# Fine-Scale Structural Information Substantially Improves mRNA Therapeutic Stability 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

[*]Correspondence should be addressed to M.M. (mihir.metkar@modernatx.com)

**ABSTRACT**

The success of COVID-19 mRNA vaccines has made the in-solution stability optimization of mRNAs a key objective. However, we still lack a complete understanding of sequence metrics that influence mRNA in-solution stability. RNA secondary structure plays a critical role in protecting against hydrolysis, the primary degradation pathway under storage conditions. Yet, the structural metrics that best guide stability-focused mRNA design remain incompletely defined. Global metrics like minimum free energy and average unpaired probability have improved mRNA stability but fail to capture local structural variation relevant to RNA degradation. We demonstrate that base-pairing log odds provide fine-scale, orthogonal insight that complements global metrics and improves stability modeling. Further, by combining local and global features into a parsimonious four-feature regression model, dubbed **St**ability **R**egression **A**nalysis using **N**ucleotide-**D**erived features (STRAND), we achieve a greater than two-fold reduction in prediction error compared to existing machine learning and deep learning approaches and demonstrate robust generalization across diverse transcript contexts. This compact and interpretable model provides an accurate and reliable framework for predicting mRNAs in-solution stability.

## 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 mRNA-based vaccine development 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 between mRNA sequence variants 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[21]. 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 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[24,25], which is bounded between 0 and 1 and compresses extreme values, the log-odds transformation expands the dynamic range[26], particularly for high and low probabilities, thereby amplifying meaningful structural differences that might otherwise be obscured. LO reflects the local base-pairing probability distribution, offering finer resolution than global metrics like FE or AUP. By analyzing mRNAs with similar global

77 scores but divergent in-solution half-lives, we show that LO captures functional
78 differences not apparent in traditional metrics. Finally, we demonstrate that incorporating
79 LO into a compact four-feature regression model, Stability Regression Analysis using
80 Nucleotide-Derived features (STRAND), significantly improves mRNA stability prediction
81 over existing approaches and generalizes across diverse transcript contexts, providing
82 an interpretable framework for stability-focused mRNA design.
83

84 **RESULTS**

85 **Traditional RNA Structural Metrics Partially Explain In-Solution Stability**

86 A major goal in RNA therapeutic design is to extend mRNA half-life in-vial[27].
87 Existing studies have attempted to improve stability using sequence-derived structural
88 metrics, such as FE and AUP, which provide approximations of RNA secondary
89 structure[18]. However, whether these metrics alone are sufficient to predict mRNA
90 degradation rates remains an open question[9].

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

97 Across the NLuc dataset, we observed moderate negative correlations between
98 half-life and FE (Pearson r = -0.52) as well as AUP (Pearson r = -0.47) (red and blue
99 points in **Figure 1B, 1C**). These correlations suggest that mRNAs with lower FE and AUP,

100 indicative of more stable secondary structures, tend to degrade more slowly. However,
101 while these metrics broadly capture structural trends, they fail to fully account for the
102 variability in observed degradation rates. For example, the NLuc sequence variants with
103 FE around -400 kcal/mol (N = 21, NLuc-21 set) showed a lower correlation than the global
104 comparison, with Pearson correlation values of -0.38 and 0.29, respectively (**Figure 1B,**
105 **1C**, highlighted in red). Moreover, linear regression models using FE and AUP as
106 predictors exhibited high residual errors, particularly at higher values of either metric
107 (**Figure 1D, 1E**). We also noted similar observations for eGFP CDS variants (N = 13;
108 **Figure S1**), in which eGFP sequences with FE near –500 kcal/mol show approximately
109 1.7-fold differences in their measured half-lives and sequences with AUP near 0.40 show
110 approximately 2-fold differences in their half-lives. These findings suggest that the
111 observed discrepancy between the traditional metrics and degradation rates is a general
112 phenomenon across different coding sequence sets. While global structure formation
113 plays a role in stability, additional sequence-dependent factors influence degradation in
114 ways that traditional metrics fail to capture.

115 To further investigate this variability, we applied Uniform Manifold Approximation
116 and Projection (UMAP) to reduce the three-dimensional structural feature space into a
117 lower-dimensional representation.[28] The resulting UMAP embedding, based on traditional
118 metrics, including minimum FE normalized by transcript length (FE/length), average
119 unpaired probability (AUP), and GC content, separated the NLuc sequences into distinct
120 clusters (**Figure 1F**). Normalizing FE by transcript length enabled comparison across
121 sequences of varying lengths, ensuring that the energy values reflect general structural
122 stability rather than absolute sequence length. The resulting UMAP clusters were

identified as high (H), medium (M), and low (L) based on the average in-solution half-life. 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, and aimed to determine whether additional features could explain their stability discrepancies.

**Local Structural Features Differentiate Sequences with Similar Global Metrics**

The two sequences, X320004D1 and X930007D1, had FEs 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 three-fold (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**)[29]. A heatmap visualization of per-position base-pairing probabilities (BPPs) (**Figure 2B**) 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. Specifically, X320004D1 formed a more branched secondary structure, with a higher number of multiloops (41) and hairpin loops (24), which could introduce greater structural

146 flexibility and accessibility. In contrast, X930007D1 contained fewer multiloops (34) and
147 hairpin loops (17) but a higher number of inner loops (45 vs. 32 in X320004D1), forming
148 more extended helical regions (**Figure 2C**). To quantify these differences, we compared
149 the BPP distributions of the two sequences (**Figure 2D**). The BPP distribution of
150 X320004D1 showed a sharp bimodal pattern, with clear peaks near 0.0 (unpaired) and
151 1.0 (paired), suggesting more well-defined structural elements. In contrast, X930007D1
152 exhibited a broader, more continuous distribution, with a dominant peak near 1.0 but more
153 intermediate BPP values, indicating a less rigid structure.

154 These divergences between the global metrics and stability were not unique to this
155 sequence pair. Among sequences with higher stability (FE near -475 kcal/mol), we
156 observed similar disconnect between global metrics and measured half-lives. For
157 example, X310055T3 and X310005T1 showed highly distinct predicted structures
158 (**Figure S2A**) and measured half-lives (1.08 vs 2.03 h), despite having nearly identical
159 FE (–476.09 vs –499.53 kcal/mol), GC content (53.56 vs 52.16%), and AUP of 0.28. In
160 addition, the two sequences showed varying structural compositions (**Figure S2B**) and
161 BPP distributions (**Figure S2C, D**). Together, these examples demonstrate that global
162 metrics alone do not capture the local structural variation that may underlie differences in
163 RNA stability, motivating a more systematic analysis of BPPs across a broader sequence
164 set.

165

166 **Base-Pairing Probability Alone Does Not Fully Explain mRNA Stability**

167 To determine whether the structural patterns observed in X320004D1 and
168 X930007D1 extend to a broader dataset, we analyzed 21 NLuc variants (NLuc-21; red

169 points in **Figure 1B, C**) with similar FE and AUP values but differing in-solution half-lives.
170 Despite having nearly identical global structural metrics, these sequences displayed
171 distinct patterns of local base-pairing interactions (**Figure 3A, 3B**).

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

180 To quantify these differences, we examined whether the number of unpaired
181 positions (BPP < 0.2) or the presence of long unpaired patches (≥5 nucleotides)
182 correlated with half-life. Pearson correlation analysis showed that neither total unpaired
183 positions (r = -0.47, **Figure 3C**) nor long unpaired patches (r = -0.32, **Figure 3D**) provided
184 a significantly stronger correlation than FE or AUP. Moreover, our analysis showed no
185 significant improvement in Pearson correlation with different lengths of unpaired patches
186 (≥3 or ≥10 nucleotides; **Figure S3**).

187 These findings indicate that BPP alone does not account for mRNA stability
188 differences among sequences with similar global metrics. Although BPP provides a
189 continuous measure of pairing likelihood at each position, its bounded range (0 to 1)
190 compresses differences at the extremes, making it difficult to resolve functionally
191 meaningful distinctions, especially among positions with high or low pairing confidence.

192 Additionally, when summarized (e.g., by averaging in AUP), BPP can obscure the
193 contribution of unpaired positions, as zero-probability sites are effectively treated as
194 absence of signal rather than informative features. This suggests that a metric with
195 greater dynamic range may be needed to explain the stability variation that global metrics
196 and BPP miss.
197
198 **Log-Odds of Base-Pairing Captures RNA Stability Differences Beyond BPP**
199 To address this limitation, we employed a base-pairing log-odds (LO) metric, which
200 transforms BPP values into an unbounded scale (-∞ to +∞). Highly paired positions (BPP
201 > 0.5) yield positive LO values, while strongly unpaired positions (BPP < 0.5) yield
202 negative values. This log transformation expands the dynamic range at both extremes of
203 the BPP distribution, amplifying differences among positions with very high or very low
204 pairing confidence that are otherwise compressed on the 0-to-1 BPP scale.
205 To test whether LO captures RNA degradation patterns better than BPP, we first
206 examined per-position LO values for X320004D1 and X930007D1 (shown in **Figure 2**).
207 While both sequences had similar global structural metrics, LO values were distributed
208 differently, particularly in weakly structured regions (**Figure 4A)**. Notably, X320004D1
209 contained a greater number of strongly negative LO positions, indicating unpaired regions
210 that are likely more susceptible to degradation. In contrast, X930007D1 exhibited fewer
211 such sites, suggesting fewer vulnerable positions, consistent with its longer half-life.
212 A heatmap visualization of LO across NLuc-21 sequences (**Figure 4B, C**)
213 demonstrated that LO more effectively differentiated stable and unstable sequences
214 compared to BPP. Sequences with shorter half-lives exhibited more widespread strongly

negative LO values (red regions in **Figure 4C**), consistent with a greater number of confidently unpaired positions. To quantify this, we focused on weakly paired regions (LO < -1.363, derived from BPP curvature analysis) and found that the average LO (ALO) of these positions showed a strong positive correlation with in-solution half-life (Pearson r = 0.78, **Figure 4D**), indicating that sequences with less extreme negative LO values in their unpaired regions tend to be more stable.

However, when we computed the ALO across all positions, the correlation with half-life was negative (Pearson r = -0.63, **Figure 4E**). Although this opposing direction appeared counterintuitive, closer analysis of the heatmap in **Figure 4C** revealed that sequences with shorter half-lives also contained more positions with extremely positive LO values (blue regions; LO > 20). These sequences thus exhibit a more polarized LO distribution, with greater representation at both extremes, while longer-lived sequences show more moderate LO profiles overall. When averaged across all positions, the extreme positive values in shorter-lived sequences inflate their ALO, producing the observed negative correlation. Despite this reversal in sign, the magnitude of the ALO correlation still exceeds that of FE or AUP, and the strong positive correlation in the weakly paired regions (r = 0.78) provided a clear link between local structural features and mRNA stability. Overall, these results demonstrate that the LO transformation reveals structurally meaningful variation among sequences that appear similar to traditional metrics.

236 **Multi-feature Regression with ALO Improves mRNA Stability Prediction**

237 Previous work by Wayment-Steele et al.[21] evaluated statistical and deep learning
238 models on these full-length transcripts (**Figure 1A**), achieving Spearman correlations of
239 0.36–0.48 on orthogonal degradation datasets[9,21]. Despite the advantage of large-scale,
240 high-resolution input, predictive performance on full-length sequences remained modest,
241 highlighting the challenge of generalizing across design contexts.

242 To test whether LO complements existing structural metrics, we incorporated ALO
243 alongside FE/length, AUP, and GC content in a unified UMAP embedding (**Figure 5A**).
244 This integrated model revealed improved separation between NLuc sequence clusters,
245 with all clusters achieving statistically significant differences in in-solution half-life (**Figure**
246 **5B**). Notably, the p-value between the high half-life (red) and medium half-life (green)
247 clusters improved from 0.62 without ALO (**Figure 1G**) to 0.023 with ALO (**Figure 5B**),
248 based on a Mann-Whitney test representing approximately a 27-fold increase in
249 significance. This demonstrates that LO contributes orthogonal information, enabling
250 better resolution of stability differences that traditional global metrics alone fail to capture.

251 Since no single metric fully explains RNA stability, we next sought to construct a
252 multi-feature regression model integrating key structural features. An ideal model should
253 balance explanatory power with generalizability while minimizing overfitting, particularly
254 given the small dataset size (N = 69). Given the strong correlations among features, we
255 selected a minimal, non-redundant feature set that captures complementary aspects of
256 RNA structure. FE (normalized by length) and AUP represent the global secondary
257 structure stability, which influences degradation. GC content captures sequence
258 composition effects on stability, as GC-rich regions tend to form stronger base-pairs. ALO

259 reflects local structural differences missed by traditional metrics, providing finer resolution
260 of base-pairing variability. This feature set ensures that we account for thermodynamic
261 stability, sequence composition biases, and fine-scale structural variability.

262 We next used Partial Least Squares Regression (PLSR)[30] to assess the linear
263 relationships between the feature sets and in-solution half-lives of NLuc sequences, as it
264 is well-suited for modeling datasets with few observations and collinear predictors. The
265 PLSR biplot revealed partially overlapping contributions from featured traditional
266 metrics—FE normalized by length, AUP, and GC content (**Figure 5C**). GC content
267 emerged as the dominant contributor, projecting strongly along both principal components
268 and aligning with samples exhibiting longer half-lives. In contrast, FE/length and AUP
269 were collinear and pointed in the opposite direction of GC, suggesting potential
270 redundancy in the information they provide. The ALO loading vector was clearly
271 separated from traditional metrics (both GC and FE/length), confirming its orthogonal
272 contribution.

273 To further interpret the contributions of each feature in the PLSR model, we
274 examined both the component weights and Variable Importance in Projection (VIP)
275 scores. Component 1 weights revealed that GC content and ALO contributed most
276 strongly to the positive direction of the latent axis associated with increasing half-life, while
277 FE/Length and AUP contributed in the opposite direction (**Figure 5D, top**). VIP scores,
278 which reflect the overall importance of each feature across all components, identified GC
279 content and FE/Length as the most influential predictors (VIP = 1.05 and 1.11,
280 respectively), with AUP and ALO contributing moderately (VIP = 0.96 and 0.86,
281 respectively) (**Figure 5D, bottom**).

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). We refer to this sequence and structure feature-based model as Stability Regression Analysis using Nucleotide-Derived features (STRAND). STRAND integrates global (FE, AUP), local structural (ALO) and sequence features (GC) into a compact, interpretable framework for predicting mRNA degradation.

We next 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 (**Figure S5A**)**.** STRAND achieved strong predictive performance 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 mRNA stability prediction (**Figure 5E**).

To compare STRAND's performance with state-of-the-art published approaches, we evaluated half-life predictions using the benchmark DegScore-XGBoost model

305 described by Wayment-Steele et al.[21], along with the two top-performing deep learning
306 models from that study: Nullrecurrent, the overall first-place model, and Kazuki2, the
307 second-place model which achieved the best performance on the specific dataset we
308 used. For the same eGFP test set, all three models predicted degradation scores at the
309 nucleotide level, and the average per-position degradation scores (AUP) showed strong
310 correlation with experimental degradation rates (**Figure S5B**). Predicted half-lives for the
311 eGFP sequences for each model were then calculated using the relation half-life =
312 ln(2)/AUP and compared to the experimentally measured values.

313 The three models achieved comparable performances for the eGFP sequences,
314 with the DegScore-XGBoost model achieving a Spearman correlation of 0.56 (Pearson r
315 = 0.58) and RMSE of 0.46 (**Figure 5F**). NullRecurrent and Kazuki2 showed similar
316 performance, with a Spearman correlation of 0.58 and 0.60, respectively (Pearson r =
317 0.40 and 0.62), and RMSE of 0.42 and 0.48 (**Figure S5C**). By comparison, STRAND
318 achieved a Spearman correlation of 0.82, a Pearson r of 0.72, and RMSE of 0.19,
319 outperforming all three ML models with approximately a two-fold reduction in prediction
320 error relative to the best-performing comparison model. These results demonstrate that
321 integrating local and global structural features in a compact model like STRAND improves
322 both the accuracy and generalizability of mRNA degradation prediction.

323 To further assess STRAND's generalizability, we applied the model to sequences
324 derived from an independent study, VZV gE and SARS-CoV-2 Spike mRNA variants[19].
325 These differed substantially from the NLuc training set in both sequence length and
326 structural feature ranges (**Figure 6A**). Moreover, in-solution half-lives of VZV and Spike
327 sequences, measured under both 20 mM $Mg^{2+}$ (**Figure 6A)** and 10 mM $Mg^{2+}$ (**Figure**

**S6A**) buffer conditions, were substantially longer than those of the NLuc training set. The LO distributions and per-position LO profiles of both VZV and Spike sequences showed a clear trend: sequences with more positions at strongly positive LO values tended to have longer half-lives (**Figures 6B, C**), consistent with the relationship between local structural confidence and stability observed in the NLuc dataset. When STRAND predictions were compared to experimentally measured half-lives, the model preserved strong rank-order accuracy for both VZV (Spearman r = 0.86) and Spike (Spearman r = 0.95) sequences, although absolute prediction errors were large (RMSE = 20.36 and 5.25, respectively) (**Figure 6D**). Notably, the FE values for both VZV and Spike sequences fall well outside the range observed in the NLuc training set, and their GC content values lie near the extremes of the NLuc distribution (**Figure 6A**). The large RMSE likely reflects extrapolation beyond the training domain and the substantial difference in half-life scale between these transcripts and the NLuc training data. This prediction trend was also maintained for the half-lives measured in 10 mM $Mg^{2+}$ buffer condition (**Figure S6B**). Overall, these results show that STRAND generalizes across proteins in predicting stability rank order, although precise half-life prediction will require training data within the relevant distribution.

## 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 ALO of base-pairing probability provides orthogonal information that complements traditional metrics such as FE, 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 multi-feature model integrating ALO with FE, AUP, and GC content outperforms previously published statistical and deep learning models in predicting mRNA stability across variants encoding multiple proteins, 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[12,31,32], it provides only a single, optimal conformation[13,33] 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[34,35], yet are not captured by MFE alone.

ALO and AUP are both derived from 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

370 extremes. Moreover, when BPP values are averaged across a sequence, positions with
371 BPP near zero, which represent confidently unpaired and degradation-prone sites, are
372 effectively treated as an absence of pairing signal rather than as informative structural
373 features. ALO addresses this by expanding BPP values into a logarithmic scale, ranging
374 from $-\infty$ to $+\infty$, thereby amplifying differences near 0 and 1 and improving interpretability
375 of pairing confidence[26]. For example, BPP values of 0.9, 0.99, and 0.999 are all close to
376 1 and difficult to distinguish on the original scale, but map to log-odds of 2.20, 4.60, and
377 6.91, respectively, offering clearer resolution of high pairing confidence. Similarly, low
378 BPP values (e.g., 0.01 vs. 0.1) become more distinguishable once transformed, allowing
379 better identification of confidently unpaired regions. This enhanced contrast allows ALO
380 to highlight structural differences between sequences that appear similar using traditional
381 metrics, as demonstrated by our binning analysis. These findings led us to explore
382 whether combining ALO with other established metrics could enhance predictive
383 modeling of mRNA stability.

384 Building on these insights, we developed STRAND, a multi-feature regression
385 model incorporating ALO, FE, AUP, and GC content. Despite its compact four-feature
386 design, this model performs competitively with previously published statistical and deep
387 learning approaches while maintaining greater interpretability. Importantly, it generalizes
388 well to independent data, including eGFP, VZV gE, and SARS-CoV-2 Spike constructs,
389 with strong rank-order performance even when applied to transcripts well outside the
390 training domain. The model's simplicity makes it especially suitable for early-stage
391 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.

Beyond its predictive utility, STRAND and the LO metric can be readily incorporated into existing mRNA design workflows (**Figure S7**). During the initial sequence design stage, LO or ALO can be an additional objective or constraint within optimization algorithms, complementing traditional metrics to guide the sequence generation with favorable local structural profiles. Alternatively, for a given candidate sequence pool, STRAND can be used to predict in-solution half-lives, or ALO can serve as a filter to prioritize candidates with structural features associated with greater stability prior to experimental validation. In either application, STRAND's computational simplicity and its reliance on readily computed *in silico* features make it practical for integration into high-throughput design pipelines.

While STRAND performs well overall, there are limitations that offer opportunities for further improvement. As expected, prediction accuracy declines at the extremes of the half-life range, likely due to limited representation of highly stable or unstable sequences in the training set (**Figure 1A**). One of the most influential biological variables affecting degradation is transcript length[7,36–38], yet STRAND was trained using variants of a single open reading frame (ORF), as it was the only transcript in the dataset with a sufficient number of sequence variants for training. While STRAND performed reasonably well even for sequence sets of different length and feature values outside of the training regime, expanding the training dataset to include transcripts encoding proteins of varying lengths and sequence features would improve the model's robustness and generalizability. For example, STRAND predictions for MEV sequences were poor

415 (**Figure S8**). This is likely attributable to the short transcript length of MEV variants, for
416 which FE normalization by length becomes unreliable due to the non-linear relationship
417 between FE and sequence length at shorter lengths[39], combined with their narrow FE
418 distributions that fall outside the training regime. Increasing the diversity of the dataset
419 may uncover new features that contribute independently of both ALO and MFE, providing
420 further gains in predictive accuracy and biological insight. Moreover, STRAND currently
421 models CDS-level structural features and does not explicitly account for other important
422 dimensions of mRNA design, such as co-optimization of UTR features, or incorporating
423 structure predictions specific for relevant modified nucleotides. Investigating how these
424 factors interact with the structural features captured by STRAND represents an important
425 direction for future work.

426 For any given value of conventional metrics such as MFE, sequences often share
427 similar range of LO values but differ in their distributions, making it difficult for LO alone
428 to distinguish stability differences across broader structural contexts. LO effectively
429 resolves differences between sequences that share similar global metrics but differ in
430 degradation rates, capturing subtle structural features that traditional approaches often
431 overlook. However, it lacks the broad discriminatory power needed to classify sequences
432 across a wider spectrum of structural contexts. Therefore, the strength of LO lies in
433 complementing existing metrics, providing finer-scale insights particularly valuable in
434 cases where traditional metrics offer limited differentiation.

435 Additionally, a key limitation of our approach stems from inherent inaccuracies in
436 *in silico* RNA secondary structure predictions, which rely on imperfect thermodynamic
437 parameters and do not always fully capture biologically relevant structural ensembles[40–

438 [42]. As our stability model relies on these predicted structures, its accuracy is likely
439 influenced by the quality of the folding algorithms. Continued improvements in RNA
440 structure prediction methods could therefore further enhance the performance and
441 reliability of STRAND.

442 Across the NLuc, eGFP, VZV, and Spike datasets, we observed distinct patterns
443 of LO distributions that are associated with differences in degradation. Understanding
444 how the balance between highly structured and weakly structured regions contributes to
445 degradation across different sequence contexts remains an open question. Nevertheless,
446 STRAND demonstrates that well-chosen structural features can substantially reduce
447 prediction error for mRNA stability, and its compact, interpretable design makes it readily
448 adaptable as new training data and design contexts become available.
449

450 **METHODS**

451 **RNA Structure Prediction and Metric Calculation**

452 Data Acquisition

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

456

457 Feature Calculation using LinearPartition

458 Acquired RNA sequences were used as input for the LinearPartition[29] software to
459 generate structural metrics, including the FE predictions, the BPP for each nucleotide
460 position, the average base pairing probability across the sequence, and the AUP across
461 the sequence. The BPP at each nucleotide position is calculated as the summed
462 probability of that nucleotide forming a base pair with any other nucleotide across the
463 entire predicted structural ensemble derived from the partition function.

464

465 Structure Visualization using RNAfold

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

469

470 Log Odds Calculation

471 Using the BPP calculated by LinearPartition[29], the Log Odds Per Position (LOPP) was
472 determined for each nucleotide using the formula: LOPP = ln(p / (1-p)), where p is the

473 BPP. To handle probabilities of exactly 0 or 1, a pseudocount of $1e^{-10}$ was added to p
474 when BPP = 0 and subtracted from p when BPP = 1 prior to the log odds calculation.
475
476 BPP Curvature Analysis and Thresholding
477 The relationship between BPP and LO was examined by calculating the mathematical
478 curvature of the BPP vs. LO curve. Curvature was determined by numerically calculating
479 the second derivative of LO with respect to BPP ($d^2(\ln(p/(1-p)))/dp^2$) within the boundary
480 conditions $0 \leq p \leq 1$. This analysis identified points of maximum curvature at BPP values
481 of approximately 0.20 and 0.80. Based on this, a BPP threshold of 0.2 was selected to
482 define positions considered likely to be unpaired for downstream analyses.
483
484 **Sequence Clustering Analysis**
485 Uniform Manifold Approximation and Projection (UMAP)[28] was employed to explore
486 potential clustering based on sequence features. Input features included FE normalized
487 by sequence length, Average Unpaired Probability (AUP), and GC content. Analyses
488 were performed both including and excluding the Average Log Odds (ALO) as an
489 additional feature. UMAP was performed using the UMAP package (Konopka, T. ; DOI:
490 10.32614/CRAN.package.umap) in R (v 4.4.0 ; https://www.r-project.org) utilizing default
491 parameters (including n_neighbors=15, min_dist=0.1, n_epochs=200). Due to the
492 stochastic nature of the algorithm, UMAP analysis was repeated 20 times for each feature
493 set; cluster consistency was assessed by visually confirming the presence of three distinct
494 clusters and the stability of sequence identities within each cluster across runs.

495

496 **Linear Regression Modeling**

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

499

500 **PLS Regression Modeling**

501 To model the relationship between sequence-derived features and in-solution half-life,
502 Partial Least Squares Regression (PLSR)[30,44] was implemented using the pls package in
503 R (v 4.4.0). The final set of predictor variables included FE normalized by sequence length,
504 Average Unpaired Probability (AUP), GC content, and Average Log Odds (ALO).
505 Predictor variables were centered and scaled prior to modeling. Given the limited training
506 dataset size (N = 69), Leave-One-Out Cross-Validation (LOOCV) was utilized on the
507 training set for model parameter tuning by minimizing the Root Mean Squared Error of
508 Prediction (RMSEP) (Fig. S5A). Variable Importance in Projection (VIP) scores were
509 calculated using the plsVarSel package[45] in R (v 4.4.0 ; https://www.r-project.org) to
510 assess the contribution of each predictor variable to the final PLSR model trained on the
511 full training set. Final model performance was evaluated on a separate, independent
512 EGFP dataset (N = 13) by calculating the Pearson and Spearman correlation coefficients,
513 and Root Mean Square Error (RMSE) between predicted and actual half-life values.

514

## DATA AVAILABILITY

The STRAND model described in this paper is publicly available to support reproducibility and broad use. The raw training data used in this study were obtained from the OpenVaccine: COVID-19 mRNA Vaccine Degradation Prediction dataset (Leppek et al., 2022). Any additional data required to verify and extend the research are available from the corresponding author upon request.

## AUTHOR CONTRIBUTIONS

**S.Y.:** Conceptualization (Lead), Formal analysis (Lead), Investigation (Lead), Validation (Lead), Visualization (Lead), Writing – original draft (Lead), Writing – review & editing (Lead). **S.E.A.:** Data curation (Supporting), Formal analysis (Supporting), Investigation (Supporting), Writing – original draft (Supporting), Writing – review & editing (Supporting). **Y.J.:** Formal analysis (Supporting), Investigation (Supporting), Validation (Supporting), Writing – review & editing (Supporting). **J.W.D.:** Conceptualization (Lead), Supervision (Supporting), Writing – review & editing (Supporting). **M.M.:** Conceptualization (Lead), Investigation (Lead), Project administration (Lead), Resources (Lead), Supervision (Lead), Writing – original draft (Lead), Writing – review & editing (Lead).

## KEYWORDS

RNA Secondary Structure, RNA Stability, Base Pairing Probability, Base Pairing Log Odds, Computational Prediction, RNA Therapeutics, mRNA Vaccines

## 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., J.W.D., Y.J. are currently employees of and shareholders in Moderna Inc.

## Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the author(s) used Claude and ChatGPT to assist with drafting and editing portions of the manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

552 **REFERENCES**


553 1. Baden, L. R. *et al.* Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *N.*
554 *Engl. J. Med.* **384**, 403–416 (2021).
555 2. Crommelin, D. J. A., Anchordoquy, T. J., Volkin, D. B., Jiskoot, W. & Mastrobattista,
556 E. Addressing the Cold Reality of mRNA Vaccine Stability. *J. Pharm. Sci.* **110**, 997–
557 1001 (2021).
558 3. Li, Y. & Breaker, R. R. Kinetics of RNA Degradation by Specific Base Catalysis of
559 Transesterification Involving the 2‘-Hydroxyl Group. *J. Am. Chem. Soc.* **121**, 5364–
560 5372 (1999).
561 4. Garneau, N. L., Wilusz, J. & Wilusz, C. J. The highways and byways of mRNA decay.
562 *Nat. Rev. Mol. Cell Biol.* **8**, 113–126 (2007).
563 5. Pogocki, D. & Schöneich, C. Chemical stability of nucleic acid–derived drugs. *J.*
564 *Pharm. Sci.* **89**, 443–456 (2000).
565 6. Mikkola, S., Kaukinen, U. & Lönnberg, H. The effect of secondary structure on
566 cleavage of the phosphodiester bonds of RNA. *Cell Biochem. Biophys.* **34**, 95–119
567 (2001).
568 7. Chheda, U. *et al.* Factors Affecting Stability of RNA – Temperature, Length,
569 Concentration, pH, and Buffering Species. *J. Pharm. Sci.* **113**, 377–385 (2024).
570 8. Mauger, D. M. *et al.* mRNA structure regulates protein expression through changes in
571 functional half-life. *Proc. Natl. Acad. Sci.* **116**, 24075–24083 (2019).
572 9. Leppek, K. *et al.* Combinatorial optimization of mRNA structure, stability, and
573 translation for RNA-based therapeutics. *Nat. Commun.* **13**, 1536 (2022).

574 10. Ward, M., Courtney, E. & Rivas, E. Fitness functions for RNA structure design.
575 *Nucleic Acids Res.* **51**, e40 (2023).
576 11. Dirks, R. M., Lin, M., Winfree, E. & Pierce, N. A. Paradigms for computational
577 nucleic acid design. *Nucleic Acids Res.* **32**, 1392–1403 (2004).
578 12. Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. Expanded sequence
579 dependence of thermodynamic parameters improves prediction of RNA secondary
580 structure1. *J. Mol. Biol.* **288**, 911–940 (1999).
581 13. Mathews, D. H. *et al.* Incorporating chemical modification constraints into a
582 dynamic programming algorithm for prediction of RNA secondary structure. *Proc.*
583 *Natl. Acad. Sci.* **101**, 7287–7292 (2004).
584 14. Bicknell, A. A. *et al.* Attenuating ribosome load improves protein output from
585 mRNA by limiting translation-dependent mRNA decay. *Cell Rep.* **43**, (2024).
586 15. Cohen, B. & Skiena, S. Natural selection and algorithmic design of mRNA. *J.*
587 *Comput. Biol. J. Comput. Mol. Cell Biol.* **10**, 419–432 (2003).
588 16. Ward, M., Richardson, M. & Metkar, M. mRNA Folding Algorithms for Structure
589 and Codon Optimization. Preprint at https://doi.org/10.48550/arXiv.2503.19273
590 (2025).
591 17. Terai, G., Kamegai, S. & Asai, K. CDSfold: an algorithm for designing a protein-
592 coding sequence with the most stable secondary structure. *Bioinformatics* **32**, 828–
593 834 (2016).
594 18. Wayment-Steele, H. K. *et al.* Theoretical basis for stabilizing messenger RNA
595 through secondary structure design. *Nucleic Acids Res.* **49**, 10604–10617 (2021).

596 19. Zhang, H. *et al.* Algorithm for optimized mRNA design improves stability and
597 immunogenicity. *Nature* **621**, 396–403 (2023).
598 20. Gu, X., Qi, Y. & El-Kebir, M. DERNA Enables Pareto Optimal RNA Design. *J.*
599 *Comput. Biol. J. Comput. Mol. Cell Biol.* **31**, 179–196 (2024).
600 21. Wayment-Steele, H. K. *et al.* Deep learning models for predicting RNA
601 degradation via dual crowdsourcing. *Nat. Mach. Intell.* **4**, 1174–1184 (2022).
602 22. Kaukinen, U., Lyytikäinen, S., Mikkola, S. & Lönnberg, H. The reactivity of
603 phosphodiester bonds within linear single-stranded oligoribonucleotides is strongly
604 dependent on the base sequence. *Nucleic Acids Res.* **30**, 468–474 (2002).
605 23. Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E. & Weeks, K. M. RNA
606 motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods* **11**,
607 959–965 (2014).
608 24. Lu, Z. J., Gloor, J. W. & Mathews, D. H. Improved RNA secondary structure
609 prediction by maximizing expected pair accuracy. *RNA* **15**, 1805–1813 (2009).
610 25. Hofacker, I. L., Bernhart, S. H. F. & Stadler, P. F. Alignment of RNA base pairing
611 probability matrices. *Bioinformatics* **20**, 2222–2227 (2004).
612 26. Norton, E. C. & Dowd, B. E. Log Odds and the Interpretation of Logit Models.
613 *Health Serv. Res.* **53**, 859–878 (2018).
614 27. Metkar, M., Pepin, C. S. & Moore, M. J. Tailor made: the art of therapeutic mRNA
615 design. *Nat. Rev. Drug Discov.* **23**, 67–83 (2024).
616 28. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and
617 Projection for Dimension Reduction. Preprint at
618 https://doi.org/10.48550/arXiv.1802.03426 (2020).

619 29. LinearPartition: linear-time approximation of RNA folding partition function and
620 base-pairing probabilities | Bioinformatics | Oxford Academic.
621 https://academic.oup.com/bioinformatics/article/36/Supplement_1/i258/5870487.
622 30. Wold, S., Ruhe, A., Wold, H. & Dunn, III, W. J. The Collinearity Problem in Linear
623 Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses.
624 *SIAM J. Sci. Stat. Comput.* **5**, 735–743 (1984).
625 31. Flamm, C., Fontana, W., Hofacker, I. L. & Schuster, P. RNA folding at
626 elementary step resolution. *RNA* **6**, 325–338 (2000).
627 32. Hofacker, I. L. *et al.* Fast folding and comparison of RNA secondary structures.
628 *Monatshefte Für Chem. Chem. Mon.* **125**, 167–188 (1994).
629 33. Dowell, R. D. & Eddy, S. R. Evaluation of several lightweight stochastic context-
630 free grammars for RNA secondary structure prediction. *BMC Bioinformatics* **5**, 71
631 (2004).
632 34. Zarringhalam, K., Meyer, M. M., Dotu, I., Chuang, J. H. & Clote, P. Integrating
633 Chemical Footprinting Data into RNA Secondary Structure Prediction. *PLOS ONE* **7**,
634 e45160 (2012).
635 35. Cordero, P. & Das, R. Rich RNA Structure Landscapes Revealed by Mutate-and-
636 Map Analysis. *PLOS Comput. Biol.* **11**, e1004473 (2015).
637 36. Agarwal, V. & Kelley, D. R. The genetic and biochemical determinants of mRNA
638 degradation rates in mammals. *Genome Biol.* **23**, 245 (2022).
639 37. Cheng, F. *et al.* Research Advances on the Stability of mRNA Vaccines. *Viruses*
640 **15**, 668 (2023).

641 38. Feng, L. & Niu, D.-K. Relationship Between mRNA Stability and Length: An Old
642 Question with a New Twist. *Biochem. Genet.* **45**, 131–137 (2007).
643 39. Trotta, E. On the Normalization of the Minimum Free Energy of RNAs by
644 Sequence Length. *PLOS ONE* **9**, e113380 (2014).
645 40. Chaturvedi, M., Rashid, M. A. & Paliwal, K. K. RNA structure prediction using
646 deep learning — A comprehensive review. *Comput. Biol. Med.* **188**, 109845 (2025).
647 41. Akiyama, M. & Sato, K. RNA Secondary Structure Prediction Based on Energy
648 Models. in *RNA Structure Prediction* (eds Kawaguchi, R. K. & Iwakiri, J.) 89–105
649 (Springer US, New York, NY, 2023). doi:10.1007/978-1-0716-2768-6_6.
650 42. Doshi, K. J., Cannone, J. J., Cobaugh, C. W. & Gutell, R. R. Evaluation of the
651 suitability of free-energy minimization using nearest-neighbor energy parameters for
652 RNA secondary structure prediction. *BMC Bioinformatics* **5**, 105 (2004).
653 43. Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R. & Hofacker, I. L. The
654 Vienna RNA Websuite. *Nucleic Acids Res.* **36**, W70–W74 (2008).
655 44. Wold, S., Sjöström, M. & Eriksson, L. PLS-regression: a basic tool of
656 chemometrics. *Chemom. Intell. Lab. Syst.* **58**, 109–130 (2001).
657 45. Mehmood, T., Liland, K. H., Snipen, L. & Sæbø, S. A review of variable selection
658 methods in Partial Least Squares Regression. *Chemom. Intell. Lab. Syst.* **118**, 62–69
659 (2012).
660
661
662

## LIST OF FIGURE LEGENDS AND CAPTIONS

**Figure 1. In-solution half-life distributions across protein mRNA variants and structure–function analysis in the Nano Luciferase set. (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.

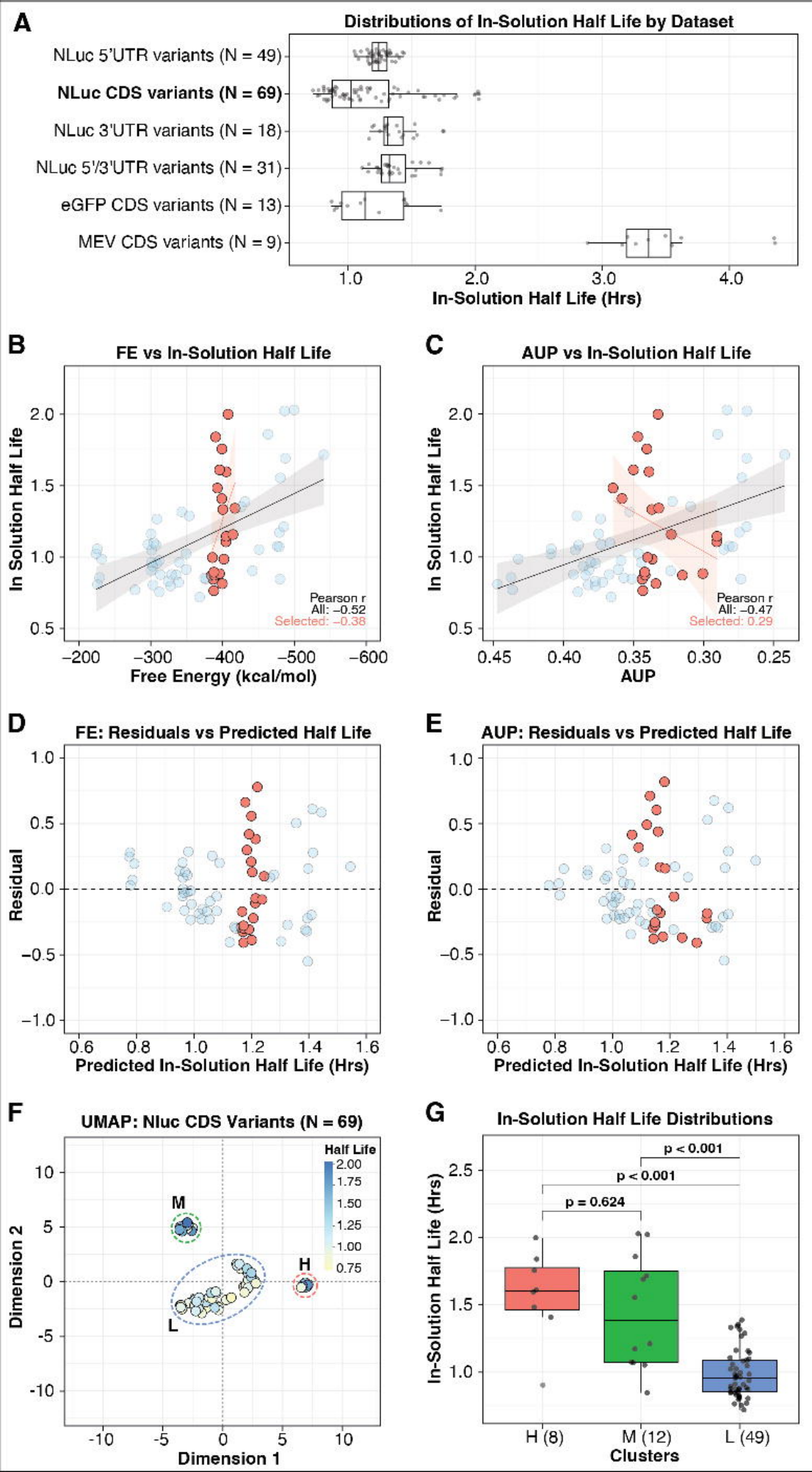
**Figure 2. Structure comparison between two RNA sequences with similar global metrics but substantial differences in in-solution half-life. (A)** Predicted secondary structures of two NLuc sequences, X320004D1 and X930007D1, which exhibit similar FE and AUP but differ substantially in in-solution half-life. Global structure metrics and half-life values are shown. **(B)** Heatmap of local BPP values by position for the two selected 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 comparison of BPPs binned at 0.05 resolution for the two selected sequences. Color scale indicates $\log_{10}$ frequency of positions in each bin.

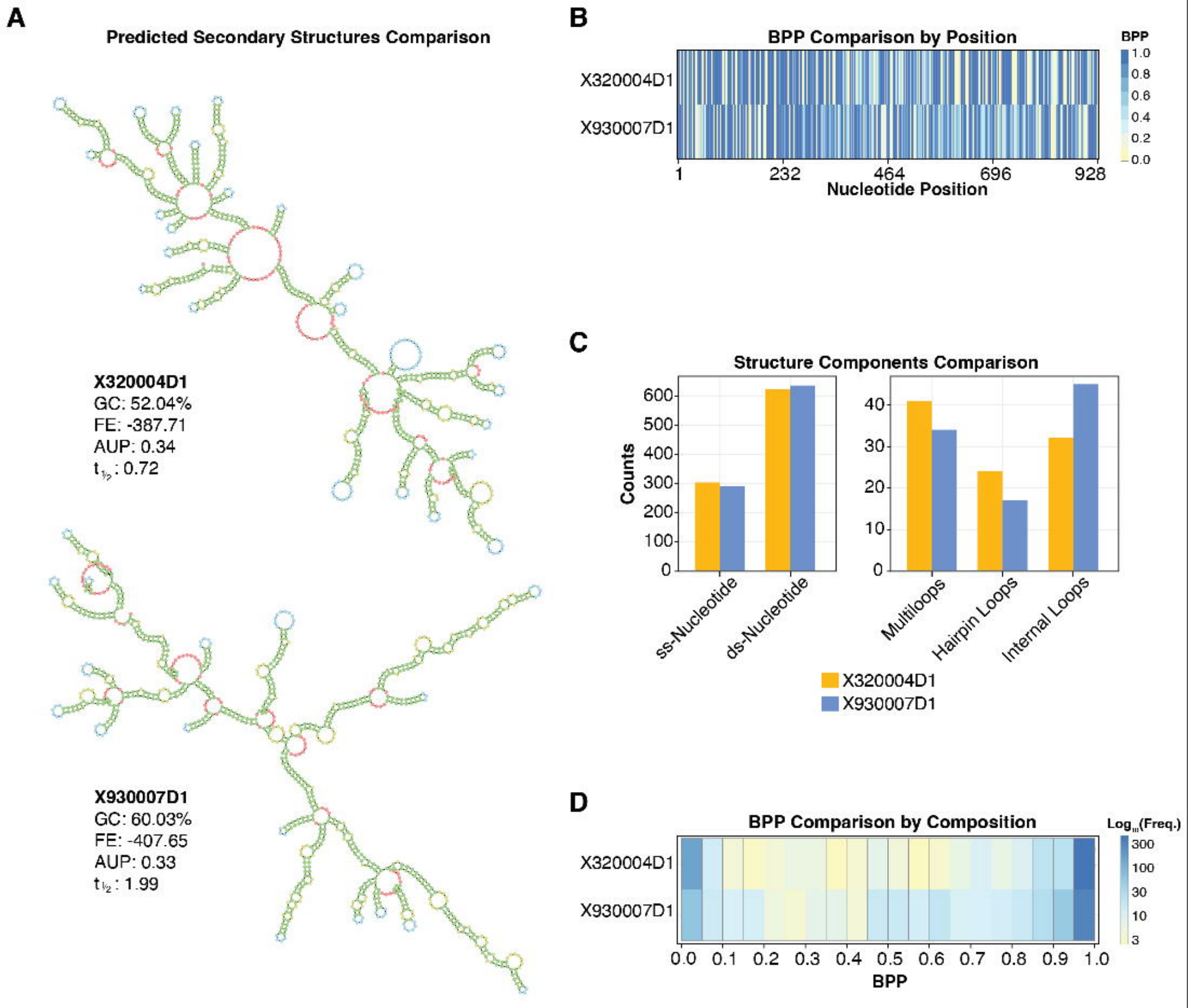
**Figure 3. Local RNA base-pairing patterns differ among similarly structured sequences and associate with in-solution half-life. (A)** Heatmap of local BPP values by position for sequences with similar global metrics (FE and AUP) but differing in-solution half-lives. Sequences are ordered by increasing half-life. **(B)** Heatmap comparisons of BPPs binned at 0.05 resolution across the same sequences as in **(A)**. Color scale indicates $\log_{10}$ frequency of positions in each bin. Numbers flanking each row indicate the count of positions in the lowest and highest BPP bins. **(C)** Correlation between the number of unpaired positions (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.

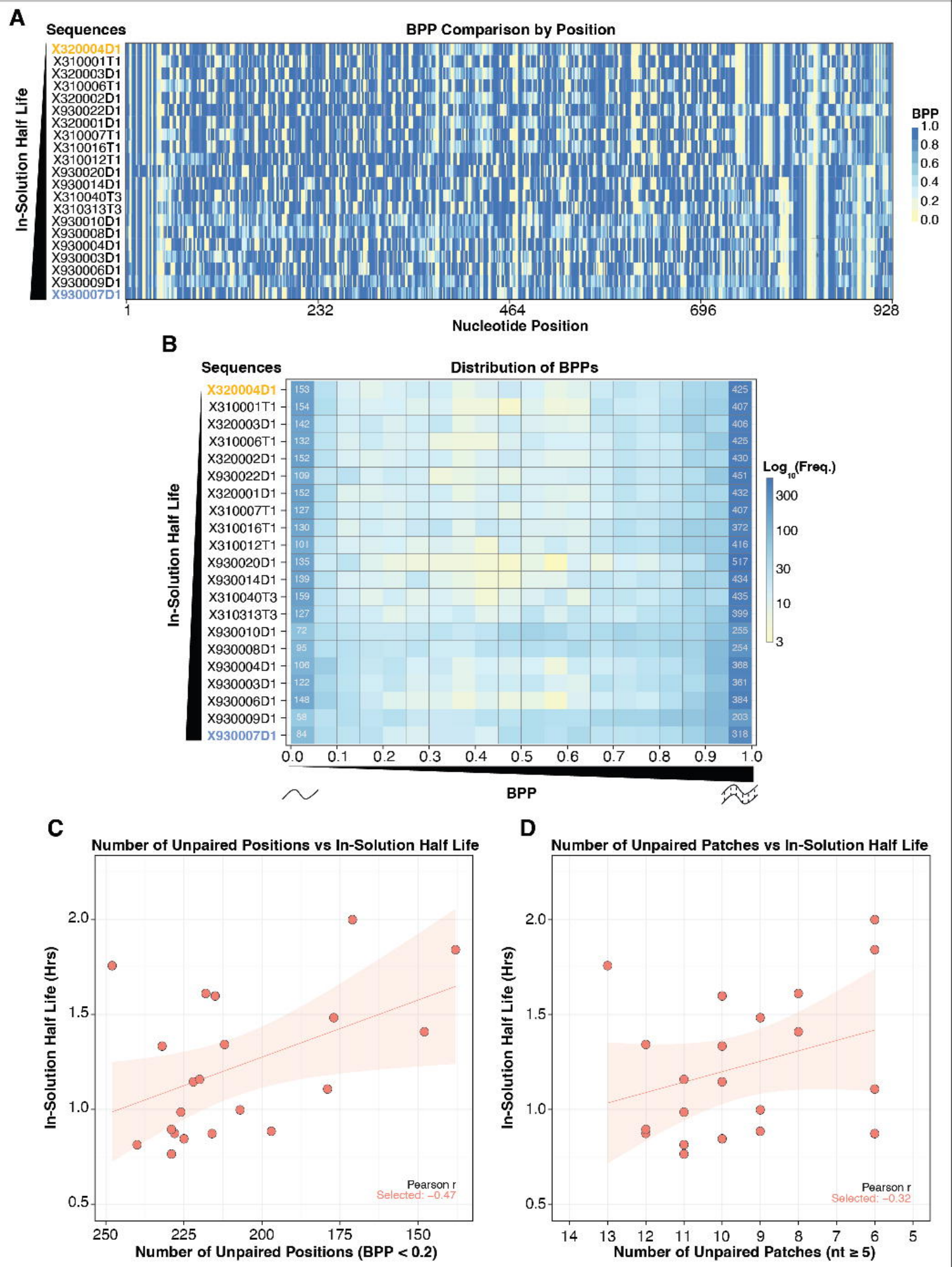
**Figure 4. Base-pairing probability Log-odds (LO) capture local structure features predictive of RNA stability. (A)** Histogram of LO values (binwidth = 0.5) for the two sequences (X320004D1 and X930007D1) with similar global structure metrics but divergent half-lives. **(B)** Heatmap comparisons of binned LO values across the subset of NLuc sequences (**Figure 3**), sorted by increasing in-solution half-life. Color scale indicates the frequency of positions in each bin. **(C)** Heatmap comparisons of LO values by position for the same sequences as in **(B)** with color scale indicating LO values and highlighting localized structural differences along the transcript. **(D)** Correlation of average LO for weakly paired regions (LO < -1.36, equivalent to BPP < 0.2) and in-solution half-life. Pearson r for the NLuc-21 sequences is shown. **(E)** Same as **(D)** but for average LO across all positions.

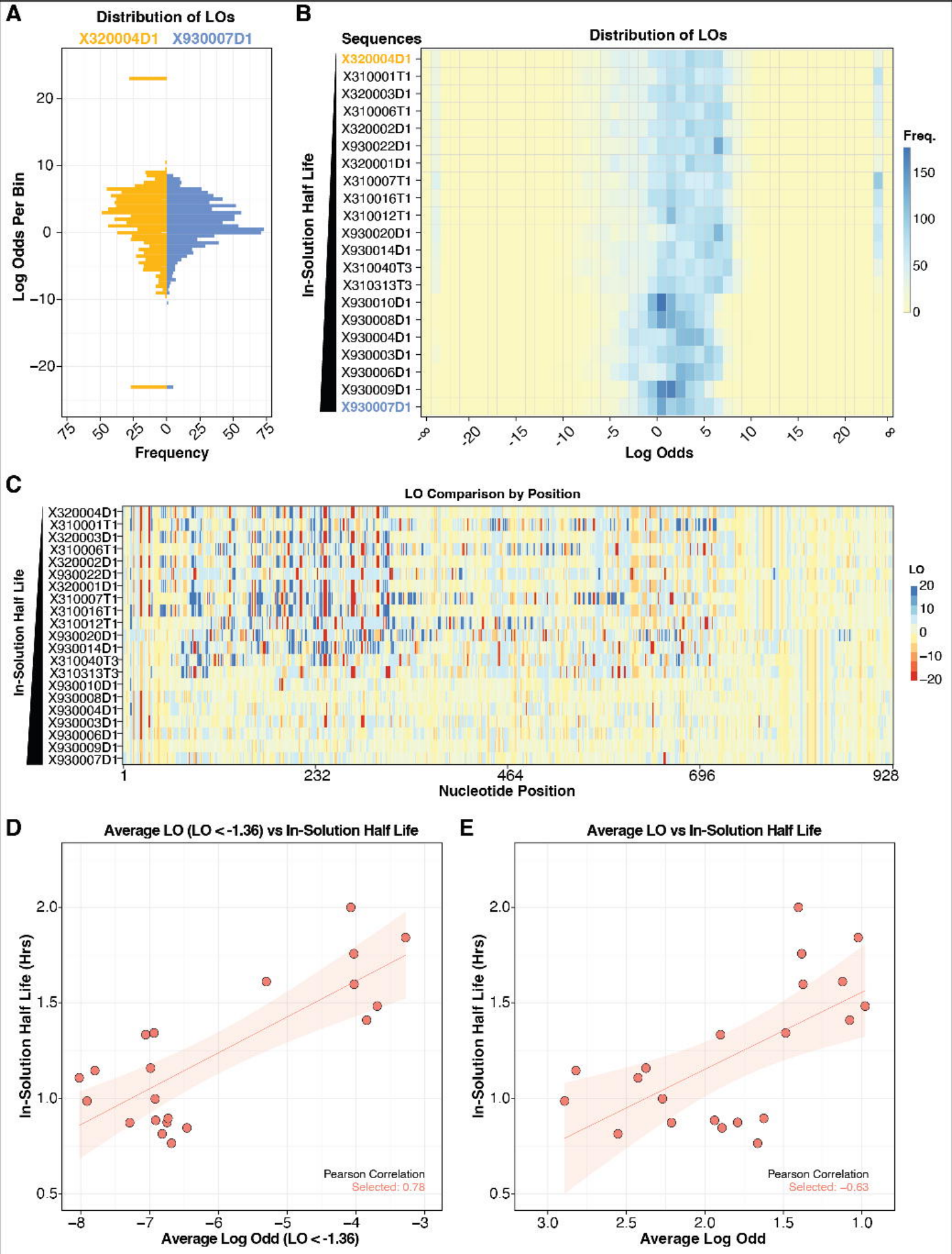
**Figure 5. Multi-feature regression model of RNA degradation shows improved prediction accuracy through orthogonal contribution from Average Log Odds (ALO). (A)** UMAP projection of NLuc sequences based on a multi-feature model including FE/length, AUP, GC content, and ALO, showing separation into three clusters. **(B)** Boxplot distributions of in-solution half-lives 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

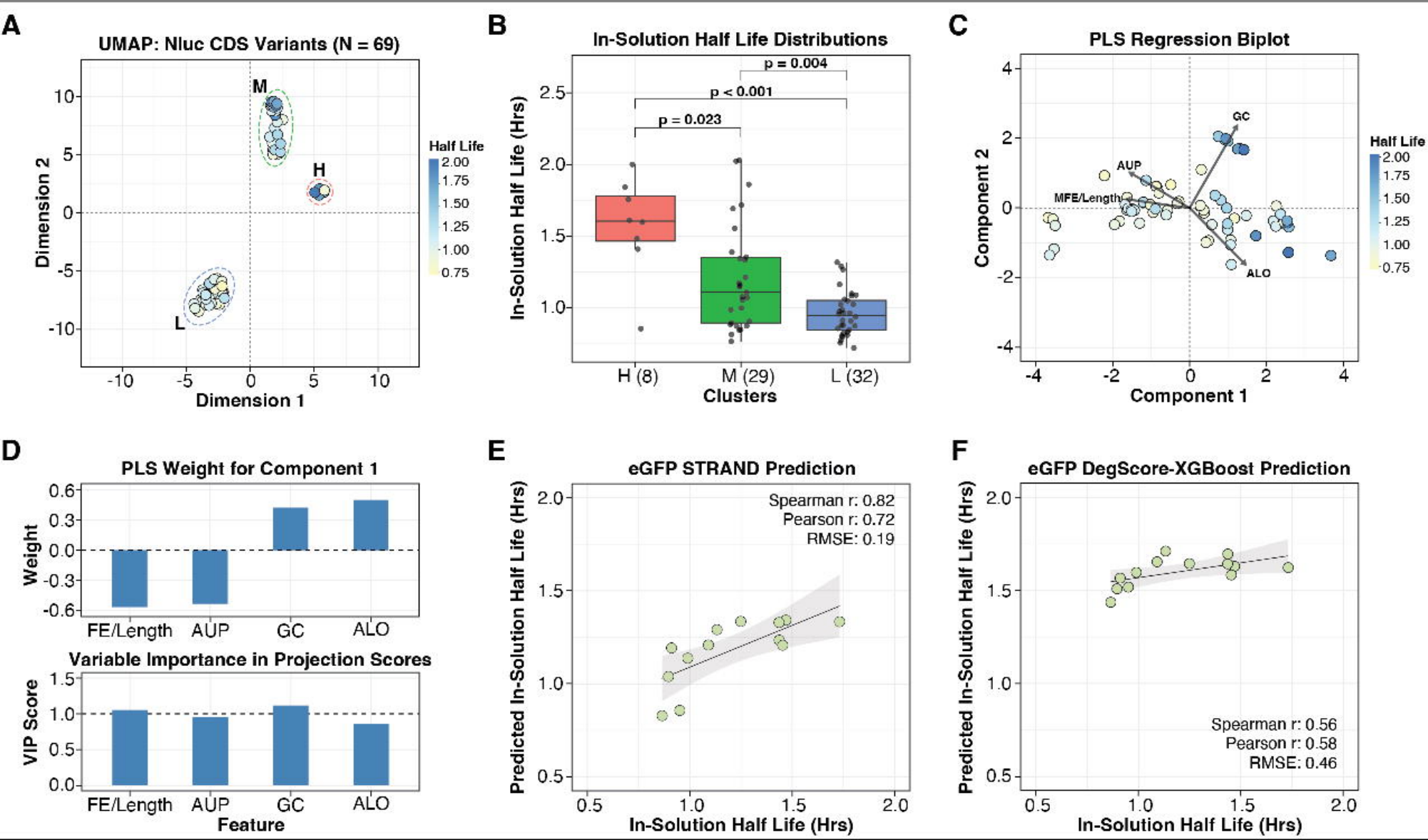
724 cluster is noted in parentheses. **(C)** PLSR biplot of NLuc sequences using FE/length, AUP,
725 GC content, and ALO as features, plotted along Component 1 and Component 2. Arrows
726 indicate loading vectors for each feature. Points are colored by in-solution half-life. **(D)**
727 PLSR Component 1 weight of each feature (Top). Positive and negative weights indicate
728 the direction and magnitude of each feature's contribution to the first latent component.
729 Variable Importance in Projection (VIP) scores across all components (Bottom). The
730 dashed line indicates a canonically used threshold to define important variables. **(E)**
731 PLSR model predictions of in-solution half-life for eGFP sequences using models trained
732 on NLuc variants. A linear regression fit is shown with a 95% confidence interval
733 represented by the shaded region. **(F)** DegScore-XGBoost model predictions for the
734 same eGFP sequences as in **(E)**. A linear regression fit is shown with a 95% confidence
735 interval represented by the shaded region.
736
737 **Figure 6. STRAND generalizes to mRNAs encoding viral antigens with structural**
738 **features outside the training range. (A)** Distributions of features (AUP, FE, GC content)
739 and in-solution half-lives for Spike, VZV, and NLuc sequence sets, illustrating the
740 substantial differences in feature ranges between the training (NLuc) and test (VZV,
741 Spike) datasets. **(B)** Heatmaps of binned LO distributions (binwidth = 1) for VZV gE (top)
742 and Spike (bottom) sequences, sorted by increasing half-life. Color scale indicates
743 frequency of positions in each bin. **(C)** Per-position LO heatmaps for VZV gE (top) and
744 Spike (bottom) sequences. Color scale indicates LO values. **(D)** STRAND predicted half-
745 life versus experimentally measured half-life for VZV (left) and Spike (right) sequences.
746 Spearman r, Pearson r, and RMSE are shown.
747

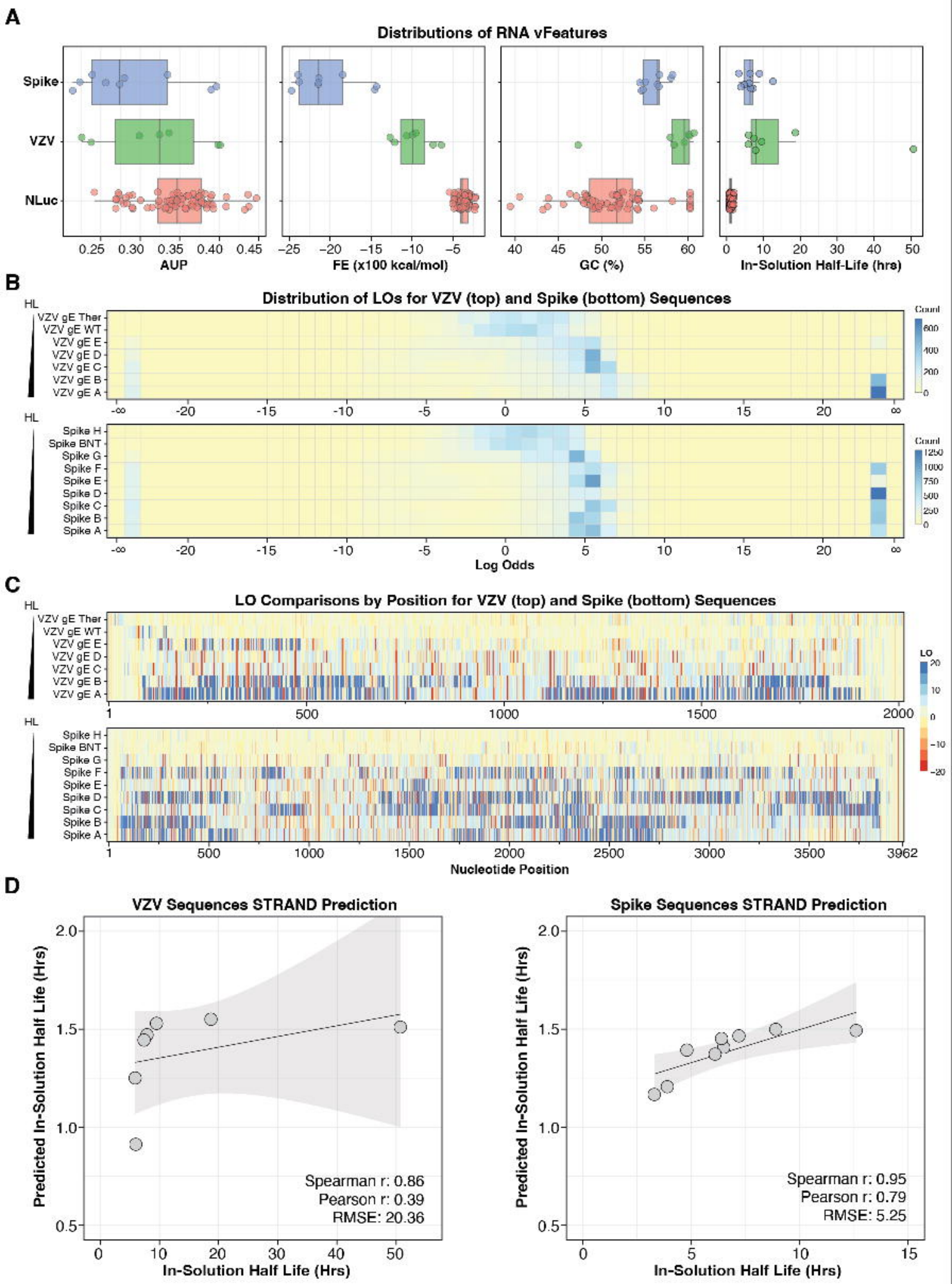

A
Distributions of In-Solution Half Life by Dataset
NLuc 5'UTR variants (N = 49)
NLuc CDS variants (N = 69)
NLuc 3'UTR variants (N = 18)
NLuc 5'/3'UTR variants (N = 31)
eGFP CDS variants (N = 13)
MEV CDS variants (N = 9)
1.0
2.0
3.0
4.0
In-Solution Half Life (Hrs)
B
FE vs In-Solution Half Life
In Solution Half Life
Pearson r
All: −0.52
Selected: −0.38
Free Energy (kcal/mol)
C
AUP vs In-Solution Half Life
In Solution Half Life
Pearson r
All: −0.47
Selected: 0.29
AUP
D
FE: Residuals vs Predicted Half Life
Residual
Predicted In-Solution Half Life (Hrs)
E
AUP: Residuals vs Predicted Half Life
Residual
Predicted In-Solution Half Life (Hrs)
F
UMAP: Nluc CDS Variants (N = 69)
Half Life
M
H
L
Dimension 2
Dimension 1
G
In-Solution Half Life Distributions
p < 0.001
p < 0.001
p = 0.624
In-Solution Half Life (Hrs)
H (8)
M (12)
L (49)
Clusters


A
Predicted Secondary Structures Comparison
X320004D1
GC: 52.04%
FE: -387.71
AUP: 0.34
t½: 0.72
X930007D1
GC: 60.03%
FE: -407.65
AUP: 0.33
t½: 1.99
B
BPP Comparison by Position
X320004D1
X930007D1
1
232
464
696
928
Nucleotide Position
BPP
1.0
0.8
0.6
0.4
0.2
0.0
C
Structure Components Comparison
Counts
600
500
400
300
200
100
0
40
30
20
10
0
ss-Nucleotide
ds-Nucleotide
Multiloops
Hairpin Loops
Internal Loops
X320004D1
X930007D1
D
BPP Comparison by Composition
X320004D1
X930007D1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
BPP
Log10(Freq.)
300
100
30
10
3


A
Sequences
BPP Comparison by Position
In-Solution Half Life
X320004D1
X310001T1
X320003D1
X310006T1
X320002D1
X930022D1
X320001D1
X310007T1
X310016T1
X310012T1
X930020D1
X930014D1
X310040T3
X310313T3
X930010D1
X930008D1
X930004D1
X930003D1
X930006D1
X930009D1
X930007D1
1
232
464
696
928
Nucleotide Position
BPP
1.0
0.8
0.6
0.4
0.2
0.0
B
Sequences
Distribution of BPPs
In-Solution Half Life
X320004D1 153 425
X310001T1 154 407
X320003D1 142 406
X310006T1 132 425
X320002D1 152 430
X930022D1 109 451
X320001D1 152 432
X310007T1 127 407
X310016T1 130 372
X310012T1 101 416
X930020D1 135 517
X930014D1 139 434
X310040T3 159 435
X310313T3 127 399
X930010D1 72 255
X930008D1 95 254
X930004D1 106 368
X930003D1 122 361
X930006D1 148 384
X930009D1 58 203
X930007D1 84 318
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
BPP
Log10(Freq.)
300
100
30
10
3
C
Number of Unpaired Positions vs In-Solution Half Life
In-Solution Half Life (Hrs)
2.0
1.5
1.0
0.5
250 225 200 175 150
Number of Unpaired Positions (BPP < 0.2)
Pearson r
Selected: −0.47
D
Number of Unpaired Patches vs In-Solution Half Life
In-Solution Half Life (Hrs)
2.0
1.5
1.0
0.5
14 13 12 11 10 9 8 7 6 5
Number of Unpaired Patches (nt ≥ 5)
Pearson r
Selected: −0.32


A
Distribution of LOs
X320004D1
X930007D1
Log Odds Per Bin
Frequency
B
Sequences
Distribution of LOs
In-Solution Half Life
Freq.
Log Odds
C
LO Comparison by Position
In-Solution Half Life
LO
Nucleotide Position
D
Average LO (LO < -1.36) vs In-Solution Half Life
In-Solution Half Life (Hrs)
Pearson Correlation
Selected: 0.78
Average Log Odd (LO < -1.36)
E
Average LO vs In-Solution Half Life
In-Solution Half Life (Hrs)
Pearson Correlation
Selected: −0.63
Average Log Odd


A
UMAP: Nluc CDS Variants (N = 69)
M
H
L
Dimension 1
Dimension 2
Half Life
2.00
1.75
1.50
1.25
1.00
0.75
B
In-Solution Half Life Distributions
p = 0.004
p < 0.001
p = 0.023
In-Solution Half Life (Hrs)
H (8)
M (29)
L (32)
Clusters
C
PLS Regression Biplot
GC
AUP
MFE/Length
ALO
Component 1
Component 2
Half Life
D
PLS Weight for Component 1
Weight
FE/Length
AUP
GC
ALO
Variable Importance in Projection Scores
VIP Score
Feature
E
eGFP STRAND Prediction
Spearman r: 0.82
Pearson r: 0.72
RMSE: 0.19
Predicted In-Solution Half Life (Hrs)
In-Solution Half Life (Hrs)
F
eGFP DegScore-XGBoost Prediction
Spearman r: 0.56
Pearson r: 0.58
RMSE: 0.46
Predicted In-Solution Half Life (Hrs)
In-Solution Half Life (Hrs)


A
Distributions of RNA vFeatures
Spike
VZV
NLuc
AUP
FE (x100 kcal/mol)
GC (%)
In-Solution Half-Life (hrs)
B
Distribution of LOs for VZV (top) and Spike (bottom) Sequences
HL
VZV gE Ther
VZV gE WT
VZV gE E
VZV gE D
VZV gE C
VZV gE B
VZV gE A
Spike H
Spike BNT
Spike G
Spike F
Spike E
Spike D
Spike C
Spike B
Spike A
Count
Log Odds
C
LO Comparisons by Position for VZV (top) and Spike (bottom) Sequences
LO
Nucleotide Position
D
VZV Sequences STRAND Prediction
Predicted In-Solution Half Life (Hrs)
In-Solution Half Life (Hrs)
Spearman r: 0.86
Pearson r: 0.39
RMSE: 20.36
Spike Sequences STRAND Prediction
Spearman r: 0.95
Pearson r: 0.79
RMSE: 5.25