Challenges In the Prediction of Half-life of lncRNAs

by
Shaikh Shuhail
(Department of Computational Biology)
(MT23237)

Under the Supervision of Prof. Gajendra Pal Singh Raghava

Indraprastha Institute of Information Technology Delhi
Jan, 2026

Challenges In the Prediction of Half-life of lncRNAs

by
Shaikh Shuhail
(Department of Computational Biology)
(MT23237)

Submitted
in partial fulfillment of the requirements for the degree of
Master of Technology

to

Indraprastha Institute of Information Technology Delhi
Jan, 2026

Certificate

This is to certify that the thesis titled "Challenges in the prediction of Half-life of lncRNAs" being submitted by (Shaikh Shuhail) to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

January, 2026

Prof. Gajendra Pal Singh Raghava

Department of Computational Biology

Indraprastha Institute of Information Technology Delhi

New Delhi 110 020

Acknowledgements

The project has provided an opportunity to go beyond theoretical understanding and to learn and explore the practical applications of the concepts learned in the classroom.

First and foremost, I am grateful to the Almighty, to help me learn and dedicate myself to the project.

I extend my heartfelt gratitude to Prof. Gajendra Pal Singh Raghava, who has been the backbone supporting me throughout the project helping me learn and grow, for his guidance and patience throughout this work and for his encouragement and valuable lessons.

I express my gratitude to Shubham Choudhary, PhD Scholar, for his insights and help at various stages of the project.

Last but not the least, I thank all my friends and batchmates, who were present around to teach me, help me out during critical situations, keep me motivated, standing beside and aiding me complete this project. I acknowledge and am grateful for each one's contribution.

Shaikh Shuhail

Department of Computational Biology

MT23237

Abstract

Long non-coding RNAs (lncRNAs) are key regulators of gene expression, and their stability, commonly quantified as half-life, plays a critical role in cellular function. Recent computational efforts have attempted to predict RNA half-life from sequence, with limited success. For instance, Shi et al. applied deep learning models and initially reported spearman correlations of 0.7–0.8, but performance dropped to 0.06–0.09 after five-fold validation.

In this study, we developed machine learning and deep learning models using sequence-derived features to predict lncRNA half-life. Among the approaches tested, Random Forest based on nucleotide composition features performed best, achieving a spearman correlation of 0.9862 on the training dataset but only 0.0592 on the validation dataset. Furthermore, clustering analysis revealed that different transcript groups exhibited nearly identical mean half-life distributions, indicating that sequence-derived features alone do not meaningfully stratify lncRNAs by stability.

These results, consistent with prior studies, demonstrate the persistent difficulty of predicting RNA half-life in silico. Further, inclusion of features such as RNA–binding protein motifs, structure–based minimum free energy and sub–cellular localization did not improve the model performances. This suggests that RNA stability is regulated by features beyond those included. Therefore, in this paper, we outline the approaches studied and the challenges to predict the RNA stability, further highlighting the need to integrate multi-omic strategy or design an algorithm to predict RNA half-life.

Contents

List of Figures

List of Tables

List of Abbreviations

RNA: Ribonucleic Acid

DNA: Deoxyribonucleic Acid

mRNA: messenger RNA

lncRNA: long non-coding RNA

tRNA: transfer RNA

AI: Artificial Intelligence

ML: Machine Learning

DL: Deep Learning

RBP: RNA-Binding Protein

CNN: Convolution Neural Network

ANN: Artificial Neural Network

RNN: Recurrent Neural Network

NLP: Natural Language Processing

MLP: Multilayer Perceptron

BiLSTM: Bidirectional Long Short Term Memory

SVR: Support Vector Regression

MAE: Mean Absolute Error

MSE: Mean Squared Error

RMSE: Root Mean Squared Error

ReLU: Rectified Linear Unit

UMAP: Uniform Manifold Approximation and Projection

MFE: Minimum Free Energy

Graphical Experimental Architecture

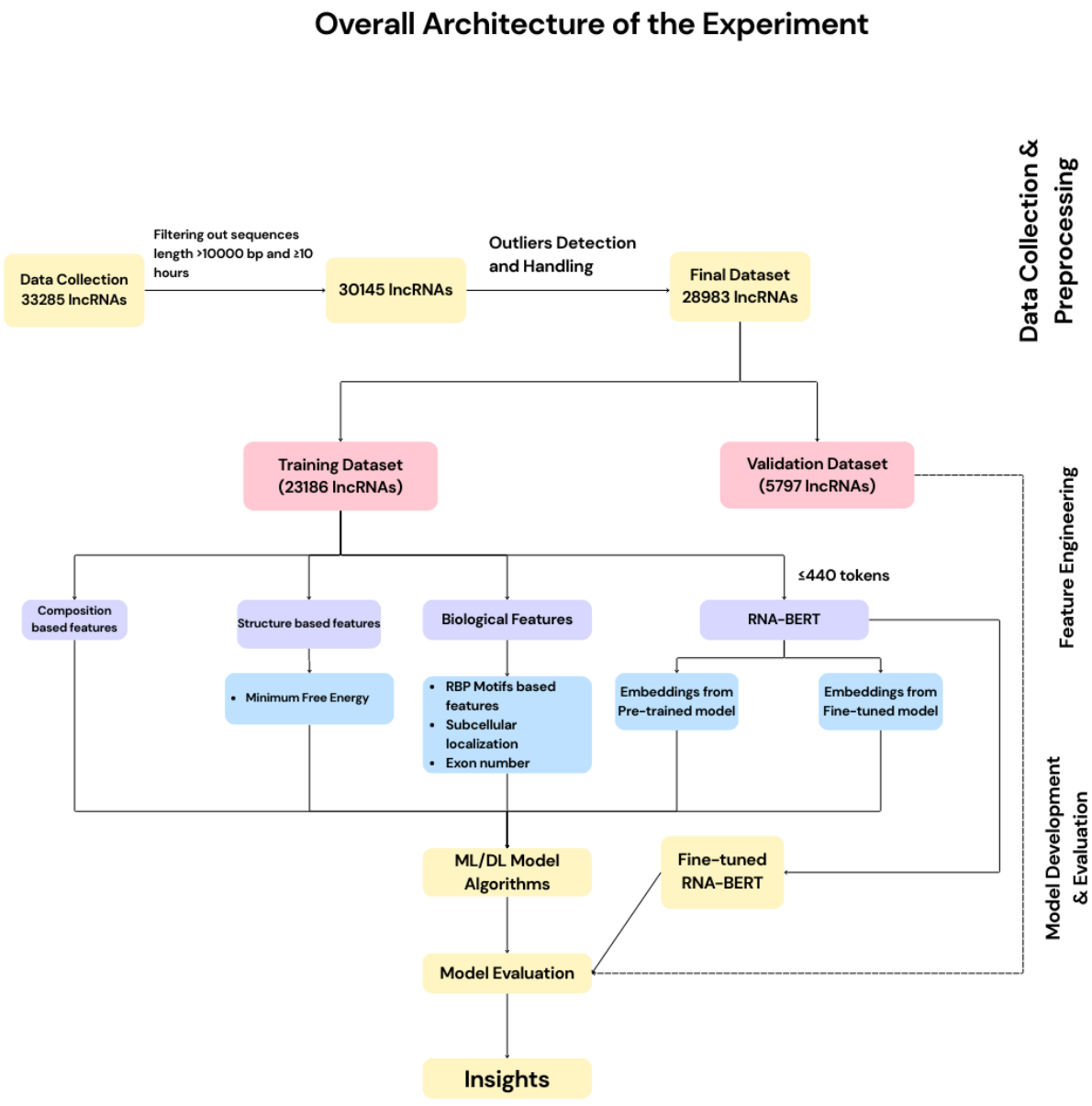


Figure 1: Overall architecture of the experiment

1. Introduction

RNA molecules are essential for cellular function. They not only play a major role through transcription and translation, but are also responsible for the different functions due to their stability and degradation. RNA half-life is the time taken for an RNA molecule to degrade into half of its original amount and is a crucial determinant of gene expression levels and post transcriptional regulation. With all different types of RNA classes, the long non-coding RNAs (lncRNAs) are widely recognized for their diverse regulatory structure and function such as chromatin remodeling, cellular differentiation, splicing etc. (Mattick et al., 2023). However, the mechanisms governing the stability still remain enigmatic.

As of now, RNA half-life is estimated through an intensive experimental process involving transcriptional inhibition of the cell followed by RNA sequencing at multiple time points. These techniques are resource and condition-specific, cost consuming and often limited in throughput. With the advancement in the field of computational sciences, this problem has led to look for a computational based solution and design a model capable of predicting the RNA half-life from intrinsic features such as nucleotide sequence or structure properties (Shi et al., 2021).

The advancements in machine learning (ML) and deep learning (DL) has helped to model and understand the biological sequences better. The pre-trained language models such as RNA-BERT (Akiyama & Sakakibara, 2022) has also enabled the contextual learning from the raw sequences, providing opportunities to capture the patterns of the given sequence and determine their behaviour and functional insights. Despite the process, there aren't any such models or tools available that could predict the RNA half-life, especially for lncRNAs, which are less conserved and variable structurally. (Mattick et al., 2023)

In this study, we look forward to exploring and building the model using the ML and DL approach to predict the half-life of lncRNA, using a large-scale dataset of around 33,000 human lncRNAs with their experimentally measured half-lives (Shi et al., 2021). We constructed a pipeline that incorporates, (i) Traditional sequence features; (ii) Predicted secondary structure features; (iii) Embedding features from RNA-BERT; (iv) Other features (such as subcellular localization, RBP binding motif based features, exon number etc.).

Through various models and their results, we observed that all the models performed poorly. The regression (R^2) values were close to zero or negative. The embedding model based on RNA-BERT was also unable to correctly represent the contextual features resulting in poor model. This suggested the fundamental limitation to use the sequences alone to capture the biology of RNA stability.

In our work, we highlight the challenges of modeling the half-life of RNA molecules using the primary features, minimum free energy, motifs based features and further the need to incorporate other biological features such as cellular context, epitranscriptome modifications, and other approaches for precise model building. It also suggests building a new algorithmic approach focusing on half-life prediction.

2. Literature Review

2.1 RNA Half-Life and Its Biological Importance

RNA half-life determines the stability of the RNA transcript, which measures how long a molecule will be present in the cell before it degrades. The transcript stability is essential to determine the mRNA abundance, protein expression and dynamics of cellular processes and responses (Wang & Liu, 2022). Various studies on RNA half-life show that they are highly regulated and differ based on the transcript type, cell type and different experimental conditions. (Ietswaart et al., 2024). RNA molecules aren't just involved in transcription and translational roles. Different RNA molecules are associated with different roles in cellular functioning. Similarly, the long non-coding RNAs (lncRNAs) are associated and responsible for gene regulation and chromatin stability (Mattick et al., 2023).

Though the mRNA stability and functioning is widely studied, the lncRNAs are still enigmatic to be completely understood, and the half-life of these molecules are important to understand their roles and functions that could further help use these properties for biological applications of disease treatment, drug discovery, vaccine development and others.

Several experimental approaches have been employed to measure RNA stability, including: (i) Transcriptional inhibition followed by time-series RNA-sequencing (eg: using Actinomycin D),

(ii) Metabolic labeling with 4sU or SLAM-seq and RNA decay profiling under different stress conditions. These techniques, while powerful, are often labor-intensive, low-throughput, and context-dependent, prompting the need for predictive computational models.

2.2 Computational Efforts for RNA Half-Life Prediction

There have been several attempts to computationally predict the RNA based on sequences as studied by Li & Liang (2017). With the computational advancements, the RNA degradation kinetics and transcript stability has been studied and experimented. In 2012, Tani et. al, determined the genome-wide RNA stability in mammals, revealing non-coding RNAs and mRNA into two classes of long and short lived RNA.

In 2021, Shi et al., performed half life prediction through both experimental method and sequence based ML/DL modeling. The experimental results were obtained through a transcriptional inhibition method. But the machine learning and deep learning sequence based approach was non-resilient, suggesting incorporating other features and multi-omics to produce relevant insights and prediction. Similarly, in more recent times, deep learning models (CNNs, RNNs) have been employed for transcript degradation kinetics (Wayment-Steele et al., 2022).

In order to reduce the cost and artifacts of experimental approaches, Conte et al. (2022) developed an in-silico method StaRTrEK (STAbility Rates ThRough Expression Kinetics) to estimate the RNA half-lives without transcriptional inhibition.

Though, there have been several computational efforts towards the prediction of RNA, some of which are stated above, there has been no robust study to predict lncRNA half-life directly from their sequences. This has been more complex and challenging due to the non-conservative nature of the molecules unlike mRNA, tRNA and other molecules. This urges the need to study extensively and develop a computational method which could fulfill the gap towards prediction of half-life of lncRNAs in-silico.

2.3 Emerging Role of Pretrained Models (RNA-BERT)

Inspired by breakthroughs in NLP (Natural Language Processing), pretrained models like DNA-BERT and RNA-BERT have been developed to encode biological sequences in a contextualized manner(Ji et al., 2021 and Akiyama & Sakakibara, 2022). RNA-BERT, specifically designed for RNA sequences for effective embedding of RNA bases, particularly for non-coding RNAs. It captures contextual and structural information and generates informative base embeddings, which can be then used for RNA classification (Akiyama & Sakakibara, 2022).

However, its application in regression tasks such as RNA half-life prediction remains largely unexplored. While pretrained models capture higher-order sequence semantics, their ability to model decay kinetics: a process influenced by both intrinsic and extrinsic factors is still an open question.

2.4 Knowledge Gaps and Contribution

The experimental half-life data has been produced in abundance through the course of time. Though, the current computational approaches are non-robust to develop a model to predict the half-life of the RNAs. The current models show low predictive accuracy ($R^2 \leq 0.3$) and depend on additional information, such as RBP motifs, subcellular localization, exon numbers etc, which are not available at all conditions. These approaches are more determined and extensively studied in mRNA compared to any other classes of RNA (lncRNAs) which, unlike mRNA, differ in structure, function and are non-conservative in nature. As well as, there are no such models which are generalized across the dataset or cell types.

Moreover, the majority of studies have not benchmarked sequence-only models on large lncRNA datasets using modern representation learning tools.

In this work, we aim to address these following gaps by:

- Building a large-scale pipeline to predict lncRNA half-life from sequence-only features, including k-mer composition, structural features, and RNA-BERT embeddings.
- Evaluating a diverse set of ML and DL models to benchmark performance.

- Reporting and analyzing consistently poor R^2 values, highlighting the limitations of sequence-based models for this task and emphasizing the need for integrative, multi-modal approaches.

3. Materials and Methods

3.1 Data Collection

To develop models for predicting RNA half-life based solely on sequence information, we utilized transcriptomic half-life data from the study by Shi et al. (2021). The authors performed a time-course RNA sequencing experiment in human A549 lung adenocarcinoma cells, using actinomycin D to inhibit transcription and sampling RNA at ten distinct time points post-treatment. Through this, they experimentally determined and calculated the transcript decay profile and half-life values.

We availed this public data, shared and submitted by Shi et al which consisted of transcript IDs and their corresponding half-life values for lncRNA. These transcript IDs were extracted from the NONCODE v6 database (Zhao et al., 2021), following complete sequence retrieval of all 33,285 nucleotide sequences.

After the sequence retrieval, exploratory data analysis (EDA) was performed and since the half-life was measured at 10 different time points. The mean half-life was calculated and used as the primary label for the regression analysis.

Finally, the 33,285 lncRNA sequences and the calculated half-life of each sequence were used as the base for feature extraction and model building.

3.2 Feature Engineering

Feature engineering is the next step after the data collection. To predict the half-life of lncRNA molecules, the features were extracted from the primary sequence. Broadly categorized features such as sequence based features, structure based features and embeddings from pre-trained

RNA-BERT model and other features (including RBP motifs based features, cellular localization, exon numbers) were extracted.

3.2.1 Sequence-Based Features

Features were directly extracted from the nucleotide sequence composition for each lncRNA molecule.

- **k-mer Frequencies (k = 1 to 3)**: The frequencies of all the subsequences eg: (A, UG, GUC) were determined. These k-mer capture the compositional pattern associated with RNA decay and stability.

- **GC Content**: Percentage of guanine and cytosine were computed since, GC content influences transcripts stability.

- **Sequence Length**: The total nucleotide count of each transcript was recorded as a feature.

- **Nucleotide-Entropy features:** This calculates the shannon entropy for whole nucleotide sequences.

- **Distance Distribution of Nucleotides:** This calculates the distance distribution of nucleotides from the whole nucleotide sequences.

3.2.2 Structure-Based Features

To incorporate RNA folding information, we utilized the secondary structure based features extracted in the study by Shi et al. (2021) for all the RNA sequences, given in the supplementary information. They predicted the secondary structure using RNAfold and extracted:

- **Minimum Free Energy (MFE)**: Reflects the thermodynamic stability of the RNA structure. More negative MFE values typically indicate more stable transcripts.

3.2.3 Other features

- **RNA Binding Protein (RBP) Motifs**: The RBP motifs were obtained from Shi et al. 2021, which was extracted from ATtRACT database (Giudice et al. 2016). A total of 2297 RBP motifs were extracted, of which 466 motifs (FDR<0.1) were used as features for predictive analysis.

- **Localization based features**: Among the given lncRNA, experimental half-life has been calculated for those sequences specific to nucleus or cytoplasm or present in both locations. This was obtained only for 7764 lncRNAs(5758 - nucleus specific, 1515 - cytoplasm specific, 491 - present in both nucleus and cytoplasm). The one-hot encoding was used to feature the subcellular localization of lncRNAs.

- **Exon number features**: In the study by Shi et al. 2021, they showed differences in the lncRNAs half life based on the exon numbers. Therefore, these exon numbers were leveraged as features.

In order to assess the relevance among the features and their relationship with the half-life, Pearson correlation coefficient was computed (Figure 10b, 10c, 10d).

3.2.4 Embedding-Based Features: RNA-BERT

To learn contextual representations of RNA sequences, we used **RNA-BERT**, a pretrained transformer-based language model for nucleotide sequences.

- **Sequence Selection**: RNA-BERT is trained with a fixed input size of 440 tokens. Therefore, only transcripts with sequence lengths ≤440 nucleotides and those with half-life ≤10 hours were retained for this step (4397 sequences).

- **Fine-tuning and Embedding Extraction**: The RNA-BERT model was fine-tuned on the task-specific dataset and embedding vectors were generated.

- **Embedding Usage**: The embedding vectors were used as input for training the downstream models for half-life prediction, in order to capture contextual, semantic and structural aspects of the sequences.

3.3 Model Development

The traditional machine learning (ML) and deep learning (DL) models, as well as a separate RNA-BERT based regression model was implemented and evaluated to predict and estimate the RNA half-life. The modeling flow includes, data pre-processing, followed by feature normalisation and outlier handling and finally model training and evaluation.

3.3.1 Data Preprocessing and Outlier Handling

- **Sequence Filtering:** For the traditional ML/DL models, we initially filtered and retained those sequences with length <10000 nucleotides and half-life ≤10 hrs (30145 sequences). The following cut-off filter was applied in order to generate more relevant analysis and focus on the majority of the dataset, instead of extreme end points conditions.

- **Missing Values**: Any sequences with missing values for sequence, structure, or half-life were removed during initial curation.

- **Outlier Detection**: Exploratory data analysis revealed a subset of transcripts with extremely high or low half-life values.

    - We defined outliers as transcripts using the inter-quartile method, and excluded those sequences which lie outside the inter-quartile range, from the training dataset to prevent skewing model training.

- **Sequence Filtering (for RNA-BERT)**: For BERT-based modeling, only sequences with lengths ≤440 nucleotides were retained, as required by the pretrained tokenizer architecture.

3.3.2 Feature Standardization

Feature standardization was performed to prevent differences in feature scales from influencing the learning algorithms disproportionately, all numerical features such as k-mer frequencies, GC content, sequence length, and minimum free energy (MFE), motif counts etc were standardized using **StandardScaler** from the **scikit-learn** library. This method transforms each feature to have a **mean of 0** and **standard deviation of 1**, following Z-score normalization:

$$x_{scaled} = \frac{x - \mu}{\sigma}$$

In order to prevent data leakage and biases in test data, the standardization process needs to be performed separately for train and test data. This preprocessing step is essential for optimizing the performance of models sensitive to feature scaling, such as SVR and neural networks.

3.3.3 Feature Selection and Dimensionality Management

Our analysis incorporated distinct feature sets, including k-mer compositions and motif-based features. Given the relatively low dimensionality of these feature sets (approximately 150 composition-based features and under 500 motif-based features) in comparison to the large number of RNA sequences (approximately 30,000), explicit dimensionality reduction techniques such as Principal Component Analysis (PCA) were deemed unnecessary for these features in order to preserve the individual features and loss of information through dimensionality reduction.

3.3.4 Machine Learning Models

A range of machine learning models were implemented in order to predict the half-life of lncRNAs from the engineered features, using scikit-learn library. The selected model incorporated both linear and ensemble methods to evaluate the predictive performance. The following models were used:

- **Linear Regression (LR):** This model was used as the baseline to assess the linear separability of the features.

- **Support Vector Regression (SVR):** The model was used due to its strength to identify and capture the complex non-linear dependencies that simple models overlook.

- **Random Forest Regressor:** This is a decision tree based ensemble learning model that manages the feature interactions through multiple decision trees and prevents overfitting, being robust.

- **XGBoost Regressor:** This is a gradient boost model known for its higher efficiency and accuracy.

For optimal performance, hyperparameter tuning was done for these models rather than relying on the default parameters to ensure that models were neither underfit nor overfit.

The performance were evaluated upon the following regression metrics:

- **Spearman Correlation Coefficient** (**R/ρ**): This measures the correlation that assess the strength and direction of association between two ranked variables.

$$\rho = 1 - \frac{6\Sigma d_{i2}}{n(n^2 - 1)}$$

- **Mean Absolute Error (MAE):** Measures the average magnitude of the errors in a set of predictions, without considering their direction.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - y'_i\right|$$

- **Root Mean Squared Error (RMSE):** Measures the square root of average of squared errors.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y'_i)^2}$$

3.3.5 Deep Learning Models

Deep learning models were also used to leverage their ability in order to predict the transcript stability. The following models were developed using TensorFlow and Keras architecture:

- **One-Dimensional Convolutional Neural Network (1D-CNN)**: 1D CNNs were implemented for sequence-based feature extraction, particularly effective at identifying local patterns and motifs within RNA sequences. This architecture is well-suited for processing both one-hot encoded sequence data and dense RNA-BERT embeddings.

- **Multilayer Perceptron (MLP)**:
    - Architecture: Fully connected layers with decreasing neuron sizes (from 256 → 128 → 64→ 32)
    - Activation: **ReLU**
    - Regularization: **Dropout (0.3)** and **Batch Normalization** between layers
    - Loss function: **Mean Squared Error (MSE)**
    - Optimizer: **AdamW**

The MLP was trained on both handcrafted features and RNA-BERT embeddings, with early stopping based on validation loss to prevent overfitting.

- **Bidirectional Long Short-Term Memory (BiLSTM)**:
    - BiLSTM networks are specifically utilized for processing sequential inputs, including raw RNA sequence data and contextual RNA-BERT embeddings.

Hyperparameters critical to the training and performance of these deep learning models, such as batch size, learning rate, and dropout rates, were optimized in order to achieve an optimal prediction performance, though the models were consistently closer to zero or negative.

3.3.6 RNA-BERT Feature Extraction, Fine-Tuning and Regression

RNA-BERT models are pre-trained on a vast amount of RNA datasets to learn the contextual representations of the RNA sequences. This transformer based model was utilized to generate embeddings, evaluate models, fine-tune these models as per the training dataset, to leverage its potential for our dataset. The following steps were performed to obtain the embeddings and evaluate the models:

- **Leveraging Pre-trained Embeddings:** RNA-BERT are pre-trained on rich, contextual RNA datasets. The embeddings are generated from this pre-trained model to generate high quality and biologically meaningful representations.
- **Evaluate different models:** Both traditional ML models and DL models were used on the extracted embeddings.
- **Fine-tuning**: RNA-BERT was fine-tuned on the dataset using average half-life as the regression target. We used a learning rate scheduler and Adam optimizer. This enables the model to learn the nuances of the RNA stability regression task.

Since, the models performance was consistent to those of traditional feature based approach. The next step of cross-validation was skipped for these conditions.

3.3.7 K-mer based clustering

In order to explore the grouping of RNA sequences based on compositional similarity, the k-mer frequency vectors were computed for each sequence using 3-mer. The obtained frequency matrix was normalized using StandardScaler, followed by unsupervised clustering using k-means (k=3). UMAP, a non-linear dimensionality reduction technique was used to qualitatively assess the clusters. Each cluster was then evaluated for average RNA half-life to investigate the relation and relevancy of sequence composition for RNA stability insights.

4. Results and Evaluation

We evaluated the performance of all models: traditional ML, deep learning, and RNA-BERT-based approaches on the task of predicting lncRNA half-life from sequence derived features and all other features including RBP motifs, cellular localization, etc. Despite extensive feature engineering and optimization, all models demonstrated consistently poor regression performance, revealing fundamental challenges in this predictive task.

4.1 Performance of Classical Machine Learning Models

The ML models (Linear Regression, SVR, Random Forest, and XGBoost) were trained on different engineered features. Across all models, the spearman correlation (R/ρ) on the test set was close to zero (Table 1).

Features	Model	Training Dataset			Validation Dataset		
		R	**MAE**	**RMSE**	**R**	**MAE**	**RMSE**
All comp	Linear Regression	0.0649	1.4413	1.8223	0.0337	1.4603	1.8496
	SVR	0.4518	1.2465	1.7423	0.0349	1.4263	1.8997
	Random Forest	**0.9862**	**0.4978**	**0.6213**	**0.0592**	**.1.3290**	**1.6420**
	XGBoost	0.6855	1.1975	1.5257	0.0378	1.4719	1.8664
	Gradient Boosting	0.2667	1.4074	1.7784	0.0421	1.4618	1.8490

Table 1: ML models and their performance on features

These models struggled to learn meaningful patterns from the sequence features, indicating that linear and weakly non-linear relationships are insufficient to model RNA half-life using standard engineered features.

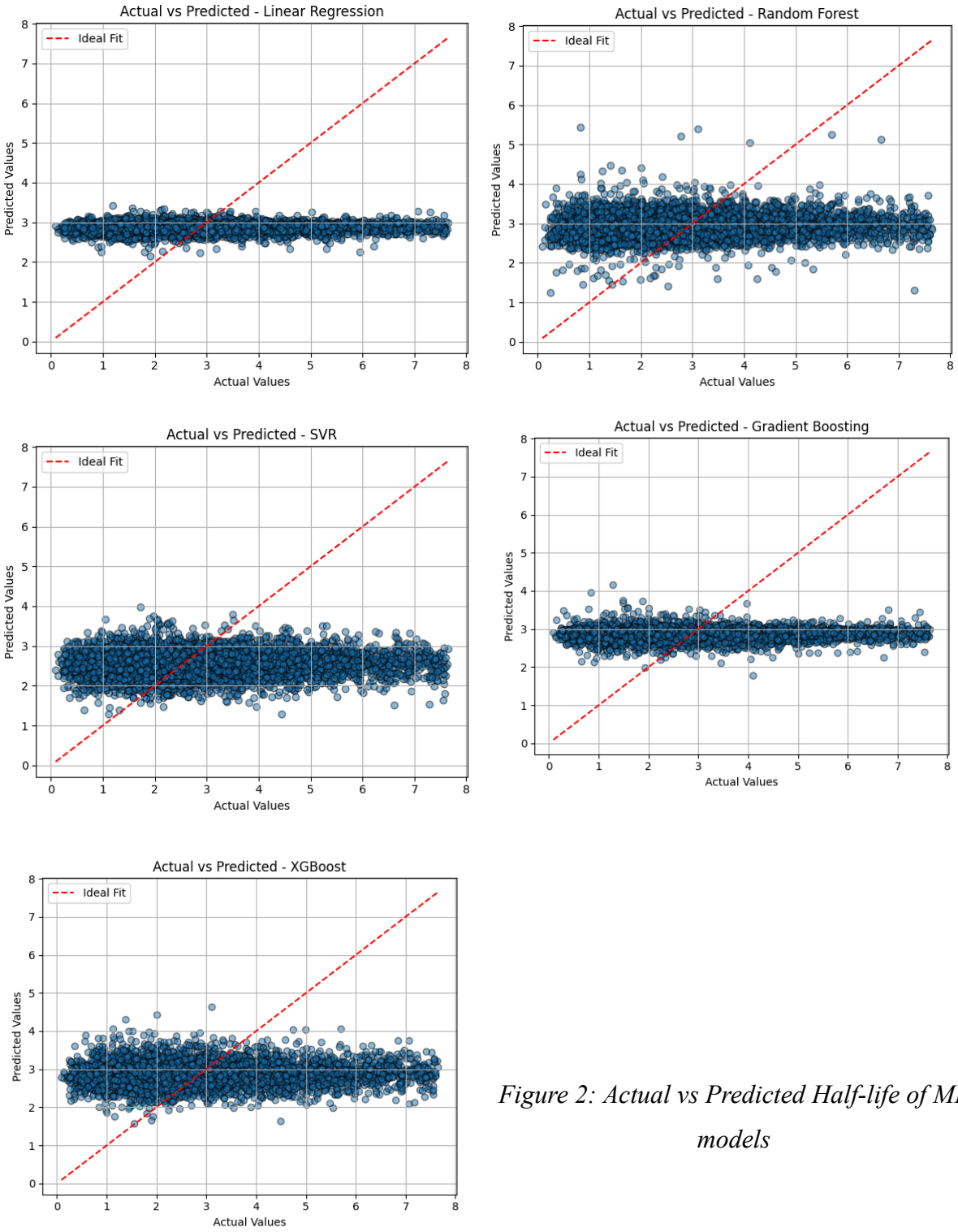


Figure 2: Actual vs Predicted Half-life of ML models

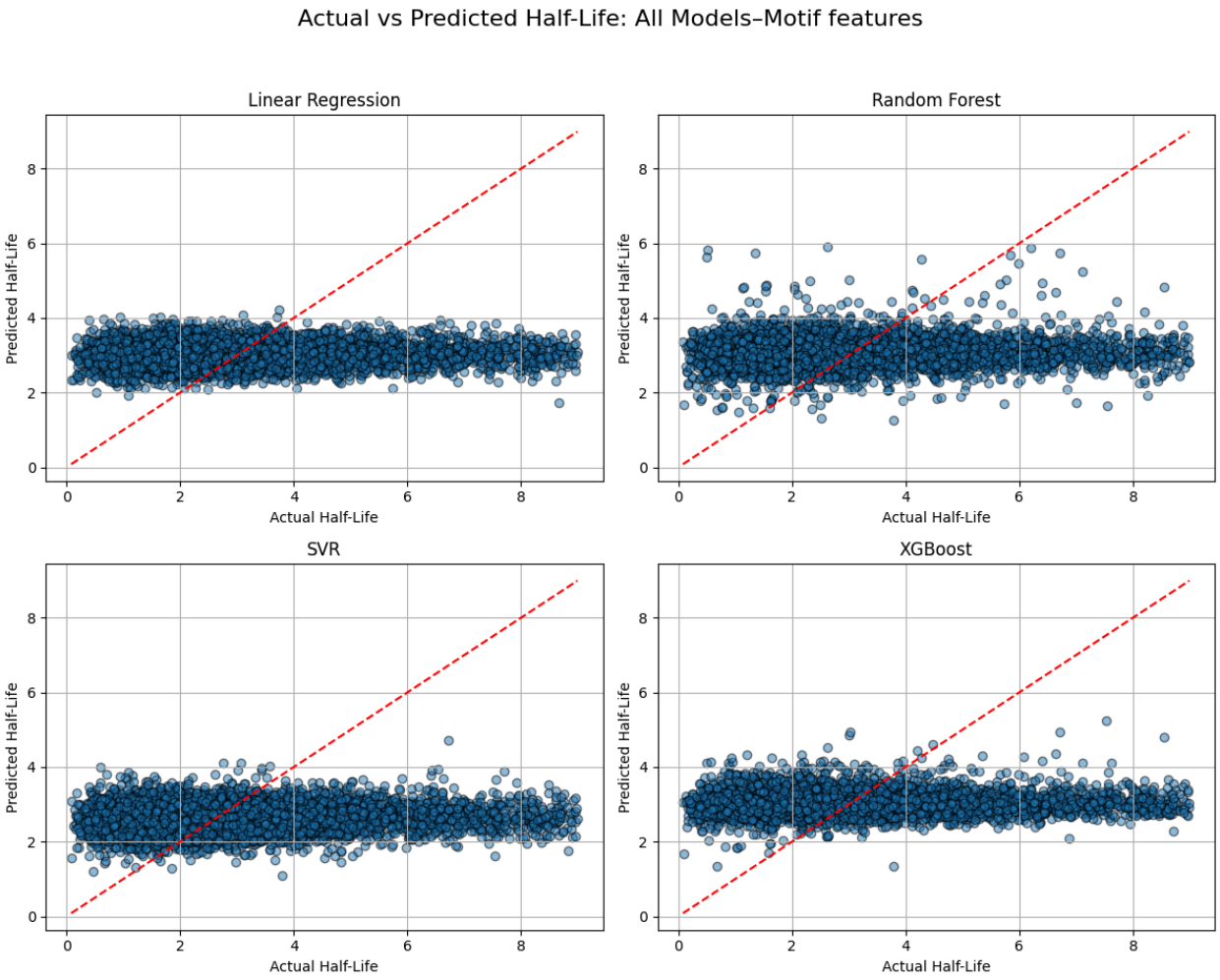


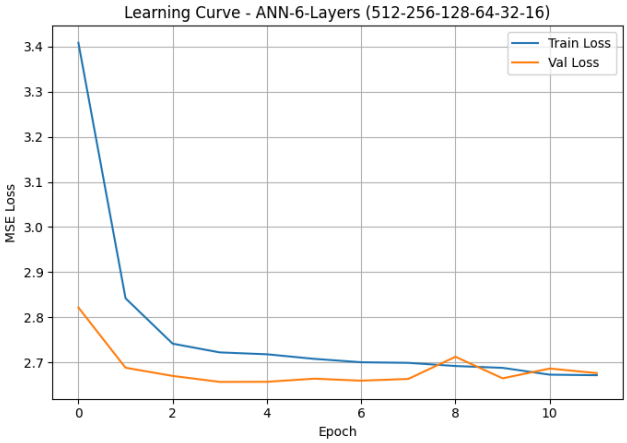
Figure 3: Actual vs Predicted Half-life of ML models using RBP motifs based features

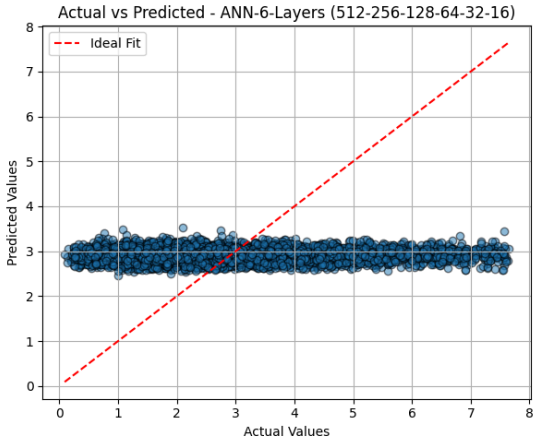
4.2 Performance of Deep Learning Models

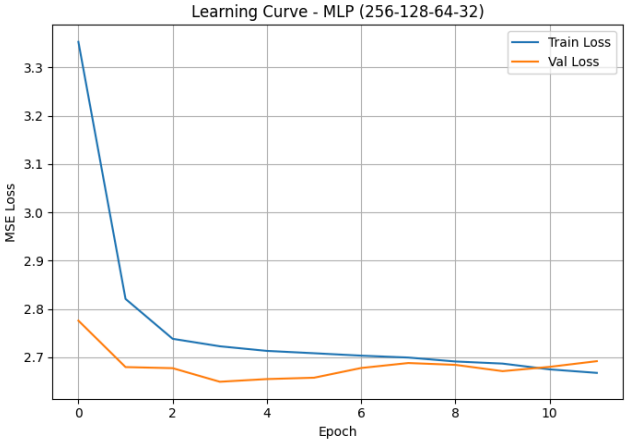
The DL models such as 1D-CNN, MLP and BiLSTM were also used with the full engineered feature set as well as one-hot or numerical encodings. While these models captured more complex representations, the R^2 value remained consistent with results of the ML model. The optimization and regularization of the model also failed to substantiate better results (Table 2).

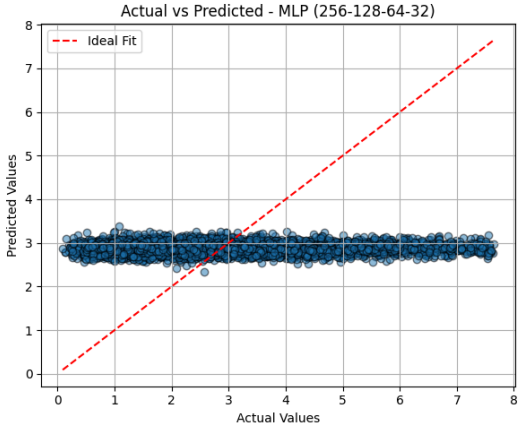
Features	Model	Training Dataset			Validation Dataset		
		R	MAE	RMSE	R	MAE	RMSE
All comp	ANN (6 layers)	0.1658	1.3050	1.6226	0.0545	1.3109	1.6301
	MLP	0.1319	1.3125	1.6353	0.0346	1.3081	1.6226
	1D–CNN	0.3711	1.2038	1.5185	0.0521	.1.3144	1.6569
	BiLSTM	0.0443	1.3247	1.6467	0.0390	1.3107	1.6295

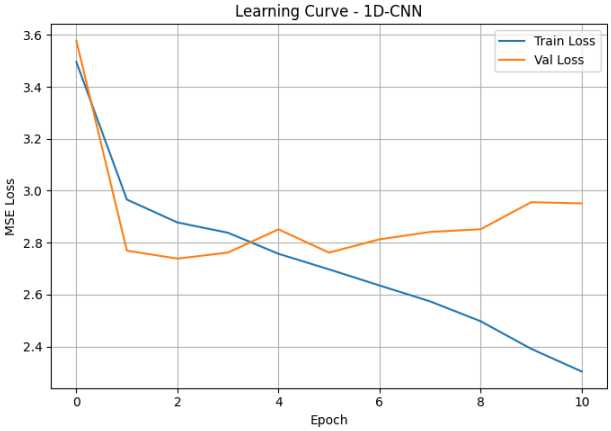
Table 2: DL models and their performance on composition features and sequence feature

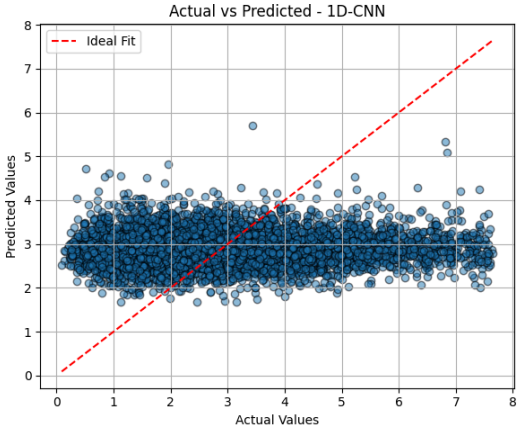

Learning Curve - ANN-6-Layers (512-256-128-64-32-16)
Train Loss
Val Loss
MSE Loss
Epoch


Actual vs Predicted - ANN-6-Layers (512-256-128-64-32-16)
Ideal Fit
Predicted Values
Actual Values


Learning Curve - MLP (256-128-64-32)
Train Loss
Val Loss
MSE Loss
Epoch


Actual vs Predicted - MLP (256-128-64-32)
Ideal Fit
Predicted Values
Actual Values


Learning Curve - 1D-CNN
Train Loss
Val Loss
MSE Loss
Epoch


Actual vs Predicted - 1D-CNN
Ideal Fit
Predicted Values
Actual Values

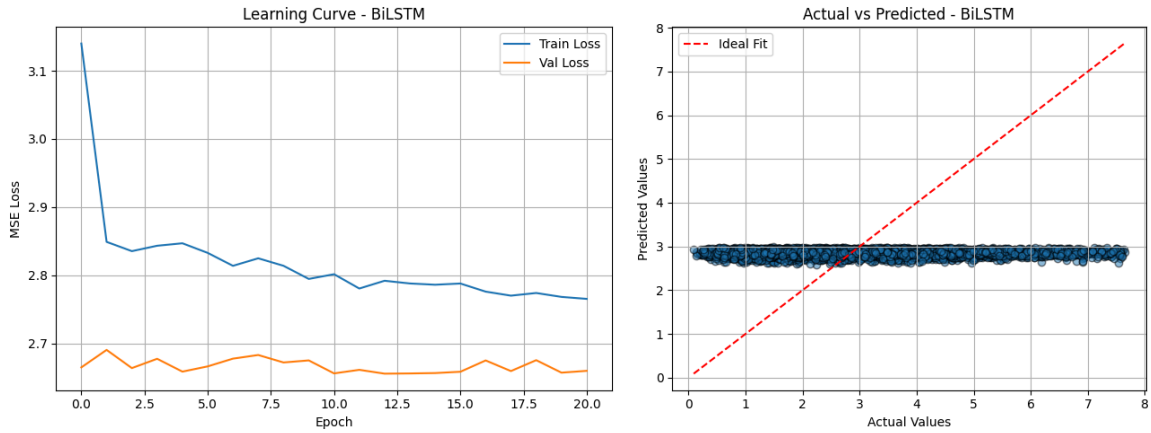


Figure 4: Learning Curve and Actual vs Predicted values for all the Deep Learning models

4.3 RNA-BERT-Based Regression Results

RNA-BERT generated embeddings were used for the downstream analysis for regression. The high dimensional captured tokens obtained from the pre-trained model were then modelled with a suite of both ML and DL models. The models remained consistent as the traditional approaches above. RNA-BERT was also fine tuned to enable the model to learn the nuances of RNA stability and to obtain a better insight on the training dataset.

However, despite the contextual power of BERT embeddings, the model failed to generalize, hinting that half-life determinants may not be captured by sequence context alone even with deep representations.

Features	Model	Training Dataset			Validation Dataset		
		R	MAE	RMSE	R	MAE	RMSE
Embeddings	Linear Regression	0.1483	2.5133	4.3583	0.0407	2.8541	5.0685
	SVR	0.3516	2.1342	4.5267	0.0881	2.5264	5.2409
	Random Forest	**0.9062**	**1.0792**	**1.8535**	**0.0619**	**2.9780**	**5.0606**
	XGBoost	0.8148	1.1778	1.6907	0.0460	2.9285	5.1271
	Gradient Boosting	0.6038	1.7586	2.5065	0.0385	2.9214	5.1451

Table 3: ML models and their performance in RNA-BERT

Features	Model	Training Dataset			Validation Dataset		
		R	MAE	RMSE	R	MAE	RMSE
Embeddings	ANN (6 layers)	0.1297	2.5441	4.3510	0.0428	2.8242	4.9987
	MLP	0.1077	2.6944	4.4058	0.0369	2.9379	4.9929
	1D–CNN	**0.4966**	**1.8377**	**3.0375**	**0.0540**	**2.9769**	**5.2231**
	BiLSTM	-0.0243	2.6547	4.4444	-0.0040	2.8787	5.0045

Table 4: DL models and their performance in RNA-BERT

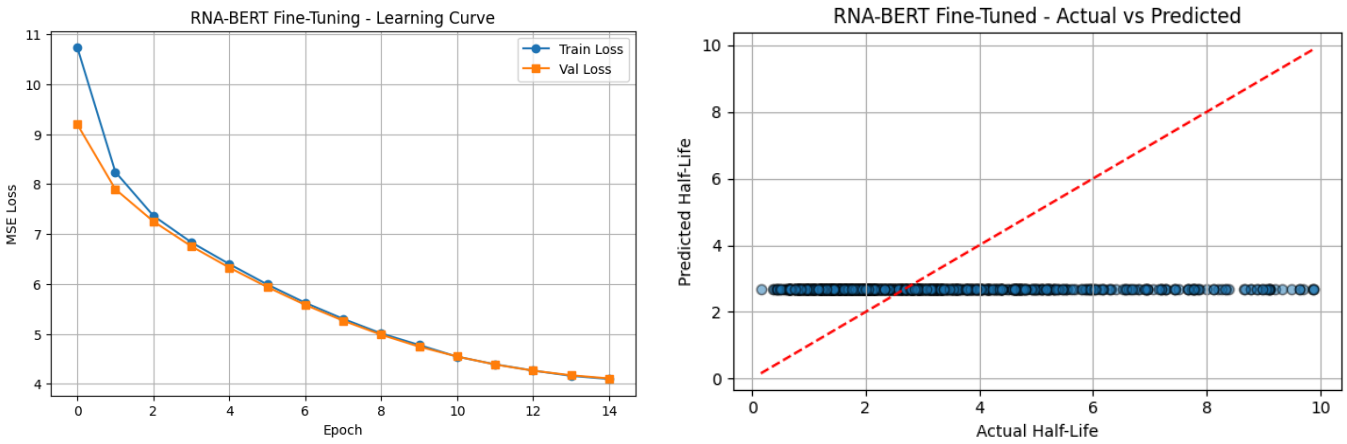


Figure 5: Fine-tune RNA-BERT Learning curve and Predictions

Model	R	MAE	RMSE
RNABert Fine Tune (10 epochs)	0.0113	1.5152	2.1785
RNABert Fine tune (15 epochs)	-0.0425	1.4917	2.0267

Table 5: Fine-tune RNA-BERT model results

4.4 K-mer Clustering Results

The distinct clusters were obtained from the sequence features of the 3-mer based clustering approach (Figure 6-7) demonstrated compositional differences captured by k-means clustering. Still, the average half-life across all the three clusters remained nearly the same (Figure 8-9).

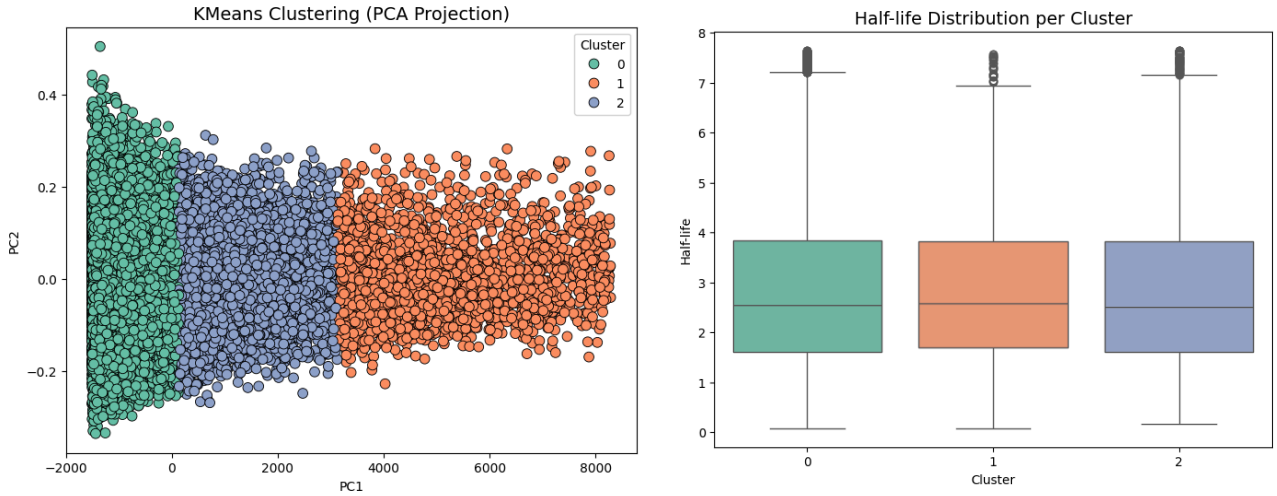


Figure 6-7: PCA Projection of K-means clustering; Half-life distribution of each cluster

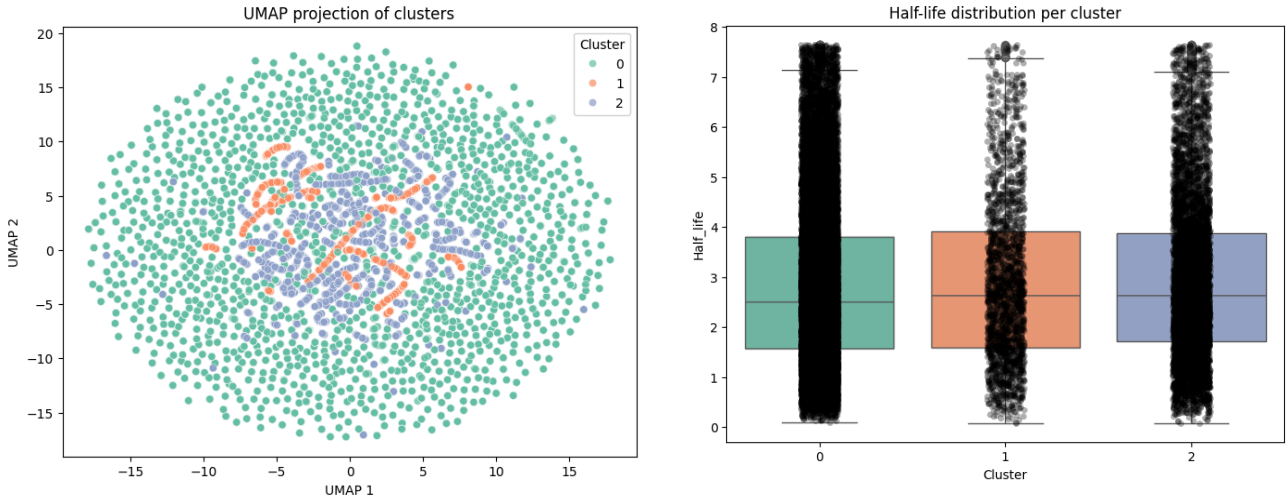


Figure 8-9: UMAP Projection of K-means clustering; Half-life distribution of each cluster

4.5 Summary of Findings

All the models demonstrated poor regression performance, with spearman correlation values consistently < 0.06, indicating no better than mean-prediction. Even powerful representations like RNA-BERT embeddings did not substantially improve predictive power. The observed weak correlations among various features and half-life suggests that lncRNAs half-life is governed by complex, multifactorial biological processes not easily inferred from sequence and other related features.

4.6 Observational Insights from EDA

The distribution of half-life values was skewed toward short half-lives, with ~80% of lncRNAs having half-lives < 5 hours (Figure 8a). Sequence length, GC content, MFE, RBP motifs and localization based features also showed very low or no correlation with half-life ($|r| < 0.1$) (Figure 8b, 8c, 8d).

These weak relationships emphasize that traditional compositional features as well as available motif information poorly captures the decay dynamics.

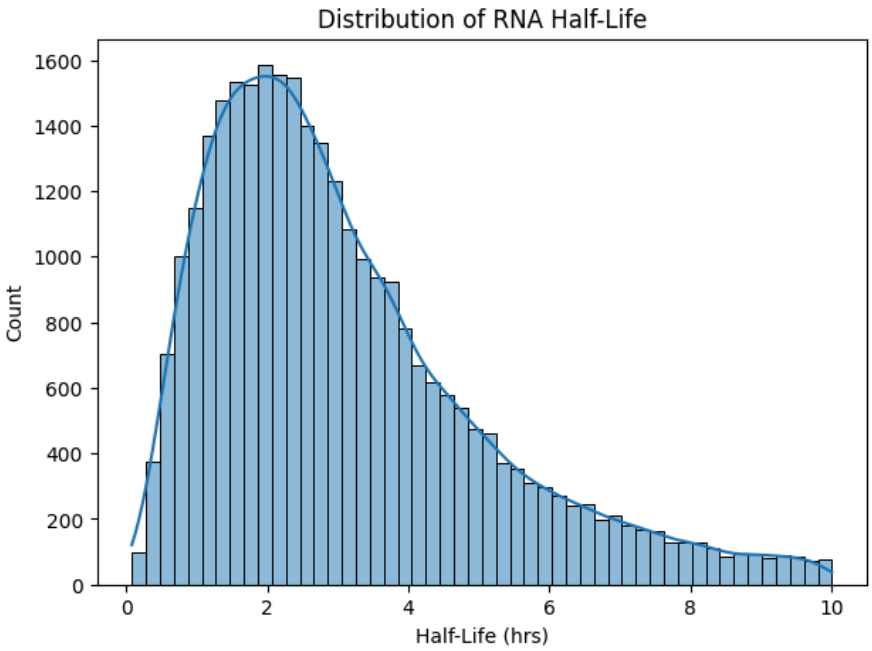


Figure 10a: Distribution of RNA Half-life for lncRNAs

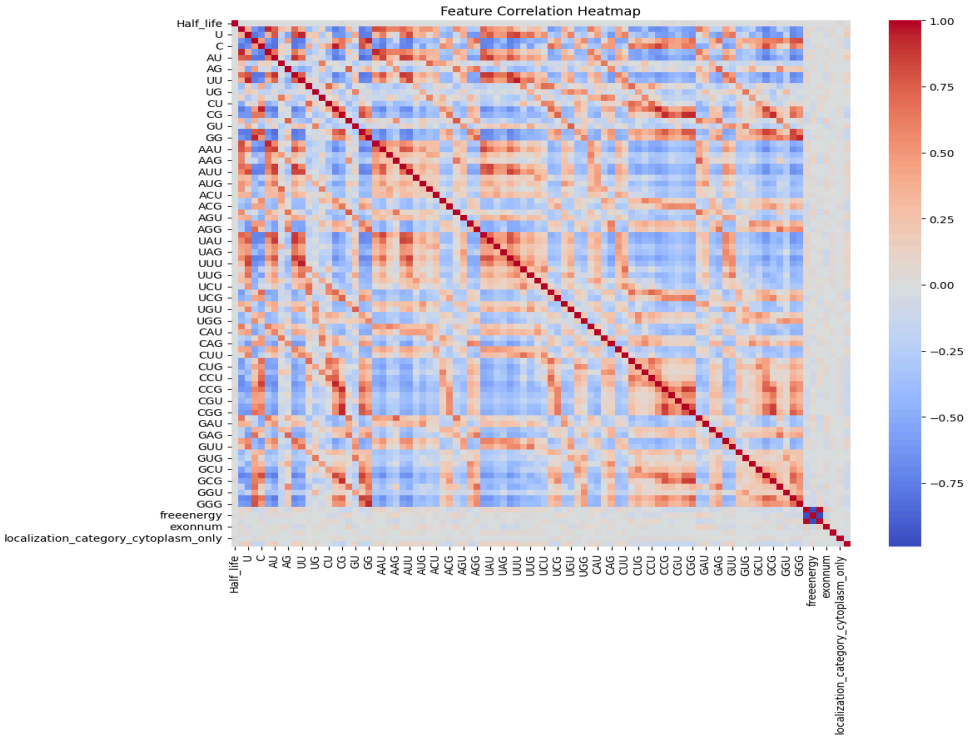


Figure 10b: Heatmap of compositional feature correlation

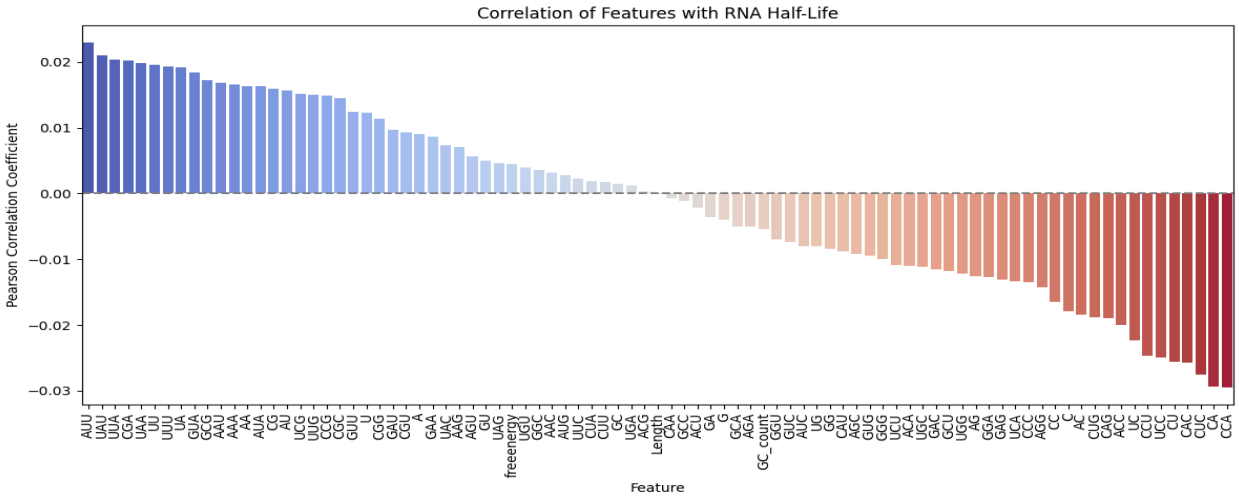


Figure 10c: Correlation of Composition based features with RNA Half-life

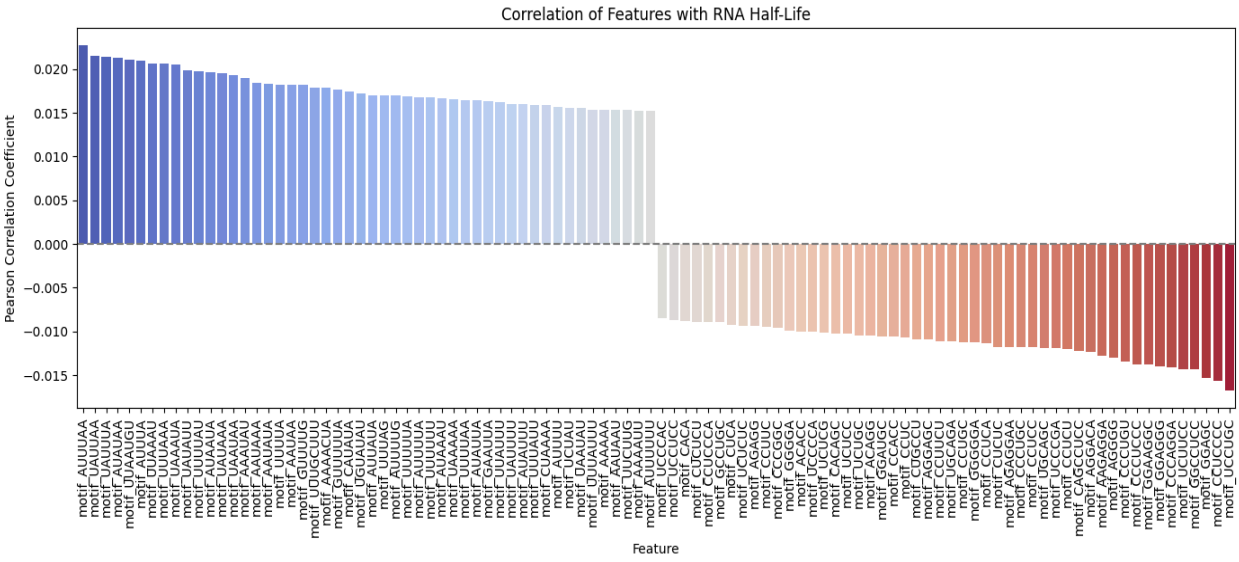


Figure 10d: Correlation of top motifs based features with RNA Half-life

5. Discussion

5.1 Interpretation of Poor Model Performance

Despite employing a wide array of machine learning and deep learning models, including advanced contextual embeddings from RNA-BERT, our results showed consistently poor regression performance. Across all approaches, Spearman correlation values were near zero or negative, and half-life predictions deviated minimally from the global mean.

Our findings clearly indicate that sequence based features are insufficient as well as the high-dimensional RNA-BERT features are inadequate to predict the RNA half-life, leading to failure in building models. This also suggests that the RNA stability is governed by features and factors beyond the primary sequence.

5.2 Biological Complexity Beyond Sequence

The poor predictive performance highlights that RNA stability is not just an intrinsic property encoded within nucleotide sequence, rather half-life is a product of dynamic cellular environment that involves other layers to it as well such as:

- RBP Motifs Problem: We included RBP motifs, but this doesn't imply that it will be used in the cell, since RNA folding could hide such motifs or two proteins would compete for the same binding sites.
- Epitranscriptome modification: Chemical modifications of m^6A methylation are added post transcriptionally which aren't present at RNA sequence level, and significantly influence the transcript stability.
- Sub-cellular localization: We used these localization features, but the degradation also depends on whether it is currently being translated or performing its function or whether it is dormant. Both change the fate of half-life completely.
- Cell-State Dependency: RNA half-life is dynamic depending on the cell types, developmental stages and external stress conditions.

Though the sequences provide valuable baseline insights about the RNA sequences, the post-translational modifications, cell state, subcellular localization and RBP play a crucial role in the regulation of these lncRNAs. Despite taking some of the features other than sequence based, such as RBP motifs, subcellular localization and exon numbers per sequence, the models failed to capture the relevant context since the metadata of such features are limited to fewer sequences and not available in abundance, specially for the lncRNAs. Thus, making it difficult to understand the context of the complex regulatory mechanism of lncRNA stability and decay rates.

5.3 Insights from Clustering Analysis

Interestingly, our unsupervised 3-mer UMAP + K-means clustering showed that while sequence features can separate lncRNAs into distinct compositional clusters, the mean half-life across clusters remained nearly identical (2.85, 2.91, and 2.93 hours). This confirms that k-mer patterns vary across sequences but are not informative of half-life, reinforcing our conclusion that primary sequence features have low biological signal with respect to stability.

5.4 Comparison with Prior Work

Shi et al. (2021) together with experimental determination of both mRNA and lncRNA, built a deep learning model to predict the half-life of RNA, but the results were consistent to be low (R^2 < 0.2). Due to the importance of the RNA stability and its degradation, several models have been built to understand the degradation model such as for mRNA sequences (highly conserved) by He et al. (2023). It incorporates deep learning and biological insights for understanding the relation of each nucleotide irrespective of their position, outperforming other previous models, but were trained on a small dataset of sequence length <110 bp. In our study, we built a predictive model on a larger dataset (>30,000 lncRNAs), for lncRNAs, which are less conserved and structurally diverse compared to mRNAs and show even lower predictive performance.

Furthermore, we have also evaluated a pretrained model–RNA-BERT for half-life regression. While RNA-BERT has shown promise in classification tasks, its lack of performance in this regression task suggests that it does not encode sufficient decay-relevant information—perhaps due to its training objective being unrelated to transcript stability.

6. Conclusion

RNA stability prediction in this investigation highlights both the progression and persistent challenge that necessitates to continue the work. This work highlights that only sequence and static features based approaches are limited to understand the dynamics of RNA stability. The pre-trained model (RNA-BERT) also failed to bring the contextual outcome for our dataset and regression task.

The current aspect of the study reveals that all possible features currently utilized or available are either too sparse, or when present, not sufficient to yield complete insights from RNA sequences necessary for robust half-life prediction.

As the current models are limited, it is important to address the challenges and advance future efforts to explore or innovate approaches for predictive analysis. The incorporation of interaction networks, cellular context, epitranscriptome marks would be essential and could possibly improve half-life prediction. Also, the less conserved and structurally diverse nature of the lncRNAs being complex to capture biological insights, opens the doors to explore the entirely new algorithm based prediction of half-life rather than relying on existing features or superficial high dimensional features without deeper insights.

References:

1. Shi, K., Liu, T., Fu, H., Li, W., & Zheng, X. (2021). Genome-wide analysis of lncRNA stability in human. *PLoS computational biology*, *17*(4), e1008918. https://doi.org/10.1371/journal.pcbi.1008918
2. Mattick, J. S., Amaral, P. P., Carninci, P., Carpenter, S., Chang, H. Y., Chen, L. L., Chen, R., Dean, C., Dinger, M. E., Fitzgerald, K. A., Gingeras, T. R., Guttman, M., Hirose, T., Huarte, M., Johnson, R., Kanduri, C., Kapranov, P., Lawrence, J. B., Lee, J. T., Mendell, J. T., … Wu, M. (2023). Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nature reviews. Molecular cell biology*, *24*(6), 430–447. https://doi.org/10.1038/s41580-022-00566-8
3. Ietswaart, R., Smalec, B. M., Xu, A., Choquet, K., McShane, E., Jowhar, Z. M., Guegler, C. K., Baxter-Koenigs, A. R., West, E. R., Fu, B. X. H., Gilbert, L., Floor, S. N., & Churchman, L. S. (2024). Genome-wide quantification of RNA flow across subcellular compartments reveals determinants of the mammalian transcript life cycle. *Molecular cell*, *84*(14), 2765–2784.e16. https://doi.org/10.1016/j.molcel.2024.06.008
4. Wang, C., & Liu, H. (2022). Factors influencing degradation kinetics of mRNAs and half-lives of microRNAs, circRNAs, lncRNAs in blood in vitro using quantitative PCR. *Scientific reports*, *12*(1), 7259. https://doi.org/10.1038/s41598-022-11339-w
5. Tani, H., Mizutani, R., Salam, K. A., Tano, K., Ijiri, K., Wakamatsu, A., Isogai, T., Suzuki, Y., & Akimitsu, N. (2012). Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome research*, *22*(5), 947–956. https://doi.org/10.1101/gr.130559.111
6. Li, M., & Liang, C. (2022). LncDC: a machine learning-based tool for long non-coding RNA detection from RNA-Seq data. *Scientific reports*, *12*(1), 19083. https://doi.org/10.1038/s41598-022-22082-7
7. Wayment-Steele, H. K., Kladwang, W., Watkins, A. M., Kim, D. S., Tunguz, B., Reade, W., Demkin, M., Romano, J., Wellington-Oguri, R., Nicol, J. J., Gao, J., Onodera, K., Fujikawa, K., Mao, H., Vandewiele, G., Tinti, M., Steenwinckel, B., Ito, T., Noumi, T., He, S., … Das, R. (2022). Deep learning models for predicting RNA degradation via dual crowdsourcing. *Nature machine intelligence*, *4*(12), 1174–1184. https://doi.org/10.1038/s42256-022-00571-8

8. Conte, F., Papa, F., Paci, P., & Farina, L. (2022). StaRTrEK:in silico estimation of RNA half-lives from genome-wide time-course experiments without transcriptional inhibition. *BMC bioinformatics*, *23*(1), 190. https://doi.org/10.1186/s12859-022-04730-x
9. Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics (Oxford, England)*, *37*(15), 2112–2120. https://doi.org/10.1093/bioinformatics/btab083
10. Akiyama, M., & Sakakibara, Y. (2022). Informative RNA base embedding for RNA structural alignment and clustering by deep representation learning. *NAR genomics and bioinformatics*, *4*(1), lqac012. https://doi.org/10.1093/nargab/lqac012
11. Zhao, L., Wang, J., Li, Y., Song, T., Wu, Y., Fang, S., Bu, D., Li, H., Sun, L., Pei, D., Zheng, Y., Huang, J., Xu, M., Chen, R., Zhao, Y., & He, S. (2021). NONCODEV6: an updated database dedicated to long non-coding RNA annotation in both animals and plants. *Nucleic acids research*, *49*(D1), D165–D171. https://doi.org/10.1093/nar/gkaa1046
12. Giudice, G., Sánchez-Cabo, F., Torroja, C., & Lara-Pezzi, E. (2016). ATtRACT-a database of RNA-binding proteins and associated motifs. *Database : the journal of biological databases and curation*, *2016*, baw035. https://doi.org/10.1093/database/baw035
13. Hugging face: RNABert: https://huggingface.co/multimolecule/rnabert
14. Mathur, M., Patiyal, S., Dhall, A., Jain, S., Tomer, R., Arora, A., & Raghava, G. P. (2021). Nfeature: A platform for computing features of nucleotide sequences. *BioRxiv*, 2021-12. https://doi.org/10.1101/2021.12.14.472723
15. He, S., Gao, B., Sabnis, R., & Sun, Q. (2023). RNAdegformer: accurate prediction of mRNA degradation at nucleotide resolution with deep learning. *Briefings in bioinformatics*, *24*(1), bbac581. https://doi.org/10.1093/bib/bbac581
16. scikit learn library for feature normalization, train test split and model