

1 RNASSTOP: A Deep Learning Framework for mRNA Chemical

2 Stability Prediction and Optimization

3 Shenggeng Lin (林圣庚)^{1,2,3}, Junwei Chen (陈俊炜)^{1,3}, Heqi Sun (孙鹤淇)¹, Yufang
4 Zhang (张育芳)¹, Wenyi Yang (杨文艺)⁴, Hong Tan (谭洪)^{1,3}, Dong-Qing Wei (魏冬
5 青)¹, Qinghua Jiang (蒋庆华)^{4,5,*}, Yi Xiong (熊毅)^{1,2,3,*}

6 ¹ State Key Laboratory of Microbial Metabolism, School of Life Sciences and
7 Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China

8 ² Shanghai Innovation Institute, Shanghai 200231, China

9 ³ Artificial Intelligence Biomedical Center, Zhangjiang Institute for Advanced Study,
10 Shanghai Jiao Tong University, Shanghai 200240, China

11 ⁴ School of Life Science and Technology, Harbin Institute of Technology, Harbin 150001,
12 China

13 ⁵ School of Interdisciplinary Medicine and Engineering, Harbin Medical University,
14 Harbin 150081, China

15
16 * Corresponding authors: xiongyi@sjtu.edu.cn (Xiong Y), qhjiang@hit.edu.cn (Jiang
17 Q).

18
19 **Running title:** Lin S et al / RNASSTOP: mRNA Stability Prediction

20
21 Main text word count (from Introduction to Discussion): 4166

22 Number of references: 34

23 Number of figures: 6

24 Number of tables: 0

25 Number of supplementary figures: 11

26 Number of supplementary tables: 1

27 Number of supplementary files: 1

28 Character count of the article title (including spaces): 90

29 Character count of the running title (including spaces): 48

30 Number of keywords: 5

31 Abstract word count: 146

Abstract

Messenger RNA (mRNA) vaccines offer promising therapeutic potential for combating various diseases, yet their inherent chemical instability hampers long-term efficacy. Although several methods have been developed to predict mRNA degradation, they often suffer from limited accuracy and lack the capability for rational codon optimization. Here, we propose RNASTOP, a novel framework that integrates deep learning with heuristic search to simultaneously predict and optimize mRNA chemical stability. RNASTOP achieves a 13% improvement in accuracy over the top-performing model on the Stanford OpenVaccine competition dataset and demonstrates robust generalization in predicting full-length mRNA degradation. When applied to mRNA codon optimization, RNASTOP reduces the minimum free energy of the varicella-zoster virus vaccine sequence by 75.73% while maintaining high translation efficiency. Overall, RNASTOP serves as a powerful tool for predicting and optimizing mRNA chemical stability, accelerating the development of mRNA therapeutics. The source code of RNASTOP is available at <https://github.com/xlab-BioAI/RNASTOP>.

KEYWORDS: mRNA; Chemical stability; Codon optimization; Deep learning; Heuristic search.

Introduction

Messenger RNA (mRNA)-based therapeutics have emerged as highly promising medical interventions [1–3], exemplified by the rapid deployment of mRNA vaccines [4–6]. However, the inherent chemical instability of mRNA poses a significant limitation to the durability of these therapeutics [7–9]. Hydrolysis within lipid nanoparticle formulations diminishes the amount of intact mRNA remaining throughout shipping and storage, while *in vivo* degradation subsequent to vaccine administration limits protein synthesis and compromises therapeutic efficacy [2,10]. Developing mRNA therapeutics with enhanced chemical stability is therefore critical for mitigating degradation during transit and improving overall clinical efficacy [10–12].

Codon optimization has proven to be a highly effective strategy for engineering chemically stable mRNA therapeutics [3,9,13–18]. Current computational optimization methods generally fall into four categories: traditional statistical approaches, heuristic search strategies, deep learning-based models, and multi-factor integration tools.

Traditional approaches, such as host-preferred codon substitution, replace codons based on host frequency statistics. However, they prioritize translation efficiency while often compromising mRNA chemical stability. Heuristic search-based methods, such as LinearDesign [3], utilize dynamic programming to search for chemically stable nucleotide sequences by translating key features like minimum free energy (MFE) into computable scores. Deep learning-based models [17,18] are trained on massive natural sequences to capture underlying evolutionary information and intricate sequence patterns, thereby generating novel sequences not found in nature. Multi-factor integration tools, such as the commercial software OptimumGene (GenScript), perform global computational optimization by integrating multiple factors, including codon preference, mRNA secondary structure, and GC content, to generate optimized mRNA sequences.

Despite significant algorithmic advancements, identifying mRNA sequences with enhanced chemical stability within the vast synonymous sequence space remains a formidable challenge [19]. For example, a viral spike protein antigen can be encoded by approximately 10^{633} potential mRNA sequences [3]. When utilizing heuristic search algorithms to navigate this massive sequence space, the mRNA degradation scoring function becomes crucial, as it dictates the trajectory of the optimization process [3]. A lower degradation score correlates with enhanced chemical stability [20,21]. However, current predictive models that serve as these essential scoring functions exhibit limited accuracy and lack the capability to guide rational codon optimization [9]. Early predictive models overly simplified the process, postulating that the cleavage probability of an mRNA linkage is strictly proportional to the unpairing probability of the 5' nucleotide [20], neglecting the highly variable local sequence and structural contexts. Although recent deep learning architectures, such as Nullrecurrent [9] and RNAdegformer [19], have advanced predictive capabilities, they have not yet enabled practical sequence design guided by prediction. Therefore, there is an urgent need for an integrated framework capable of both highly accurate prediction and practical codon optimization.

In this study, we introduce RNASSTOP (Prediction and Optimization of mRNA chemical STability), a framework that can accurately predict mRNA degradation at both single-nucleotide and full-length levels and enhance mRNA chemical stability through codon optimization. On the Stanford OpenVaccine Kaggle (SOVK) competition dataset [9], RNASSTOP achieved a 13% improvement in degradation prediction accuracy over

the state-of-the-art model. Furthermore, its robust generalization capability was validated on an independent dataset of full-length mRNAs [22], where RNAS TOP exhibited the highest correlation with experimental degradation values among all evaluated models. Crucially, RNAS TOP provides biological interpretability by intrinsically capturing the sequence and structural patterns that govern degradation. Finally, RNAS TOP was applied to optimize BNT-162b2 and varicella-zoster virus (VZV) mRNA vaccine sequences, successfully reducing their MFE by 20.96% and 75.73%, respectively. Collectively, these findings demonstrate the utility of RNAS TOP in advancing rational vaccine design and mRNA therapeutics.

Method

Overview of the RNAS TOP framework

RNAS TOP is a deep learning-based framework designed to accurately predict mRNA degradation and enhance chemical stability. As illustrated in **Figure 1**, RNAS TOP primarily consists of an mRNA degradation prediction model (Figure 1A) and an mRNA codon optimization module (Figure 1B). The mRNA degradation prediction model primarily comprises three modules: the mRNA feature embedding module, the dual-branch feature decoupling-and-aggregating network, and the prediction layer. Specifically, the predictive model integrates nucleic acid large language models (LLMs) and a structural embedding module to extract multi-source feature embeddings from mRNA sequences. These embeddings are then processed by a dual-branch feature decoupling-and-aggregating network for feature fusion. Finally, the fused feature embeddings are fed into the prediction layer to predict mRNA degradation. The mRNA codon optimization module couples this predictive model with a beam search algorithm. It utilizes both the MFE and the codon adaptation index (CAI) as optimization criteria to concurrently improve mRNA chemical stability and translation efficiency.

Dataset

Stanford OpenVaccine Kaggle competition dataset

The mRNA sequences utilized in the SOVK competition were sourced from the Eterna platform [9]. The dataset comprises 3029 mRNA sequences of length 107 nt and 3005 mRNA sequences of length 130 nt. For the 107 nt sequences, the competition organizers provided nucleotide-level degradation profiles for the first 68 nucleotides under four

accelerated degradation conditions, alongside reactivity profiles obtained via the In-line-seq method [22]: deg_pH10, deg_Mg_pH10, deg_50°C, deg_Mg_50°C, and reactivity. For the 130 nt sequences, degradation profiles for the first 102 nucleotides were provided under two conditions (deg_Mg_pH10, deg_Mg_50°C) along with reactivity profiles. To ensure data quality, these 6034 mRNA sequences were filtered based on three criteria (Note 1 in File S1). Following this screening process, 2218 sequences of length 107 nt and 1172 sequences of length 130 nt were retained. These 3390 high-quality sequences were subsequently partitioned into a training dataset, a public test dataset, and a private test dataset (Figure S1).

The SOVK competition aimed to develop models capable of accurately predicting three specific target types: deg_Mg_pH10, deg_Mg_50°C, and reactivity, given the mRNA sequences and their secondary structures as inputs (Figure 1A) [9]. Model performance was evaluated using the Mean Column Root Mean Squared Error (MCRMSE) across the three target types, defined as:

MCRMSE

$$= \frac{1}{N_t} \sum_{j=1}^{N_t} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_{ij} - \hat{y}_{ij})^2} \quad (1)$$

where N_t is the number of scoring data types, n is the number of nucleotide linkages in the dataset, y_{ij} represents the experimental ground-truth value, and $\hat{y}_{ij}$ denotes the predicted value.

Full-length mRNA dataset

To assess the model's generalization capability in predicting overall degradation of full-length mRNAs, we utilized an independent dataset comprising 188 mRNAs encoding four classes of protein targets: a short multi-epitope vaccine, the model protein nano-luciferase (with one class consisting of varied untranslated regions and a second consisting of varied coding sequences), and enhanced green fluorescent protein [22]. The overall degradation values for these full-length mRNAs were experimentally characterized using the PERSIST-seq method.

The mRNA feature embedding module

The ViennaRNA Package [23] was used to compute the loop type, secondary structure, and base-pairing probability (BPP) matrix for each mRNA sequence. Secondary

structures, represented in “dot-bracket” notation, were subsequently converted into adjacency and distance matrices (incorporating both first-order and second-order distances). For sequence embedding, RNASTOP incorporates representations from two nucleic acid LLMs: RNA-FM [24] and DNABERT [25] (Notes 2–4 in File S1). Loop types were embedded using a standard PyTorch embedding layer. Additionally, a matrix embedding module was employed to process the BPP, adjacency, and distance matrices (Note 5 in File S1).

The dual-branch feature decoupling-and-aggregating network

The dual-branch feature decoupling-and-aggregating network [26] consists of a feature decoupling module and a feature aggregating module (Figure 1A). The feature decoupling module contains a shared encoder and four dual-branch private encoders. The shared encoder is structured as a stack of four transformer encoder blocks (Figure 1C). Each private encoder consists of a global encoder and a local encoder. The structure of the global encoders is identical to that of the shared encoder. The local encoder utilizes Conv1d, BatchNorm1d, ReLU activation, dropout, and ConvTranspose1d layers (Figure 1C). The feature aggregating module includes a global feature fusion module, a local feature fusion module, and a multi-feature fusion layer. The global feature fusion module has the same structure as the shared encoder. The local feature fusion module and the multi-feature fusion layer both have the same structure as the local encoder.

Model training

The RNASTOP model was trained using the Adam optimizer [27] coupled with a cosine annealing learning rate scheduler. To maximize data utilization and enhance the diversity of training samples, we retained 811 mRNA sequences (length 107 nt, SN_filter = 0) that contained noisy labels. Early stopping was implemented to halt training if model performance on the public test dataset failed to improve for 50 consecutive epochs. Following hyperparameter tuning (Note 6 in File S1; Figure S2), we selected a learning rate of 4.0×10^{-4} , a batch size of 64, a convolutional kernel size of 7, a dropout rate of 0.3, an embedding dimension of 128, and 4 transformer encoder blocks.

Estimation of overall full-length mRNA degradation and sequence similarity

To estimate the overall degradation of full-length mRNAs, we aggregated the single-nucleotide degradation values predicted by RNASTOP into a mean degradation score:

$$k_{deg}^{mRNA} = \frac{\sum_{i=1}^L k_{deg}^i}{L} \quad (2)$$

where k_{deg}^i is the degradation of nucleotide i , and L is the length of the mRNA. The model's performance was assessed by calculating the correlation coefficient between k_{deg}^{mRNA} and the experimentally measured full-length mRNA degradation values. In addition, ten alternative strategies for aggregating single-nucleotide predictions were systematically evaluated (Note 7 in File S1; Figure S3).

Sequence similarity between mRNAs was quantitatively evaluated using a local sequence alignment approach, formulated as:

$$Similarity = \frac{Score_{local}(seq1, seq2)}{\max(len(seq1), len(seq2))} \quad (3)$$

where $Score_{local}$ represents the optimal local alignment score computed using the *pairwise2.align.localxx* algorithm from the Biopython library. Identical nucleotides at aligned positions were awarded one point, with no penalties applied for mismatches or gaps.

The mRNA codon optimization module

RNASTOP employs a beam search algorithm for mRNA codon optimization, which consists of four main steps (Figure 1B).

Codon degradation prediction

The predictive model calculates a degradation value for each codon in the sequence. This value is the sum of the degradation values of the codon's three constituent nucleotides.

Codon degradation ranking

The codons are sorted in descending order based on their degradation values to identify the top K codons most susceptible to degradation.

Synonymous codon substitution

Synonymous mutations are introduced to the top K codons to generate mutant sequences. Because only single-codon mutations are made, each new sequence differs from its parent by exactly one codon. Mutant sequences that show improved biophysical properties compared to their parent sequence are retained as candidates.

Candidate sequence selection and iteration

Candidate sequences are evaluated and ranked based on a comprehensive scoring function:

$$\begin{aligned} \text{score} \\ = -\frac{MFE}{L} \\ + CAI \end{aligned} \quad (4)$$

where L is the length of the mRNA sequence. The K sequences with the highest scores are selected for the subsequent iteration. This iterative process continues until the score converges and no further improvement is observed. The sequence with the highest final score is considered the optimal mRNA design. Following the study by Huang et al. [3], we set the beam width K to 500. Additionally, the performance of a strictly degradation-based scoring function was also evaluated for comparison (Note 8 in File S1; Figure S4; Table S1).

Results

RNASTOP is a strong predictor of mRNA degradation

The performance of the RNASTOP model in predicting mRNA degradation at single-nucleotide resolution was evaluated on the SOVK competition dataset, which comprises both public and private test datasets [9]. In this competition, 1636 teams submitted 35,806 models, scored using the MCRMSE. Among these models, the Nullrecurrent model (**Figure 2A**) ranked first, followed by the Kazuki2 model [9]. The DegScore model [22] served as the competition's baseline (Note 9 in File S1). RNASTOP outperformed all 35,806 models on both the public and private test datasets, as illustrated in Figure 2B. On the public test dataset, RNASTOP achieved the lowest MCRMSE of 0.198, outperforming both Nullrecurrent and Kazuki2 (0.228) by approximately 0.03. More importantly, on the private test dataset, RNASTOP yielded an MCRMSE of 0.296, representing a 13% improvement over the leading Nullrecurrent model (0.342). Compared to the baseline DegScore model, the MCRMSE of

RNASTOP was 49% and 37% lower on the public and private test datasets, respectively. Notably, RNASTOP contains approximately 8.5 million trainable parameters, significantly fewer than the 30 million in Nullrecurrent, further underscoring its efficiency. Furthermore, runtime analysis demonstrated that RNASTOP achieves a balance between predictive precision and computational speed (Note 10 in File S1; Figure S5).

Beyond single-nucleotide evaluation, we assessed RNASTOP's capability to predict overall degradation on an independent full-length mRNA dataset [22]. The lengths of these mRNAs range from 504 to 1588 nt (Figure 2C), nearly tenfold longer than the SOVK training sequences. As illustrated in Figure 2D, RNASTOP exhibited the highest correlation with experimentally determined overall degradation values, achieving a Spearman correlation coefficient of 0.50 ($P = 1.7 \times 10^{-13}$). The Spearman correlation coefficients for the Kazuki2 model and the Nullrecurrent model were 0.48 ($P = 3.3 \times 10^{-12}$) and 0.43 ($P = 9.5 \times 10^{-10}$), respectively. Notably, all three deep learning models significantly outperformed ViennaRNA RNAfold v. 2.4.14 [23] ($R = 0.25$, $P = 5.4 \times 10^{-4}$), the DegScore linear regression model [22] ($R = 0.36$, $P = 2.9 \times 10^{-7}$), and the DegScore-XGBoost (DegScore-XG) model ($R = 0.42$, $P = 1.8 \times 10^{-9}$). These experimental results underscore RNASTOP's exceptional ability to accurately predict the overall degradation of full-length mRNAs and demonstrate its robust generalization capacity.

Nucleic acid LLMs and the dual-branch architecture enable robust transfer learning

The SOVK training dataset comprises 1189 mRNA sequences of 107 nt, whereas the private test dataset includes 629 sequences of 107 nt and 1172 sequences of 130 nt (Figure S1). We visualized the sequence distributions of the training dataset and the private test dataset using *t*-distributed stochastic neighbor embedding (*t*-SNE) [28] (Figure 3A). The visualization revealed that the distribution of the 130-nt private test sequences significantly diverged from that of the 107-nt training sequences. The lowest sequence similarity score between the two datasets was 0.2923. Despite this distributional shift, RNASTOP accurately predicted the degradation values of the 130-nt sequences at single-nucleotide resolution, demonstrating strong transfer learning capabilities.

To determine whether nucleic acid LLMs underlie this transfer learning performance,

we conducted an ablation study evaluating RNASTOP with and without nucleic acid LLM embeddings across three specific subsets of the 130-nt private test dataset: the sequence with the highest similarity to the training set, the sequence with the lowest similarity, and the entire dataset (Figure 3B). Across all three subsets, the nucleic acid LLM-integrated RNASTOP model demonstrated superior predictive performance. This advantage was most pronounced in the lowest-similarity subset, where the inclusion of nucleic acid LLMs reduced the MCRMSE by 0.071 (an improvement of approximately 20%). These findings confirm that pre-trained nucleic acid LLMs significantly enhance the transfer learning capabilities of the RNASTOP model.

Additionally, we investigated the contribution of the dual-branch network architecture by evaluating two ablated models: one retaining only the attention-based module (RNASTOP_ATT) and another retaining only the convolutional module (RNASTOP_CNN). As shown in Figure 3C, both ablated models suffered substantial performance drops on the private test dataset and the full-length mRNA dataset. On the full-length dataset, the RNASTOP_ATT model exhibited severely degraded performance in predicting overall degradation. This demonstrates that the dual-branch architecture is vital for the model's transfer learning capabilities.

To further explore how the dual-branch network architecture improves model performance, we utilized the *t*-SNE technique to visualize the mRNA embeddings extracted by RNASTOP_ATT, RNASTOP_CNN, and RNASTOP on the SOVK dataset, as shown in Figure 3D. The embeddings from RNASTOP_ATT and RNASTOP_CNN occupied distinct regions within the latent space, whereas the full RNASTOP model successfully integrated both representational spaces. Heatmap visualizations of these embeddings (Figure 3E and F) revealed that RNASTOP_CNN prioritizes specific local positional contexts (e.g., positions 20-40 and 70-80), while RNASTOP_ATT captures global attention across specific embedding dimensions (e.g., dimensions 22, 61, and 103). By concatenating these branches, the full model expands its embedding dimension from 128 to 256 (Figure 3G). This fused embedding effectively synthesizes information from both sub-networks, simultaneously capturing local positional motifs and critical global features. Thus, the dual-branch architecture integrates complementary feature-extraction perspectives, equipping the RNASTOP model with robust predictive accuracy and generalization capabilities.

RNASTOP is capable of capturing sequence and structural patterns governing

degradation

To better evaluate RNASSTOP's ability to capture underlying biophysical rules, we analyzed its predictions on representative mRNA sequences. **Figures 4A–E** highlight sequences from the private test dataset that achieved the lowest RMSE for the deg_Mg_pH10 condition (see Figure S6 for the top 10 sequences). The RNASSTOP predictions closely mirrored the experimental ground truth, successfully identifying diverse chemical stability patterns. For example, paired nucleotides forming stem regions are known to exhibit high chemical stability. RNASSTOP consistently predicted low degradation values for these regions (Figure 4C). Furthermore, five-nucleotide hairpin loops are thermodynamically optimal and highly stable [22]. For the sequence in Figure 4B, which contains hairpins of 4, 5, and 6 nucleotides, RNASSTOP accurately predicted the lowest degradation value for the 5-nucleotide hairpin. Similarly, symmetrical internal loops are chemically more stable than asymmetrical ones [22], a pattern correctly identified by RNASSTOP in the sequences shown in Figure 4A and E. Conversely, Figure 4F and G illustrate sequences with the highest RMSE (see Figure S7 for the top 10 worst-predicted sequences). The sequence in Figure 4F features an exceedingly complex secondary structure, with only 130 nucleotides yet comprising four internal loops, seven stem regions, two hairpin loops, and one bulge, posing a significant challenge for model prediction. In Figure 4G, the mRNA's secondary structure includes longer single strands, with discrepancies between RNASSTOP's predictions and experimental measurements predominantly occurring in nucleotides on these single strands. The RNASSTOP model tends to predict higher degradation values for single-stranded nucleotides. These divergences in predicting highly complex structures (Figure 4F) and extended single-stranded regions (Figure 4G) highlight current model limitations, which are likely due to insufficient examples of these structural patterns in the training data. To address these limitations and enhance prediction accuracy, future improvements will focus on enriching the dataset with such challenging samples to better capture complex spatial dependencies.

To further investigate whether RNASSTOP captures the influence of sequence patterns on mRNA degradation, we performed validation using 1280 mRNA sequences, with results illustrated in Figure 4H. Each visualization in Figure 4H represents the average of normalized degradation values. These 1280 mRNA sequences share the same secondary structure, differing only in the types of nucleotides in their hairpin loops. Prior experimental findings have indicated that linkages that lead to a 3' uridine are

particularly prone to degradation [22]. According to the results in Figure 4H, RNASTOP successfully captures this trend. To delve deeper into whether RNASTOP discerns the impact of secondary structures on mRNA degradation, validation was performed on 1000 mRNA sequences, with results shown in Figure 4I. Each visualization in Figure 4I also represents the average of normalized degradation values. Symmetrical internal loops are more stable than asymmetrical ones [22], and the results in Figure 4I suggest that the RNASTOP model also captures this trend. Furthermore, the quantitative analysis of structural pattern recognition on the public test dataset confirmed that RNASTOP detects specific secondary structure patterns affecting mRNA degradation (Note 11 in File S1; Figure S8).

Additionally, we demonstrated that RNASTOP can reveal important features that affect mRNA degradation using a Leave-One-Feature-Out approach. The results suggest that mRNA sequences and BPP matrices may be key features affecting mRNA chemical stability (Note 6 in File S1; Figure S9).

Optimizing mRNA chemical stability and translation efficiency with RNASTOP

RNASTOP was applied to optimize the BNT-162b2 and VZV glycoprotein E (gE) mRNA vaccine sequences. Initially, we evaluated two heuristic search algorithms: beam search and Monte Carlo tree search (MCTS). Preliminary experiments indicated that beam search was computationally more efficient (Note 12 in File S1; Figures S10 and S11). Consequently, RNASTOP employed the beam search algorithm for all subsequent codon optimization tasks.

Given that a lower MFE typically drives RNA into a more compact, secondary structure-rich conformation that resists in-solution degradation [22], we established MFE as a primary optimization objective. In addition to MFE, GC content, the proportion of unpaired nucleotides, and RNASTOP's overall degradation score serve as crucial indicators of chemical stability. Furthermore, because mRNA vaccine efficacy relies on both stability and translation efficiency, we incorporated the CAI as a co-optimization objective to improve sequence translation efficiency.

The progressive impact of the optimization process on RNA sequence properties is illustrated in **Figure 5A–E**, using the widely adopted BNT-162b2 mRNA vaccine [29] from BioNTech-Pfizer as a baseline. The MFE exhibited a consistent decline throughout the optimization iterations, dropping from an initial -1265.90 kcal/mol to -1531.20 kcal/mol. For comparison, the state-of-the-art RNA pre-trained models

GEMORNA [18] and CodonTransformer [17] achieved MFE values of only -1187.6 kcal/mol and -1338.8 kcal/mol, respectively, while the commercial tool OptimumGene reached -1252.0 kcal/mol. This highlights RNASSTOP's superiority over existing methods in optimizing chemical stability. Concurrently, other chemical stability-related metrics improved: GC content increased from 0.5695 to 0.5852, the unpaired nucleotide proportion decreased from 0.3967 to 0.3443, and the overall degradation score dropped from 0.4575 to 0.4345. Crucially, the original BNT-162b2 sequence already possessed a high CAI of 0.9516, making further enhancement challenging. Nevertheless, RNASSTOP still achieved a slight increase to 0.9556. In contrast, optimization by GEMORNA, CodonTransformer, and OptimumGene reduced the CAI to 0.8313, 0.8762, and 0.9049, respectively. Notably, although the LinearDesign algorithm achieved a low MFE of -2547.1 kcal/mol, its CAI decreased to 0.7325. This indicates that while LinearDesign can rigorously minimize MFE to its theoretical limit, prioritizing extreme chemical stability often comes at a severe cost to translation efficiency. These results confirm that RNASSTOP achieves a superior and simultaneous improvement in both chemical stability and translation efficiency.

To validate RNASSTOP's generalizability, we applied it to the wild-type VZV gE mRNA vaccine sequence. As shown in Figure 5F–J, all evaluated metrics exhibited improvement. Following RNASSTOP optimization, the MFE decreased from -527.4 kcal/mol to -926.8 kcal/mol, and the CAI increased from 0.6687 to 0.8601. Additionally, the GC content increased from 0.4650 to 0.5907, the unpaired nucleotide proportion dropped from 0.3858 to 0.2830, and the overall degradation score decreased from 0.6238 to 0.4380. Overall, RNASSTOP enhanced the MFE of the BNT-162b2 and VZV mRNA vaccines by 20.96% and 75.73%, respectively, and improved their CAI by 0.42% and 28.62%. These findings demonstrate RNASSTOP's robust capability and broad generalizability in concurrently enhancing RNA sequence chemical stability and translation efficiency.

We further analyzed the secondary structures of the mRNA sequences before and after optimization (**Figure 6**). As shown in Figure 6A and B, the MFE structures exhibit significant topological changes following RNASSTOP optimization. The original mRNAs (left) exhibit highly branched topologies characterized by numerous single-stranded loops and short, interrupted helices. In contrast, the RNASSTOP-optimized sequences (right) adopt highly compact architectures dominated by extended, continuous double-stranded stems, as highlighted in the magnified insets. This increase

in base-pairing density and reduction in structural branching indicate a lower free energy state, suggesting enhanced chemical stability and potentially higher resistance to endonucleolytic degradation *in vivo*.

This macroscopic stabilization is fundamentally driven by a reshaping of the RNA thermodynamic folding landscape, as evidenced by mountain plots and positional entropy profiles (Figures 6C and D). For the wild-type sequences (left), the divergence among the MFE, partition function, and centroid curves implies a complex energy landscape characterized by multiple competing suboptimal conformations. This structural ambiguity is further reflected by the widespread high-density spikes in the positional entropy profiles, indicating significant flexibility at the single-nucleotide level. Conversely, RNASTOP optimization (right) successfully condenses the folding trajectory into a single, well-defined global minimum. The near-perfect overlap of the MFE, partition function, and centroid curves in the optimized mountain plots, coupled with a reduction in positional entropy, demonstrates that the optimized sequences are highly deterministic. Overall, these data highlight the robust capability of RNASTOP to rationally navigate the synonymous sequence space, transforming dynamic, heterogeneous mRNA sequences into chemically stable vaccine sequences.

Discussion

In this work, we introduce RNASTOP, a framework that can accurately predict mRNA degradation and enhance its chemical stability. The meticulously crafted network architecture empowers RNASTOP to harness the strengths of various underlying neural network architectures, extracting and fusing feature embeddings of mRNAs from multiple perspectives, thereby endowing RNASTOP with superior prediction performance. By integrating nucleic acid LLMs and the dual-branch network architecture, RNASTOP can predict the degradation of mRNA sequences outside the distribution of the training dataset, which greatly enhances its practical utility.

Additionally, RNASTOP can capture patterns in sequence and structure that affect mRNA degradation. This is a critical attribute that not only attests to its high biological interpretability but also implies that mRNA sequences designed with RNASTOP will inherently acquire these patterns, conferring enhanced chemical stability. Moreover, we used RNASTOP to optimize the BNT-162b2 and VZV gE mRNA vaccine sequences, demonstrating its effective optimization performance and practicality.

Despite its strengths, RNASSTOP has some limitations. While RNASSTOP employs a straightforward yet effective beam search algorithm for mRNA sequence optimization, various other computational methods have been developed with the growing interest in mRNA therapeutics [3,14,30–34]. These methods employ diverse strategies such as deep learning, partition function calculations, and quantum computing to refine mRNA sequences and enhance their therapeutic efficacy. Furthermore, RNASSTOP does not account for nucleotide modifications, which also impact mRNA chemical stability. With access to training data on modified nucleotides, RNASSTOP could be extended to predict and optimize the chemical stability of modified mRNAs.

In conclusion, the prediction and optimization of mRNA chemical stability are pivotal to the development of mRNA therapeutics. Our research offers a novel framework that can accurately predict mRNA degradation at both single-nucleotide and full-length levels via deep learning and enhance mRNA chemical stability through codon optimization. The RNASSTOP model could shorten the development time and reduce the costs associated with developing mRNA therapeutics.

Code availability

The source code used in this study is available on GitHub (<https://github.com/xlab-BioAI/RNASSTOP>). The code has also been submitted to BioCode at the NGDC, China National Center for Bioinformation (CNCB) (BioCode: BT007947), which is publicly accessible at <https://ngdc.cncb.ac.cn/biocode/tools/BT007947>.

CRediT author statement

Shenggeng Lin: Investigation, Methodology, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing. **Junwei Chen:** Investigation, Writing – original draft. **Heqi Sun:** Investigation, Methodology. **Yufang Zhang:** Investigation, Methodology. **Wenyi Yang:** Visualization, Writing – review & editing. **Hong Tan:** Investigation, Methodology, Visualization. **Dong-Qing Wei:** Project administration. **Qinghua Jiang:** Project administration, Writing – review & editing. **Yi Xiong:** Investigation, Conceptualization, Project administration, Funding acquisition, Supervision, Writing – review & editing. All authors have read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

This work was supported by National Key Research and Development Program of China (Grant Nos. 2023YFC2506400, 2023YFC2506402 and 2024YFA1306902), National Natural Science Foundation of China (Grant No. 32570779), and Shanghai Municipal Education Commission (Grant No. 2024AIZD008). The computational experiments were partially run at the Center for High-Performance Computing, Shanghai Jiao Tong University. We thank Jianmin Wang for helpful discussions regarding the manuscript.

Supplementary material

Supplementary material is available at *Genomics, Proteomics & Bioinformatics* online (<https://doi.org/10.1093/gpbjnl/qzaxxxx>).

ORCID

0009-0004-4925-7402 (Shenggeng Lin)

0009-0002-3614-1553 (Junwei Chen)

0009-0000-4979-8242 (Heqi Sun)

0009-0002-8063-7695 (Yufang Zhang)

0009-0008-1176-0343 (Wenyi Yang)

0009-0008-8030-6670 (Hong Tan)

0000-0003-4200-7502 (Dong-Qing Wei)

0000-0002-1827-0389 (Qinghua Jiang)

0000-0003-2910-6725 (Yi Xiong)

References

[1] Kramps T, Elbers K. Introduction to RNA vaccines. *Methods Mol Biol* 2017;1499:1–11.

- [2] Crommelin DJA, Anchordoquy TJ, Volkin DB, Jiskoot W, Mastrobattista E. Addressing the cold reality of mRNA vaccine stability. *J Pharm Sci* 2021;110:997–1001.
- [3] Zhang H, Zhang L, Lin A, Xu C, Li Z, Liu K, et al. Algorithm for optimized mRNA design improves stability and immunogenicity. *Nature* 2023;621:396–403.
- [4] Corbett KS, Flynn B, Foulds KE, Francica JR, Boyoglu-Barnum S, Werner AP, et al. Evaluation of the mRNA-1273 vaccine against SARS-CoV-2 in nonhuman primates. *N Engl J Med* 2020;383:1544–55.
- [5] Baden LR, El Sahly HM, Essink B, Kotloff K, Frey S, Novak R, et al. Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *N Engl J Med* 2021;384:403–16.
- [6] Szabo GT, Mahiny AJ, Vlatkovic I. COVID-19 mRNA vaccines: platforms and current developments. *Mol Ther* 2022;30:1850–68.
- [7] Zhang NN, Li XF, Deng YQ, Zhao H, Huang YJ, Yang G, et al. A thermostable mRNA vaccine against COVID-19. *Cell* 2020;182:1271–83.e16.
- [8] Verbeke R, Hogan MJ, Lore K, Pardi N. Innate immune mechanisms of mRNA vaccines. *Immunity* 2022;55:1993–2005.
- [9] Wayment-Steele HK, Kladwang W, Watkins AM, Kim DS, Tunguz B, Reade W, et al. Deep learning models for predicting RNA degradation via dual crowdsourcing. *Nat Mach Intell* 2022;4:1174–84.
- [10] Schoenmaker L, Witzigmann D, Kulkarni JA, Verbeke R, Kersten G, Jiskoot W, et al. mRNA-lipid nanoparticle COVID-19 vaccines: structure and stability. *Int J Pharm* 2021;601:120586.
- [11] Kon E, Elia U, Peer D. Principles for designing an optimal mRNA lipid nanoparticle vaccine. *Curr Opin Biotechnol* 2022;73:329–36.
- [12] Li S, Moayedpour S, Li R, Bailey M, Riahi S, Kogler-Anele L, et al. CodonBERT large language model for mRNA vaccines. *Genome Res* 2024;34:1027–35.
- [13] Chu Y, Yu D, Li Y, Huang K, Shen Y, Cong L, et al. A 5' UTR language model for decoding untranslated regions of mRNA and function predictions. *Nat Mach Intell* 2024;6:449–60.
- [14] Gong H, Wen J, Luo R, Feng Y, Guo J, Fu H, et al. Integrated mRNA sequence optimization using deep learning. *Brief Bioinform* 2023;24:bbad001.
- [15] Blakney AK. A tool for optimizing messenger RNA sequence. *Nature* 2023;621:262–4.
- [16] Goulet DR, Yan Y, Agrawal P, Waight AB, Mak AN, Zhu Y. Codon optimization using a recurrent neural network. *J Comput Biol* 2023;30:70–81.
- [17] Fallahpour A, Gureghian V, Filion GJ, Lindner AB, Pandi A. CodonTransformer: a multispecies codon optimizer using context-aware neural networks. *Nat Commun* 2025;16:3205.
- [18] Zhang H, Liu H, Xu Y, Huang H, Liu Y, Wang J, et al. Deep generative models design mRNA sequences with enhanced translational capacity and stability. *Science* 2025;390:eadr8470.
- [19] He S, Gao B, Sabnis R, Sun Q. RNAdegformer: accurate prediction of mRNA degradation at nucleotide resolution with deep learning. *Brief Bioinform* 2023;24:bbac581.
- [20] Wayment-Steele HK, Kim DS, Choe CA, Nicol JJ, Wellington-Oguri R, Watkins AM, et al. Theoretical basis for stabilizing messenger RNA through secondary structure design. *Nucleic Acids Res* 2021;49:10604–17.
- [21] Bae H, Collier J. Codon optimality-mediated mRNA degradation: linking translational elongation to mRNA stability. *Mol Cell* 2022;82:1467–76.
- [22] Leppek K, Byeon GW, Kladwang W, Wayment-Steele HK, Kerr CH, Xu AF, et al. Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics. *Nat Commun* 2022;13:1536.

- [23] Lorenz R, Bernhart SH, Honer Zu Siederdisen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA package 2.0. *Algorithms Mol Biol* 2011;6:26.
- [24] Chen J, Hu Z, Sun S, Tan Q, Wang Y, Yu Q, et al. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *arXiv* 2022;2204.00300.
- [25] Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* 2021;37:2112–20.
- [26] Zhao Z, Bai H, Zhang J, Zhang Y, Xu S, Lin Z, et al. CDDFuse: correlation-driven dual-branch feature decomposition for multi-modality image fusion. *Proc IEEE/CVF Conf Comput Vis Pattern Recognit* 2023:5906–16.
- [27] Kingma DP, Ba J. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980* 2014.
- [28] Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;9:2579–605.
- [29] Polack FP, Thomas SJ, Kitchin N, Absalon J, Gurtman A, Lockhart S, et al. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *N Engl J Med* 2020;383:2603–15.
- [30] Zhang H, Sarkar A, Bertels K. A resource-efficient variational quantum algorithm for mRNA codon optimization. *arXiv* 2024;2404.14858.
- [31] Fu H, Liang Y, Zhong X, Pan Z, Huang L, Zhang H, et al. Codon optimization with deep learning to enhance protein expression. *Sci Rep* 2020;10:17617.
- [32] Sumi S, Hamada M, Saito H. Deep generative design of RNA family sequences. *Nat Methods* 2024;21:435–43.
- [33] Dai N, Tang WY, Zhou T, Mathews DH, Huang L. Messenger and non-coding RNA design via expected partition function and continuous optimization. *arXiv* 2023;2401.00037.
- [34] Fox DM, Branson KM, Walker RC. mRNA codon optimization with quantum computers. *PLoS One* 2021;16:e0259101.

Figure legends

Figure 1 Overview of RNASTOP framework

A. Overview of the mRNA degradation prediction model, including the feature embedding module, the dual-branch feature decoupling-and-aggregating network, and the prediction layer. **B.** Overview of the mRNA codon optimization module based on the mRNA degradation prediction model and beam search. **C.** Sub-modules of the mRNA degradation prediction model, including the attention-based neural network, the convolution-based neural network and the prediction layer.

Figure 2 RNASTOP achieves accurate prediction of mRNA degradation

A. Simple schematic of the model architecture used by the Nullrecurrent. This model integrates two sets of features, combining elements of classic recurrent neural networks and convolutional neural networks. **B.** RNASTOP's MCRMSE is lower than that of state-of-the-art models on both the public and private test datasets of the SOVK competition. **C.** Distribution of mRNA length and degradation values on the full-length

mRNA dataset ($n = 188$). **D.** RNASSTOP's prediction results exhibit the highest correlation with experimentally determined degradation values on the full-length mRNA dataset ($n = 188$).

Figure 3 The nucleic acid large language models and dual-branch network architecture improve the prediction performance of RNASSTOP model

A. mRNA distribution on the training dataset and private test dataset of the SOVK competition ($n = 2990$). **B.** The nucleic acid LLMs improve the transfer learning performance of RNASSTOP. **C.** The performance of different model architectures on the SOVK dataset and the full-length mRNA dataset, with MCRMSE serving as the evaluation metric for the SOVK dataset and the Spearman correlation coefficient for the full-length mRNA dataset. **D.** *t*-SNE visualization of mRNA embeddings extracted by different model architectures on the SOVK dataset ($n = 2000$). **E.–G.** Heatmaps of mRNA embeddings extracted by the RNASSTOP_CNN model (E), RNASSTOP_ATT model (F), and the multi-feature fusion layer (G) on the SOVK dataset ($n = 2000$).

Figure 4 RNASSTOP is capable of capturing sequence and structure patterns affecting mRNA degradation

A.–E. Representative structures of the best-predicted mRNA sequences under degradation conditions at 10 mM Mg^{2+} , pH 10, 1 day, 24°C. Different colors represent different degradation values after normalization. **F. and G.** Representative structures of the worst-predicted mRNA sequences under degradation conditions of 10 mM Mg^{2+} , pH 10, 1 day, 24°C. **H. and I.** RNASSTOP reveals degradation patterns based on both sequence (*i.e.*, linkages ending at 3' uridine are particularly reactive) and structure (symmetric internal loops, circled, have suppressed hydrolytic degradation compared to asymmetric internal loops). Different colors represent different degradation values after normalization.

Figure 5 RNASSTOP is capable of optimizing the mRNA sequences to enhance their chemical stability and translation efficiency

A.–E. Changes in key metrics during the iterative optimization of the BNT-162b2 mRNA vaccine sequence. **F.–J.** Corresponding changes for the VZV gE mRNA vaccine sequence.

Figure 6 Secondary structure analysis of mRNA vaccine sequences before and after optimization

A. Secondary structures of BNT-162b2 (left) and the optimized sequence (right). **B.** Secondary structures of VZV gE (left) and the optimized sequence (right). **C.** Mountain plots (top) and positional entropy profiles (bottom) for BNT-162b2, comparing the original (left) and optimized (right) sequences. **D.** Corresponding plots for the VZV gE sequences, comparing the original (left) and optimized (right) versions.













