Fine scale structural information substantially improves multivariate regression model for mRNA in-vial degradation prediction

Soon Yi[1,2], Sara E. Ali[3], Yashrajsinh Jadeja[3], J. Wade Davis[3], Mihir Metkar[3,*]

[1]Center for RNA Science and Therapeutics, School of Medicine, Case Western Reserve University, Cleveland, OH, USA
[2]Department of Biochemistry, School of Medicine, Case Western Reserve University, Cleveland, OH, USA
[3]Moderna, Inc., Cambridge, MA, USA

Abstract

The success of COVID-19 mRNA vaccines has made optimizing mRNAs for in-vial stability a key objective. However, we still lack a complete understanding of the sequence metrics that influence mRNA stability in solution. RNA secondary structure plays a central role in protecting against hydrolysis, the primary degradation pathway under storage conditions. Yet, the structural metrics that best guide stability-focused mRNA design remain unclear. Global metrics like minimum free energy and average unpaired probability have improved mRNA stability but fail to capture local structural variation relevant to degradation. We show that base-pairing probability, in terms of log odds, provide fine-scale, orthogonal insight that complements global metrics and improves stability modeling. By combining local and global features into a four-feature regression model, dubbed STRAND (Stability Regression Analysis using Nucleotide-Derived features), we achieve substantial gains in predictive performance over current methods. This compact and interpretable model provides a practical framework for designing mRNAs with enhanced in-solution stability.

*Correspondence: mihir.metkar@modernatx.com

Introduction

Optimizing therapeutic mRNA requires not only effective protein expression but also in-vial stability that resists degradation during storage and delivery. Although the rapid development of mRNA-based vaccines has highlighted the platform's potential[1], mRNA instability remains a critical barrier to broader deployment, particularly in settings lacking reliable cold-chain infrastructure such as developing countries[2]. Degradation primarily occurs through hydrolysis and transesterification[3,4] reactions in-solution during storage in-vial, typically initiating at structurally accessible nucleotides within unpaired loops or flexible regions[5,6]. Consequently, solution conditions such as temperature, pH, and ionic strength can drastically reduce mRNA half-life, thereby limiting therapeutic efficacy and significantly increasing manufacturing complexity and costs[7].

A key mRNA degradation determinant is its secondary structure, which arises from intramolecular base pairing. Base-paired regions can shield vulnerable sites from chemical or enzymatic attacks, extending RNA stability[6]. Accordingly, rational mRNA design has focused on engineering sequences that adopt favorable structural conformations. However, this design process requires reliable metrics that are predictive of RNA stability[8,9].

Traditional metrics such as free energy (FE), average unpaired probability (AUP), and GC content have been widely used to guide mRNA optimization[10,11]. Free energy minimization identifies the most thermodynamically stable structure an RNA can adopt[12,13]; AUP captures the average likelihood that nucleotides remain unpaired; and GC content correlates with duplex stability due to stronger G-C base pairing[8,14]. These metrics are computationally efficient and biologically relevant, and they have been integrated into mRNA design tools[15,16] such as CDSFold[17], RiboTree[18], LinearDesign[19] and DERNA[20]. However, despite their usefulness, these metrics do not fully explain observed degradation differences as reflected by their modest Spearman correlation coefficient (<0.6)[9], suggesting they fail to capture important aspects of RNA structural ensembles[9].

Large-scale efforts have attempted to model mRNA degradation using both statistical and deep learning approaches. For example, the dual crowd-sourcing study by Wayment-Steele et al.[21] combined human- and machine-designed sequences to

benchmark predictive models of degradation, including those incorporating experimental selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) data[6,22,23]. Despite access to thousands of synthetic constructs and high-resolution structural probing data, models achieved only moderate predictive accuracy when applied to full-length mRNAs, with Spearman correlation coefficients (rho) ranging from 0.36 to 0.48[24]. These results reflect the inherent difficulty of modeling mRNA degradation, which likely depends on a combination of global and local structural features not fully captured by existing metrics.

To address this limitation, we propose a new metric: the base pairing probability in terms of log-odds (LO), which quantifies the base-pairing likelihood at each position across the structural ensemble by comparing the odds of being paired to unpaired. Unlike base-pairing probability[25,26], which is bounded between 0 and 1 and compresses extreme values, the log-odds transformation expands the dynamic range[27], particularly for high and low probabilities, thereby amplifying meaningful structural differences that might otherwise be obscured. LO reflects the local base-pairing probabilities distribution, offering finer resolution than global metrics like free energy or AUP. By analyzing mRNAs with similar global scores but divergent in-solution half-lives, we show that LO captures functional differences not apparent in traditional metrics. Finally, we demonstrate that incorporating LO into a multivariate regression model significantly improves the in-vial mRNA stability prediction, providing an orthogonal feature that complements existing structural descriptors.

Results

Traditional RNA Structural Metrics Partially Explain In-Solution Stability

Figure 1

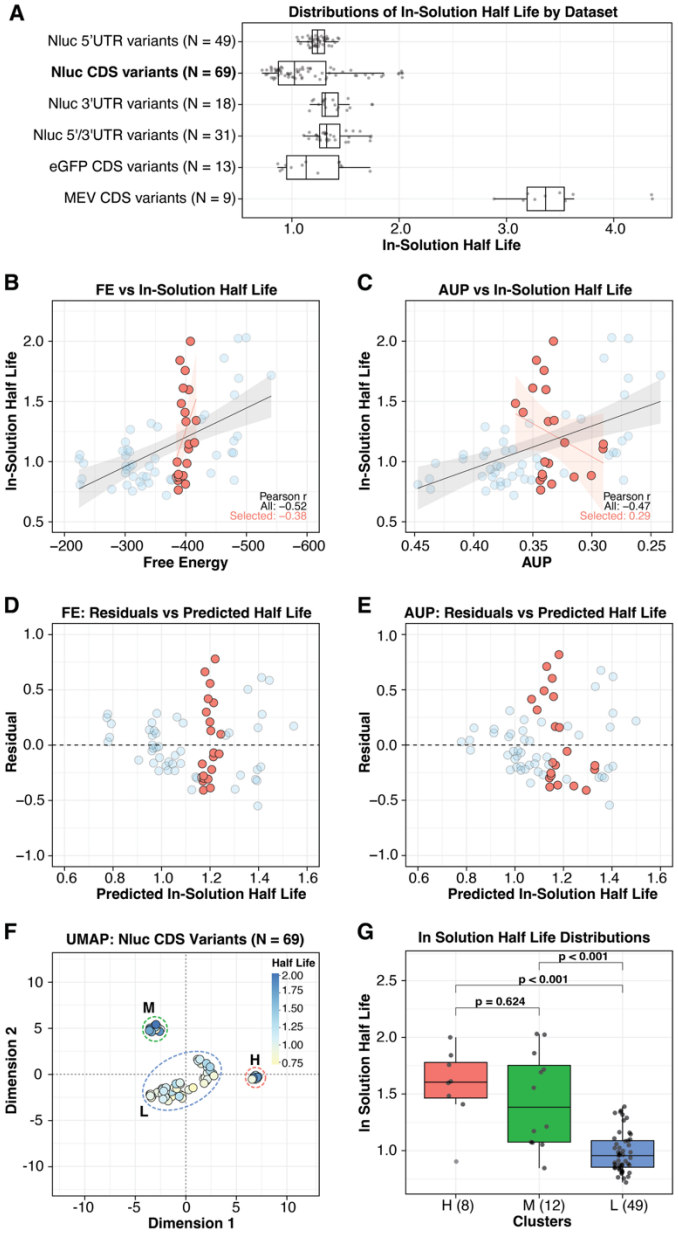

Figure 1. In-solution half-life distributions across protein mRNA variants and structure–function analysis in the Nano Luciferase set.
We analyzed a curated dataset of Nluc, eGFP, and MEV mRNA variants[9] to examine the relationship between RNA structural features—such as free energy (FE), average unpaired probability (AUP), and experimentally measured in-solution half-lives.

(A) A boxplot distribution of in-solution half-lives for mRNA variants of different proteins. The box represents the interquartile range, the horizontal line indicates the median, and whiskers denote 1.5X the interquartile range; individual sequences are plotted as points. The NanoLuc (NLuc; bold) CDS variants were selected for further analysis due to their wide range of half-lives and large number of sequences.

(B) Correlation of FE with half-life in the NLuc dataset. Red-highlighted sequences show similar FE but divergent half-lives. Pearson r for the entire dataset (black) and for sequences with similar FE (red) are shown.

(C) Same as (B) but for AUP.

(D) Residual plots for FE showing increased error for sequences with higher half-lives.

(E) Same as (D) but for AUP.

(F) UMAP projection of NLuc sequences based on global structural features (FE, AUP, and GC content), revealing three clusters. Clusters are labeled as high (H), medium (M), and low (L) based on their in-solution half-lives.

(G) Boxplot distribution of half-lives for sequences in each UMAP-defined cluster in (F), with pairwise statistical comparisons shown. The Mann-Whitney test was used to calculate p-values. The number of sequences within each cluster is noted in parentheses.

A major goal in RNA therapeutic design is to extend mRNA half-life in-vial[28]. Existing studies have attempted to improve stability using sequence-derived structural metrics, such as FE and AUP, which provide approximations of RNA secondary structure.[18] However, whether these metrics alone are sufficient to predict mRNA degradation rates remains an open question.[9]

To assess the relationship between RNA structure and in-solution degradation, we examined a dataset of Nano Luciferase (NLuc), enhanced green fluorescence protein (eGFP), and multi-epitope vaccine (MEV) mRNAs with measured half-lives in-solution (**Figure 1A**)[9]. NLuc sequences were selected for further analysis due to their broad half-life range and the largest number of available sequence variants (N = 69), making them an ideal test set for evaluating structural predictors.

Across the NLuc dataset, we observed moderate negative correlations between half-life and FE (Pearson r = -0.52) as well as AUP (Pearson r = -0.47) (red and blue points in **Figure 1B, 1C**). These correlations suggest that mRNAs with lower FE and AUP, indicative of more stable secondary structures, tend to degrade more slowly. However, while these metrics broadly capture structural trends, they fail to fully account for the

variability in observed degradation rates. For example, the NLuc sequence variants with free energy around -400 kcal/mol (N = 21, NLuc-21 set) showed a lower correlation than the global comparison, with Pearson correlation values of -0.38 and -0.29, respectively (**Figure 1B, 1C**, highlighted in red). Moreover, linear regression models using FE and AUP as predictors exhibited high residual errors, particularly at higher values of either metric (**Figure 1D, 1E**). We also noted similar observations for eGFP CDS variants (N = 13; **Figure S1**), in which eGFP sequences with similar FE (–500 kcal/mol) show approximately 1.7-fold differences in their measured half-lives or the sequences with similar AUP (~0.40) show approximately 2-fold differences in their half-lives. These findings suggest that the observed discrepancy between the traditional metrics and degradation rates is a general phenomenon across different coding sequence sets. While global structure formation plays a role in stability, additional sequence-dependent factors influence degradation in ways that traditional metrics fail to capture.

To further investigate this variability, we applied Uniform Manifold Approximation and Projection (UMAP) to reduce the three-dimensional structural feature space into a lower-dimensional representation.[29] The resulting UMAP embedding, based on traditional metrics—including minimum FE normalized by transcript length (free energy/length), average unpaired probability (AUP), and GC content—separated the NLuc sequences into distinct clusters (**Figure 1F**). These clusters were identified as high (H), medium (M), and low (L) based on the average in-solution half-life. Normalizing FE by transcript length enabled comparison across sequences of varying lengths, ensuring that the energy values reflect general structural stability rather than absolute sequence length. However, despite the apparent separation in UMAP space, clusters H and M exhibited no statistically significant difference in their half-life distributions (**Figure 1G**), underscoring the inability of traditional metrics to resolve functionally distinct stability groups. This suggests that additional sequence-derived features may be necessary to explain mRNA degradation more accurately. To explore this further, we selected two NLuc sequences with nearly identical FE and AUP values but extreme differences in half-life. By analyzing their structural characteristics, we aimed to determine whether additional features could explain these discrepancies.

Local Structural Features Differentiate Sequences with Similar Global Metrics

Figure 2

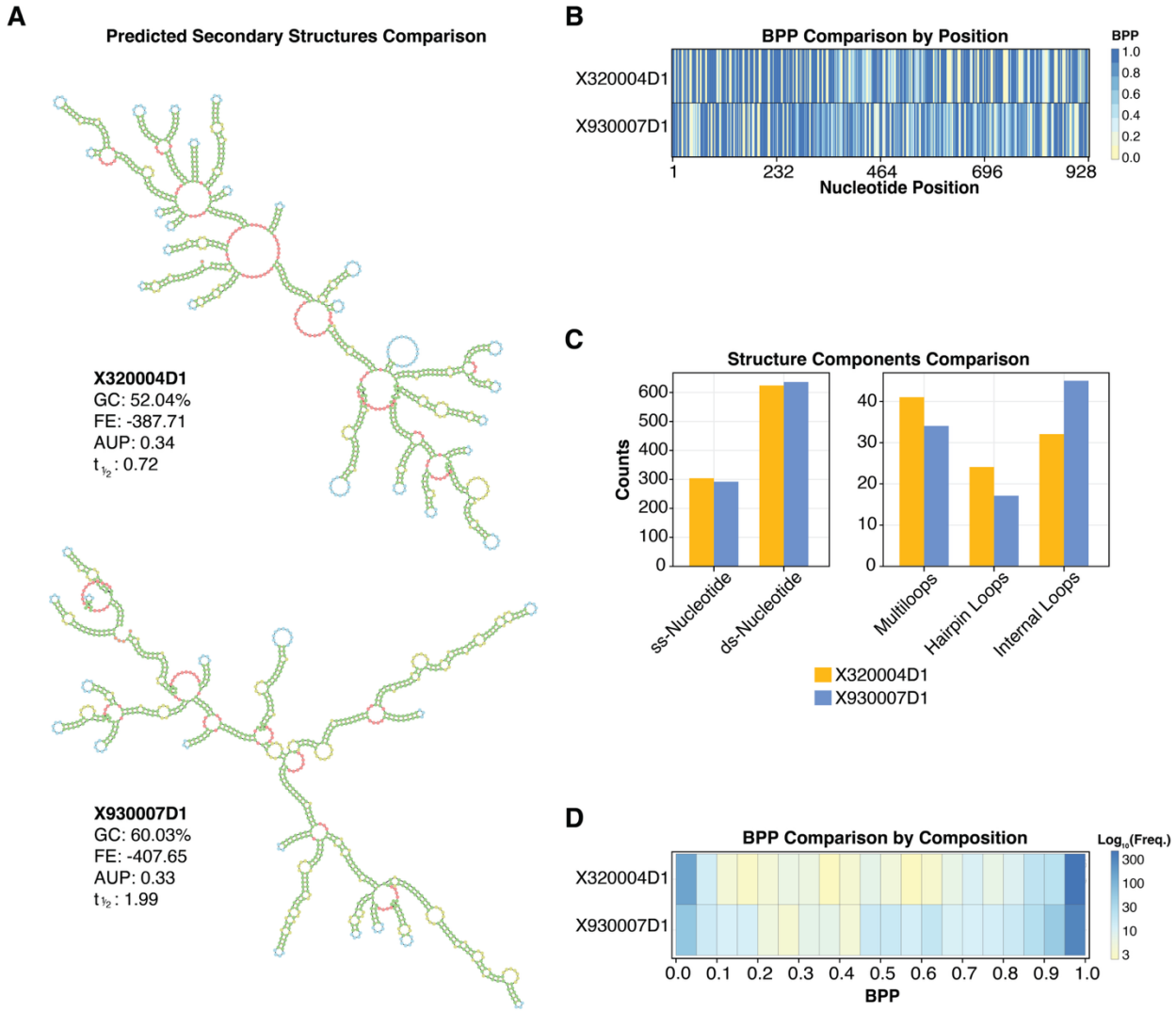

Figure 2. Structure comparison between two RNA sequences with similar global metrics but substantial differences in in-solution half-life.
We predicted RNA secondary structures for sequences X320004D1 and X930007D1, and analyzed their structural components and base-pairing probabilities (BPPs).

(A) Predicted secondary structures of two NLuc sequences, X320004D1 and X930007D1, exhibit similar global structure metrics (FE and AUP) but differ substantially in in-solution half-life.
(B) Heatmap comparing local BPP values by position, revealing fine-scale structural differences between the two sequences.
(C) Barplot comparing structural components between the predicted maximum expected accuracy structures, including counts of unpaired nucleotides, hairpin loops, interior loops, and multiloops.

(D) Heatmap comparing BPPs binned at 0.05 resolution, showing clear differences in the distribution of local pairing strength between the two sequences.

The two sequences, X320004D1 and X930007D1, had free energies of -387.71 and -407.65 kcal/mol, a ~5% difference, while their AUP values differed by less than 3% (0.34 vs. 0.33). Notably, X930007D1 had a higher GC content (60.03%) than X320004D1 (52.04%), though the direct effect of GC content on stability in this context remains uncertain. Their measured half-lives varied by nearly threefold (0.72 vs. 1.99 hours, respectively) (**Figure 2A**).

To investigate the structural basis for this difference, we predicted secondary structures of both mRNAs using LinearPartition (**Figure 2A**)[30]. Despite having similar global structural metric values, visual inspection of the predicted structures revealed striking differences. While both sequences contained a similar proportion of single- vs. double-stranded regions, they differed in secondary structure composition. A heatmap visualization of BPPs (**Figure 2B**) further illustrated that, despite similar global metrics, these sequences exhibited distinct base-pairing patterns along the sequence and BPP compositions that may contribute to their degradation differences.

X320004D1 formed a more branched secondary structure, with a higher number of multiloops (24) and hairpin loops (41), which could introduce greater structural flexibility and accessibility. In contrast, X930007D1 contained fewer multiloops (17) and hairpin loops (34) but a higher number of inner loops (45 vs. 32 in X320004D1), forming more extended helical regions (**Figure 2C**). These differences suggest that while both sequences achieve similar global stability, their local structural features differ substantially. To quantify these differences, we analyzed per-position base-pairing probabilities (BPPs) along the sequence (**Figure 2D**). The BPP distribution of X320004D1 showed a sharp bimodal pattern, with clear peaks near 0.0 (unpaired) and 1.0 (paired), suggesting more well-defined structural elements. In contrast, X930007D1 exhibited a broader, more continuous distribution, with a dominant peak near 1.0 but more intermediate BPP values, indicating a less rigid structure. The sequences with higher stability (free energy < -475 kcal/mol) exhibited similar differences between their structure metrics and measured half-lives (**Figure S2**). We chose two sequences with extreme in-solution half-lives, X310055T3 and X310005T1, which showed highly distinct predicted structures and

measured half-lives (1.08 vs 2.03), despite having nearly identical FE (–476.09 vs –499.53 kcal/mol), GC content (53.56 vs 52.16%), and identical AUP at 0.28. This suggested that finer-scale variations in base-pairing probability, rather than just global structure, could help explain the observed differences in RNA stability.

Base-Pairing Probability Alone Does Not Fully Explain mRNA Stability

Figure 3

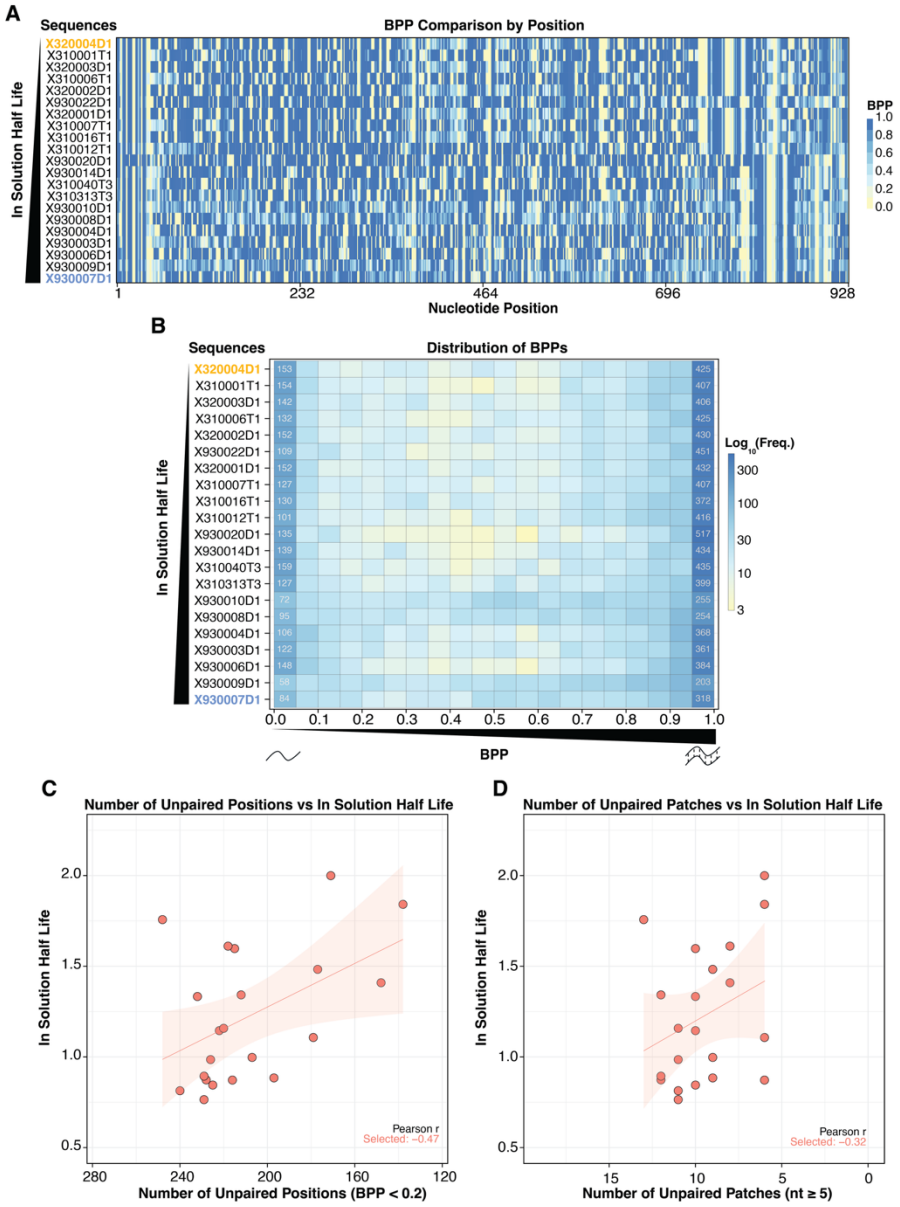

Figure 3. Local RNA base-pairing patterns differ among similarly structured sequences and associate with in-solution half-life.
We predicted the secondary structures of 21 NLuc variants (NLuc-21; red points in Fig. 1B) with similar FE and AUP values, and further analyzed their calculated base-pairing probabilities (BPPs) to investigate whether local structural differences could explain their divergent in-solution half-lives.

(A) Heatmap of BPPs by nucleotide position for sequences with similar global metrics (FE and AUP) but differing in-solution half-lives. Sequences are ordered by increasing half-life.
(B) Heatmap showing the distribution of BPPs (binned at width = 0.05) across the same sequences as in (A), highlighting differences in local pairing composition.
(C) Correlation between the number of unpaired nucleotides (defined as BPP < 0.2) and in-solution half-life. Pearson r for the NLuc-21 sequences (red) is shown.
(D) Correlation between the number of unpaired patches (≥5 consecutive unpaired positions with BPP < 0.2) and in-solution half-life. Pearson r for the NLuc-21 sequences (red) is shown.

To determine whether the structural patterns observed in X320004D1 and X930007D1 extend to a broader dataset, we analyzed 21 NLuc variants (NLuc-21; red points in **Figure 1B, C**) with similar FE and AUP values but differing in-solution half-lives. Despite having nearly identical global structural metrics, these sequences displayed distinct patterns of local base-pairing interactions (**Figure 3A, 3B**).

A heatmap of base-pairing probabilities (BPPs) across these NLuc-21 sequences (**Figure 3A**) revealed that, in general, each sequence possessed unique patterns of paired and unpaired regions. Some sequences exhibited more extended single-stranded regions (X320004D1), while others contained higher proportions of stable paired regions (X310012T1). Interestingly, while some sequences with higher stability displayed consistent base-pairing throughout the sequence (X930004D1), others with similar stability had more localized variations in paired and unpaired regions (X930009D1 vs X930007D1).

To quantify these differences, we examined whether the number of unpaired positions (BPP < 0.2, derived from BPP curvature analysis) or the presence of long unpaired patches (≥5 nucleotides) correlated with half-life. Pearson correlation analysis showed that neither total unpaired positions (r = -0.47, **Figure 3C**) nor long unpaired patches (r = -0.32, **Figure 3D**) provided a significantly stronger correlation than free energy or AUP. Moreover, our analysis showed no significant improvement in Pearson correlation with different lengths of unpaired patches (≥3 or ≥10 nucleotides; **Figure S3**). Thus, while local base-pairing interactions contribute to RNA stability, they likely do not fully explain the observed differences in degradation rates.

These findings indicate that BPP alone does not account for mRNA stability differences among sequences with similar global metrics. The inability of traditional metrics and simple base-pairing analysis to fully resolve these variations suggests that additional structural descriptors are needed. To further refine our understanding of RNA degradation, we next explored whether base-pairing log-odds (LO), a bidirectional metric such that it symmetrically distinguishes between confidently paired vs unpaired position, could capture the structural nuances that BPP alone could not.

Log-Odds of Base-Pairing Captures RNA Stability Differences Beyond BPP

Figure 4

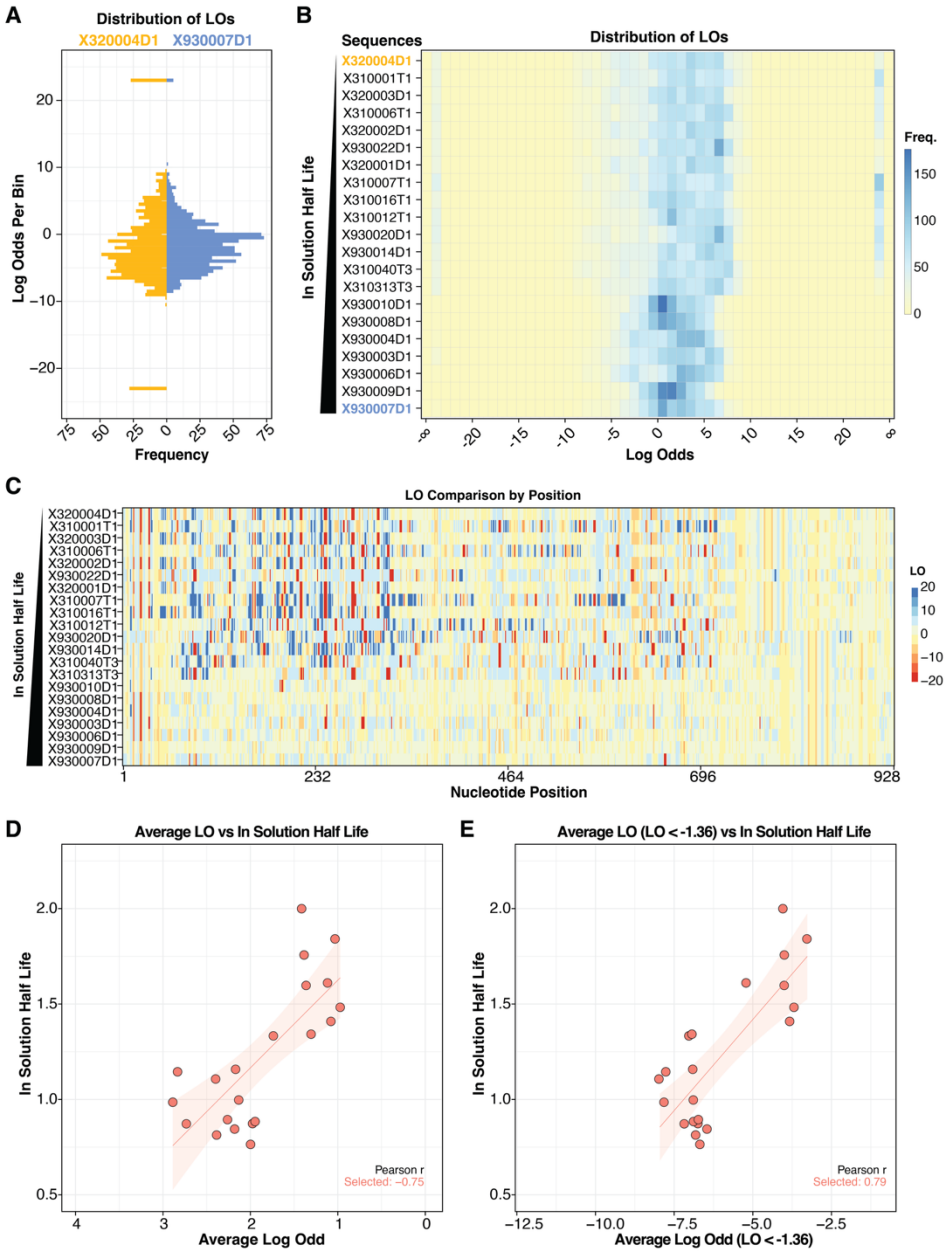

Figure 4. Base-pairing probability Log-odds (LO) capture local structure features predictive of RNA stability.

We calculated positional log-odds (LOs) from the base-pairing probabilities (BPPs) of the NLuc-21 sequences and analyzed them to identify local structural patterns associated with stability differences.

(A) Histogram of binned LO values (binwidth = 0.5) for two sequences (X320004D1 and X930007D1) with similar global structure metrics but divergent half-lives.; X-

axis shows the number of positions with LO values in each bin while Y-axis represents LO bins.

(B) Heatmap showing the distribution of LO values (binned) across the subset of NLuc sequences (Figure 3), sorted by increasing in-solution half-life. Color scale indicates frequency of positions in each bin.

(C) Heatmap of LO values by nucleotide position for the same sequences in (B) with color scale indicating LO values and highlighting localized structural differences along the transcript.

(D) Correlations of Average LO and in-solution half-life. Pearson r for the NLuc-21 sequences (red) is shown.

(E) Same as (D) but for LO for weakly paired regions (equivalent to BPP < 0.2).

We hypothesized that the weak explanatory power of BPP stems from its unidirectional nature, going from 0 to 1, thus rendering it unable to capture the tendency of the RNA sequences to be either stable or unstable.

To address this limitation, we developed a base-pairing probability log-odds (LO) metric to capture subtle but functionally important structural differences. LO transforms BPP values into a bidirectional scale ($-\infty$ to $+\infty$), where highly paired positions have positive LO values, and strongly unpaired positions have negative LO values. Unlike BPP, which ranges between 0 (unpaired) and 1 (paired), LO spans negative infinity to positive infinity, allowing for better resolution between weakly and strongly paired nucleotides (**Figure 4A, 4B**).

To test whether LO captures RNA degradation patterns better than BPP, we first examined per-position LO values for X320004D1 and X930007D1 (shown in **Figure 2**). While both sequences had similar global structural metrics, LO values were distributed differently, particularly in weakly structured regions (**Figure 4A; first and last row in Figure 4B)**. X320004D1 contained a greater number of strongly negative LO positions, indicating confidently unpaired regions, whereas X930007D1 exhibited fewer such sites and a higher frequency of positions with LO values near zero, aligning with its longer half-life.

A heatmap visualization of LO across NLuc-21 sequences (**Figure 4B, 4C**) demonstrated that LO more effectively differentiated stable and unstable sequences compared to BPP. Sequences with longer half-lives had fewer regions with strongly

negative LO values (red regions in **Figure 4C**), while sequences with shorter half-lives exhibited more widespread negative LO values.

To quantify this, we compared average LO (ALO) across sequences to in-solution half-life. ALO showed a significantly stronger correlation with stability than BPP-based metrics in the NLuc-21 set, with a Pearson correlation of -0.75 (**Figure 4D**)—approximately twice the correlation strength of free energy and three times that of AUP. Furthermore, when we focused on only weakly paired regions (LO < -1.363, derived from BPP curvature analysis), the correlation further improved to -0.79 in the NLuc-21 set (**Figure 4E**).

These results indicate that LO and ALO better capture functionally relevant RNA structural differences, distinguishing sequences with similar global structural metrics but varying degradation rates. By providing a more dynamic range than BPP, LO serves as a more accurate predictor of mRNA stability.

Multivariate regression Model to Improve mRNA Stability Prediction

Figure 5

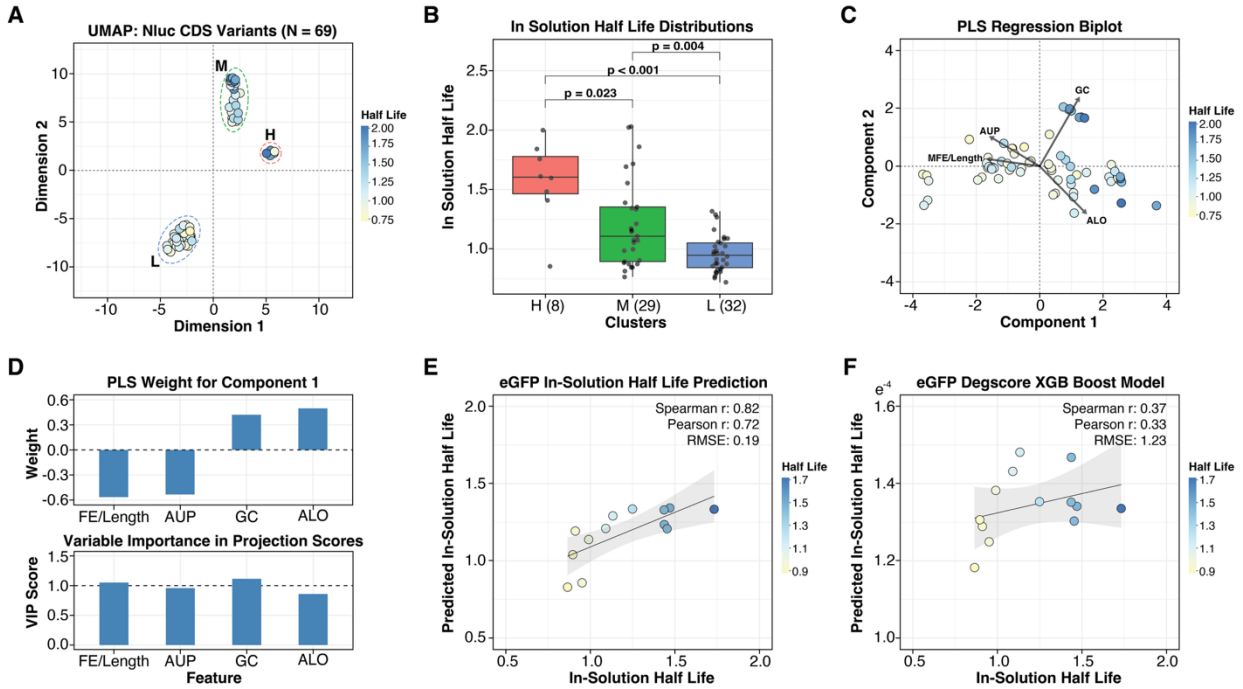

Figure 5. Multivariate regression model of RNA degradation shows improved predictive value with orthogonal contribution from Average Log Odds (ALO). The curated NLuc and eGFP mRNA variants[9] were utilized for multivariate regression modeling. For the eGFP mRNA variants, in-solution half-life from previously published deep learning models[21] were used as a comparison.

(A) UMAP projection of NLuc sequences based on a multi-feature model including FE/length, AUP, GC content, and ALO, showing separation into three clusters.

(B) Cumulative distribution of in-solution half-life distributions for each UMAP-derived cluster, with statistical comparisons indicating significant differences between groups. The Mann-Whitney test was used to calculate p-values. The number of sequences within each cluster is noted in parentheses.

(C) Partial Least Squares Regression (PLSR) biplots of NLuc sequences using feature sets ALO, plotted along Component 1 and Component 2. Arrows indicate loading vectors for each feature. Points are colored by in-solution half-life.

(D) PLSR Component 1 weighs each feature (Top). Positive and negative weights indicate the direction and magnitude of each feature's contribution to the first latent component. Variable Importance in Projection (VIP) scores across all components (Bottom). The dashed line indicates a canonically used threshold to define important variables.

(E) PLSR model predictions of in-solution half-life for eGFP sequences using models trained on NLuc variants. A linear regression fit is shown with a 95% confidence interval represented by the shaded region.

(F) DegScore-XGBoost model predictions for the same eGFP sequences as in (D). A linear regression fit is shown with a 95% confidence interval represented by the shaded region.

Previous work by Wayment-Steele et al.[21] evaluated statistical and deep learning models[7] and tested on these full-length transcripts, achieving Spearman correlations of 0.36–0.48 on orthogonal degradation datasets[9,21]. Despite the advantage of large-scale, high-resolution input, predictive performance on full-length sequences remained modest, highlighting the challenge of generalizing across design contexts.

To test whether LO complements existing structural metrics, we incorporated ALO alongside FE/length, AUP, and GC content in a unified UMAP embedding (**Figure 5A**). This integrated model revealed improved separation between NLuc sequence clusters, with all clusters exhibiting statistically significant differences in in-solution half-life (**Figure 5B**). Notably, the p-value between the high half-life (H; red) and medium half-life (M; green) clusters improved from 0.62 without ALO (**Figure 1G**) to 0.023 with ALO (**Figure 5B**), based on a Mann-Whitney test, representing approximately a 27-fold increase in significance. These findings demonstrate that LO contributes orthogonal information, enabling better resolution of stability differences that traditional global metrics alone fail to capture.

Since no single metric fully explains RNA stability, we next sought to construct a multivariate regression model integrating key structural features. An ideal model should balance explanatory power with generalizability while minimizing overfitting due to the small dataset (N = 69). Given the strong correlations among features, we selected a minimal, non-redundant feature set that captures complementary aspects of RNA structure. FE (normalized by length) and AUP represent the global stability of secondary structure, which influences degradation kinetics. GC content captures sequence composition effects on stability, as GC-rich regions tend to form stronger secondary structures. ALO reflects local structural differences missed by traditional metrics, providing finer resolution of base-pairing variability. This feature set ensures that we

account for thermodynamic stability, sequence composition biases, and fine-scale structural variability.

We next used Partial Least Squares Regression (PLSR)[31] to assess the linear relationships between the feature sets and in-solution half-lives of NLuc sequences, as it is well-suited for modeling datasets with few observations and collinear predictors. The PLSR biplot revealed partially overlapping contributions from featured traditional metrics—FE normalized by length, AUP, and GC content (**Figure 5C**). GC content emerged as the dominant contributor, projecting strongly along both principal components and aligning with samples exhibiting longer half-lives. In contrast, FE/length and AUP were collinear and pointed in the opposite direction of GC, suggesting potential redundancy in the information they provide. In contrast, the ALO loading vector was clearly separated from traditional metrics (both GC and FE/length), contributing orthogonal information.

To further interpret the contributions of each feature in the PLSR model, we examined both the component weights and Variable Importance in Projection (VIP) scores. Component 1 weights revealed that GC content and ALO contributed most strongly to the positive direction of the latent axis associated with increasing half-life, while FE/Length and AUP contributed in the opposite direction (**Figure 5D, top**). In contrast, VIP scores, which reflect the overall importance of each feature across all components, identified GC content and FE/Length as the most influential predictors (VIP = 1.05 and 1.11, respectively), with AUP and ALO contributing moderately (VIP = 0.96 and 0.86, respectively) (**Figure 5D, bottom**). We refer to this sequence and structure feature-based model as STRAND (Stability Regression Analysis using Nucleotide-Derived features). STRAND integrates global (FE, AUP), local structural (ALO) and sequence features (GC) into a compact, interpretable framework for predicting mRNA degradation. Notably, ALO exhibited the highest weight along Component 1, the primary latent axis associated with increasing half-life. Although its overall contribution across all components was below one, the strong Component 1 loading suggests that ALO captures orthogonal structural information that helps differentiate sequences not resolved by traditional metrics.

The bivariate correlations between the variables further support the unique informational content provided by ALO (**Figure S4**). While ALO exhibited strong negative

correlations with FE/Length (r = -0.87) and AUP (r = -0.90), it showed moderate positive correlation with GC (r = 0.37) accompanied by a broad distribution of values. In contrast, FE/Length and AUP both showed stronger, more consistent negative correlations with GC (r = –0.66 and –0.47, respectively). This divergent relationship between ALO and GC demonstrated a new pattern of covariance that was not captured by the initial set of predictors (GC, AUP and FE/length).

Next, we aimed to evaluate the generalizability and predictive utility of STRAND. Given the limited size of the NLuc dataset (N = 69), we did not perform a traditional train-test split, as doing so would risk underpowering the model and obscuring true structure-function relationships. Instead, we trained the models on the full NLuc dataset and tested their ability to generalize to the eGFP variants (N=13). To maximize the data utilized in each training fold given our small dataset, we performed Leave-One-Out Cross-Validation (LOOCV) for model tuning and evaluation. The model trained using FE/length, AUP, GC, and ALO resulted in strong predictive performance accuracy across all metrics (Spearman r = 0.82, Pearson r = 0.72, RMSE = 0.19, MAE = 0.17), highlighting the effectiveness of this compact feature set for accurate and generalizable mRNA stability prediction (**Figure 5F, right**).

To compare STRAND's performance with state-of-the-art published approaches, we evaluated half-life predictions using the benchmark DegScore-XGBoost model described by Wayment-Steele et al.[21] (**Figure 5F**), along with the two top-performing models from that study: Nullrecurrent, the overall first-place model, and Kazuki2, the second-place model which achieved the best performance on the specific dataset we used (**Figure S5B, S5C**). Although all three models predicted degradation rates at the nucleotide level, we applied the provided ansatz to convert these per-position rates into predicted half-lives[21]. Specifically, degradation rates were first obtained from the total degradation rate (SUP-k-deg). Predicted half-lives were then calculated using the relationship: half-life = ln(2)/SUP-k-deg and compared to experimentally measured values.

Using the same eGFP test set, the benchmark DegScore-XGBoost model achieved a Spearman correlation of 0.37 (Pearson r = 0.33) between observed and predicted values, and RMSE of 1.23 which was more than 6-times higher than the RMSE of STRAND predictions. The two other models, Kazuki2 and Nullrecurrent showed similar

performance, with Spearman correlation coefficients of 0.45 and 0.34 respectively (Pearson *r* of 0.35 and 0.16 respectively), and RMSE of 1.23 and 1.23 respectively. Overall, STRAND outperformed the existing models by improving the correlation between the predicted and the observed half-lives, while significantly reducing the error. Together, these results demonstrate that integrating local and global structural features in a model like STRAND improves the accuracy and mRNA degradation prediction generalizability, providing a practical framework for designing stable mRNAs.

Discussion

In this study, we examined the relationship between RNA secondary structure and mRNA in-solution half-life, focusing on the predictive power of different structural metrics. Our findings highlight two key contributions. First, we show that average log-odds (ALO) of base-pairing probability provides orthogonal information that complements traditional metrics such as folding energy (FE), average unpaired probability (AUP), and GC content. This orthogonality enables ALO to differentiate sequences that have similar traditional structure feature values but exhibit markedly different stability. Second, we demonstrate that a multivariate model integrating ALO with FE, AUP, and GC content performs better in predicting mRNA stability than previously published statistical and deep learning models, despite relying on a minimal and interpretable four-feature set. This compact feature combination balances complementary contributions while avoiding redundancy, resulting in a robust and generalizable framework for modeling RNA degradation.

While MFE has long been used as a proxy for RNA structural stability[12,32,33], it provides only a single, optimal conformation[13,34] and does not account for the diversity of structures an RNA sequence can adopt[10,11]. As illustrated in **Figure 2**, sequences with similar MFE values can exhibit vastly different structural topologies, such as the number and arrangement of stem-loops, bulges, or multibranch loops. These variations lead to distinct base-pairing patterns that can dramatically affect degradation rates[35,36], yet are not captured by MFE alone.

ALO and AUP are both derived from base-pairing probability (BPP), which ranges from 0 (unpaired) to 1 (paired). While BPP provides valuable insights, it is limited by its compressed range, making it difficult to detect subtle but functionally meaningful differences near the extremes. ALO addresses this by expanding BPP values into a logarithmic scale, ranging from $-\infty$ to $+\infty$, thereby amplifying differences near 0 and 1 and improving interpretability of pairing confidence[27]. For example, BPP values of 0.9, 0.99, and 0.999 are all close to 1 and difficult to distinguish on the original scale, but map to log-odds of 2.20, 4.60, and 6.91, respectively, offering clearer resolution of high pairing confidence. Similarly, low BPP values (e.g., 0.01 vs. 0.1) become more distinguishable once transformed, allowing better identification of confidently unpaired regions. This enhanced contrast allows ALO to highlight structural differences between sequences that

appear similar using traditional metrics, as demonstrated by our binning analysis. The binning analysis in our results demonstrates that ALO can provide quantifiable new information when AUP fails. These findings led us to explore whether combining ALO with other established metrics could enhance predictive modeling of mRNA stability.

Building on these insights, we developed STRAND, a multivariate regression model incorporating ALO, FE, AUP, and GC content. Despite its compact four-feature design, this model performs competitively with previously published statistical and deep learning approaches while maintaining greater interpretability. Importantly, it generalizes well to independent data, including eGFP constructs, with strong out-of-sample performance (**Figure 5D**). The model's simplicity makes it especially suitable for early-stage screening and optimization workflows, where interpretability and low feature complexity are critical. Moreover, its performance without reliance on experimental probing data highlights the value of well-chosen *in silico* features for robust prediction.

Looking ahead, the integration of LO into predictive pipelines opens new possibilities for rational mRNA design. As mRNA therapeutics expand across diverse delivery platforms and disease areas, there is growing demand for compact, interpretable models that balance predictive accuracy with practical utility. Future efforts that build on this framework with larger, more diverse datasets and context-aware features may yield even more robust predictors of RNA stability and functionality.

Acknowledgements

The authors thank Chris Pepin for the preliminary discussions of this work, Alicia Bicknell and Chiaowen Joyce Hsiao for critical reading of the manuscript. This work was supported in part by the NIH under the award T32 GM152319 (S.Y.)

Declaration of Interests

M.M., S.E.A., W.D., Y.J. are currently employees of and shareholders in Moderna Inc.

Methods

RNA Structure Prediction and Metric Calculation

Data Acquisition

RNA sequence and corresponding half-life data were obtained from the study by Leppek et al.[9] accompanying the OpenVaccine: COVID-19 mRNA Vaccine Degradation Prediction dataset[9,21].

Feature Calculation using LinearPartition

Acquired RNA sequences were used as input for the LinearPartition[30] software to generate structural metrics, including the free energy (FE) predictions, the base pairing probability (BPP) for each nucleotide position, the average base pairing probability across the sequence, and the average unpaired probability (AUP) across the sequence. The base pairing probability (BPP) at each nucleotide position is calculated as the summed probability of that nucleotide forming a base pair with any other nucleotide across the entire predicted structural ensemble derived from the partition function.

Structure Visualization using RNAfold

For visual representation of predicted RNA secondary structures, sequences were submitted to the RNAfold[43] web server. Default options provided by the web server were used for both structure prediction and visualization.

Log Odds Calculation

Using the base pairing probability per position (BPP) calculated by LinearPartition[30], the Log Odds Per Position (LOPP) was determined for each nucleotide using the formula: $LOPP = \ln(p / (1-p))$, where p is the BPP. To handle probabilities of exactly 0 or 1, a pseudocount of 1e-10 was added to p when BPP = 0 and subtracted from p when BPP = 1 prior to the log odds calculation.

BPP Curvature Analysis and Thresholding

The relationship between BPP and LOPP was examined by calculating the mathematical curvature of the BPP vs. LOPP curve. Curvature was determined by numerically

calculating the second derivative of LOPP with respect to BPP ($d^2(\ln(p/(1-p)))/dp^2$) within the boundary conditions $0 \leq p \leq 1$. This analysis identified points of maximum curvature at BPP values of approximately 0.20 and 0.80. Based on this, a BPP threshold of 0.2 was selected to define positions considered likely to be unpaired for downstream analyses.

Sequence Clustering Analysis

Uniform Manifold Approximation and Projection (UMAP)[29] was employed to explore potential clustering based on sequence features. Input features included free energy normalized by sequence length, Average Unpaired Probability (AUP), and GC content. Analyses were performed both including and excluding the Average Log Odds (ALO) as an additional feature. UMAP was performed using the UMAP package (Konopka, T. ; DOI: 10.32614/CRAN.package.umap) in R (v 4.4.0 ; https://www.r-project.org) utilizing default parameters (including n_neighbors=15, min_dist=0.1, n_epochs=200). Due to the stochastic nature of the algorithm, UMAP analysis was repeated 20 times for each feature set; cluster consistency was assessed by visually confirming the presence of three distinct clusters and the stability of sequence identities within each cluster across runs.

Linear Regression Modeling

Linear regression models using FE and AUP as predictors were generated using the stats package in R (v 4.4.0).

PLS Regression Modeling

To model the relationship between sequence-derived features and in-solution half-life, Partial Least Squares Regression (PLSR)[31,44] was implemented using the pls package in R (v 4.4.0). The final set of predictor variables included free energy normalized by sequence length, Average Unpaired Probability (AUP), GC content, and Average Log Odds (ALO). Predictor variables were centered and scaled prior to modeling. Given the limited training dataset size (N = 69), Leave-One-Out Cross-Validation (LOOCV) was utilized on the training set for model parameter tuning by minimizing the Root Mean Squared Error of Prediction (RMSEP) (Fig. S5A). Variable Importance in Projection (VIP) scores were calculated using the plsVarSel package[45] in R (v 4.4.0 ; https://www.r-

project.org) to assess the contribution of each predictor variable to the final PLSR model trained on the full training set. Final model performance was evaluated on a separate, independent EGFP dataset (N = 13) by calculating the Pearson and Spearman correlation coefficients, and Root Mean Square Error (RMSE) between predicted and actual half-life values.

Supplement Figures

Figure S1

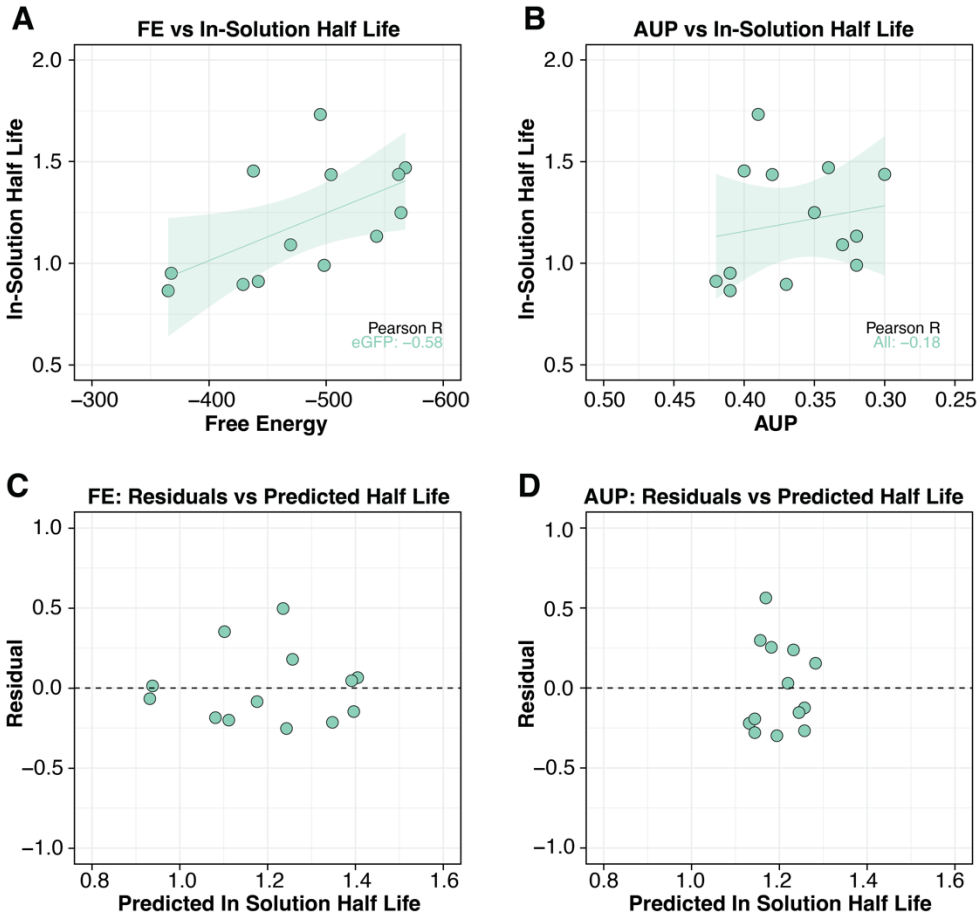

Figure S1. Structure–function analysis in the eGFP set

We analyzed a curated dataset of eGFP mRNA variants[9] to examine the relationship between RNA structural features, such as free energy (FE) and average unpaired probability (AUP), and experimentally measured in-solution half-lives.

(A) Correlation of Free Energy (FE) and half-life in the eGFP dataset.
(B) Same as (A) but for AUP.
(C) Residual plots for FE vs in-solution half-life.
(D) Same (C) but for AUP.

Figure S2

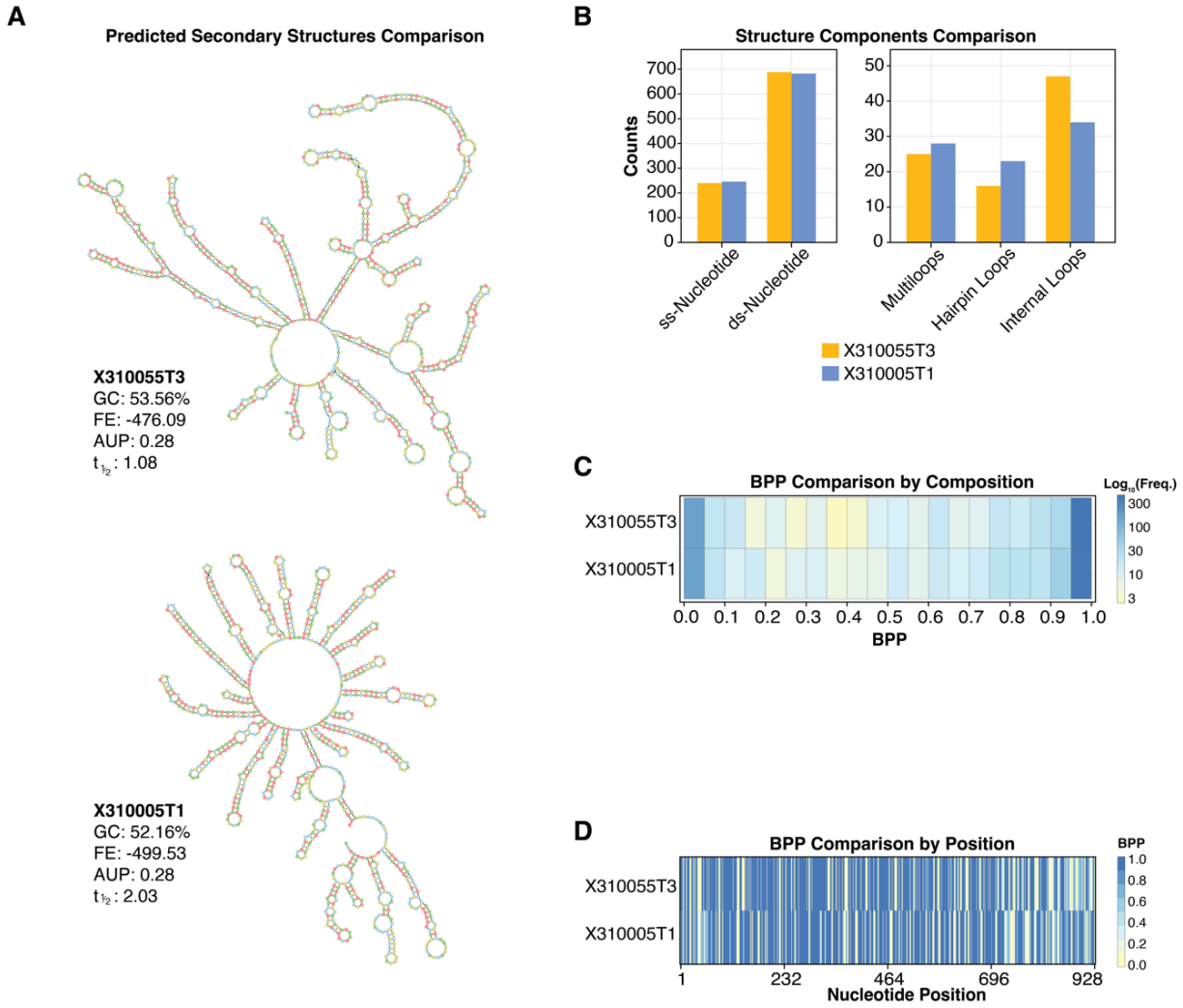

Figure S2. Structure comparison between two RNA sequences at lower FE bin with similar global metrics but different in in-solution half-life.

We predicted RNA secondary structures for sequences X310055T3 and X210005T1, and analyzed their structural components and BPPs.

(A) Predicted secondary structures of two NLuc sequences, X310055T3 and X310005T1, which exhibit similar global structure metrics (FE and AUP) but differ substantially in in-solution half-life.

(B) Barplot comparing structural components between the predicted maximum expected accuracy structures, including counts of unpaired nucleotides, hairpin loops, interior loops, and multiloops.

(C) Heatmap comparing BPPs binned at 0.05 resolution, showing clear differences in the distribution of local pairing strength between the two sequences.

(D) Heatmap comparing local BPP values by position, revealing fine-scale structural differences between the two sequences.

Figure S3

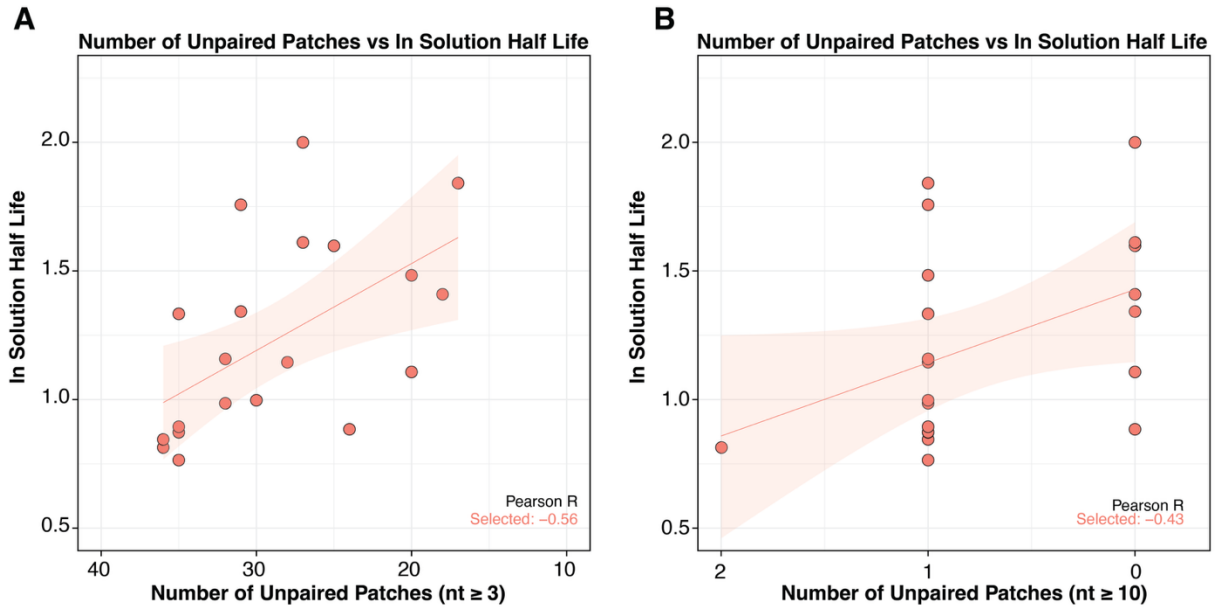

Figure S3. Correlation between the number of unpaired patches and in-solution half-life

We analyzed the predicted secondary structures of NLuc-21 to investigate whether local structural differences could explain their divergent in-solution half-lives.

(A) ≥3 consecutive unpaired positions with BPP < 0.2
(B) ≥10 consecutive unpaired positions with BPP < 0.2

Figure S4

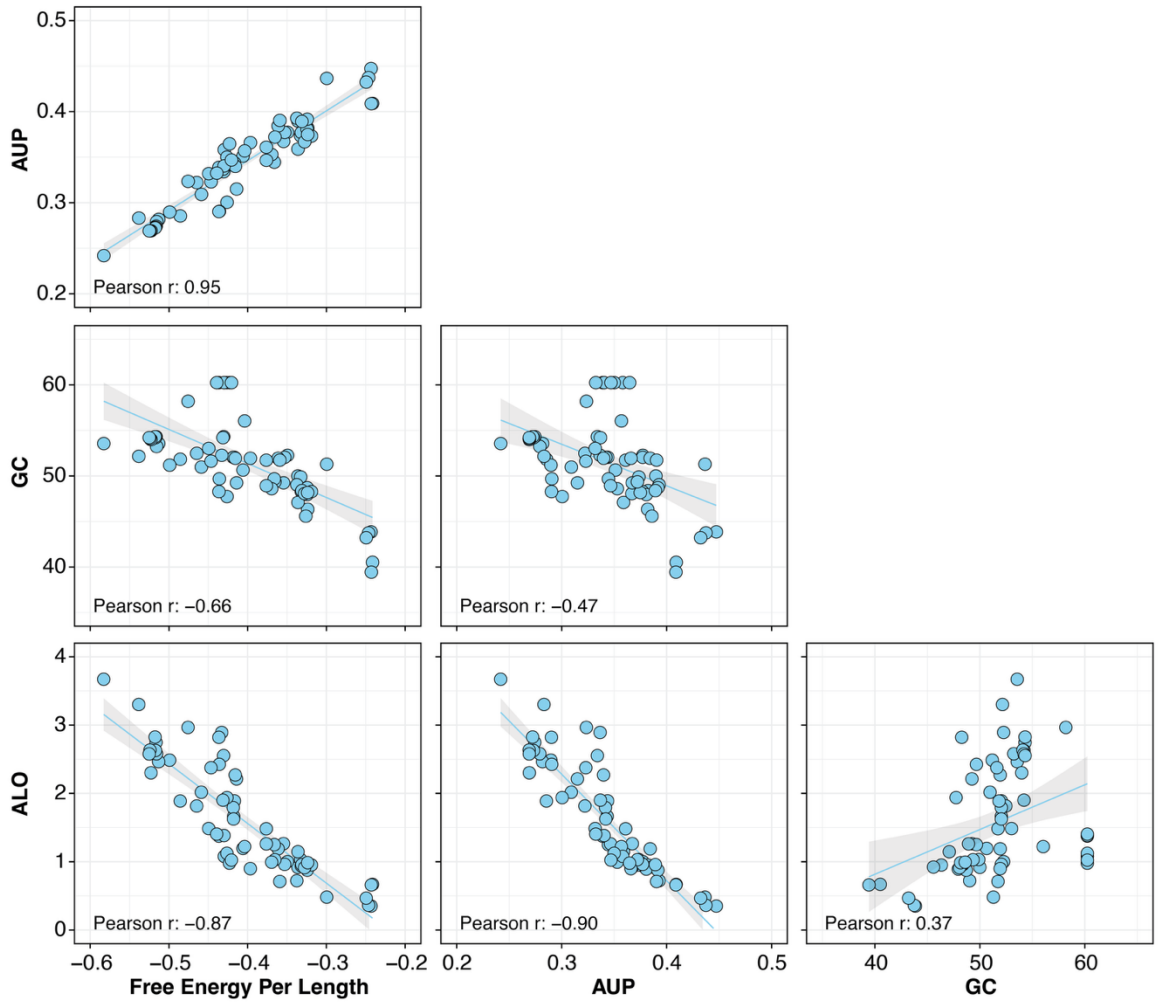

Figure S4. Bivariate correlations between independent variables used in PLSR modeling.

The relationships between the independent variables (FE/Length, AUP, GC, and ALO) for the curated Nluc mRNA variants[9] used in PLSR modeling are shown.

Figure S5

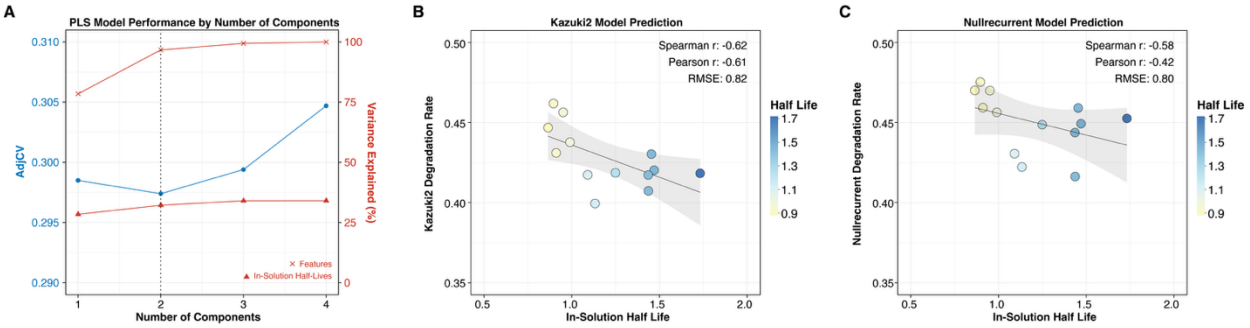

Figure S5. PLS Model Parameters and Previously Reported Half-life prediction models

Multivariate regression model was trained using the curated Nluc mRNA variants[9]. For the eGFP mRNA variants, in-solution half-lives predicted using previously published deep learning models[21] are shown.

(A) Parameters for PLS model trained using FE/length, AUP, GC, and ALO.
(B) In-solution stability prediction results for Kazuki2 model[21].
(C) In-solution stability prediction results for Nullrecurrent model[21].

References

1. Baden, L. R. *et al.* Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *N. Engl. J. Med.* **384**, 403–416 (2021).
2. Crommelin, D. J. A., Anchordoquy, T. J., Volkin, D. B., Jiskoot, W. & Mastrobattista, E. Addressing the Cold Reality of mRNA Vaccine Stability. *J. Pharm. Sci.* **110**, 997–1001 (2021).
3. Li, Y. & Breaker, R. R. Kinetics of RNA Degradation by Specific Base Catalysis of Transesterification Involving the 2‘-Hydroxyl Group. *J. Am. Chem. Soc.* **121**, 5364–5372 (1999).
4. Garneau, N. L., Wilusz, J. & Wilusz, C. J. The highways and byways of mRNA decay. *Nat. Rev. Mol. Cell Biol.* **8**, 113–126 (2007).
5. Pogocki, D. & Schöneich, C. Chemical stability of nucleic acid–derived drugs. *J. Pharm. Sci.* **89**, 443–456 (2000).
6. Mikkola, S., Kaukinen, U. & Lönnberg, H. The effect of secondary structure on cleavage of the phosphodiester bonds of RNA. *Cell Biochem. Biophys.* **34**, 95–119 (2001).
7. Chheda, U. *et al.* Factors Affecting Stability of RNA – Temperature, Length, Concentration, pH, and Buffering Species. *J. Pharm. Sci.* **113**, 377–385 (2024).
8. Mauger, D. M. *et al.* mRNA structure regulates protein expression through changes in functional half-life. *Proc. Natl. Acad. Sci.* **116**, 24075–24083 (2019).
9. Leppek, K. *et al.* Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics. *Nat. Commun.* **13**, 1536 (2022).
10. Ward, M., Courtney, E. & Rivas, E. Fitness functions for RNA structure design. *Nucleic Acids Res.* **51**, e40 (2023).
11. Dirks, R. M., Lin, M., Winfree, E. & Pierce, N. A. Paradigms for computational nucleic acid design. *Nucleic Acids Res.* **32**, 1392–1403 (2004).
12. Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure1. *J. Mol. Biol.* **288**, 911–940 (1999).
13. Mathews, D. H. *et al.* Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci.* **101**, 7287–7292 (2004).
14. Bicknell, A. A. *et al.* Attenuating ribosome load improves protein output from mRNA by limiting translation-dependent mRNA decay. *Cell Rep.* **43**, (2024).
15. Cohen, B. & Skiena, S. Natural selection and algorithmic design of mRNA. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **10**, 419–432 (2003).
16. Ward, M., Richardson, M. & Metkar, M. mRNA Folding Algorithms for Structure and Codon Optimization. Preprint at https://doi.org/10.48550/arXiv.2503.19273 (2025).

17. Terai, G., Kamegai, S. & Asai, K. CDSfold: an algorithm for designing a protein-coding sequence with the most stable secondary structure. *Bioinformatics* **32**, 828–834 (2016).
18. Wayment-Steele, H. K. *et al.* Theoretical basis for stabilizing messenger RNA through secondary structure design. *Nucleic Acids Res.* **49**, 10604–10617 (2021).
19. Zhang, H. *et al.* Algorithm for optimized mRNA design improves stability and immunogenicity. *Nature* **621**, 396–403 (2023).
20. Gu, X., Qi, Y. & El-Kebir, M. DERNA Enables Pareto Optimal RNA Design. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **31**, 179–196 (2024).
21. Wayment-Steele, H. K. *et al.* Deep learning models for predicting RNA degradation via dual crowdsourcing. *Nat. Mach. Intell.* **4**, 1174–1184 (2022).
22. Kaukinen, U., Lyytikäinen, S., Mikkola, S. & Lönnberg, H. The reactivity of phosphodiester bonds within linear single-stranded oligoribonucleotides is strongly dependent on the base sequence. *Nucleic Acids Res.* **30**, 468–474 (2002).
23. Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E. & Weeks, K. M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods* **11**, 959–965 (2014).
24. Fig. 5: Kaggle models demonstrate improved performance in independent test of degradation of full-length mRNAs. | Nature Machine Intelligence.
25. Lu, Z. J., Gloor, J. W. & Mathews, D. H. Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA* **15**, 1805–1813 (2009).
26. Hofacker, I. L., Bernhart, S. H. F. & Stadler, P. F. Alignment of RNA base pairing probability matrices. *Bioinformatics* **20**, 2222–2227 (2004).
27. Norton, E. C. & Dowd, B. E. Log Odds and the Interpretation of Logit Models. *Health Serv. Res.* **53**, 859–878 (2018).
28. Metkar, M., Pepin, C. S. & Moore, M. J. Tailor made: the art of therapeutic mRNA design. *Nat. Rev. Drug Discov.* **23**, 67–83 (2024).
29. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at https://doi.org/10.48550/arXiv.1802.03426 (2020).
30. LinearPartition: linear-time approximation of RNA folding partition function and base-pairing probabilities | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/36/Supplement_1/i258/5870487.
31. Wold, S., Ruhe, A., Wold, H. & Dunn, III, W. J. The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM J. Sci. Stat. Comput.* **5**, 735–743 (1984).
32. Flamm, C., Fontana, W., Hofacker, I. L. & Schuster, P. RNA folding at elementary step resolution. *RNA* **6**, 325–338 (2000).
33. Hofacker, I. L. *et al.* Fast folding and comparison of RNA secondary structures. *Monatshefte Für Chem. Chem. Mon.* **125**, 167–188 (1994).

34. Dowell, R. D. & Eddy, S. R. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics* **5**, 71 (2004).
35. Zarringhalam, K., Meyer, M. M., Dotu, I., Chuang, J. H. & Clote, P. Integrating Chemical Footprinting Data into RNA Secondary Structure Prediction. *PLOS ONE* **7**, e45160 (2012).
36. Cordero, P. & Das, R. Rich RNA Structure Landscapes Revealed by Mutate-and-Map Analysis. *PLOS Comput. Biol.* **11**, e1004473 (2015).
37. Agarwal, V. & Kelley, D. R. The genetic and biochemical determinants of mRNA degradation rates in mammals. *Genome Biol.* **23**, 245 (2022).
38. Cheng, F. *et al.* Research Advances on the Stability of mRNA Vaccines. *Viruses* **15**, 668 (2023).
39. Feng, L. & Niu, D.-K. Relationship Between mRNA Stability and Length: An Old Question with a New Twist. *Biochem. Genet.* **45**, 131–137 (2007).
40. Chaturvedi, M., Rashid, M. A. & Paliwal, K. K. RNA structure prediction using deep learning — A comprehensive review. *Comput. Biol. Med.* **188**, 109845 (2025).
41. Akiyama, M. & Sato, K. RNA Secondary Structure Prediction Based on Energy Models. in *RNA Structure Prediction* (eds. Kawaguchi, R. K. & Iwakiri, J.) 89–105 (Springer US, New York, NY, 2023). doi:10.1007/978-1-0716-2768-6_6.
42. Doshi, K. J., Cannone, J. J., Cobaugh, C. W. & Gutell, R. R. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics* **5**, 105 (2004).
43. Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R. & Hofacker, I. L. The Vienna RNA Websuite. *Nucleic Acids Res.* **36**, W70–W74 (2008).
44. Wold, S., Sjöström, M. & Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **58**, 109–130 (2001).
45. Mehmood, T., Liland, K. H., Snipen, L. & Sæbø, S. A review of variable selection methods in Partial Least Squares Regression. *Chemom. Intell. Lab. Syst.* **118**, 62–69 (2012).