Flores: mRNA stability prediction based on RNA language model

Seong Woo Han
Computer and Information Science
University of Pennsylvania
seonghan@seas.upenn.edu

Farica Zhuang
Computer and Information Science
University of Pennsylvania
farica@seas.upenn.edu

Abstract

mRNA stability is a crucial factor in gene expression regulation and has significant implications for cellular function and therapeutic applications, making it an important area of study. Accurate prediction of mRNA stability can provide insights into the mechanisms underlying gene regulation and assist in designing mRNA-based therapeutics with enhanced stability and efficacy. Here, we developed an mRNA stability prediction model named Flores, using a foundation model initially pre-trained on the human transcriptome. The pre-trained model was fine-tuned with a dataset comprising a compendium of experimentally determined mRNA sequences and their corresponding half-life values. By leveraging the extensive information captured in the pre-trained model and the specific stability data, we aimed to enhance the model's ability to accurately predict mRNA half-lives. We explored the application of Flores in predicting mRNA half-life using different embedding techniques to handle long sequences. Despite hypothesizing that the transformer-based Flores would outperform hybrid convolutional and recurrent deep neural networks, the results indicated otherwise. The current state-of-the-art model, Saluki, performed better, likely due to its effective incorporation of biological sequence features and absence of input length limitations to process entire sequences without the information loss associated with embeddings in transformer models. By directly incorporating crucial mRNA features, Saluki can capture details that a nucleotide-only strategy might miss, underscoring the importance of preserving original sequence information and carefully integrating sequence features for optimal performance. Hence, our findings suggest that while transformer-based models like Flores hold promise, the incorporation of biological sequences and their features is more critical than merely using more powerful models.

1 Introduction

The steady-state level of mRNA is influenced by the rate of transcription and the rate of decay. Significant progress has been made in predicting steady-state mRNA abundances by focusing on DNA-encoded features that impact transcription rates. However, less is known about mRNA-encoded determinants that affect decay rates. Better understanding mRNA decay rates would enable the design of more stable and effective mRNA therapeutics. Experimentally, the rate of RNA decay is measured by its half-life, which is the time required for the RNA concentration to decrease by half.

Predicting mRNA stability has been studied in the past [1, 2, 3]. In this work, we focus on comparing our transformer-based mRNA stability prediciton model, Flores, against the state-of-the-art computational model, Saluki [4]. Saluki utilizes a hybrid convolutional and recurrent deep neural network architecture to predict mRNA half-life from three features: spliced mRNA sequences, an encoding of the first position of each codon, and $5'$ splice site junctions. The nucleotide sequence is one-hot encoded into four input tracks. Additionally, to account for codon composition, the first position of each codon in the CDS is encoded as 1, while all other positions, including those in the $5'$ UTR and $3'$ UTR, are encoded as 0. This encoding implicitly differentiates between the $5'$ UTR, CDS, and $3'$ UTR regions. Furthermore, splicing sites are annotated by marking each $5'$ splice site at the exon's $5'$ nucleotide with a 1, while all other positions are encoded as 0.

However, while splice site information is deemed to be an important feature of the model, this information can only be easily incorporated for annotated natural transcripts. When working with synthetic mRNA sequences designed for vaccines and therapeutics, there would be no splice site information. Consequently, models like Saluki, which rely heavily on these annotations, may not perform optimally for synthetic mRNAs. This limitation highlights the need for a more versatile model that can handle both natural and synthetic mRNA sequences.

Despite the advancements in models like Saluki, there has yet to be a large language model (LLM) based mRNA stability model. In the era of LLMs, models such as the Transformer have demonstrated their ability to capture both local and global dependencies in sequences [5]. Additionally, the Hyena architecture has shown potential for significantly increasing context length in sequence models by combining long convolutions and gating mechanisms [6]. Moreover, the application of LLMs to biological data has become a trend for various downstream analyses

[7, 8, 9, 10]. By moving beyond the constraints of splice site annotations, transformer models have the potential to provide more accurate and reliable predictions based on sequences and other relevant biological information, ultimately improving the design and efficacy of mRNA-based interventions. We hypothesize that by leveraging the powerful attention mechanisms of transformer models, this approach can capture complex dependencies and interactions within the mRNA sequence, even in the absence of explicit splice site information.

To explore whether transformer-based models can outperform other deep learning methods such as CNN/RNN in predicting mRNA stability, we developed an mRNA stability prediction model named Flores, using a foundation model initially pre-trained on the hu- man transcriptome for sequence inputs and compared its performance against Saluki. By doing so, we aim to assess the effectiveness of transformer-based models in this domain and identify potential improvements in predicting mRNA stability.

2 Data

2.1 mRNA half-lives data

The mRNA half-life per gene was determined based on studies collected from Saluki. Gene annotations for protein-coding genes were derived from Ensembl v83 (hg38 genome).

2.2 Preprocessing

To align with Saluki's training and testing datasets, we selected the representative transcript for each gene by choosing the transcript with the longest ORF, followed by the longest $5'$ UTR, and then the longest $3'$ UTR among all transcripts corresponding to that gene. The half-life values were z-score normalized by subtracting their respective mean values and dividing by their standard deviations to standardize the scale (Figure 1).

Figure 1 shows the distribution of mRNA half-life values for both the training and test datasets. The histogram includes the original half-life values and their z-score normalized counterparts. The purple shaded area represents the distribution of the original half-life values in the training data, while the light blue shaded area represents the original half-life values in the test data. The dark purple shaded area shows the distribution of the normalized half-life values in the training data, and the dark blue shaded area shows the normalized half-life values in the test data.

Upon obtaining 13,663 mRNA transcript sequences, we applied 10-fold cross-validation to those sequences. In each fold, 12,295 sequences were used for training, and 1,368 sequences were used for testing. This approach ensures that our model is evaluated on different subsets of the data, enhancing its generalizability and robustness. Each fold serves as a validation set once, while the remaining nine folds are used for training.

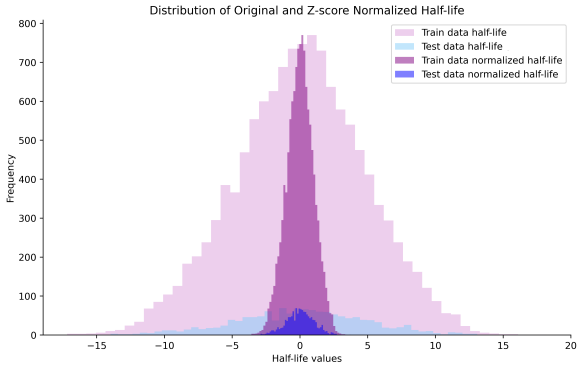

Figure 1: Distribution of Original and Z-score Normalized Half-life

3 Methods

3.1 Handling long sequences

The median length of human transcripts is roughly 3000 base pairs (bp) or more [11]. This poses a challenge for transformer-based models like Flores, which have a limited input size of 512 tokens. To effectively utilize Flores for these long sequences, we implemented the following three embedding technique. These embedding techniques allow us to handle long mRNA transcript sequences by segmenting them and effectively capturing their essential features using transformer-based models. The resulting vector representations are then used to predict the mRNA half-life through a neural network model, enabling overcoming the initial sequence length challenges. All trainings were set to 10 epochs with early stopping. Patience is set to 3, meaning that training will stop if there is no improvement after three consecutive epochs.

1. **Mean Pooling**: We divided each long mRNA sequence into several chunks, each within the 512-token limit. For each chunk, we passed the sequence through Flores and extracted the hidden states from the last layer. We then computed the average of these hidden states across all chunks. This method, known as mean pooling, creates a single vector representation for the entire transcript by averaging the hidden states of all chunks.

2. **[CLS] Token Embedding**: We divided the long sequences into chunks, each within the 512-token limit. For each chunk, we used the embedding of the [CLS] token from the last layer. The [CLS] token, which stands for "classification" token, is designed to capture the overall context of the sequence. We either concatenated or averaged the [CLS] token embeddings from all chunks to form a comprehensive vector representation of the transcript.

3. **Hierarchical Attention**: We divided the long sequences into smaller chunks. For each chunk, we obtained the hidden states from the last layer of Flores. We applied a hierarchical attention mechanism where:

- **Chunk-Level Attention**: Attention weights are computed for each chunk to highlight the most relevant parts of the transcript.
- **Aggregation**: These weights are used to combine the chunk embeddings into a single vector representation. Hierarchical attention helps capture both local and global dependencies in the transcript. This method is particularly useful for effectively capturing long-range dependencies and structural information, which is crucial in our case as RNA structures far from the 5′ UTR or CDS can potentially affect mRNA stability.

Each resulting vector representation from these techniques is then fed into a Multi-Layer Perceptron (MLP) to predict the mRNA half-life. We validated our model using 10-fold cross-validation, ensuring that each transcript was used for both training and validation, thereby providing a robust assessment of the model's performance.

3.2 Input sequences

To investigate which parts of the mRNA are most important for stability, we experimented with four different input sequences:

1. **Entire 3′ UTR**: The full length of the 3′ UTR was used as input. The 3′ UTR is known to harbor various regulatory motifs and binding sites for RNA-binding proteins, which can impact mRNA degradation and stability. This sequence was embedded to create a vector representation for input to the model.
2. **CDS + 3′ UTR**: We combined the coding sequence (CDS) with the 3′ UTR. The CDS can reflect codons and splice sites, providing the sequence for protein synthesis. Its interplay with the 3′ UTR may [**?**]reveal insights into how coding regions influence mRNA stability through co-translational and post-transcriptional mechanisms. This combined sequence was embedded for model input.
3. **5′ UTR + CDS + 3′ UTR**: The entire mRNA transcript, including the 5′ UTR, CDS, and 3′ UTR, was used. The 5′ UTR plays a role in translation initiation, and including it along with the CDS and 3′ UTR provides a comprehensive view of how different regions of the mRNA contribute to overall stability. This entire sequence was also embedded for model input.
4. **Last 512 tokens of the 3′ UTR**: We extracted the last 512 base pairs of the 3′ UTR from each mRNA transcript. Unlike the other methods above, this approach uses the discrete sequences directly without embedding them, preserving the original information. The advantage of this method is that it retains the detailed sequence information, which might be lost during the embedding process. However, a limitation is that it cannot capture information from sequences longer than 512 base pairs.

By testing these different input sequences, we aim to identify the regions of the mRNA that are most influential in determining its stability.

4 Results

Table 1 presents the Pearson correlation coefficients (r) and Spearman's rank correlation coefficients (rho) for different training techniques and input sequences used in predicting mRNA half-life using Flores. The training techniques compared are Mean Pooling, [CLS] Token Embedding, and Hierarchical Attention, with three different learning rates for each method. The input sequences evaluated include the 3' UTR, the combination of the 3' UTR and CDS, the combination of the 5' UTR, CDS, and 3' UTR, and a method using the last 512 sequences from the end of the 3' UTR without embedding.

The table shows that Hierarchical Attention consistently achieves higher correlation coefficients, particularly with the combination of 3' UTR and CDS input sequences, reaching a Pearson correlation of 0.48 and a Spearman correlation of 0.46 at a learning rate of 2e-4. This indicates its superior ability to capture relevant features for mRNA half-life prediction. On the other hand, the Mean Pooling method also performs well, especially for the 3' UTR input sequence with a Pearson correlation of 0.44 and a Spearman correlation of 0.42 at a learning rate of 1e-3. The CLS Token Embedding method shows lower correlation values across different input sequences and learning rates.

Interestingly, the method using 512 sequences from the end of the 3' UTR without embedding shows a Pearson correlation of 0.38 and a Spearman correlation of 0.36 at a learning rate of 1e-5. This suggests that preserving the original sequence information can be nearly as effective as using embedding techniques, emphasizing the potential benefits of using raw sequences directly for specific mRNA features. The learning rate of 1e-3 didn't converge, unable to show Pearson (r) and Spearman (rho) correlation. Overall, the table highlights the effectiveness of Hierarchical Attention for the embedding technique and potentially suggests that using the original sequence information can be as effective, highlighting the potential benefits of not embedding sequences when dealing with specific mRNA features. The corresponding scatter plots for Table 1 can be found in Figures 1-4.

Table 2 presents the Pearson (r) correlation coefficients from an ablation analysis in which Saluki was evaluated after training it with different combinations of input tracks: sequence track (S), sequence and coding frame tracks (SC), sequence and splice site tracks (Ss), and sequence, coding frame, and splice site tracks (SCs). This analysis helps determine the impact of each track combination on the model's performance in predicting mRNA half-life. Saluki utilizes a hybrid convolutional and recurrent deep neural network architecture to predict mRNA half-life from spliced mRNA sequences. Unlike transformer models, Saluki does not have length restrictions and incorporates key gene structure annotations. The nucleotide sequence is one-hot encoded into four input tracks. Additionally, to account for the impact of splicing and codon composition on RNA stability, binary tracks were added to mark exon junctions and codon start positions, implicitly labeling the 5' and 3' UTRs. This approach ensures the capture of important mRNA features that would otherwise require substantial auxiliary training information. In the table, the combinations of input tracks show varying degrees of effectiveness, with the sequence track alone

Input Type	Embedding Technique	Learning Rate	Pearson (r)	Spearman (rho)
3' UTR	Mean Pooling	1e-3	0.44	0.42
		2e-4	0.43	0.40
		1e-5	0.32	0.30
	Hierarchical Attention	1e-3	0.35	0.34
		2e-4	0.35	0.33
		1e-5	0.22	0.22
	CLS	1e-3	0.27	0.25
		2e-4	0.27	0.26
		1e-5	0.27	0.26
3' UTR + CDS	Mean Pooling	1e-3	0.39	0.36
		2e-4	0.40	0.37
		1e-5	0.38	0.36
	Hierarchical Attention	**1e-3**	**0.48**	**0.46**
		2e-4	**0.48**	**0.46**
		1e-5	0.30	0.30
	CLS	1e-3	0.39	0.36
		2e-4	0.40	0.37
		1e-5	0.38	0.36
3' UTR + CDS + 5' UTR	Mean Pooling	1e-3	0.40	0.38
		2e-4	0.39	0.37
		1e-5	0.29	0.28
	Hierarchical Attention	1e-3	0.46	0.44
		2e-4	0.45	0.43
		1e-5	0.28	0.27
	CLS	1e-3	0.35	0.32
		2e-4	0.35	0.33
		1e-5	0.10	0.09
Last 512 sequences of the 3' UTR	No embedding	1e-3	0.00	0.00
		2e-4	0.31	0.29
		1e-5	0.38	0.36

Table 1: Pearson Correlation Coefficients for Different Embedding Techniques and Input Sequences

	Pearson (r)
S	0.59
SC	0.68
Ss	0.74
SCs	0.77

Table 2: Pearson Correlation Coefficients for Different Input Tracks in Saluki. S: Sequence track, C: Coding frame tack, s: 5' splice site track

achieving a Pearson correlation of 0.59, the sequence and coding frame tracks achieving 0.68, the sequence and splice site tracks achieving 0.74, and the combination of sequence, coding frame, and splice site tracks achieving the highest correlation of 0.77. These results demonstrate the importance of including spatial positioning of splice sites and codons in improving the accuracy of mRNA half-life predictions, highlighting the usage of raw sequences and feature integration.

5 Discussion

In this study, we explored the application of Flores on predicting mRNA half life using different embedding techniques to deal with long sequences. This was done by training pre-trained model with different range of transcript sequences and their corresponding mRNA half life. We hypothesized that transformer-based Flores would perform better than hybrid convolutional and recurrent deep neural network. However, Saluki performed better, likely due to its effective incorporation of raw sequences and their features. The hybrid CNN/RNN model's lack of input length limitations allows it to process entire sequences without the information loss associated with embeddings in transformer models. By incorporating crucial mRNA features directly, the Saluki can capture details that a nucleotide-only strategy might miss, emphasizing the importance of preserving original sequence information and carefully incorporating sequence features to achieve optimal performance.

In future work, we aim to explore embedding techniques for capturing long sequences without losing information and to integrate mRNA features to avoid the limitations of a nucleotide-

only strategy. There is also room to optimize hyperparameters of the model, such as learning rate and others, to further enhance performance. Additionally, having a better foundation model could potentially lead to improved results. For example, pretraining a Hyena architecture [6] with mRNA sequences, which can accept up to 1 million tokens, would allow for the retention of complete sequence information and potentially provide better performance in predicting mRNA half-life. Alternatively, we could potentially fine-tune published RNA foundation models such as BigRNA and RNAernie [12, 13]. Additionally, we propose the integration of sequence data with other biologically relevant features that are also able to be extracted from synthetic mRNAs, such as codon usage patterns and secondary structure predictions using codon adaptation index (CAI) and minimum free energy (MFE) of the secondary structure.

Another avenue of improvement is the training data used for the model. Currently, the training data obtained from Saluki is made up of an integrated data from multiple experiments with varying measurement units by using the first principal component (PC1) values as labels. This approach, while attempting to standardize and harmonize diverse datasets, may introduce certain biases and obscure specific biological signals due to the inherent variability in experimental conditions and measurement techniques. The reliance on PC1 values as a universal label might oversimplify the complex nature of mRNA stability, potentially limiting the model's ability to accurately predict stability across different contexts. Hence, there needs to be discussions of designing a different training dataset with more relevant conditions and uniform experimental measurements. Furthermore, the experiments were done a mix of measurement methods that introduce significant technological and methodological biases [4] but were combined in the compendium of half lives used for training. Hence, the labels might not properly reflect the stability behaviors. There might be a need for condition specific or experiment specific predictions for a more accurate design of mRNA sequences for healthy patients or otherwise.

For further application, we can potentially use an enhanced version of Flores, incorporating the lessons learned, as an oracle to generate stable synthetic 3' UTR sequences using generative models for mRNA vaccine optimization. By utilizing these advanced models to create optimized mRNA sequences, we can enhance the stability and efficacy of mRNA-based vaccines, contributing to more effective and reliable mRNA vaccine development.

References

[1] J. Cheng, K. C. Maier, Ž. Avsec, P. Rus, and J. Gagneur, "Cis-regulatory elements explain most of the mrna stability variation across genes in yeast," *Rna*, vol. 23, no. 11, pp. 1648–1659, 2017.

[2] N. Spies, C. B. Burge, and D. P. Bartel, "3 utr-isoform choice has limited influence on the stability and translational efficiency of most mrnas in mouse fibroblasts," *Genome research*, vol. 23, no. 12, pp. 2078–2090, 2013.

[3] L. V. Sharova, A. A. Sharov, T. Nedorezov, Y. Piao, N. Shaik, and M. S. Ko, "Database for mrna half-life of 19 977 genes obtained by dna microarray analysis of pluripotent and differentiating mouse embryonic stem cells," *DNA research*, vol. 16, no. 1, pp. 45–58, 2009.

[4] V. Agarwal and D. R. Kelley, "The genetic and biochemical determinants of mrna degradation rates in mammals," *Genome Biology*, vol. 23, no. 1, p. 245, 2022.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[6] M. Poli, S. Massaroli, E. Nguyen, D. Y. Fu, T. Dao, S. Baccus, Y. Bengio, S. Ermon, and C. Ré, "Hyena hierarchy: Towards larger convolutional language models," in *International Conference on Machine Learning*, pp. 28043–28078, PMLR, 2023.

[7] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri, "Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome," *Bioinformatics*, vol. 37, no. 15, pp. 2112–2120, 2021.

[8] Z. Zhou, Y. Ji, W. Li, P. Dutta, R. Davuluri, and H. Liu, "Dnabert-2: Efficient foundation model and benchmark for multi-species genome," *arXiv preprint arXiv:2306.15006*, 2023.

[9] Ž. Avsec, V. Agarwal, D. Visentin, J. R. Ledsam, A. Grabska-Barwinska, K. R. Taylor, Y. Assael, J. Jumper, P. Kohli, and D. R. Kelley, "Effective gene expression prediction from sequence by integrating long-range interactions," *Nature methods*, vol. 18, no. 10, pp. 1196–1203, 2021.

[10] E. Nguyen, M. Poli, M. Faizi, A. Thomas, M. Wornow, C. Birch-Sykes, S. Massaroli, A. Patel, C. Rabideau, Y. Bengio, *et al.*, "Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution," *Advances in neural information processing systems*, vol. 36, 2024.

[11] I. Lopes, G. Altab, P. Raina, and J. P. De Magalhães, "Gene size matters: an analysis of gene length in the human genome," *Frontiers in genetics*, vol. 12, p. 559998, 2021.

[12] A. Celaj, A. J. Gao, T. T. Lau, E. M. Holgersen, A. Lo, V. Lodaya, C. B. Cole, R. E. Denroche, C. Spickett, O. Wagih, *et al.*, "An rna foundation model enables discovery of disease mechanisms and candidate therapeutics," *bioRxiv*, pp. 2023–09, 2023.

[13] N. Wang, J. Bian, Y. Li, X. Li, S. Mumtaz, L. Kong, and H. Xiong, "Multi-purpose rna language modelling with motif-aware pretraining and type-guided fine-tuning," *Nature Machine Intelligence*, pp. 1–10, 2024.

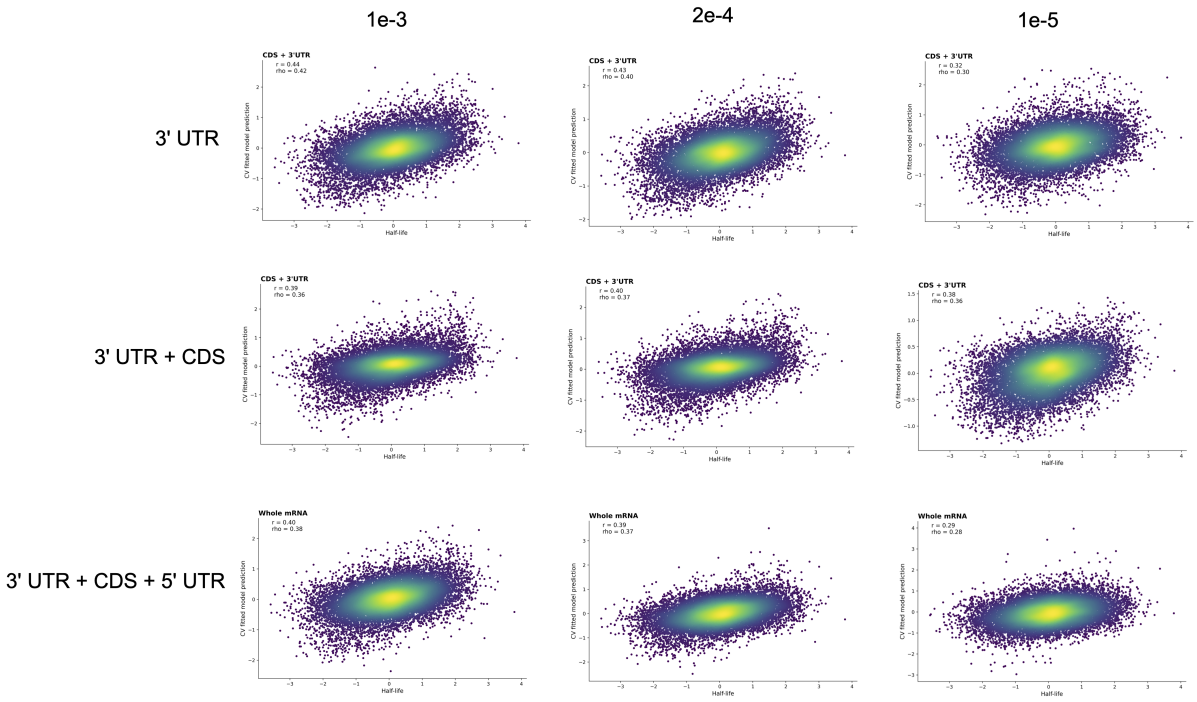

Figure 2: Mean pooling on three Sequence Inputs across three learning rates

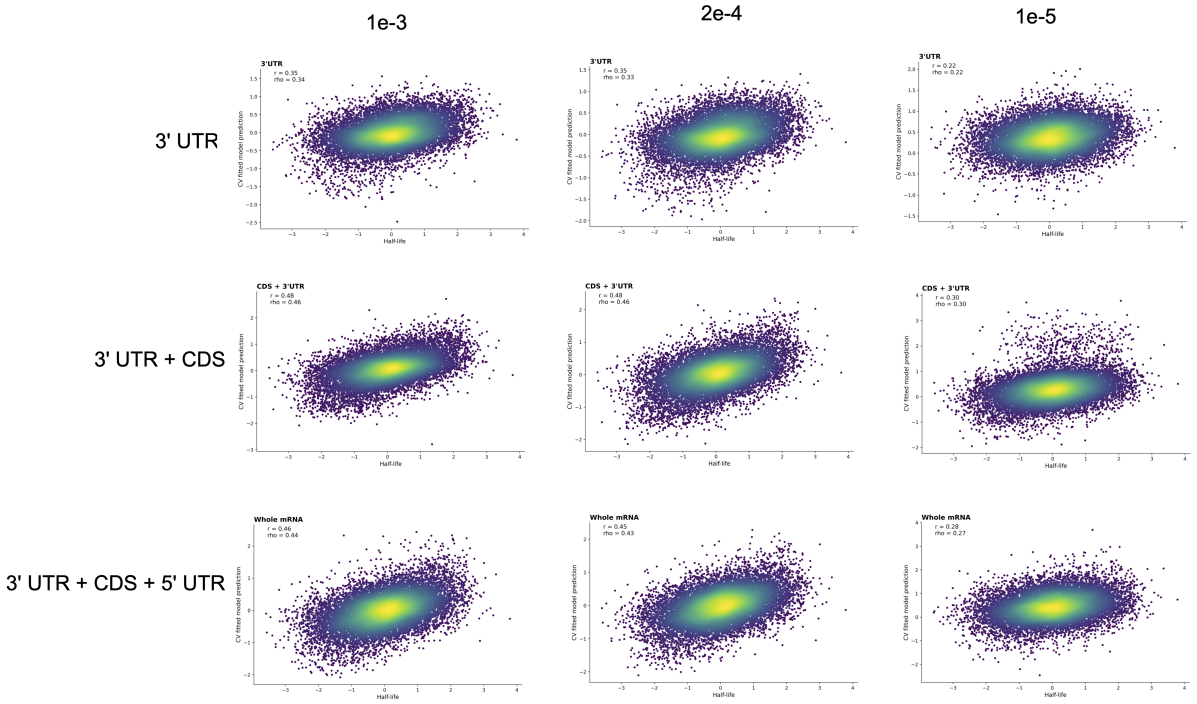

Figure 3: Hierarchical Attention on three Sequence Inputs across three learning rates

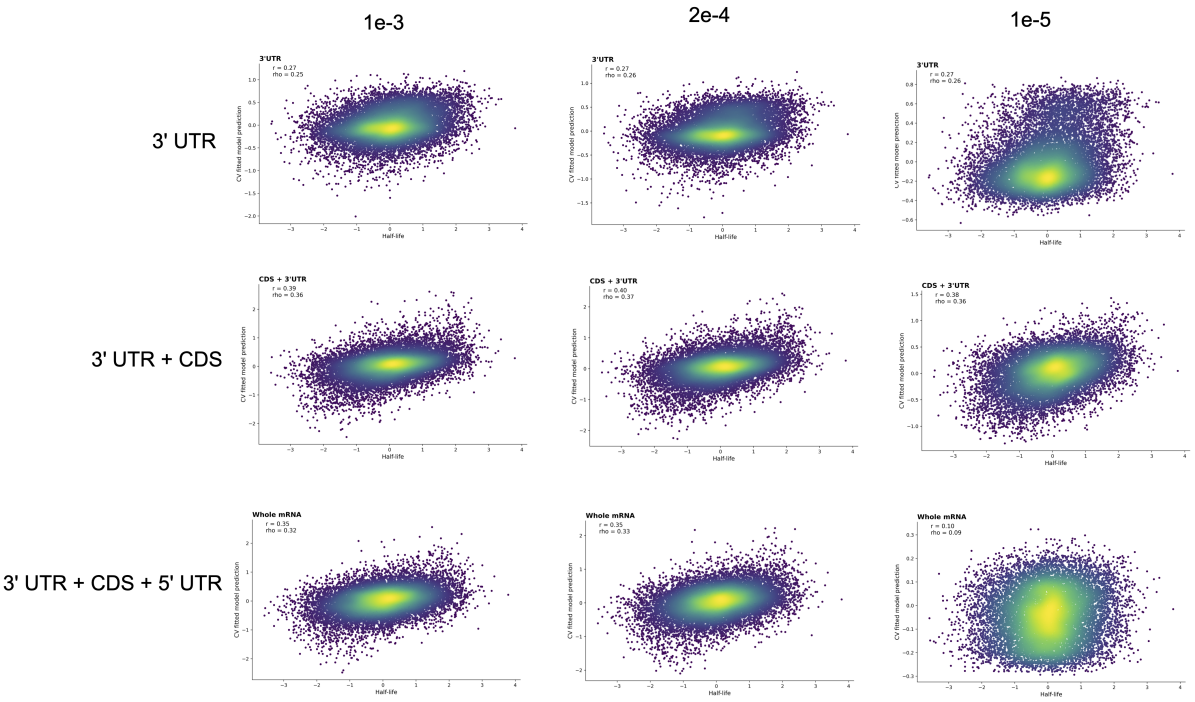

Figure 4: [CLS] Token Embedding on three Sequence Inputs across three learning rates

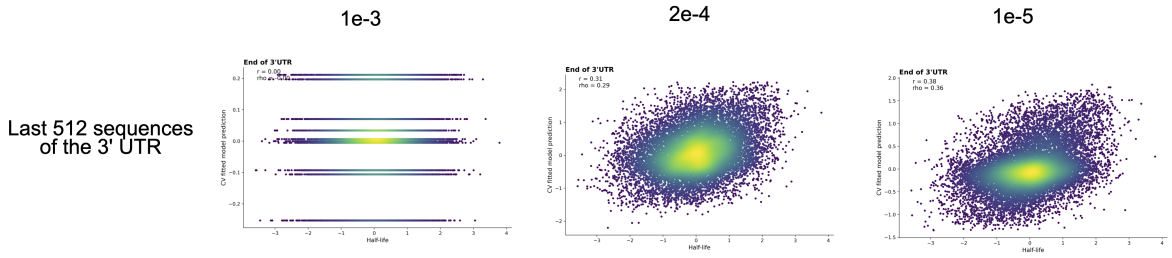

Figure 5: 512 sequences from end of 3' UTR across three different learning rates