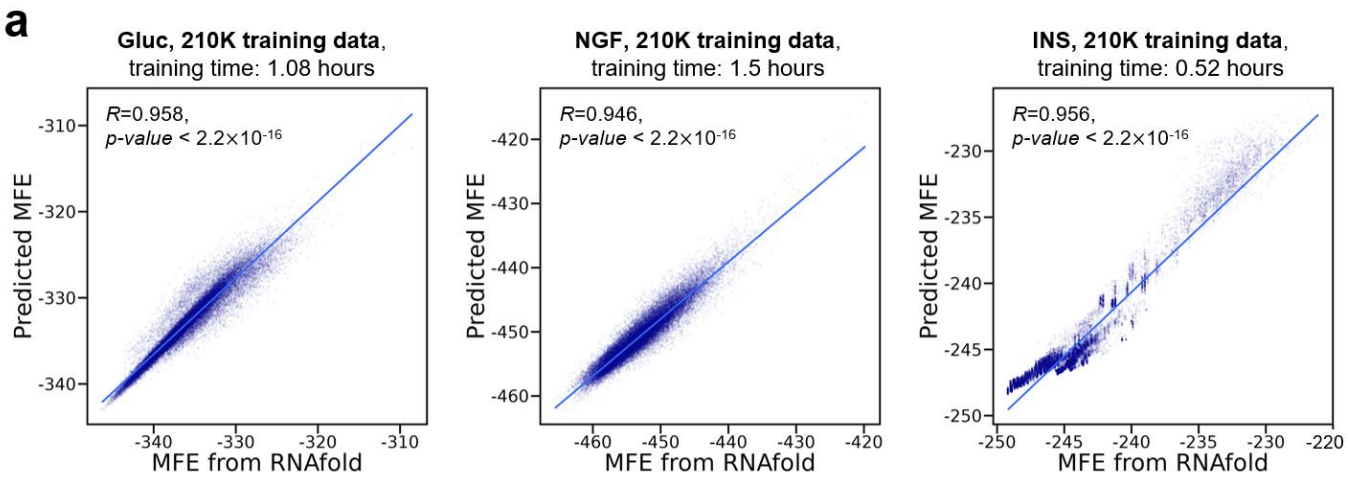
adding design details for LD3 and LD4 to Table S3 (e.g., CAI values and predicted translation level). A minor point: the authors mentioned that RD1 and RD2 in Figure S11 are identical to those in Fig. 4a, yet the values seem different. In Figure 4a, RD1 is ~5 at the 6 h time point, whereas in Figure S11 it is ~2.5. Could the authors double-check these data?



Thank you for the suggestion. We have added the requested design details (e.g., CAI values and predicted translation levels) for LD3 and LD4 to Table S3.
Regarding the apparent discrepancy in the RD1 protein expression level at the 6-hour time point between Figure 4a (~5) and Figure S11 (~2.5), we double-checked the raw data and confirmed the measurement values. This may be because the transfection of in vitro transcribed mRNA typically takes several hours, and analysis at the 24-hour time point is recommended (PMID: 40494942).

3. Figure S13 shows a strong correlation between the MFE model and RNAfold predictions, but the absolute scales are quite different. For INS sequences, RNAfold MFE values range from –250 to –220, whereas the MFE model spans only –207.25 to –206.25 (30-fold difference). Could the authors comment on the source of this discrepancy? This might not affect the sequence optimization part since the weight parameter w could compensate for this effect. But additional clarification of the MFE model's training and benchmarking protocol would be helpful.

We appreciate the reviewer for their insightful observation regarding the discrepancy in the absolute scale of MFE predictions between our model (− 207.25 to − 206.25) and RNAfold (− 250 to − 220) for INS sequences, as shown in Figure S13. Upon careful examination, we identified that in the previous version's evaluation of the MFE model, we inadvertently used *padded* CDS sequences (fixed at 4,500nt) as input, whereas the model had been trained on *unpadded* original sequences. This discrepancy in input sequence lengths caused the observed deviation in the scale of predicted MFE values. We have now corrected the evaluation protocol by using the original unpadded CDS sequences. With this correction, the MFE model now produces predictions that match both the expected value range and maintain strong correlation with RNAfold labels. We have updated the corrected figure in the revision.



4. I successfully installed the software package. To make RiboDecode more broadly useful, it would help to allow optimization of user-defined coding sequences, rather than limiting inputs to the three predefined examples (gluc, NGF, H1N1).

Thank you for your suggestion. We have uploaded the latest version and it allows optimization of user-input codon sequence.

some minor errors:
1) Line 917: "Gaussian luciferase" should be "Gaussia luciferase".
2) Table S5: both columns for 24h protein expression are labeled “In HEK293T”. Column label "In AREP19" should be "In ARPE19".

Thank you very much. We have modified them.

Reviewer #4 (Remarks to the Author):

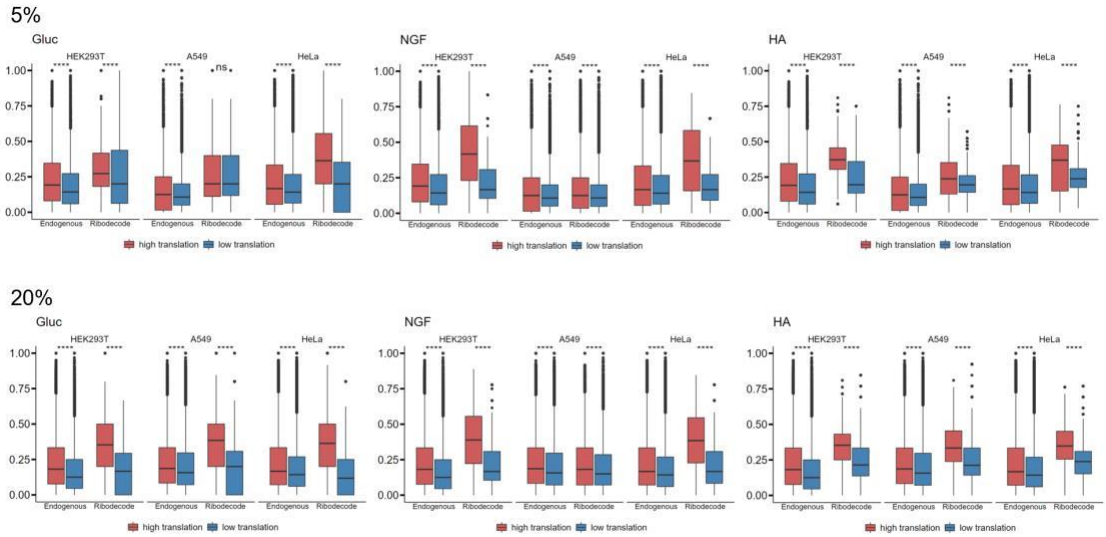
We thank the reviewer for raising new concerns based on the previous issues. To keep the document concise and clear, we have extracted the points requiring response and included the prior questions and corresponding replies (with non-essential content omitted) before addressing the new points and responses.

1. Previous comment: "New Genes" Definition: The term "new genes" in the study may be misleading if the sequences come from the same gene families or are not entirely novel. A clarification or additional analysis is necessary to validate this claim.
Previous response: We are sorry for the misleading word. We meant genes which were not seen during training. We have renamed it as "unseen genes" in our revised manuscript.

New point: Thank you for renaming. We would appreciate if the authors could clarify, somewhere in the manuscript, that some of the genes in the unseen dataset may be related to those in the seen dataset, and that no filtering was applied to separate them.
New response: We appreciate your comment regarding this matter. We added the following content to Methods: Furthermore, potential associations between genes in the unseen dataset and those in the seen dataset may lead to overestimation of model performance. To evaluate this potential issue, we obtained gene family

annotations from the HUGO Gene Nomenclature Committee (HGNC; https://www.genenames.org/) and excluded genes in the unseen dataset that belonged to the same gene families as those in the seen dataset (resulting in the removal of 130 genes). The results demonstrated only a marginal decrease in prediction accuracy on the "unseen gene test set" (with R^2 decreasing from 0.81329 to 0.81295). Therefore, although some of the genes in the unseen dataset are related to those in the seen dataset, the performance of the model was not overestimated.

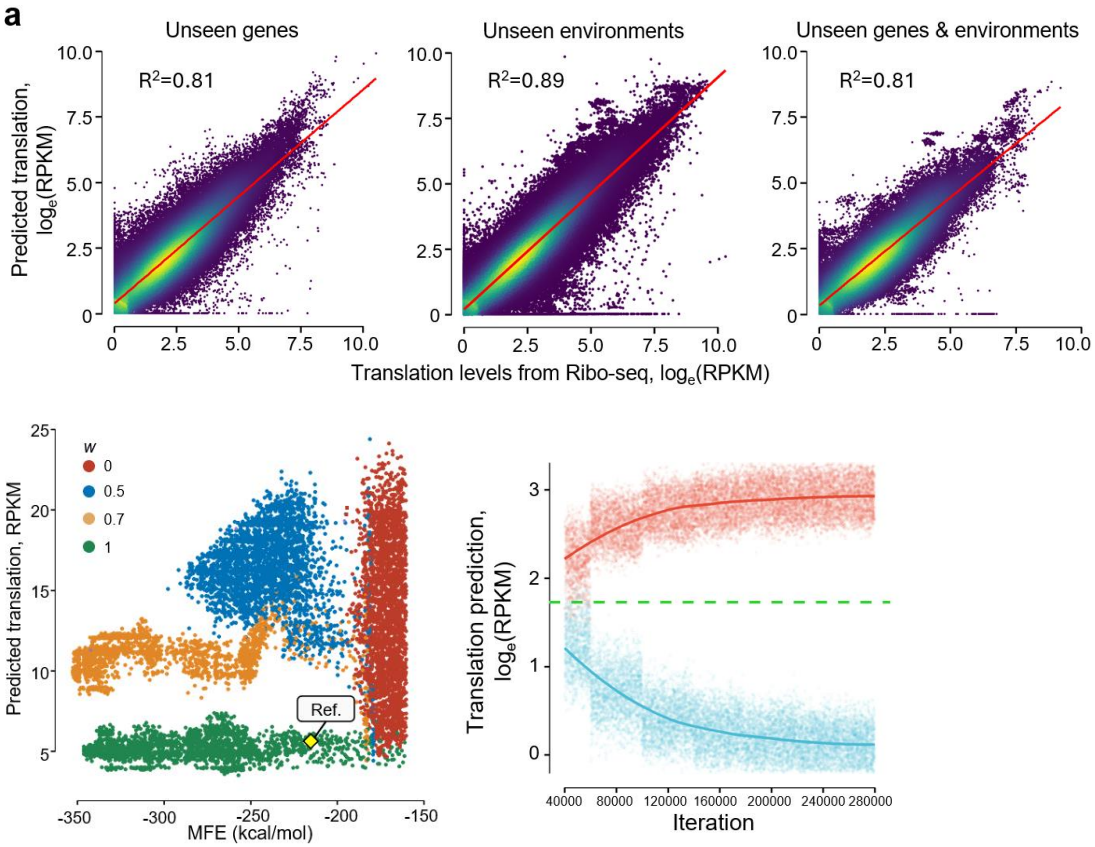
2. Previous point: CAI Threshold (Fig 3c): The threshold of 10% for CAI calculation is arbitrary. Since the differences between "Endo" and "RD" are small but statistically significant, a smaller threshold (e.g., 1%) may make the result statistically not significant. The authors should explain the rationale or explore other thresholds.
Previous response: ……



New point: We are satisfied with this analysis. For the benefit of readers, it would be helpful if the authors included this figure in the supplementary materials.
New response: Thank you for the suggestion. We added this figure in the supplementary materials (Figure S16).

3. Previous point: Model Extrapolation (Figs 2 & 3): Figure 2 shows a maximum predicted translation of 10, while Figure 3 reports values up to 25. This discrepancy suggests potential extrapolation by the model, which should be addressed and clarified.
Previous response: The expression values in the original manuscript were log converted. We appreciate your attention to this and have accordingly revised the labels in figure 2a: …The translation levels from Ribo-seq were log-transformed (see Methods).

New point: We thank the authors for clarifying. Given the values, we are assuming the authors used a log2 transformation, if yes, could the authors add this as well as the label the Y-axis unit as "RPKM" for ease of interpretation?
New response: Thank you for the suggestion. We modified the axis labels of Figure 2a, 3b, and S5, such as:

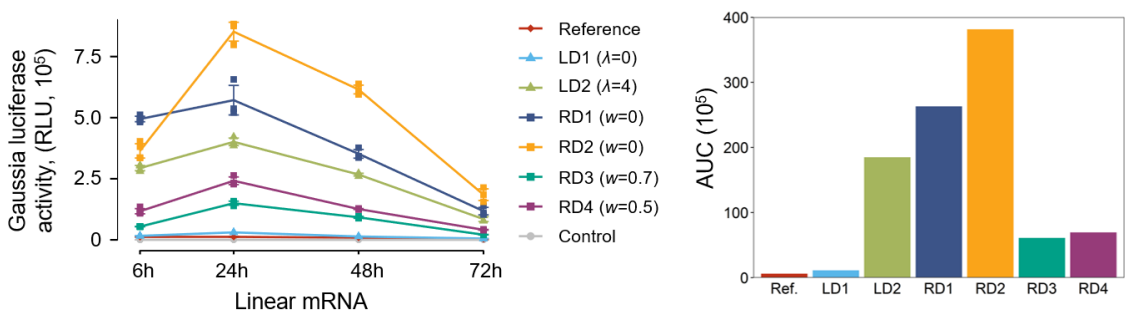


4. Previous comment: Definition of translation: The authors do not provide a clear definition of "translation level" in either the result or methods section. Is 'translation level' the RPKM from Ribo-Seq or is it 'translation efficiency' defined for each transcript, RPKM from Ribo-Seq divided by RPKM from paired RNA-Seq? authors need to explicitly specify the definition for better interpretability of the results.
Previous response: ……

New point: We are satisfied with the authors' edits and clarification. We would recommend the authors add RPKM to plot axes labels or in the legend for better readability.
New response: Thank you for the suggestion. We modified the axis labels of Figure 2a, 3b, and S5.

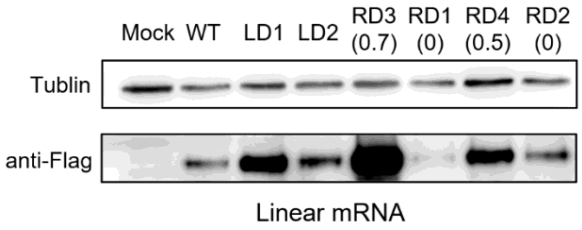
5. Previous comment: ……
Previous response: ……

New point: The updated figure/data, along with the authors' clarification, adequately addresses our concern. Additionally, it would be ideal to include AUC plots alongside the line plots to aid in the interpretation of the data.
New response: Thank you for the suggestion. After consideration, we have decided to include AUC plots alongside the line plots in the supplementary figures, such as:



6. Previous comment: ……

Previous response: ……



New point: The updated figure/data, along with the authors' clarification, adequately addresses our concern. Additionally, it would be ideal to include WB band quantification on the WB itself rather than Figure S15 for easier data interpretation.

New response: Thank you for the suggestion. We removed the previous Figure S15 and include the quantification result on the below the WB figures, such as:

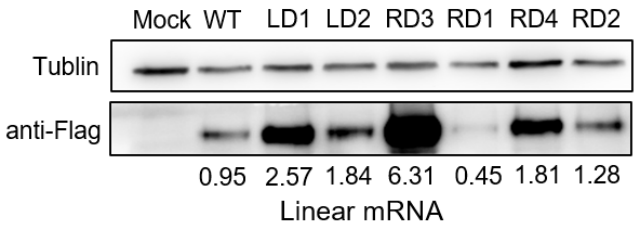


Figure 5a. Western blot analysis shows protein expression of the HA variants in HEK293T cells 24 hours after transfection. RD sequences were designed by RiboDecode (w parameter were set to 0, 0, 0.7, and 0.5 for RD1-4). LD sequences were designed by LinearDesign (λ parameter were set to 0 and 4 for LD1 and LD2). The expression values were quantified using GelAnalyzer.

7. Previous comment: ……

Previous response: ……

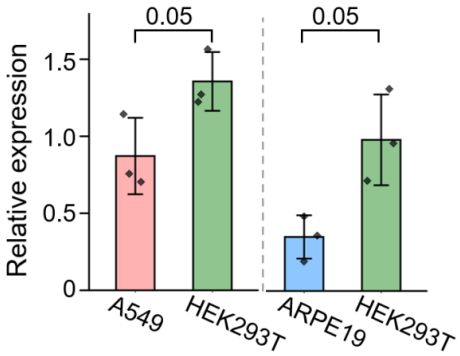


Figure 4b. Relative expression of experimentally measured protein expression values of mRNA variants designed by RiboDecode at 24 hours in different cells. The error bars denote standard variation. One-sided Wilcoxon test was used to calculate p-values shown in the figure.

New point: We are satisfied with the authors' explanation and the corresponding edits to the text as well as the updated figure. However, we would like to draw attention to a recurring phrase in several figure captions: "standard variation." If this refers to a novel metric, we kindly ask the authors to clarify its definition. Otherwise, we assume this is intended to be "standard deviation" and request that the terminology be corrected accordingly.

New response: Thank you for pointing this out. "Standard deviation" is the most accurate and proper expression. We have corrected all the statements.

8. Previous comment: Fluc Performance (Fig S8): The 17-fold improvement for Fluc (vs. 72-fold for Gluc) suggests the model's translation rules are not well-generalized. Sequence feature comparisons could be performed between the RD and reference sequences for these 2 related proteins to explain this discrepancy. Missing CAI control here as well.
Previous response: ……

New point: We thank the authors for their explanation and the corresponding edits to the text. We agree with the authors that "…optimal codon strategy is gene specific." While we agree that context-aware approaches like RiboDecode are valuable for capturing complex features, simpler metrics such as CAI may still be effective for certain genes (e.g., Fluc in this manuscript). So, instead of "…may even lead to counterproductive optimization strategies….", we would recommend use "ineffective" (something to this effect) to better present this result in context of data from wider literature.

New response: We agree with your point and have made the following revisions: we revised the statement from "…indicating that CAI is not a reliable predictor of protein expression levels in this context and may even lead to counterproductive optimization strategies" to "…indicating that CAI is not a reliable predictor of protein expression levels and ineffective optimization strategy in this context."

9. Previous comment: Missing Controls (Fig. 5 and 6): No codon-optimized (e.g., high CAI) sequences were used as reference. As Kudla et al., (PMID: 16700628) have demonstrated that WT sequences with low GC3% usually do not express well.
Previous response: Thank you for this comment. We would like to clarify the sequences included as controls in Figures 5 and 6. In Figure 5, the HA-LD2 sequence had a CAI of 0.96 and GC3% of 88%. In Figure 6, the NGF-LD2 sequence had a CAI of 0.95 and GC3% of 82%. These sequences can effectively serve as the requested 'codon-optimized. These LD2 sequences, along with our RiboDecode sequences, consistently outperformed the wild-type (WT) sequences, which likely aligns with the expectation from Kudla et al. that WT sequences with potentially lower GC3% might exhibit lower expression.

New point: We are satisfied with the authors' explanation and the corresponding edits to the text. We think referring to the conclusions from Kudla et al. (PMID: 16700628) in their manuscript will add some context to these results. So, we recommend adding it as a reference in this manuscript.
New response: Thank you for the suggestion. We have incorporated this reference in the manuscript: "Additionally, other indices have been used to guide sequence optimization. For instance, higher GC content (GC%) has been associated with enhanced gene expression[13]".

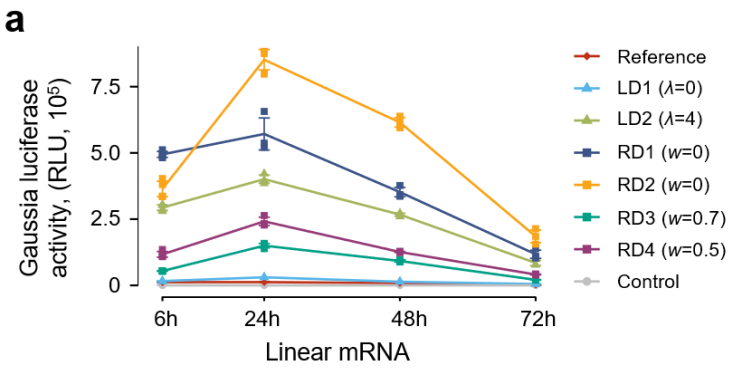
10. Previous comment: Minor Comment #1: Line 74-76. MFE is discussed only in relation to mRNA stability. Lower MFE/higher mRNA structure can also be a hurdle for translation elongation and needs to be acknowledged here.
Previous response: Thank you for your reminder. We emphasized that MFE also affects translation in Discussion: …incorporating MFE optimization played a complementary role in modulating mRNA secondary structure stability and translation.

New point: We thank the authors for their explanation and the corresponding edits to the text. However, we recommend clarifying what is meant by the term "complementary." Are the authors referring to a secondary or supporting role?
New response: We appreciate the opportunity to clarify. The term 'complementary' might be misleading. We did not suggest that MFE optimization was secondary in importance. To avoid ambiguity, we have revised the text to

emphasize the synergistic coordination: "...incorporating MFE optimization synergistically modulating mRNA secondary structure stability and translation efficiency".

11. Previous comment: Minor Comment #6: Lines 1050-1051, 1065-1066, 1081-1082. The w values for the RD sequences need to be shown either in the figure legends or at least in the figure captions.
Previous response: Thank you for your suggestion. We have assigned the w parameters to the sequences in the Figures.

New point: We are satisfied with the authors' explanation and the corresponding edits to the text. We would recommend to also add the parameter for LD designs so readers have a better idea about LD1 and LD2 whether they have higher weight for CAI or MFE.
New response: Thank you for the suggestion. We added the parameters of LD sequences in figures, such as:



REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The authors have addressed most of our comments. On seeing this revision, however, it was surprising to read that the authors' neural MFE model is specific to particular sequences. There already exist models which can predict MFE for any sequence without needing to be re-trained for each individual mRNA, like the Krueger/Ward model. The manuscript mentions that this prior model isn't available, but it seems that it has been out for a while: https://github.com/rkruegs123/jax-rnafold. The authors should test the model in their pipeline, to evaluate if their MFE model provides any benefit compared to it.

Response: We sincerely thank the reviewer for providing the link to the Krueger/Ward model (JAX-RNAfold) and prompting this valuable comparison. We apologize for our previous oversight. Following your recommendation, we have evaluated the Krueger/Ward model and compared it to our MFE optimization approach. Our findings confirm that our sequence-specific model provides some advantages for our research application, primarily concerning sequence length and computational efficiency.

Sequence Length and Scalability

The most significant issue with the JAX-RNAfold model is its scalability, with its reported memory complexity of $O(n^3)$ (PMID: 38038257). In addition, it reports that JAX-RNAfold limits optimization to 1,250 nucleotides (nt) on an A100 GPU (80GB video RAM, VRAM). We tested it on our machine (V100 GPU with 32GB VRAM) and found it can only optimize sequences up to 600 nt. This makes it unsuitable for optimizing most full-length human mRNAs, which are the focus of our therapeutic applications and typically have coding sequences of 1,500–2,000 nt. In contrast, our model is capable of optimizing sequences up to 4,500 nt on the same hardware, making it uniquely suited for our study's goals.

Optimization Performance

To perform a direct comparison, we optimized a sequence short enough for both models to process: the 558 nt Gaussia luciferase (Gluc) mRNA. In this head-to-head test, our model outperformed the JAX-RNAfold model on the following metrics:

- Better MFE: Our model achieved a lower (more stable) MFE of -345.46 kcal/mol compared to -327.1 kcal/mol.
- Faster Optimization: Our model was faster, completing the optimization in 1.08 hours versus 1.55 hours.
- Lower Resource Use: Our model required significantly less GPU memory (0.99 GB) compared to the JAX-RNAfold model (26.36 GB).

In summary, while the Krueger/Ward model is a valuable MFE tool with general purpose, its length restrictions and higher resource demands limit its applicability in our study. Our model offers longer sequence compatibility, making it better suited for our research needs.

Gene tested	Properties	JAX-RNAfold	Our MFE model
Gluc (558 nt)	Time-cost	1.55 hours	1.08 hours
	GPU memory	26.36 GB	0.99 GB
	MFE	-327.1 kcal/mol	-345.46 kcal/mol

Table S12. Gluc MFE optimization by the JAX-RNAfold and our model. The MFE values were calculated by RNAfold.

We have also revised the methods in the manuscript.

We evaluated our MFE optimization framework against existing MFE optimization methods, including the general-purpose, differentiable model (JAX-RNAfold)[72]. Our analysis revealed that the high computational complexity of JAX-RNAfold severely limits its use to short sequences (under 600 nt on our V100 GPU with 32 GB video RAM), making it unsuitable for our primary goal of optimizing full-length human mRNAs. To perform a direct, head-to-head comparison, we used a compatible sequence (Gaussia luciferase, 558 nt). Our model achieved a lower (more favorable) MFE, completing the optimization faster and with substantially less GPU memory than JAX-RNAfold (Table S12). While LinearDesign is another powerful tool, it is not a differentiable model and is thus incompatible with our gradient-based optimization framework. Given the critical need for scalability to therapeutically relevant sequence lengths and the efficiency observed in direct comparison, we proceeded with our sequence-specific MFE optimization approach.

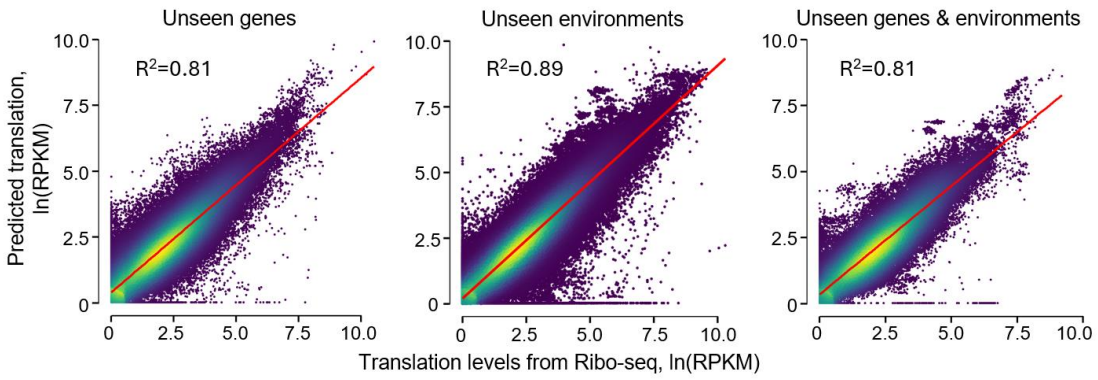
Reviewer #3 (Remarks to the Author):

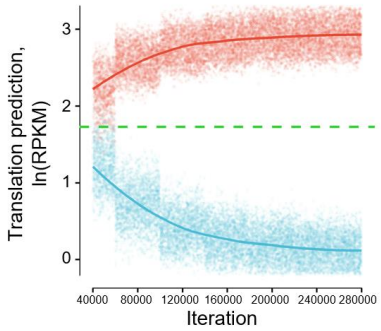
The authors have addressed my previous concerns.

A few minor points:

(1) It is unusual to use loge(RPKM) in the axis label in Fig. 2a. If the base of the logarithm is e, it is more conventional to use ln() instead.

Thank you for the suggestion. We have modified the axis labels in Fig. 2a and Fig. S5:





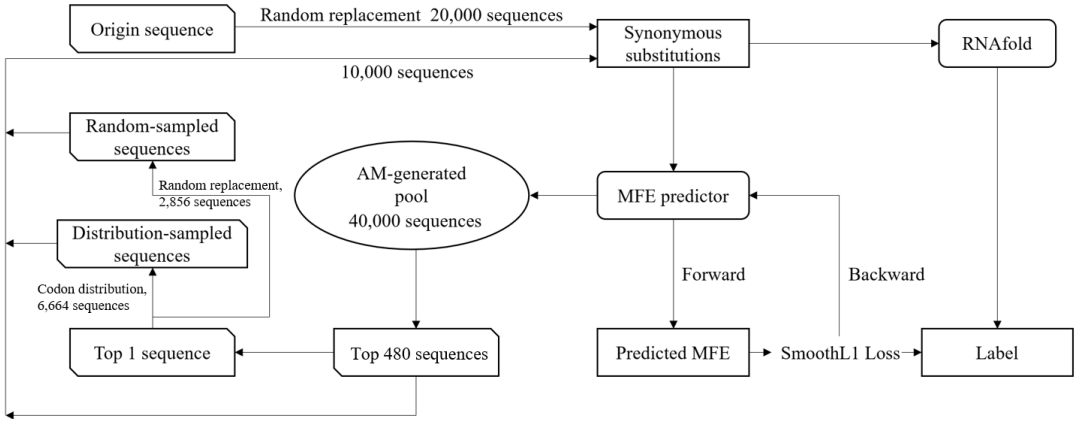
(2) The manuscript mentions that the MFE model needs to be trained for new CDS inputs. It might be helpful if the authors could provide additional guidance in the GitHub repository on how the MFE model is trained for a new sequence. For example, how many training samples are typically generated? This information would facilitate broader adoption of the method.

The "Integrative Training of MFE Model" section of the manuscript describes the MFE model training process. In response to your suggestions, we have added detailed explanations of the MFE model training procedure on GitHub:

MFE Model Training and Optimization

The software **automatically** trains the MFE model using an active learning approach that integrates model training with sequence optimization for any input mRNA sequence (where $w>0$). The process consists of the following steps:

1. **Initial Sampling**: Generate 20,000 sequences via random synonymous substitutions (≤10% codons) of the original sequence.
2. **Initial Training**: Predict MFE for these sequences using the model, with RNAfold providing ground truth labels.
3. **Sequence Generation**:
    - The codon optimizer generates new sequences using the current model.
    - Select the top 480 sequences by predicted MFE (lower is better).
    - Apply random and codon-distribution-based replacements to the lowest-MFE sequence, yielding 10,000 new sequences.
4. **Model Retraining**: Retrain the model on the new sequences, using both MFE predictions and RNAfold labels.


Origin sequence
Random replacement 20,000 sequences
Synonymous substitutions
RNAfold
10,000 sequences
Random-sampled sequences
Random replacement, 2,856 sequences
AM-generated pool 40,000 sequences
MFE predictor
Distribution-sampled sequences
Codon distribution, 6,664 sequences
Forward
Backward
Top 1 sequence
Top 480 sequences
Predicted MFE
SmoothL1 Loss
Label

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

The authors have satisfactorily addressed our comments. As a minor suggestion to help readers, the authors should include the time to train their model in the time taken to perform the optimization (since this must be done for every new sequence, unlike JAX-RNAFold).

Thank you for your reminder. We have updated our manuscript to provide a more detailed description of the model training and sequence optimization times in the Methods section:

"Training the translation model required approximately 24 hours. The training time for the MFE model varied depending on the input sequence length. For example, for the Gluc, NGF, and HA mRNA codon sequences, the MFE model training took 1.08, 1.5, and 3.13 hours, respectively. After the MFE model training was completed, the optimization phase was carried out, which required about 1 hour for any sequence."

Reviewer #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #6 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #4 (Remarks to the Author):

Peer Review Report for NCOMMS-24-67515-T:

The authors introduce RiboCode, a novel deep learning-based framework designed to optimize mRNA codon sequences to enhance protein production and therapeutic efficacy. Unlike traditional pre-defined metric-based codon optimization, they claim their deep learning-based method better learned complex relationships governing translation and explored a bigger sequence space. RiboCode leveraged large-scale ribosome profiling (Ribo-seq), cellular environment (e.g., RNA-seq), and RNA secondary structure prediction datasets to better capture the interplay between codon usage, cellular context, translation efficiency, and stability.
RiboCode's sequence optimizer framework combined a trained translation prediction model, an MFE active learning model, and a codon optimizer that employed activation maximization to explore and optimize mRNA sequences for high translation and/or stability. The authors validated its effectiveness through in vitro experiments and in vivo studies, demonstrating increased protein expression, and improvements in vaccine immunogenicity and therapeutic efficacy. Additionally, they highlighted its compatibility with chemical modifications and mRNA formats, including m1Ψ-modified and circular mRNAs which are highly relevant to the field of mRNA therapeutics.
This study underscored the advantages of using deep learning approaches to mRNA codon optimization. While the manuscript presents a novel and promising approach, the points raised below highlight significant gaps that need to be addressed to ensure clarity, rigor, and reliability. Overall, we recommend major revisions to improve the robustness of the findings and alignment with the claims made. With these revisions, the manuscript could make a valuable contribution to the field of mRNA therapeutics.

Major comments:

Major Comment #1: Additional in silico validation is needed to support the claims

1. Sequence Space Exploration: The manuscript emphasizes the exploration of sequence space as a key strength of RiboCode, but no direct comparison with LinearDesign or other existing tools (Ribotree (PMID: 34520542), CDSFold (PMID: 26589279)) are provided to demonstrate this claim. Including such comparisons is essential to validate the broader sequence exploration ability of RiboCode.

Thank you for your suggestion.
We added the following content to a new section in Methods, titled 'Comparative analysis of sequence space across design Methods': We generated 1,000 full-length Gluc codon sequences using Ribotree, LinearDesign, CDS-fold, and RiboDecode, respectively. High-dimensional sequence embeddings were extracted using CodonBERT, followed by t-SNE dimensionality reduction for visualization.

In addition, we added the following content to 'RiboDecode's Optimization Strategies for Enhanced mRNA Translation' section in Results: Moreover, RiboDecode-generated sequences spanned a broader embedding space compared to those produced by Ribotree, CDSfold, and LinearDesign (Figure S6, see Methods), suggesting enhanced sequence diversity.

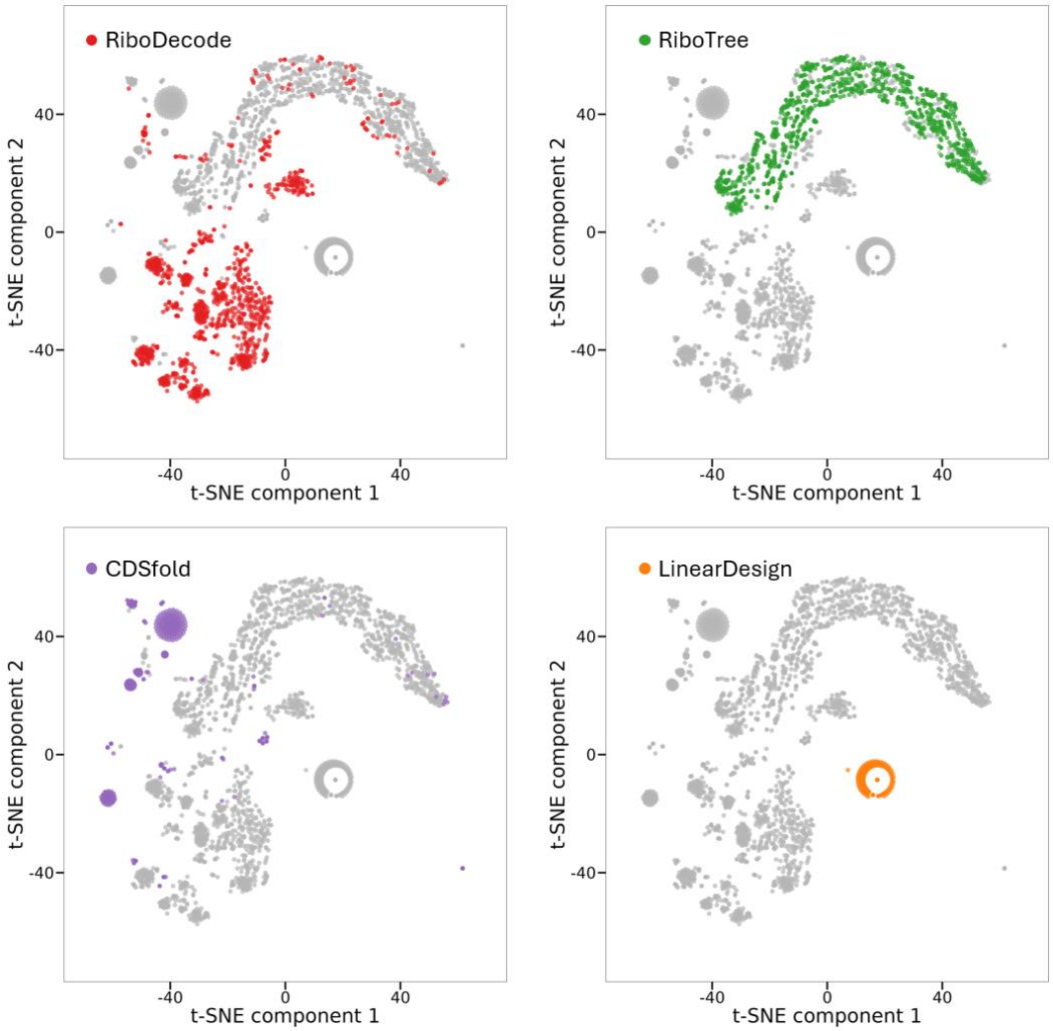


Figure S6. t-SNE distribution of Gluc codon sequences generated by RiboDecode, RiboTree, CDSfold, and LinearDesign (see Methods). The results show that the sequence space of RiboDecode is larger than other methods.

Reviewers 4 and 5: The updated figure/data, along with the authors' clarification, adequately addresses our concern.

2. MFE Model Learning: The manuscript lacks evidence for how well the MFE model learns to predict MFE values. Results on unseen mRNAs (separated by families) and their associated metrics would provide confidence in the generalizability of the MFE model. A recent article from Szikszai highlights this issue (PMID: 35748706).

Thanks for the suggestion. We benchmarked our model to RNAfold and found that our the MFE value highly agreed with the ones from RNAfold, showing the reliability of our MFE model. We showed the result in the Methods section of the revision. We evaluated the MFE values of mRNAs between our model and RNAfold. We found that our MFE value highly agreed with the one from RNAfold (Figure S13), showing the reliability of our MFE model.

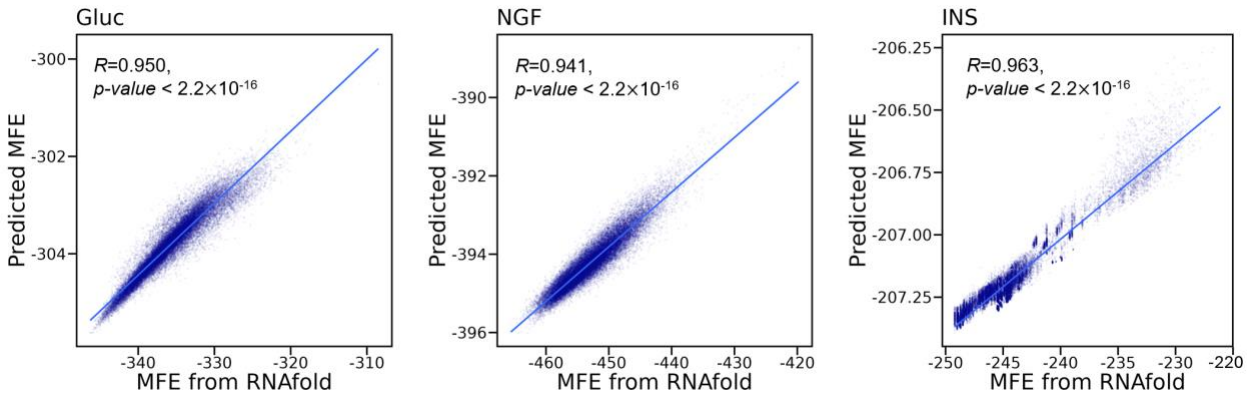


Figure S13. Evaluation of the Predictive Performance of the MFE Model to RNAfold. Pearson's correlation coefficients are provided. Blue lines denote the linear fit. All values are reported in kcal/mol.

For the second question, our model cannot predict MFEs for unseen mRNAs. For an unseen mRNA, our model must first train the sequences with MFEs labelled by RNAfold. Our requirement for the MFE model is that it be differentiable which will be compatible with our optimizer. By the time we started our study, there was no differentiable MFE optimization model. We therefore developed our own differentiable MFE model. As another reviewer suggested, we found that the recently developed a differentiable general MFE model (Paper name: "Scalable Differentiable Folding for mRNA Design") also could efficiently optimize MFE. Once this model is publicly available (it is in biorxiv right now), we will comprehensively evaluate it and incorporate it into our integrative framework. We have discussed this limitation in Discussion in the revision.
Third, our MFE model cannot predict MFEs for unseen mRNAs. For an unseen mRNA, our model must first train the sequences with MFEs labelled by RNAfold. A general MFE model should be developed and used in future.

Reviewers 4 and 5: The updated figure/data, along with the authors' clarification, adequately addresses our concern.

3. "New Genes" Definition: The term "new genes" in the study may be misleading if the sequences come from the same gene families or are not entirely novel. A clarification or additional analysis is necessary to validate this claim.
We are sorry for the misleading word. We meant genes which were not seen during training. We have renamed it as "unseen genes" in our revised manuscript.

Reviewers 4 and 5: Thank you for renaming. We would appreciate if the authors could clarify, somewhere in the manuscript, that some of the genes in the unseen dataset may be related to those in the seen dataset, and that no filtering was applied to separate them.

4. Energy Landscape Visualization: Since a local optimizer is employed, the manuscript should analyze the energy landscape of different sequences. This would help demonstrate whether the optimizer avoids local minima and, if not, why this is acceptable in the given context.

Thank you for the suggestions. We agree that our model uses a gradient ascent optimization based on activation maximization (AM), which is a local optimizer. We provided several lines of evidence that our method effectively avoided local optima or it was acceptable in our case.

1. We present the translation optimization trajectory (Figure S4) and the sequence space of translation level & MFE (Figure 3a, 3b) for the Gluc codon sequences generated by RiboDecode. The results

demonstrate that RiboDecode achieves highly optimized sequences, with the predicted translation level increasing from 5.9 to 25 and the MFE decreasing from -216 to -350 kcal/mol—approaching the lowest MFE achieved by LinearDesign, which is more like a global optimizer (-346.2 kcal/mol).

2. Per the suggestion, we now include the fitness score trajectory during optimization (*w*=0.5; Figure S3), which shows a consistent improvement in fitness score. Based on these results, we believe our method effectively avoids poor local optima and the optima it finds are sufficient and acceptable.

We added the fitness score trajectory to 'RiboDecode is a Deep Learning Framework for mRNA Codon Optimization' section in Results: Using a gradient ascent optimization approach based on activation maximization (AM), the optimizer adjusts the codon distribution to maximize the fitness score (Figure S3).

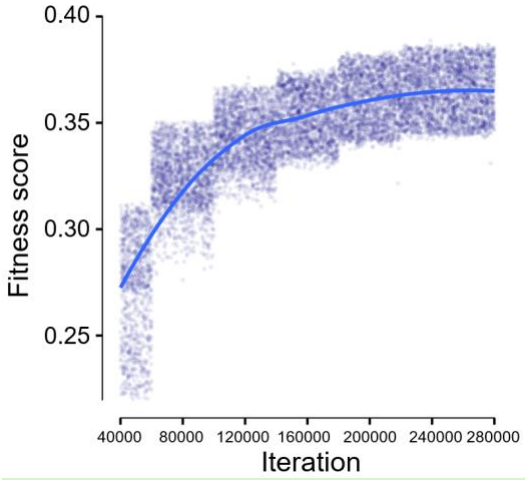
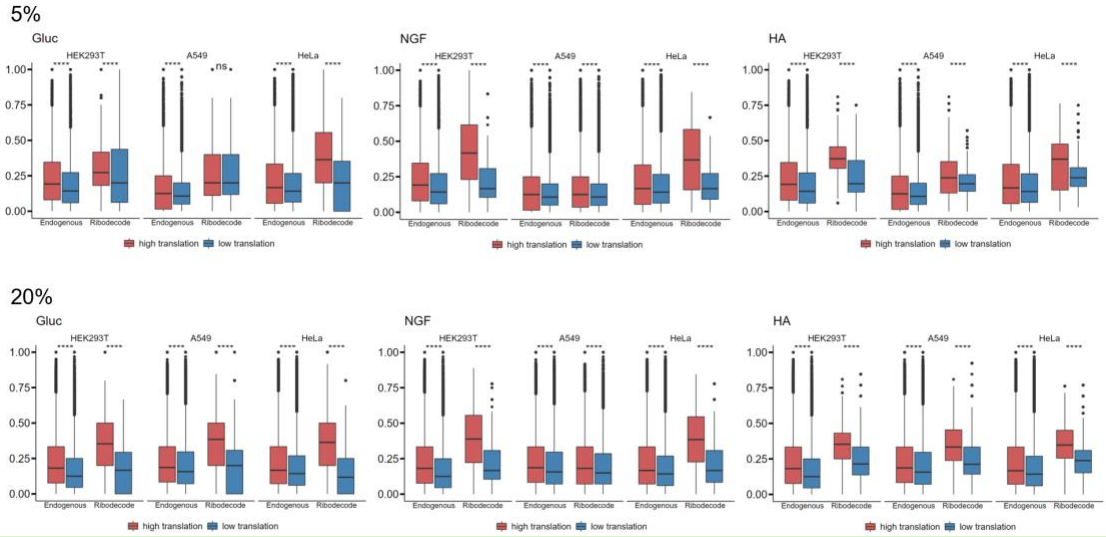


Figure S3. Fitness score trajectory during the iterative optimization of Gluc codon sequences (*w*=0.5). The fitness score is defined as the inverse of the loss function in our codon frame work (see Methods).

Reviewers 4 and 5: The updated figure/data, along with the authors' clarification, adequately addresses our concern.

5. CAI Threshold (Fig 3c): The threshold of 10% for CAI calculation is arbitrary. Since the differences between "Endo" and "RD" are small but statistically significant, a smaller threshold (e.g., 1%) may make the result statistically not significant. The authors should explain the rationale or explore other thresholds.

1. We conducted the same analysis at thresholds of 5% and 20%. The results were consistent with those obtained at the 10% threshold, except Gluc in A549 cells at the 5% threshold. We have added the following content to 'Codon usage analysis' section in Methods: We further evaluated the impact of varying expression thresholds (top/bottom 5%, 10%, and 20%), with consistent results observed across these cutoff values, except Gluc in A549 cells at the 5% threshold.
2. A threshold of 1% would yield an inadequate sample size of approximately 20 genes, rendering the subsequent analysis statistically unreliable.



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

Reviewers 4 and 5: We are satisfied with this analysis. For the benefit of readers, it would be helpful if the authors included this figure in the supplementary materials.

6. Model Extrapolation (Figs 2 & 3): Figure 2 shows a maximum predicted translation of 10, while Figure 3 reports values up to 25. This discrepancy suggests potential extrapolation by the model, which should be addressed and clarified.

The expression values in the original manuscript were log converted. We appreciate your attention to this and have accordingly revised the labels in figure 2a: …The translation levels from Ribo-seq were log-transformed (see Methods).

Reviewers 4 and 5: We thank the authors for clarifying. Given the values, we are assuming the authors used a log2 transformation, if yes, could the authors add this as well as the label the Y-axis unit as “RPKM” for ease of interpretation?

7. Ablation Studies and Error Bars: Error bars are missing in the ablation studies and other analyses (e.g., different dataset splits). Including error bars and biological replicates would enhance the reliability of the findings.

Thank you for pointing this out. We presented all the results for the cross-validation (Figure S4) and ablation analysis (Figure 2b):

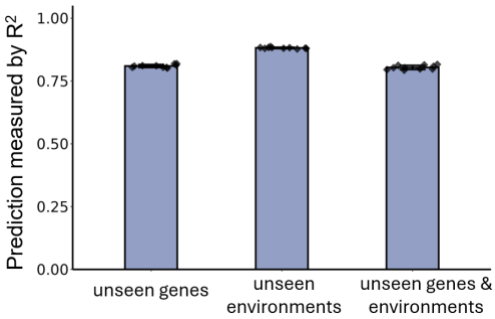


Figure S4b. 10-fold cross-validation on the test sets, where each dot on the bar represents a test. The error bars denote standard variation.

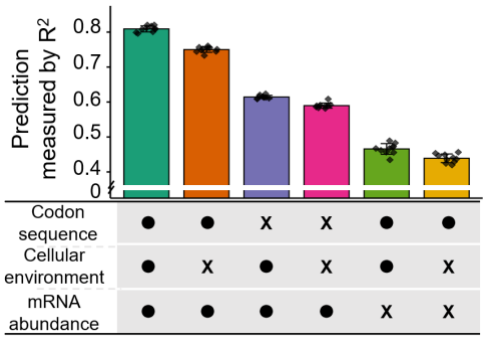


Figure 2b. …10-fold cross-validation for the ablation analyzes, where each dot on the bar represents a test. The error bars denote standard variation.

Reviewers 4 and 5: The updated figure/data, along with the authors' clarification, adequately addresses our concern.

8. Cellular Context Justification: Cellular context contributes only marginally (R2=0.06), yet it is heavily justified as a critical component of the model. This justification should be toned down or revised to align with the results.

We agree with your view that the contribution of the cellular context in the model is marginal. We revised the phrases in the manuscript and toned down the justification.

1. Removed the underlined portion in the Abstract: "RiboDecode introduces several advances, including ... ~~context-aware mRNA optimization~~ ...
2. Revised the sentence in the Abstract from "RiboDecode achieves mRNA design in different cellular context" to RiboDecode enables mRNA design with consideration of cellular context.
3. Revised the sentence in the Introduction from "RiboDecode also achieved mRNA design in different cellular context" to RiboDecode also considered cellular context.
4. Added the following content to the 'Experimental Validation Demonstrates RiboDecode's Versatility and Efficacy' section in Results: … These results confirm RiboDecode's capability for context-aware design, while the discrepancy in the ARPE19 comparison indicates potential for refining the model to better capture quantitative differences across diverse cellular environments.

Reviewers 4 and 5: We are satisfied with the authors' explanation and the corresponding edits to the text.

9. Key References Missing:
1. The authors have overlooked important contributions to the field, including Mauger et al., 2019 (PMID: 31712433) and Ribotree (PMID: 34520542). These references are foundational in mRNA translation and optimization research and should be cited with appropriate acknowledgment. The relevance of these methods (along with LinearDesign) for therapeutic mRNA design were reviewed in Metkar et al. (PMID: 38030688).
2. Other codon optimization algorithms that utilize deep learning Lines 115-117. LLM bases approaches like CodonBERT (PMID: 38951026), and ChatNT (https://doi.org/10.1101/2024.04.30.591835) have recently come out. These methods need to be acknowledged in the Introduction. While these methods do not leverage RiboSeq, the authors can point to this to indicate how their deep learning approach covers this aspect.
3. Other machine learning models for ribo-seq data: Line 129: There have been deep-learning approaches exploring Ribo-Seq data like RiboShape (PMID: 27307616), Riboformer (PMID: 38443396) and more recently RiboTIE (PMID: 38585907), RiboNN (https://doi.org/10.1101/2024.08.11.607362). The statement should be changed to acknowledge other methods so far. The authors should contrast their model and findings with a recent study from Can Cenik lab and Sanofi (RiboNN). The model uses a similar approach and also the UTR sequences in input with a much bigger compendium of datasets. While that study is not yet peer-reviewed, the authors can make their manuscript better by providing some insight on the differences between the other study and theirs to better inform the readers.

Thank you for your suggestions. We have included the following texts and these references in Introduction:

1. Mauger et al., 2019 (PMID: 31712433) and Ribotree: Previous studies have demonstrated that mRNA structure critically influences its stability *in vivo*[8] and *in solution*[9], as well as translation[17].

2. CodonBERT and ChatNT: Recent advances in codon optimization research have seen the emergence of deep learning-based algorithms, particularly large language models (LLMs) trained on cross-species nucleotide sequences. These models have been implemented for predictive modeling of mRNA translation efficiency and degradation kinetics[10,11]. Nevertheless, there persists an urgent demand for developing a rigorously validated optimization framework to improve mRNA-encoded protein expression specifically tailored for therapeutic applications.

3. RiboShape (PMID: 27307616), Riboformer (PMID: 38443396), RiboTIE (PMID: 38585907), and RiboNN：Recent studies have leveraged Ribo-seq to develop translation-focused deep learning models[12–14]. For instance, RiboNN predicts mRNA translation efficiency in mammalian cells by integrating mRNA sequences with ribosome profiling data, revealing translation-stability regulatory mechanisms[15]. However, the field critically requires a rational design framework that translates data-derived translational signatures into codon optimization strategies, enabling high-throughput exploration of sequence space and de novo generation of optimized mRNA constructs for therapeutic development.

Reviewers 4 and 5: We are satisfied with the authors' edits to the text.

Major Comment #2: The definition of 'translation level' is not clear and the relationship with mRNA abundance.

1. Definition of translation: The authors do not provide a clear definition of "translation level" in either the result or methods section. Is 'translation level' the RPKM from Ribo-Seq or is it 'translation efficiency' defined for each transcript, RPKM from Ribo-Seq divided by RPKM from paired RNA-Seq? authors need to explicitly specify the definition for better interpretability of the results.

Thanks for the comments and suggestions.
In our manuscript, the term "translation level" refers to "RPKM derived from Ribo-seq". We have added the following texts to the manuscript:

In Introduction: Ribosome profiling sequencing (Ribo-seq) is a powerful experimental technique that provides a snapshot of actively translating ribosomes on mRNA molecules, where the translation level of an mRNA can be derived from the reads per kilobase per million (RPKM) of Ribo-seq.

The first appearance in Results: T-distributed stochastic neighbor embedding (t-SNE) indicated that the model established an association between the sequence space and translation levels (similar to Ribo-seq-derived RPKM values).

Reviewers 4 and 5: We are satisfied with the authors' edits and clarification. We would recommend the authors add RPKM to plot axes labels or in the legend for better readability.

Dependence on mRNA Abundance
The definition of "translation level," based on RPKM from both Ribo-Seq and paired RNA-Seq, inherently incorporates mRNA abundance as both an input and a component of the target variable. While this approach is valid in deep learning due to the ability to model non-linear relationships, it raises the risk of overemphasizing mRNA abundance at the expense of other features, particularly given that mRNA abundance already accounts for the highest R2 among individual features. The ablation analysis indicates that the model performs slightly better than mRNA abundance alone, with codon sequence and cellular environment contributing incrementally to R2 (0.76 vs. 0.6). However, the authors should explicitly acknowledge in the Results or Discussion that the high contribution of mRNA abundance may stem from its dual role as both an input and a component of the target variable. Additionally, it would be valuable for the authors to clarify whether the model has been rigorously tested to ensure it captures meaningful relationships beyond mapping mRNA abundance to translation levels.

2. Relative Contributions of Transcription vs. Translation: The R2 of 0.6 for transcription dominance suggests that the model primarily learns transcriptional rather than translational regulation. The authors should address how this impacts the utility of the method for optimizing mRNA therapeutics, where translational rules may play a more critical role.

Thanks for the comments. As these two questions are related, we would like to respond them together. We acknowledge that mRNA abundance serves as both an input feature and is intrinsically linked to the target variable (Ribo-seq RPKM). This relationship likely contributes to its high predictive importance observed in the ablation analysis (Figure 2b).

However, several lines of evidence indicate the model captures meaningful biological relationships beyond simply mapping abundance levels, rigorously testing its capabilities:

1. Incremental Contributions: The model's predictive performance (R2) significantly improves with the addition of codon sequence and cellular context information (increasing R^2 by 0.15 and 0.06 respectively over abundance alone), demonstrating it learns from these distinct inputs.
2. Learning Complex Features: Incorporating established metrics like CAI did not improve the model's predictive accuracy. This suggests RiboDecode identifies more complex, inherent sequence features governing translation, rather than relying solely on simpler, predefined rules or just mapping abundance.
3. Experimental Validation: Crucially, the model's ability to generate sequences yielding substantial, experimentally validated increases in protein expression *in vitro* and enhanced therapeutic effects *in vivo* provides strong evidence that it learns and optimizes functional, sequence-based translational signals. This functional validation confirms it captures meaningful relationships beyond simply reflecting transcript abundance.
4. Context Specificity: The model also demonstrated the ability to design sequences for preferential expression in specific cellular contexts, requiring it to learn context-dependent translational nuances.

In summary, while the dual role of mRNA abundance influences its predictive weight, the incremental improvements from other features, the capture of complex sequence patterns, and, most importantly, the robust experimental validation of optimized outputs confirm that RiboDecode effectively learns and utilizes sequence-intrinsic translational information beyond merely mapping transcript levels.

We have added texts to Discussion in the revision acknowledging this potential influence.
While mRNA abundance serves as both an input feature and intrinsically linked to the Ribo-seq RPKM target variable, potentially explaining its high predictive weight in ablation analysis, the model still demonstrated significant improvements in prediction accuracy by integrating codon sequences and cellular context. Importantly, RiboDecode's in vitro and in vivo validation, where optimized sequences substantially enhanced protein expression and therapeutic efficacy, confirms its ability to optimize biologically meaningful translational signals rather than merely reflecting transcript abundance.

Reviewers 4 and 5: We are satisfied with the authors' explanation and the corresponding edits to the text.

Further, we recommend that the authors discuss their results in the context of the findings by Schwanhäusser et al. (PMID: 21593866), which demonstrated that post-transcriptional regulation at the level of translation played an equal or greater role in determining protein expression compared to transcriptional regulation. The current study, in contrast, emphasizes the predictive power of transcription-level measures (mRNA abundance). The authors should explore how this discrepancy impacts the interpretation of their results and the relative contributions of transcription and translation to protein expression. A deeper discussion of these points would significantly enhance the manuscript's clarity and broader relevance.

Thank you for this suggestion regarding Schwanhäusser et al.. We agree their work highlighted the major role of translation-level control in determining protein expression.

Our finding that mRNA abundance is the *strongest predictor* in our model reflects its prerequisite role as the template molecule, rather than diminishing the importance of translation efficiency itself. RiboDecode *does* explicitly consider translational control through its codon sequence and cellular environment inputs, trained on Ribo-seq data reflecting active translation. It does not, however, model post-translational regulation, as its predictions are based on ribosome occupancy prior to subsequent protein modifications or degradation.

Our experimental results, showing significantly enhanced protein expression and *in vivo* efficacy through optimizing these mRNA features, strongly validate the power of targeting the translational elements our model considers. Thus, our findings complement the Schwanhäusser study by demonstrating an effective method to leverage sequence-based translational control for enhancing protein production. We have added discussion points reflecting this context to the manuscript.

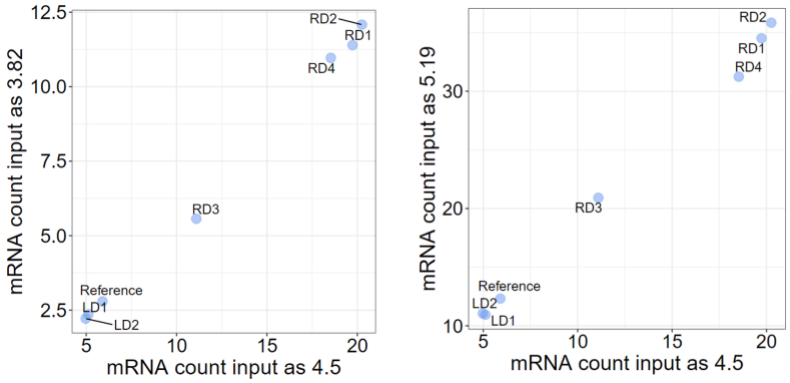
These findings also align with the broader framework of gene expression regulation: while mRNA abundance is a prerequisite for protein synthesis (consistent with its predictive dominance), our approach directly targets translational control, a critical layer highlighted by Schwanhäusser et al.43, by leveraging Ribo-seq-derived codon usage and cellular context to maximize translational efficiency. Although the model excludes post-translational events (as Ribo-seq captures ribosome activity prior to protein maturation), the achieved gains in protein output and in vivo efficacy robustly validate translational optimization as a key determinant of functional protein levels, complementing established regulatory hierarchies.

Reviewers 4 and 5: We are satisfied with the authors' explanation and the corresponding edits to the text.

3. Effect of mRNA abundance input for codon optimization: mRNA abundance is an input parameter for the translation prediction model and for every step of the codon optimizer, a constant mRNA abundance was set to 4.5 (line 647). Have the authors tested the performance of the codon optimizer by varying this input?

Thank you for this suggestion. We evaluated the effect of varying the mRNA abundance input on the translation prediction model, which informs the codon optimizer. As described in the Methods section, the input mRNA abundance is based on a log transformation (y=ln(RPKM×5+1)). Our default input value of y=4.5 corresponds to an RPKM of approximately 17.8.

Following the reviewer's suggestion, we tested the model using alternative abundance inputs corresponding to half (RPKM=8.9, y≈3.82) and double (RPKM=35.6, y≈5.19) the default RPKM value. We found that the predicted translation levels for sequences, when calculated using these different abundance inputs (y=3.82, 4.5, or 5.19), were highly correlated with each other. This indicates that while the absolute predicted translation score shifts with the abundance input, the relative scores between different sequences are largely preserved within this tested range. Since the codon optimizer primarily relies on these relative differences to guide sequence generation towards higher predicted translation, our results suggest that the optimizer's performance and the resulting optimized sequences are robust to reasonable variations in the input mRNA abundance value.



Comparation of translation prediction with different mRNA abundance (Pearson's correlation)

Reviewers 4 and 5: The updated figure/data, along with the authors' clarification, adequately addresses our concern.

4. Missing 5'UTR Context: Translation initiation is a limiting step, yet the model appears to lack information about the 5'UTR. If this is not incorporated, we think, its potential impact on translation prediction needs more discussion

Thanks for the suggestion, we added this in Discussion: " Firstly, we focused exclusively on optimizing the codon sequences while not explicitly modeling the 5'UTR. Recognizing the critical role of the 5'UTR in regulating translation initiation, our future work aims to expand RiboDecode to jointly optimize both UTRs and the codon sequence."

Reviewers 4 and 5: We are satisfied with the authors' edits to the text.

5. Importance Plot and Padding: The per-position importance plot may be biased due to 3' padding, as low importance at the 3' end could be an artifact. Additional analyses are needed to account for this.

For the results in Figure 2c, we calculated the importance scores based on the non-padding regions. We added the following content to the 'Nucleotide position contribution for translation prediction' in Methods: For the final analytical outcomes and visualization, the importance score assigned to each nucleotide position was calculated as the mean of absolute values derived from non-padding regions.

Reviewers 4 and 5: We are satisfied with the authors' edits to the text.

Major Comment #3: Incorporating MFE Optimization in the Discussion and Results

Some of the RD sequences designed for experimental validation incorporate an MFE optimization component (w>0), which is not thoroughly discussed in the manuscript. According to the methods (line 688), the weights for RD1, RD2, RD3, and RD4 are w=0,0,0.7,0.5, respectively, indicating that RD3 and RD4 were optimized for both translation level and lower MFE. This design choice warrants further clarification and discussion.

Specific Recommendations:
1. Explicit Labels for Weights (w) in Figures: Figures that compare RD sequences, particularly Figures 4, 5, and 6, should include explicit labels indicating the translation level and MFE weights (w) used for each sequence. This would improve the clarity and interpretability of the data.

Thank you for your suggestion. We have included the w parameters to the sequences in Figures 4, 5, and 6. For example:

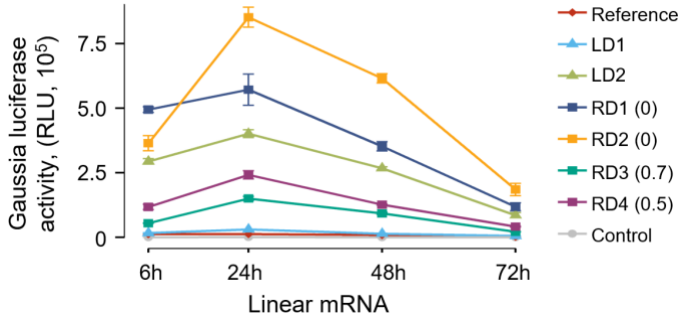


Figure 4a. Protein expression of Gluc was measured by fluorescence intensity. RD sequences were designed by RiboDecode, with w parameter indicated in parentheses…

In addition, we included the translation level and MFE weights (w) of Gluc, Fluc, HA and NGF in Table S3, S6, S7 and S8.

Reviewers 4 and 5: The updated figure/data, along with the authors' clarification, adequately addresses our concern. Additionally, it would be ideal to include AUC plots alongside the line plots to aid in the interpretation of the data.

2. Consistent Nomenclature Across RD Sequences: The order and naming of RD sequences across Figures 4, 5, and 6 are inconsistent. For example, the order of RD sequences based on w for HA differs from that of Gluc and NGF. Standardizing the nomenclature based on the weights would make data interpretation more intuitive and consistent.

We have renamed the HA sequences to make their name and w consistent with that of Gluc and NGF. For example:

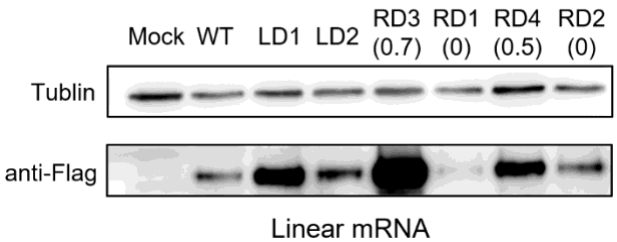


Figure 5a. Western blot analysis shows protein expression of the HA variants in HEK293T cells 24 hours after transfection. RD sequences were designed by RiboDecode, with w parameter indicated in parentheses. LD sequences were designed by LinearDesign.

Reviewers 4 and 5: The updated figure/data, along with the authors' clarification, adequately addresses our concern. Additionally, it would be ideal to include WB band quantification on the WB itself rather than Figure S15 for easier data interpretation.

3. Discussion of Results Based on Different w Values: The manuscript should explicitly discuss the varying performance of RD1 and RD2 (designed for translation only) compared to RD3 and RD4 (designed for both translation and stability). This is particularly relevant for circular RNAs, where RD3 and RD4 outperform RD1 and RD2 in protein expression. The authors should explore and discuss how optimizing for MFE contributes to higher protein expression in circular RNA contexts.

The observation that optimized codon sequences with lower MFE exhibit higher protein expression in circular RNA format aligns with a recent study (doi: https://doi.org/10.1101/2023.07.09.548293). The authors developed an optimization strategy for circRNAs and found that circRNAs with lower MFE demonstrated enhanced protein expression *in vivo*. The results suggested that this was likely because codon sequences with lower MFE facilitate the proper folding of the internal ribosome entry site (IRES), thereby enabling more efficient ribosome recruitment and translation. However, as we only experimentally tested a few numbers of circular RNAs in our study, we cannot validate the generalization of this statement. We discussed our recommended strategy of choosing *w* in Discussion. Please see the answer in the next question for details.

Reviewers 4 and 5: We are satisfied with the authors' explanation and the corresponding edits to the text.

4. Contrasting MFE and Translation-Only Models: While the Discussion notes that MFE did not have a general impact on translation, it fails to emphasize that MFE optimization significantly contributed to the performance of top-performing sequences in experimental validation. The authors should address this contrast and discuss how MFE-optimized sequences outperformed translation-only models in specific contexts.

Thanks for the suggestion, we added this in Discussion:
Second, our results indicate that while our primary goal was to enhance translation through codon optimization, incorporating MFE optimization played a complementary role in modulating mRNA secondary structure stability. However, we observed the best w value appears to vary depending on the specific mRNA and its context, suggesting that a one-size-fits-all approach may not be appropriate. Consequently, we recommend that experimental designs include the testing of multiple w values to identify the optimal balance for each mRNA target. Further research into the determinants of the optimal w value will be critical to refine this strategy and could lead to more systematic approaches for integrating MFE into mRNA design.

Reviewers 4 and 5: We are satisfied with the authors' edits to the text.

Major Comment #4: Experimental Validation: mRNA variants, cellular context, and different mRNA formats.

1. Luciferase Feature Analysis (Fig 3d): The two luciferase proteins exhibit opposite CAI importance. This variability requires further explanation, especially in relation to their sequence features.

The observed difference in CAI trends between Gluc and Fluc underscores this context-dependence. Potential contributing factors could include differences in their specific amino acid compositions, overall sequence lengths, or resulting structural elements, which interact differently with the cellular translation machinery and influence which codons are optimal in each specific case. Furthermore, we found that CAI,

as well as CPB, GC% and MFE, influenced translation but be more mRNA specific. It may be because of sequence or structure feature changes.
In the "RiboDecode's Optimization Strategies for Enhanced mRNA Translation" section of Results, we discussed this: 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, due to complex sequence or structure feature changes.

Reviewers 4 and 5: We are satisfied with the authors' explanation and the corresponding edits to the text

2. Cellular Context Claims (Fig 4): The difference between the predicted and experimental ratios for the HEK293 vs other cellular contexts is stark. This undermines the claim that the model learns cellular context. The authors should refine this claim or redesign sequences to better reflect cellular differences. Second, a barplot with error bars would be a better representation for this data.

Thanks for the suggestions. Regarding the discrepancy between the experimental and predicted ratios. We revised the texts in the revision: We next evaluated RiboDecode's ability to design for cellular context by optimizing Gluc mRNA for preferential expression in HEK293T cells over A549 and ARPE19 cells. The designed variants successfully exhibited the intended higher expression in HEK293T compared to both cell lines (Figure 4b, Table S4). Predicted expression ratios favoring HEK293T closely matched experimental results in the comparison against A549 (~1.7-fold predicted vs. ~1.55-fold experimental). Preferential expression was also achieved against ARPE19 (Figure 4b, Table S5), although the experimental fold-change (~2.8-fold) doubled the prediction (~1.4-fold). These results confirm RiboDecode's capability for context-aware design, while the discrepancy in the ARPE19 comparison indicates potential for refining the model to better capture quantitative differences across diverse cellular environments.

In addition, we have toned down the cellular context claims in multiple places in the manuscript (in response to the Major comments #1, Question 8).

Also, per the suggestion, we included error bars in the plots. In addition, we combined the three experiments in each cell line and showed two cell lines separately, which gives a more intuitive visualization. We provided the predicted values and discussion on the texts as above.

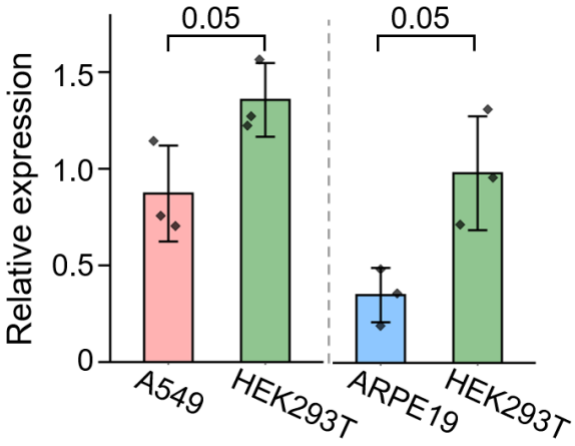


Figure 4b. Relative expression of experimentally measured protein expression values of mRNA variants designed by RiboDecode at 24 hours in different cells. The error bars denote standard variation. One-sided Wilcoxon test was used to calculate p-values shown in the figure.

Reviewers 4 and 5: We are satisfied with the authors' explanation and the corresponding edits to the text as well as the updated figure. However, we would like to draw attention to a recurring phrase in several figure captions: "standard variation." If this refers to a novel metric, we kindly ask the authors to

clarify its definition. Otherwise, we assume this is intended to be “standard deviation” and request that the terminology be corrected accordingly.

3. Performance on Modified Chemistries: With m1Ψ-modified sequences, the performance drops to a 4.6-fold improvement from 72-fold. This suggests that the model does not effectively capture rules for modified chemistries and should be discussed.

Thank you for this observation regarding the performance with m1Ψ-modified sequences.

We acknowledge that the fold-improvement for the optimized Gluc sequence compared to its reference was lower when using m1Ψ modification (up to 4.6-fold) compared to the unmodified linear version (up to 72-fold). There are two primary factors contributing to this difference in *relative* improvement:

1. Higher Baseline Expression of the m1Ψ Reference: As shown in Figure 4a and 4e, the m1Ψ modification itself significantly boosted the baseline protein expression of the Gluc reference sequence compared to the unmodified version. Since the starting expression level for the m1Ψ reference was already considerably higher, the subsequent fold-improvement achieved through RiboCode optimization appears smaller in relative terms.
2. Model Training Data: The RiboCode model was trained on large-scale Ribo-seq datasets derived from endogenous, *unmodified* mRNAs. Therefore, the sequence rules it learned primarily reflect the translation dynamics associated with standard nucleotides and do not explicitly capture rules potentially unique to m1Ψ incorporation, as well as circular RNAs.

We agree that the specific interplay with modified chemistries warrants further investigation, as mentioned in the discussion.

We acknowledged this limitation in Discussion: Finally, the current model was trained exclusively on Ribo-seq data from endogenous, unmodified mRNAs. Although our results showed significant expression enhancements for optimized sequences in both m1Ψ-modified and circular forms compared to their respective controls, the relative fold-improvement compared to the unmodified format varied between constructs. Future iterations could potentially incorporate data from modified and circular transcripts to further refine optimization rules.

Reviewers 4 and 5: We are satisfied with the authors’ explanation and the corresponding edits to the text.

4. Circular mRNA in Fig 6: Circular mRNAs perform better than linear in Figure 6 but not in previous figures. This discrepancy should be explained.

Thank you for this comment regarding the performance of circular mRNAs. Our data consistently show optimization improves expression *within* the circular format for Gluc, HA, and NGF compared to controls. However, the figures do not provide a systematic comparison of *relative* performance between linear and circular formats across these different genes, making it difficult to confirm the suggested discrepancy. Any variations in how circularization impacts expression relative to linear formats could potentially stem from gene-specific factors, such as differences in inherent mRNA stability, translation efficiency in the circular form, or interactions with the cellular machinery. Currently, our data do not support definitive conclusions about discrepancies in the relative performance of circular vs. linear formats across genes. This aspect is acknowledged in the Discussion (please see above question for this discussion).

Reviewers 4 and 5: We thank the authors for their explanation and appreciate their edits in the Discussion section.

Major Comment #5: Missing/additional controls

1. Reference Sequence CAI: The reference Gluc sequence has a CAI of 0.8, Mauger et al., 2019 showed to have very low expression. This may undermine the claim of a 72-fold improvement. A higher CAI (~1.0) control should be included.

Our original reference Gluc had a CAI of 0.8. We found that the Gluc-LD2, designed by LinearDesign, exhibits a high CAI (0.97) but demonstrated lower expression levels compared to our designed sequences, RD1 and RD2. This indicates that a high CAI was not necessarily correlated with high expression. However, as the reviewer suggested, to avoid potential overinterpretation, we did not specify 72-fold change in expression in the revised manuscript, from "72-fold increase" to "significant increase".

Reviewers 4 and 5: We are satisfied with the authors' explanation and the corresponding edits to the text

2. Fluc Performance (Fig S8): The 17-fold improvement for Fluc (vs. 72-fold for Gluc) suggests the model's translation rules are not well-generalized. Sequence feature comparisons could be performed between the RD and reference sequences for these 2 related proteins to explain this discrepancy. Missing CAI control here as well.

Thank you for the comments regarding the Firefly Luciferase (Fluc) performance and comparison to Gaussia Luciferase (Gluc).

1. Differing Fold-Improvement & Generalization: We acknowledge the difference in maximum fold-improvement observed between Fluc and Gluc. We noted the magnitude of fold-improvement can be highly dependent on the specific wild-type sequence's baseline expression and intrinsic properties. Given that RiboDecode performed well across multiple distinct proteins (Gluc, Fluc, HA, NGF), we believe the learned rules show reasonable generalization, although the achievable fold-increase varies between targets. However, As previously addressed, we have recognized that the 72-fold augmentation in Gluc expression may be attributable to the diminished expression of the reference sequence, and as such, we will refrain from highlighting this number further.

2. Sequence Feature Comparison: Figure 3d in the manuscript provides a high-level comparison of feature changes upon optimization for several proteins, including Fluc. As discussed in the manuscript, the optimal balance of features like CAI, MFE, U%, etc., appears highly gene-specific. The contrasting results between Gluc and Fluc underscore this context-dependency, suggesting a simple feature comparison might not fully capture the complex interplay determining translation efficiency for each specific sequence.

3. CAI Control & Additional Fluc Data: We appreciate the comment regarding controls and have performed the suggested additional experiments comparing WT Fluc with sequences optimized by both RiboDecode (RD1-RD4) and LinearDesign (LD1, LD2), with results shown in the figure you provided.
    - All optimized sequences significantly increased Fluc activity compared to WT.
    - The LinearDesign sequence LD2 (high CAI = 0.952) yielded the highest expression, achieving approximately a 41-fold increase over WT at 24h. LD1 (moderate CAI = 0.766) showed around a 7-fold improvement.
    - RiboDecode sequences (with lower CAIs ~0.71-0.73) achieved substantial improvements ranging from 6-fold to 16-fold over WT (e.g., RD4 achieved ~16-fold).
    - These results indicate that for Fluc, high CAI correlates well with high expression. However, RiboDecode, optimizing for complex features learned from Ribo-seq data rather than primarily maximizing CAI, still yielded significant (up to 16-fold) improvements. This contrasts sharply with our Gluc results, where high CAI correlated negatively with expression and RiboDecode significantly outperformed LinearDesign.

Conclusion: The additional experiments confirm RiboDecode effectiveness for Fluc, while the differing optimal strategies (high CAI best for Fluc, RiboDecode complex optimization best for Gluc) strongly underscore the gene-specific nature of mRNA optimization. This highlights the value of context-dependent, data-driven approaches like RiboDecode that can capture complex sequence features beyond single metrics like CAI.

We have updated the manuscript reflecting the new experiments in Results:

We additionally optimized another commonly used reporter gene, firefly luciferase (Fluc). Seven sequences, including four RiboDecode-optimized (RD1-RD4), two LinearDesign-optimized (LD1, LD2) and a WT, were transfected into HEK293T cells, and activity was measured over 72 hours (Figure S12, Table S6). All optimized sequences significantly outperformed the WT. The LD1 (CAI of 0.766) and LD2 (CAI=0.952) yielded the increased expression with about 7-and 41-fold changes, respectively. RiboDecode sequences (CAI of ~0.71-0.73) also achieved substantial improvements with 6- to 16-fold over the WT. The superior performance of the LD2 sequence with a high CAI value here—contrasting with results for Gluc—underscores that the optimal codon strategy is gene-specific. This highlights the need for context-dependent approaches like RiboDecode that capture complex features beyond single metrics like CAI.

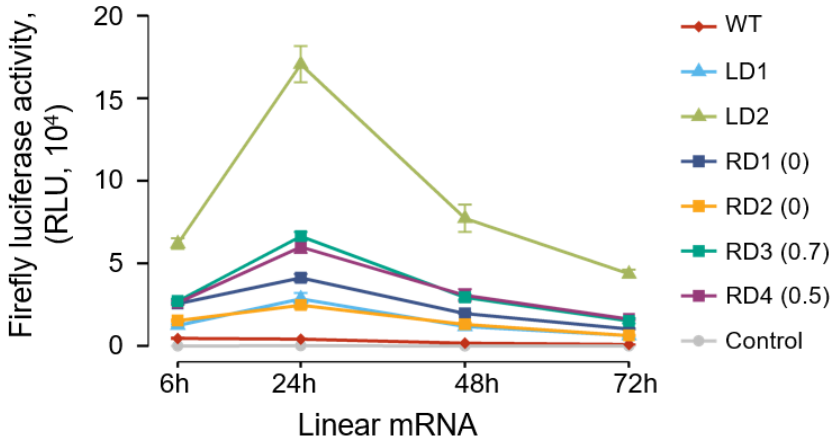


Figure S12. Fluc protein expression in transfected HEK293T cells. Protein expression was determined by fluorescence intensity. RD1, RD2, RD3, and RD4 were designed using RiboDecode, with the *w* parameters indicated in brackets. The detailed information of the sequences is provided in Table S6. "RLU": relative light unit.

Gene	Sequence ID	Parameter	MFE	CAI
Fluc	Reference	NA	-216.4	0.808
Fluc	RD1	w = 0	-195.7	0.713
Fluc	RD2	w = 0	-195.3	0.703
Fluc	RD3	w = 0.7	-342.4	0.748
Fluc	RD4	w = 0.5	-258.4	0.736
Fluc	LD1	λ=0	-346.2	0.766
Fluc	LD2	λ=4	-302.5	0.952

Table S6. This table presents sequence features for various Fluc mRNA constructs, including the wild-type sequence, RiboDecode-generated variants (RD1-RD4 with different w values), and LinearDesign-generated variants (LD1-LD2).

Reviewers 4 and 5: We thank the authors for their explanation and the corresponding edits to the text. We agree with the authors that "…optimal codon strategy is gene specific." While we agree that context-aware approaches like RiboDecode are valuable for capturing complex features, simpler metrics such as CAI may still be effective for certain genes (e.g., Fluc in this manuscript). So, instead of "…may even lead

to counterproductive optimization strategies….", we would recommend use "ineffective" (something to this effect) to better present this result in context of data from wider literature.

3. Dose Response (Fig 5): Unlike Figure 6, Figure 5 lacks a dose-response experiment. Including this would strengthen the conclusions.

Thanks for the comments. You are correct that Figure 5 does not include a dose-response experiment for the HA vaccine, whereas Figure 6 presents data for NGF at two different doses.

The rationale for this difference lies in the distinct goals and biological contexts of each application:

1. NGF Therapeutic (Figure 6): The goal was *local protein replacement therapy* in the retina. NGF exerts its therapeutic effect directly within the retina where it is expressed following intravitreal injection. Therefore, demonstrating that optimized mRNA could achieve a comparable therapeutic effect (neuroprotection) with lower *local* protein levels (resulting from a lower dose) was essential to show enhanced potency and dose-sparing potential in the target tissue.
2. HA Vaccine (Figure 5): The goal of the vaccine, administered intramuscularly, is to generate a *systemic immune response*. While the HA antigen is expressed locally (primarily in muscle and draining lymph nodes) to initiate this response, the desired outcome (protective immunity) is systemic and measured via antibody titers in the blood. High local concentration of the antigen is not the therapeutic endpoint itself. For this proof-of-concept study, demonstrating that RiboDecode optimization significantly boosted the systemic immune response (~10-fold higher neutralizing antibodies) at a single, relevant dose (10 µg) was sufficient to validate the optimization strategy's effectiveness for enhancing immunogenicity. This 10µg dose is a relevant dose level used within the range reported for preclinical influenza mRNA vaccine studies in mice (PMID: 30135514, 32164372). We have added these references into the manuscript and changed the text in Methods: BALB/c mice received two intramuscular doses of 10 µg mRNA each, with the dose determined by previous studies[73,74], administered on day 0 (prime) and day 14 (boost).

Therefore, the experiments were designed differently to best address the primary question for each application (dose-sparing potential for NGF therapeutic vs. enhanced immunogenicity for HA vaccine) within the scope of this initial validation study.

Reviewers 4 and 5: We are satisfied with the authors' explanation and the corresponding edits to the text.

4. Missing Controls (Fig. 5 and 6): No codon-optimized (e.g., high CAI) sequences were used as reference. As Kudla et al., (PMID: 16700628) have demonstrated that WT sequences with low GC3% usually do not express well.

Thank you for this comment. We would like to clarify the sequences included as controls in Figures 5 and 6. In Figure 5, the HA-LD2 sequence had a CAI of 0.96 and GC3% of 88%. In Figure 6, the NGF-LD2 sequence had a CAI of 0.95 and GC3% of 82%. These sequences can effectively serve as the requested 'codon-optimized. These LD2 sequences, along with our RiboDecode sequences, consistently outperformed the wild-type (WT) sequences, which likely aligns with the expectation from Kudla et al. that WT sequences with potentially lower GC3% might exhibit lower expression.

Reviewers 4 and 5: We are satisfied with the authors' explanation and the corresponding edits to the text. We think referring to the conclusions from Kudla *et al*. (*PMID: 16700628)* in their manuscript will add some context to these results. So, we recommend adding it as a reference in this manuscript.

Minor Comments

Minor Comment #1: Line 74-76. MFE is discussed only in relation to mRNA stability. Lower MFE/higher mRNA structure can also be a hurdle for translation elongation and needs to be acknowledged here.

Thank you for your reminder. We emphasized that MFE also affects translation in Discussion: …incorporating MFE optimization played a complementary role in modulating mRNA secondary structure stability and translation.

Reviewers 4 and 5: We thank the authors for their explanation and the corresponding edits to the text. However, we recommend clarifying what is meant by the term "complementary." Are the authors referring to a secondary or supporting role?

Minor Comment #2: Lines 91-95. The authors can be more specific and provide some examples of any trans-factors that affect translation that may not be captured by traditional methods.

Translation factors (for example, eIF4F and eIF2), RNA-binding proteins, tRNAs, and ribosomal RNAs could regulate mRNA translation (PMID: 34341537). However, predefined sequence features cannot capture the expression profiles of these factors in cells or their interactions with mRNA sequences.
We added the following content to the Introduction: Secondly, the existing methods do not adequately account for the activity of translational regulators that influence mRNA translation, such as translation factors and RNA-binding proteins.

Reviewers 4 and 5: We are satisfied with the authors' explanation and the corresponding edits to the text

Minor Comment #3: Predicted translation values below zero in Figure S4 suggest a model error. The authors should ensure that predictions are constrained to valid ranges.

The y-axis in Figure S4 represents the predicted values after log transformation, which is why negative values are present.

Reviewers 4 and 5: We are satisfied with the authors' explanation.

Minor Comment #4: The r=0.7 correlation in Figure S7 is derived from two distinct groups of points, making it a poor representation of the data. Alternative visualizations or metrics should be considered.

We acknowledge the reviewer's point that the data points in the scatter plot comparing RiboDecode's translation prediction versus experimental expression (Figure S7, left panel) appear somewhat clustered. We also agree that Pearson's correlation coefficient (R), while standard, can be influenced by such structures.

We acknowledged the shortcomings of p-value in Results: The predicted translational levels showed a positive correlation with the experimentally measured protein levels (Correlation coefficient=0.71, p-value=0.077, Pearson's correlation, Figures 4b and S7). However, the p value is marginal, potentially due to the small number of constructs tested.
In addition, a purpose of Figure S7 was to directly compare the predictive performance of RiboDecode against a traditional metric, CAI, using the same experimental data. The figure visually demonstrates that RiboDecode's predictions exhibit a positive trend with experimental expression (R=0.71, p=0.077), whereas CAI shows essentially no correlation (R=-0.072, p=0.88) for these sequences. This comparison highlights the improved predictive capability of our data-driven approach over the CAI metric in this context.

Reviewers 4 and 5: We thank the authors for their explanation and the corresponding edits to the text.

Minor Comment #5: Figure 5a requires bar plots quantifying the Western blot signals for a more accurate comparison.

Thanks for the suggestion. We have added a bar plot and text in the manuscript: Three out of four RiboDecode-optimized HA sequences and two LinearDesign-optimized sequences showed higher in vitro protein expression compared to the WT (Figures 5a, Table S7). Particularly, RD3 showed approximate 5-fold increase compared to the WT and LinearDesign-optimized sequences (Figure S15).

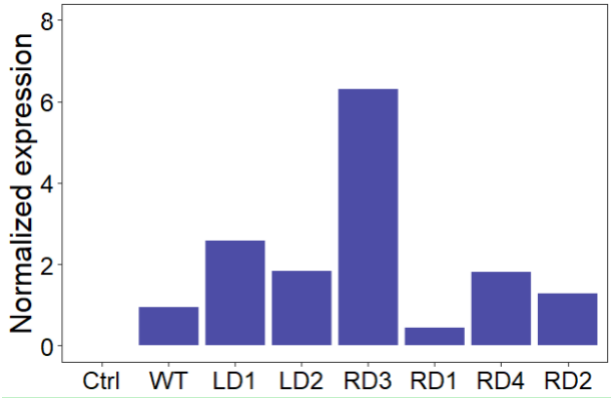


Figure S15. The gel bands of the Western blot in Figure 5a were quantified using GelAnalyzer. The expression values were normalized by Tublin.

Reviewers 4 and 5: We are satisfied with the authors' explanation and the corresponding edits to the text

Minor Comment #6: Lines 1050-1051, 1065-1066, 1081-1082. The w values for the RD sequences need to be shown either in the figure legends or at least in the figure captions.

Thank you for your suggestion. We have assigned the w parameters to the sequences in the Figures.

Reviewers 4 and 5: We are satisfied with the authors' explanation and the corresponding edits to the text. We would recommend to also add the λ parameter for LD designs so readers have a better idea about LD1 and LD2 whether they have higher weight for CAI or MFE.

Minor Comment #7: Fix grammatical errors –

1. Line 386: …to several factors.
2. Line 392: …mRNA translation whereas the previously codon optimization…

Thank you for pointing out these issues and we have corrected them.

Reviewers 4 and 5: We are satisfied with the authors' explanation and the corresponding edits to the text

Reviewer #5 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #5 (Remarks on code availability):

The code is accessible as a .whl file. The code can run the model with given parameters. I did encounter an issue with installation that I state below.

The instructions for installation should be more clear in README file. The run.py has a 'python' command. It is increasingly common for modern Linux distributions and macOS to have only python3 installed and not python. The following error was encountered which needed some troubleshooting.

Traceback (most recent call last):
File "~/.local/bin/ribo-code", line 8, in <module>
sys.exit(main())
File "~/.local/lib/python3.8/site-packages/ RiboCode/run.py", line 24, in main
result = subprocess.run(command)
File "/usr/lib/python3.8/subprocess.py", line 493, in run
*with Popen(*popenargs, **kwargs) as process:*
File "/usr/lib/python3.8/subprocess.py", line 858, in __init__
self._execute_child(args, executable, preexec_fn, close_fds,
File "/usr/lib/python3.8/subprocess.py", line 1704, in _execute_child
raise child_exception_type(errno_num, err_msg, err_filename)
FileNotFoundError: [Errno 2] No such file or directory: 'python'

I had to specify sudo ln -s /usr/bin/python3 /usr/bin/python or use alias python='python3' in my ~/.bashrc to avoid this error.

The authors should either replace python with python3 in their run.py or provide more explicit instructions.

Thank you for your feedback. We have replaced ‘python’ with ‘python3’ and uploaded the package.

Reviewers 4 and 5: We thank the authors for their edit to the code.