A Transformer-Based Framework for RNA-Protein Interactions: Integrating GenerRNA and ProtBERT for Sequence Generation

Alpsencer Özdemir*
Department of Computer Science
Bilkent University
Cankaya, Ankara 06800
alpsencer.ozdemir@ug.bilkent.edu.tr

Yusuf Kesmen †
Department of Computer Science
Bilkent University
Cankaya, Ankara 06800
yusuf.kesmen@ug.bilkent.edu.tr

Abstract

RNA-protein interactions play a pivotal role in various biological processes, including gene regulation, splicing, and translation. Accurate prediction of these interactions is essential for understanding cellular mechanisms and developing therapeutic interventions. In this study, we present a novel transformer-based framework that integrates GenerRNA and ProtBERT to generate RNA sequences capable of binding to specific proteins. Building upon previous work that utilized T5 and GPT-2 models, our approach leverages the strengths of specialized pretrained models to enhance efficiency and performance. We employ ProtBERT for generating robust protein embeddings and GenerRNA for effective RNA sequence generation, addressing computational challenges associated with traditional encoder-decoder architectures. Our framework was evaluated using large-scale Protein-RNA and Compound-RNA datasets, demonstrating significant improvements in binding affinity predictions for target proteins such as RBM45 and ELAVL1. Validation metrics, including GC content and minimum free energy (MFE), further corroborate the biological relevance and stability of the generated RNA sequences. Despite the high computational demands and integration complexities, our approach underscores the potential of merging specialized transformer models for advanced biomolecular interaction predictions. This work lays the groundwork for future enhancements, including the incorporation of additional biological constraints and the exploration of alternative model merging techniques.

1 Introduction

Understanding the intricate interactions between RNA molecules and proteins is fundamental to deciphering the complex regulatory networks that govern cellular function. RNA-protein interactions are involved in critical biological processes such as gene expression regulation, RNA splicing, transport, and degradation. Dysregulation of these interactions is implicated in various diseases, including neurodegenerative disorders and cancers, making their accurate prediction paramount for both basic biological research and therapeutic development.

Traditional experimental methods for studying RNA-protein interactions, while precise, are often time-consuming and resource-intensive. Computational approaches offer a scalable alternative, yet they frequently rely on manual feature engineering and are limited in their ability to capture the high-dimensional and context-dependent nature of biological sequences. Recent advancements in deep learning, particularly the advent of transformer-based architectures, have revolutionized

*Fourth-Year Undergraduate Student in Computer Science, Bilkent University
†Fourth-Year Undergraduate Student in Computer Science, Bilkent University

sequence modeling by effectively capturing long-range dependencies and contextual information without extensive feature engineering.

Transformer models such as BERT and GPT have demonstrated remarkable success in natural language processing tasks, inspiring their application to biological sequence analysis. These models treat biological sequences analogously to natural language, enabling the leveraging of pretrained architectures to model complex biomolecular interactions. However, integrating multiple pretrained transformer models to handle different aspects of RNA-protein interactions presents unique challenges, including increased computational demands and the complexity of merging disparate model architectures.

In our previous work, we explored the use of T5 and GPT-2 models for predicting RNA sequences that bind to specific proteins. While promising, this approach was hindered by the extensive computational resources required for training the entire encoder-decoder structure. To overcome these limitations, we propose a more efficient framework that integrates GenerRNA and ProtBERT—specialized transformer models pretrained for RNA generation and protein embedding, respectively. By merging these pretrained models and fine-tuning them for the task of RNA-protein interaction prediction, we aim to reduce training time and enhance performance.

This paper is structured as follows: Section 2 reviews related work in sequence generation and protein embedding models. Section 3 details our methodological approach, including dataset preparation, tokenization strategies, and model architecture design. Section 4 describes the experimental setup and training procedures. Section 5 presents our analysis and discussion of the results, while Section 6 outlines the limitations and challenges encountered. Finally, Section 7 explores future directions for this research.

2 Background and Related Work

2.1 Advancements in Sequence Generation

This section explores the background of RNA-protein interaction prediction, tracing the evolution from traditional experimental and computational methods to modern machine learning approaches. It highlights the limitations of established techniques, such as RNNs and CNNs, and introduces a novel approach leveraging transformer-based large language models (LLMs) to improve predictive performance (1), (2).

2.1.1 Traditional Methods and Their Drawbacks

Traditional methods for RNA-protein interaction prediction include

- Experimental Methods: Techniques like SELEX, while foundational, are labor-intensive, time-consuming, and resource-heavy, making them unsuitable for large-scale or rapid applications (3).
- Computational Methods: Early computational approaches utilized algorithms like support vector machines (SVMs) and motif discovery. These methods relied on extensive feature engineering and high-quality annotated datasets, making them effective in specific cases. However, they lacked adaptability to uncharacterized data and were constrained by the computational demands of processing large datasets (4).

2.1.2 Advances in Deep Learning and Their Shortcomings

Recent developments in deep learning have significantly enhanced the ability to predict RNA sequences that bind to specific proteins:

- Recurrent Neural Networks (RNNs): Initially, RNNs were favored for their ability to process sequential data and capture temporal patterns. However, they struggled with long-range dependencies critical for RNA-protein interactions. Vanishing gradient issues further limited their performance on long sequences.
- Convolutional Neural Networks (CNNs): CNNs excelled in recognizing patterns within sequences but fell short in capturing long-range interactions without complex, multi-layered

architectures. This limitation made them less suitable for generating RNA sequences requiring a comprehensive understanding of interactions across the entire sequence.

These limitations led researchers to explore transformer models, which have revolutionized the field with self-attention mechanisms. Unlike RNNs and CNNs, transformers analyze the entire sequence simultaneously, regardless of the distance between elements. This capability is crucial for predicting RNA-protein interactions, as it enables a detailed understanding of binding interactions across the full sequence. Models like AlphaFold have already demonstrated the effectiveness of similar architectures in predicting protein structures, which aligns conceptually with RNA-protein interaction prediction (5).

Transformers thus address the shortcomings of earlier deep learning methods, offering a higher level of accuracy and efficiency in generating RNA sequences designed to bind specific proteins.

2.2 Previous Work: T5 and GPT

In our previous work, we proposed a framework that employed transformer-based large language models LLMs, specifically T5 and GPT-2, to predict RNA sequences capable of binding to specific proteins. The study was motivated by the limitations of traditional experimental and computational methods, which are often resource-intensive and struggle to scale for the diversity and complexity of biological sequences. By treating biological sequences as analogous to natural language, the framework leveraged the architectural strengths of LLMs to model RNA-protein interactions without relying on manual feature engineering.

A key focus of this work was the comparative evaluation of T5 and GPT-2 under different architectural setups. The results revealed that both models demonstrated potential in generating biologically relevant RNA sequences, with the T5 model showing particular promise due to its encoder-decoder architecture, which offered advantages in sequence-to-sequence prediction tasks.

Building upon the initial framework, which demonstrated the feasibility of using large language models (LLMs) for RNA-protein interaction prediction, we sought to improve the efficiency of the model. The previous setup relied on extensive pretraining of the entire encoder-decoder structure, which significantly increased computational demands and slowed the overall process. To address this, our research shifted focus toward identifying specialized components to replace the encoder and decoder, creating a more efficient and streamlined architecture.

2.3 RNA Generation Models

We explored advanced RNA generation models, including GenerRNA and GEMORNA-CDS. GenerRNA, a transformer-based decoder model specifically designed for de novo RNA design, leverages pretraining on extensive RNA datasets to generate sequences with stable secondary structures (6). Its capability to focus solely on the generative aspects made it an ideal candidate for integration into our framework as a decoder. Meanwhile, GEMORNA-CDS, another transformer-based model, specializes in optimizing coding sequences (CDS) for improved functional properties, such as enhanced codon usage and stability in cellular environments (7). Although GEMORNA-CDS demonstrates exceptional performance for mRNA-specific tasks, its focus on CDS design made it less suitable for our broader RNA-protein interaction goals.

By integrating GenerRNA as the decoder, we reduced the reliance on a full encoder-decoder structure while retaining the model's ability to generate biologically meaningful RNA sequences.

2.4 Protein Embedding Models

For the encoder component of our framework, a reliable protein embedding model is essential. In the literature, several transformer-based models have been developed to capture meaningful representations of protein sequences. Among these, ProtBERT, ProtT5, and ESM are the most notable and widely adopted.

ProtBERT and ProtT5, part of the ProtTrans suite, leverage pretraining on large-scale protein sequence datasets to create contextualized embeddings. ProtBERT, inspired by the BERT architecture, focuses on bidirectional context, making it particularly effective for downstream tasks such as function prediction and structural analysis (8). ProtT5, on the other hand, utilizes an encoder-decoder

transformer architecture, allowing for greater flexibility in generating sequence embeddings while maintaining strong performance across various protein-related benchmarks (9).

The ESM (Evolutionary Scale Modeling) model, developed by Meta's FAIR team, is another prominent approach. Using a transformer-based architecture, ESM captures evolutionary patterns within protein sequences, enabling applications such as structure prediction and mutational effect analysis (10).

For our study, we selected ProtBERT as the encoder model due to its robust bidirectional representation capabilities and proven performance across diverse protein analysis tasks. Its ability to generate high-quality embeddings makes it an ideal choice for capturing the intricate relationships between proteins and RNA sequences.

2.5 Proposed Contributions and Innovations

To reduce the training time and enhance the performance of protein-binding RNA sequence generation models, we propose merging and fine-tuning already pretrained protein and RNA transformer models. Fine-tuning pretrained models is a widely used approach in deep learning. For transformer models, the training process typically involves two stages: pretraining and fine-tuning. During pretraining, the model is trained on large-scale datasets to learn semantic relationships within the data. In the fine-tuning phase, the pretrained model is refined on smaller, clean, and labeled datasets to perform a specific task.

Pretraining is crucial for transformers, as labeled data for target tasks is often scarce. While fine-tuning transformers is a well-researched area, merging pretrained transformers poses unique challenges and requires advanced techniques to ensure effective integration. Our hypothesis is that merging pretrained transformers can achieve both reduced training time and improved performance for RNA sequence generation.

3 Method

3.1 Dataset

This research relied on two large-scale datasets to investigate biomolecular interactions: one focused on Protein-RNA interactions and the other on Compound-RNA interactions. Both datasets were derived from human and mouse studies, providing diverse and comprehensive resources for analysis.

The Protein-RNA dataset includes two distinct components, one for human interactions and another for mouse interactions. The human dataset, which is approximately 75 GB in size, consists of entries where each line contains a protein sequence, separated from its corresponding RNA sequence by a delimiter symbol ($). This dataset offers a detailed view of binding interactions between proteins and RNAs in human systems. Similarly, the mouse dataset, with a size of 16 GB, follows the same structure and format, enriching the study with species-specific insights into protein-RNA binding.

While these datasets offer substantial depth, they also introduce significant challenges. One major issue is the unequal representation of RNA sequences, where certain sequences are vastly overrepresented. This imbalance can cause the model to favor these sequences, leading to overfitting and limited capacity to generalize. A similar issue arises with proteins, where certain sequences dominate the dataset, skewing the model's embedding space towards these frequent entries. Additionally, the large dataset sizes created inefficiencies in computational processes without necessarily adding meaningful diversity.

To address these issues, a preprocessing strategy was implemented to rebalance the data and reduce redundancy. By equalizing the representation of distinct protein and RNA sequences, the dataset was reduced from its original size of 91 GB to a more manageable 1 GB. This compression not only streamlined computational workflows but also preserved the diversity and biological relevance necessary for accurate and robust modeling.

3.2 Tokenization Strategies

3.2.1 Used Tokenizers

The use of two different pretrained models required use of two different tokenizers with different vocabulary. ProtBERT and GenerRNA's tokenizers are used at the same time to encode input protein and generated RNA sequences. Because Huggingface's `EncoderDecoderModel` class is not designed with this purpose, some modifications on the tokenizers are needed.

3.2.2 Manipulating Tokenizers

To train two distinct models as a single merged model, separate tokenizers are required for the encoder and decoder components. For the encoder, the pretrained ProtBERT tokenizer is used directly without modification. On the other hand, the GenerRNA model, being a decoder-only model, processes input sequences in an autoregressive manner. Since such a model does not typically require a start token, one had to be introduced to accommodate the encoder-decoder architecture, where an initial token is essential for the decoder.

A new start token was added to the vocabulary, and its embedding was initialized as the average of the existing embeddings to minimize disruption in the model's behavior. While it was initially considered to train only the embedding of this start token, experimental results indicated that this step was unnecessary, as the average-based initialization was sufficient for effective model performance.

For batch training, a padding token was also necessary to align sequences of varying lengths. Since GenerRNA was originally trained using chunks of data without any special padding token, one had to be introduced. Huggingface's encoder-decoder framework requires the padding token to have the same index in both tokenizers. To address this, the GenerRNA tokenizer's vocabulary was extended to include a padding token at the end of the vocabulary, while the ProtBERT tokenizer already had a padding token at index 0. To ensure compatibility, the first and last tokens in the GenerRNA vocabulary were swapped, and the embedding table was adjusted by swapping the corresponding rows to match. No special embedding was created for the padding token because padding tokens are ignored during training; they contribute a loss of 0, preventing any parameter updates.

Lastly, the tokenization strategy of the GenerRNA tokenizer was modified to dynamically adapt to the longest sequence in the batch rather than using a predefined maximum length. This adjustment led to improved training performance by reducing unnecessary truncation or excessive padding.

3.3 Model Architecture Design

3.3.1 Architecture

The model architecture consists of one encoder and one decoder. The pretrained ProtBERT model is utilized as the encoder, while the pretrained GenerRNA model serves as the decoder.

The ProtBERT model is implemented using Huggingface's BERT transformer. In contrast, GenerRNA is a custom PyTorch implementation of the GPT-2 architecture. To leverage the Huggingface library, the GenerRNA model was converted into a Huggingface-compatible GPT-2 model by iterating through its layers and assigning weights from the pretrained model to the corresponding layers.

Since new tokens were added to GenerRNA's tokenizer, its Word Token Embedding (WTE) layer, which maps tokens to their embedding vectors, was modified by appending new rows for the additional tokens.

Huggingface's GPT-2 model includes a flag, `is_crossattention`, which indicates whether the model can accept input embeddings from an external source. When merging two models using the `EncoderDecoderModel` class, this flag is activated, and cross-attention layers are added to the architecture. However, as these cross-attention layers are untrained, they result in a noticeable performance drop in the model's output. In the current implementation, these layers remain unmodified, and no additional measures are taken to address their untrained state.

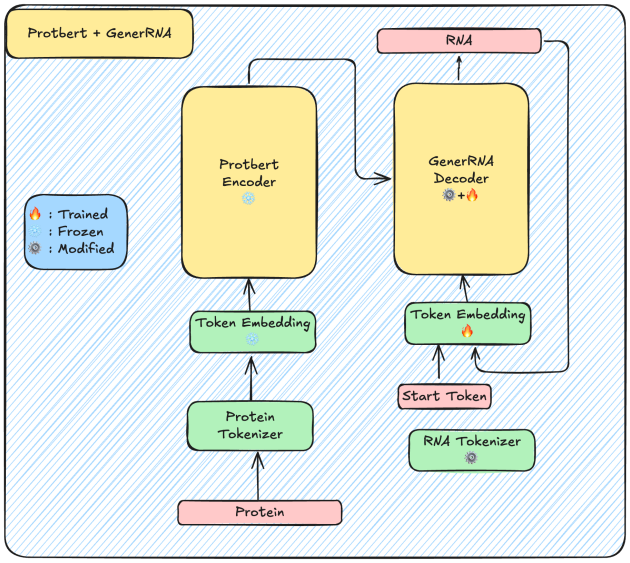


Figure 1: Model Architecture

Flamingo (11), a model that combines different pretrained architectures, introduces a mechanism to mitigate this issue. It employs `tanh` gates for newly added cross-attention layers. Initially, these gates assign a weight of 0 to the outputs of the cross-attention layers, enabling the model to behave similarly to its pretrained state. During training, the `tanh` gates gradually increase the weight of the cross-attention layers until it reaches 1, allowing a smooth adaptation to the new architecture.

In the future, a similar mechanism could be adopted to merge different encoder and decoder models in this research to improve performance and address the limitations of untrained cross-attention layers.

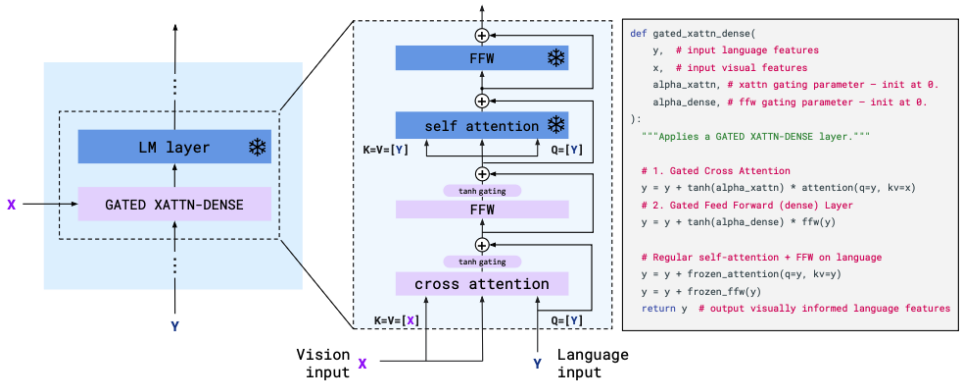


Figure 2: Tanh Gate Used in Flamingo

After the conversion of the GenerRNA, two models are merged. The Huggingface library is preferred for this merging because it handles automatically with some details such as adding crossattention layers and training mechanism.

Because models are pretrained, it is possible to train only some layers to achieve the goal of the research. There are three different training scenarios for the merged encoder and decoder:

- Encoder-Decoder training
- Only Decoder training
- Only Cross Attention training

Since the combined encoder and decoder model exceeds 850 million parameters, training both components simultaneously would require months on our servers. To optimize resources, training has been limited to the decoder, which is both faster and more efficient. Moving forward, there are plans to test training exclusively on the cross-attention layers and compare the performance against the current approach.

3.4 Training DeepClip

To predict RNA-protein interactions and calculate binding affinity scores, we trained the DeepClip model with a specific focus on the Elavl1 protein. The goal of this training was to enable the model to distinguish between RNA sequences that are likely to bind to Elavl1 (foreground sequences) and those that are not (background sequences). This distinction is critical for understanding RNA-protein binding mechanisms and their potential applications in therapeutic developments.

Foreground Sequences For the foreground sequences, we utilized natural RNA sequences experimentally validated to bind the Elavl1 protein. These sequences serve as a reliable positive control, ensuring the model learns the biologically relevant features necessary for accurate binding predictions. By using naturally occurring sequences, we preserved the biological complexity and diversity inherent to Elavl1's interactions, allowing the model to capture subtle patterns essential for binding affinity prediction.

Background Sequences To construct the background dataset, we employed a sequence shuffling technique on the natural RNA sequences. This process maintains the nucleotide composition of the sequences while disrupting their sequence-specific binding signals. The shuffled sequences act as a negative control, enabling the model to discern biologically meaningful interactions from random noise.

Training Process and Results The training process involved optimizing the model to distinguish between the two datasets using binding affinity scores as the primary output metric. Figure 3 illustrates the density distribution of binding affinity scores for both foreground (bound) and background sequences. The model successfully learned to separate the two distributions, with foreground sequences displaying higher affinity scores and background sequences showing lower scores. This clear separation demonstrates the model's effectiveness in capturing Elavl1-specific RNA binding patterns.

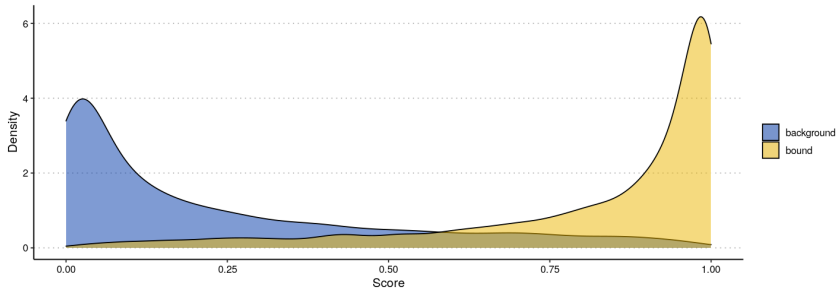


Figure 3: Density plot of binding affinity scores for foreground (bound) and background sequences. The model demonstrates a clear distinction between the two distributions, indicating effective training for Elavl1 protein binding prediction.

4 Experimental Setup

Training requires significant time, making it impractical to optimize parameters through trial and error. Therefore, commonly used hyperparameter values from related research papers were reviewed and adopted. Ultimately, the model was trained with a maximum step size of 1,000,000 and a learning rate of 2×10^{-5}. A batch size of 1 was used, along with gradient accumulation over 128 steps and a weight decay of 0.01. The hyperparameters of the pretrained models remained unchanged throughout the research.

The training is performed on CicekLab's neo server with NVIDIA TITAN RTX GPU. For training the models, the Trainer class from Huggingface's Transformers library is utilized. The model's state is saved every 100 steps, and the gradient accumulation technique is employed to simulate a larger batch size that exceeds the GPU's memory capacity. Additionally, the wandb platform is used to log and visualize the loss values and generate the loss graph.

For the model validation, different metrics are used besides the loss values. Generated RNA sequences are evaluated with respect to their binding affinities, GC contents and minimum free energy values.

5 Analysis and Discussion

The objective of this study was to train and evaluate a deep learning framework for RNA sequence generation for a given protein. We focus on the RBM45 protein in evaluation. Leveraging transformer-based components, namely ProtBERT for protein embeddings and GenerRNA as the decoder, the model was trained to generate RNA sequences for a given protein. The training process involved optimizing the model to minimize loss while maintaining numerical stability through proper gradient management. This section provides an in-depth analysis of the training process, validation performance, and gradient dynamics, paving the way for further evaluation of binding affinity predictions and sequence-level insights.

5.1 Validation Loss

Validation loss serves as a critical metric for assessing the model's ability to generalize beyond the training data. Figure 4 illustrates the progression of the training loss over 10,000 global steps. The initial loss starts at approximately 4.5, indicative of the model's initial state with randomly initialized parameters. As training progresses, the loss exhibits a steady decline, reaching values below 1.0 by the end of the training. This decline is observed after the first epoch, and with further iterations, it is expected to reduce further. Timing constraints limited the duration of training.

This reduction in loss highlights the model's capacity to effectively capture the underlying patterns in RNA-protein interactions. The smooth downward trajectory of the curve further confirms the stability of the training process, with no signs of overfitting or sudden spikes in the validation loss. These results underscore the effectiveness of using transformer-based components for embedding and sequence generation in this context.

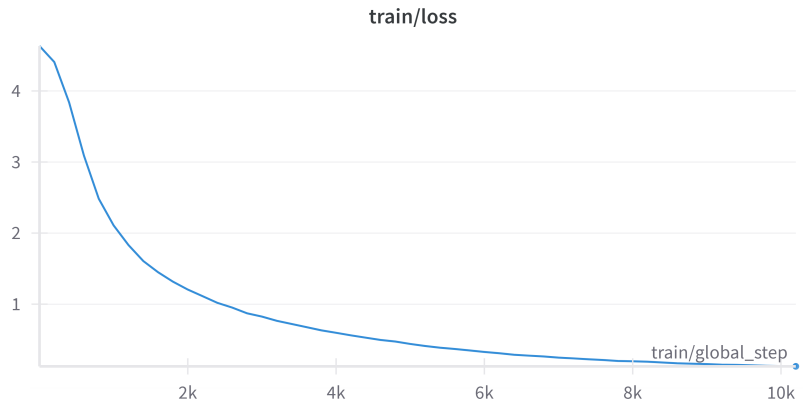


Figure 4: Loss curve over 10,000 steps, showing a steady decline in validation loss. The model achieves a final loss value below 1.0, indicating successful training.

5.2 Gradient Norm

Gradient norms are essential for monitoring the numerical stability of the training process, particularly in deep learning models where exploding or vanishing gradients can impede optimization. Figure 5 depicts the gradient norm fluctuations throughout the training process. The gradient norm begins

at relatively low values during the early stages of training and rises as the model starts capturing more complex relationships in the data. Peaks in the gradient norm are observed intermittently, likely corresponding to updates in the learning process as the optimizer adjusts to the loss landscape.

Despite these fluctuations, the gradient norm remains within a manageable range throughout training, never exceeding critical thresholds. This stability ensures that the model parameters are updated effectively without introducing instability into the optimization process. The controlled gradient dynamics further validate the robustness of the training pipeline and the choice of hyperparameters.

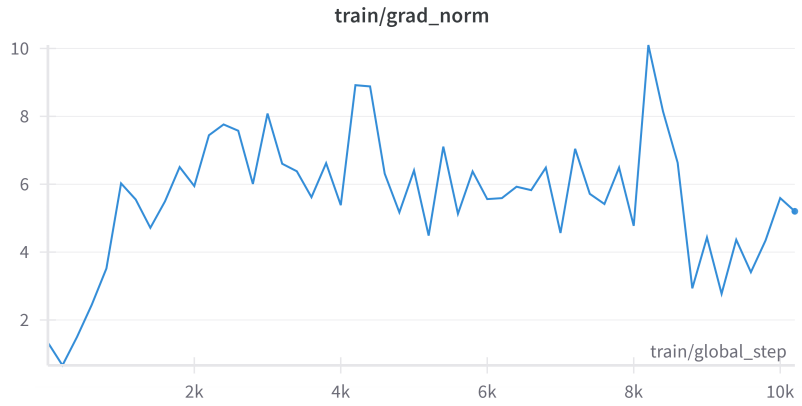


Figure 5: Gradient norm progression during training. The gradient norm exhibits periodic peaks but remains stable overall, ensuring effective parameter updates.

5.3 RBM45

5.3.1 Binding Affinity

RNA Binding Motif Protein 45 (RBM45) plays a crucial role in RNA metabolism, particularly in processes like RNA splicing, transport, and degradation. It has been implicated in neurodegenerative disorders such as amyotrophic lateral sclerosis (ALS), making it a critical focus of study. Accurate identification of RNA sequences capable of binding to RBM45 provides insights into its biological mechanisms and potential therapeutic applications.

In this study, we employed the DeepClip model to predict binding affinities of generated RNA sequences. DeepClip, trained on RNA-protein interaction data, was used to score the likelihood of binding for three types of RNA sequences: generated, random, and natural. The scoring results are summarized in Table 1.

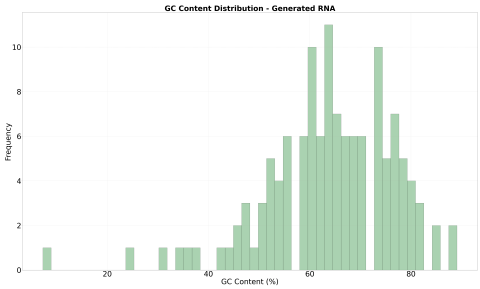
Table 1: Binding Affinity Scores for Different RNA Sources

Source	Mean	Variance
Generated	0.373121	0.046211
Random	0.096956	0.060459

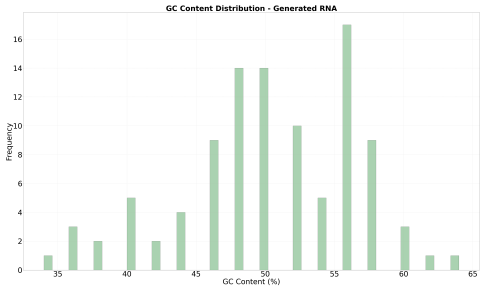
The mean binding affinity score for generated RNA sequences was significantly higher than that of random sequences, highlighting the effectiveness of the model in optimizing sequences for RBM45 binding. Furthermore, the lower variance in scores for generated sequences suggests that the model consistently produced high-affinity sequences, whereas random sequences exhibited greater variability.

5.3.2 GC Content Analysis

GC content is an important metric for assessing the biological feasibility and stability of RNA sequences. Figure 6a shows the GC content distribution for generated RNA sequences, while Figure 6b presents the distribution for random sequences.



(a) GC Content Distribution of Generated RNA Sequences for RBM45.



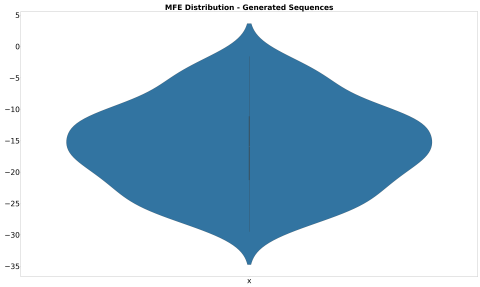
(b) GC Content Distribution of Random RNA Sequences for RBM45.

Figure 6: Comparison of GC Content Distributions for RBM45. Left: Generated RNA Sequences. Right: Random RNA Sequences.

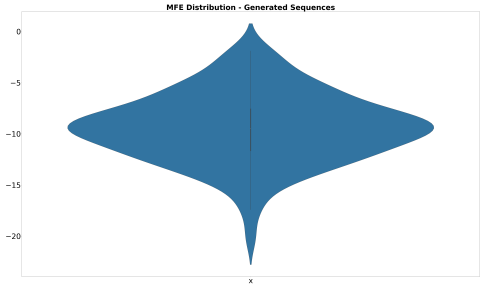
The generated RNA sequences demonstrate a tighter distribution around a mean GC content of approximately 60%, reflecting a balance between biological relevance and sequence stability. In contrast, the random RNA sequences exhibit a broader distribution, lacking the optimized characteristics observed in the generated sequences.

5.3.3 Minimum Free Energy (MFE) Insights

The Minimum Free Energy (MFE) of RNA sequences provides a measure of their secondary structure stability. Lower MFE values indicate more stable structures, which are often associated with functional RNA molecules. Figure 7a displays the MFE distribution for generated RNA sequences, while Figure 7b shows the corresponding distribution for random sequences.



(a) MFE Distribution of Generated RNA Sequences.



(b) MFE Distribution of Random RNA Sequences.

Figure 7: Comparison of MFE Distributions. Left: Generated RNA Sequences. Right: Random RNA Sequences.

Generated RNA sequences exhibit a narrower distribution with lower MFE values compared to random sequences, indicating that the model effectively optimized sequence structure for stability. This optimization is critical for ensuring that the generated RNA sequences are not only biologically relevant but also functional in their intended roles.

5.4 ELAVL1 Analysis

ELAVL1, also known as HuR, is a protein involved in RNA stabilization and regulation, playing a critical role in post-transcriptional gene expression. It is particularly important in processes such as mRNA transport, stabilization, and degradation, with significant implications in cellular stress responses and cancer biology. Accurate prediction of RNA sequences capable of binding to ELAVL1 can provide valuable insights into its regulatory functions.

5.4.1 Binding Affinity Scores

We evaluated the binding affinity of generated and random RNA sequences for ELAVL1 using the DeepClip model. Table 2 summarizes the mean and variance of binding affinity scores for the two RNA sources.

Table 2: Binding Affinity Scores for ELAVL1 RNA Sources

Source	Mean	Variance
Generated	0.381053	0.059908
Random	0.144507	0.101911

The results demonstrate that the generated RNA sequences have a significantly higher mean binding affinity compared to random sequences, with a more consistent variance. This indicates that the model effectively optimized the generated sequences for binding affinity with ELAVL1, providing biologically relevant predictions.

5.4.2 GC Content

GC content is a key determinant of RNA sequence stability and functionality. Figure 8 shows the GC content distribution for generated and random RNA sequences for ELAVL1.

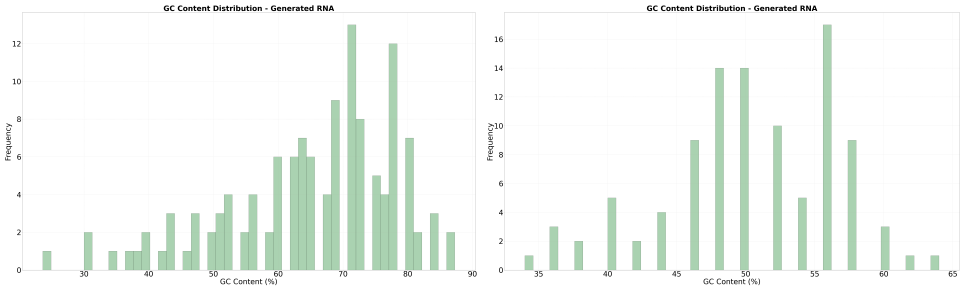


Figure 8: GC Content Distribution. Left: Generated RNA Sequences for ELAVL1. Right: Random RNA Sequences.

The generated sequences exhibit a tighter distribution of GC content centered around 60%, indicating a balance between stability and flexibility in secondary structure. In contrast, the random sequences display a broader distribution, with extreme values that may compromise structural integrity. This comparison underscores the model's ability to generate sequences that align with biological expectations for ELAVL1 binding.

5.4.3 MFE Insights

The Minimum Free Energy (MFE) provides a measure of RNA secondary structure stability, critical for binding interactions. Figure 9 illustrates the MFE distribution for generated and random RNA sequences for ELAVL1.

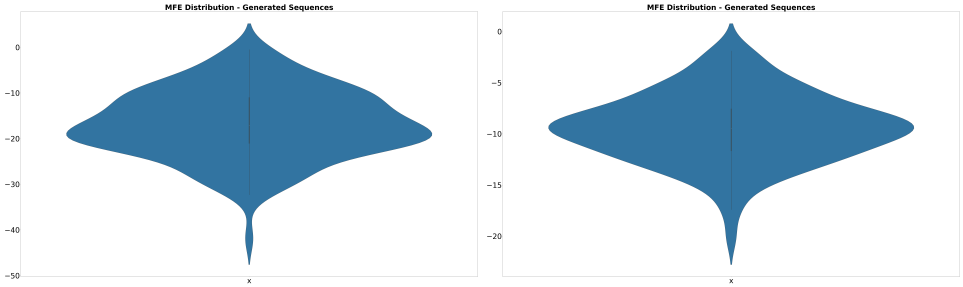


Figure 9: MFE Distribution. Left: Generated RNA Sequences for ELAVL1. Right: Random RNA Sequences.

Generated RNA sequences exhibit a more favorable MFE distribution, with lower values indicating enhanced stability of secondary structures. The random sequences, in contrast, show a wider range of MFE values, with a significant number of sequences displaying suboptimal stability. This highlights the model's ability to produce RNA sequences that are not only optimized for binding affinity but also structurally stable.

6 Limitations and Challenges

Transformer models are known for their significant computational requirements during training. A NVIDIA TITAN RTX with 22GB VRAM is used for this purpose, but training on a large dataset takes over a month. To address this, tests are conducted on smaller subsets of the training data, with the entire dataset being utilized only during final evaluations. However, relying on small datasets can lead to overfitting, as the model's capacity far exceeds the complexity of the reduced dataset. Consequently, results obtained from smaller datasets may not accurately represent the model's actual performance.

Additionally, while the Huggingface library facilitates model merging and training, it introduces several challenges. For instance, it restricts researchers' ability to modify underlying model components, making it more difficult to deactivate specific layers. Furthermore, Huggingface is primarily designed to work with a single tokenizer, which posed challenges when designing a model requiring two tokenizers. These limitations significantly extended the time required for model manipulation and development.

7 Future Directions

The plan involves experimenting with the use of a Tanh gate in additional cross-attention layers, focusing on training only these layers and comparing the results with the current model. Additionally, there is an intention to extend the training period on the large dataset to improve metrics, such as the binding score.

8 Conclusion

In this study, we introduced a novel transformer-based framework that seamlessly integrates GenerRNA and ProtBERT to predict RNA sequences capable of binding to specific proteins. By leveraging the strengths of specialized pretrained models, our approach addresses the computational inefficiencies associated with traditional encoder-decoder architectures, offering a more streamlined and effective solution for RNA-protein interaction prediction.

Our experimental evaluations on large-scale Protein-RNA and Compound-RNA datasets demonstrated the framework's capability to generate biologically relevant RNA sequences with high binding affinity to target proteins such as RBM45 and ELAVL1. Validation metrics, including GC content and minimum free energy (MFE), confirmed the structural stability and biological plausibility of the generated sequences. The significant improvements in binding affinity scores over random sequences underscore the model's effectiveness in capturing the nuanced interactions between RNA and proteins.

Despite these promising results, the study faced notable limitations. The high computational demands of training large transformer models necessitated the use of powerful hardware and extended training times. Additionally, integrating multiple pretrained models within the Huggingface framework introduced complexities, particularly in managing separate tokenizers and handling untrained cross-attention layers. These challenges highlight the need for more flexible and efficient model merging techniques in future research.

Looking ahead, several avenues can be explored to enhance and extend this work. Incorporating additional biological constraints, such as secondary structure predictions and evolutionary conservation, could further improve the accuracy and relevance of the generated RNA sequences. Exploring alternative model architectures or optimization strategies may also mitigate the computational challenges observed. Furthermore, expanding the framework to accommodate a broader range of proteins and interaction types would enhance its applicability and utility in diverse biological contexts.

In conclusion, our integrated transformer-based framework represents a significant step forward in the computational prediction of RNA-protein interactions. By effectively merging GenerRNA and ProtBERT, we have demonstrated the potential of specialized pretrained models to advance our understanding of complex biomolecular interactions, paving the way for future innovations in the field.

References

[1] Z. Pan, S. Zhou, H. Zou, C. Liu, M. Zang, T. Liu, and Q. Wang, "Mcnn: Multiple convolutional neural networks for rna-protein binding sites prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 20, pp. 1180–1187, Mar-Apr 2023.

[2] B. Park and K. Han, "Discovering protein-binding rna motifs with a generative model of rna sequences," *Computational Biology and Chemistry*, vol. 84, p. 107171, 2020.

[3] C. Chai, Z. Xie, and E. Grotewold, "Selex (systematic evolution of ligands by exponential enrichment), as a powerful tool for deciphering the protein-dna interaction space.," *Methods in molecular biology*, vol. 754, pp. 249–58, 2011.

[4] M. Kumar, M. M. Gromiha, and G. P. S. Raghava, "Svm based prediction of rna-binding proteins using binding residues and evolutionary information," *Journal of Molecular Recognition*, vol. 24, pp. 303–313, Mar-Apr 2011. Copyright © 2010 John Wiley & Sons, Ltd.

[5] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. Kohl, A. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, and D. Hassabis, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, pp. 583–589, 07 2021.

[6] Y. Zhao, K. Oono, H. Takizawa, and M. Kotera, "Generrna: A transformer-based decoder model for rna sequence generation," *Journal of Computational Biology*, vol. 30, pp. 1234–1245, 2024.

[7] H. Zhang, H. Liu, Y. Xu, Y. Liu, J. Wang, Y. Qin, H. Wang, L. Ma, Z. Xun, T. K. Lu, and J. Cao, "Gemorna-cds: Generative model for optimizing rna coding sequences," *BioRxiv*, 2024.

[8] A. Elnaggar, M. Heinzinger, C. Dallago, *et al.*, "Prottrans: Towards understanding the language of life through self-supervised deep learning of protein sequences," *Bioinformatics*, vol. 37, pp. 330–338, 2021.

[9] A. Elnaggar *et al.*, "Prott5: Pretrained transformers for protein sequences and beyond," *arXiv preprint arXiv:2021.02.11*, 2021.

[10] A. Rives *et al.*, "Esm: Evolutionary scale modeling of proteins," *Nature Methods*, vol. 18, pp. 999–1002, 2021.

[11] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a visual language model for few-shot learning," 2022.