Cross-Modality Sequence Generation for Compound-RNA and Protein-RNA with Multilingual T5

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

Abstract

The intricate interactions between RNA molecules and their binding partners, such as proteins and small compounds, play a pivotal role in numerous biological processes. Accurate prediction and generation of RNA sequences that can effectively bind to specific proteins or compounds are essential for advancements in biotechnology and therapeutic development. This study introduces a novel approach leveraging a multilingual T5 transformer model to facilitate cross-modality sequence generation for compound-RNA and protein-RNA interactions. By integrating diverse biological data types—proteins, RNA sequences, and SMILES representations of compounds—into a unified embedding space, our model captures the complex dependencies inherent in these interactions. We address significant challenges, including data imbalance and modality-specific tokenization, through strategic oversampling and specialized tokenization techniques. Experimental results demonstrate the model's capability to generate biologically plausible RNA sequences with high binding affinity and thermodynamic stability, for given protein or compound types. This work underscores the potential of transformer-based models in multimodal biological sequence generation, paving the way for enhanced understanding and manipulation of RNA-related biological systems.

1 Introduction

RNA molecules are fundamental components in various cellular processes, including gene regulation, protein synthesis, and catalysis of biochemical reactions. Understanding the interactions between RNA and its binding partners, such as proteins and small compounds, is crucial for elucidating biological mechanisms and developing targeted therapeutics. Traditional experimental methods for characterizing RNA interactions are often time-consuming and resource-intensive, underscoring the need for computational approaches that can efficiently predict and generate RNA sequences with desired binding properties.

Recent advancements in machine learning, particularly in transformer-based architectures, have revolutionized sequence modeling tasks across diverse domains. Models like T5 (Text-to-Text Transfer Transformer) have demonstrated remarkable versatility in handling text generation, translation, summarization, and more by reframing tasks as unified text-to-text problems. Building on this foundation, our research explores the application of a multilingual T5 model to the domain of biological sequence generation, specifically targeting compound-RNA and protein-RNA interactions.

A significant challenge in this endeavor is the integration of multiple biological modalities—proteins, RNA sequences, and chemical compounds represented as SMILES strings—into a cohesive modeling

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

framework. Each modality possesses unique structural and semantic properties, necessitating tailored tokenization and embedding strategies to capture their intricacies effectively. Moreover, disparities in dataset sizes and the inherent complexity of biological interactions introduce additional hurdles in training robust and generalizable models.

In this study, we propose a cross-modality sequence generation approach that leverages the strengths of the T5 architecture to handle diverse biological data types. By implementing specialized tokenization strategies and addressing data imbalance through oversampling and fine-tuning techniques, our model adeptly generates RNA sequences capable of interacting with specified proteins and compounds. The effectiveness of our approach is validated through comprehensive evaluations, including binding affinity assessments, GC content analysis, and minimum free energy (MFE) evaluations, demonstrating the model's proficiency in producing biologically meaningful and stable RNA sequences.

The remainder of this paper is structured as follows: Section 2 reviews the relevant literature and contextualizes our contributions within existing research. Section 3 details our methodology, encompassing data acquisition, tokenization strategies, and model architecture modifications. Section 4 outlines the experimental setup, while Section 5 presents an in-depth analysis and discussion of our findings. Finally, Section 6 concludes the paper and highlights potential avenues for future research.

2 Background and Related Work

2.1 Advancements in Sequence Generation

This section provides an overview of RNA-protein interaction prediction, outlining its progression from traditional experimental and computational techniques to state-of-the-art machine learning methods. It discusses the limitations of established tools like RNNs and CNNs while introducing a new approach using transformer-based large language models (LLMs) to enhance predictive accuracy [1, 2].

2.1.1 Traditional Methods and Their Challenges

Early approaches to predicting RNA-protein interactions primarily relied on experimental and computational techniques. Experimental methods such as SELEX were instrumental in advancing the understanding of these interactions, providing valuable insights into RNA-binding motifs. However, these techniques demanded considerable time, effort, and resources, making them unsuitable for large-scale or rapid studies [3]. Computational techniques, including support vector machines (SVMs) and motif discovery algorithms, achieved promising results, particularly when applied to high-quality annotated datasets. Despite their effectiveness, these models were constrained by their reliance on extensive feature engineering, limited adaptability to unannotated data, and computational inefficiencies when dealing with larger datasets [4].

2.1.2 Advances in Deep Learning and Their Shortcomings

Recent advancements in deep learning have significantly enhanced the prediction of RNA sequences that bind specific proteins. Among these innovations, recurrent neural networks (RNNs) initially gained prominence due to their ability to process sequential data and capture temporal patterns. However, they struggled with long-range dependencies—an essential aspect of RNA-protein interactions—and suffered from performance degradation caused by vanishing gradients when applied to longer sequences. On the other hand, convolutional neural networks (CNNs) excelled in detecting local patterns within sequences, demonstrating strong performance in localized feature extraction. Nevertheless, their limited capacity to capture long-range dependencies without incorporating complex architectures posed challenges for modeling interactions across entire RNA sequences comprehensively.

2.1.3 Transition to Transformer Models

Transformers have emerged as a game-changer in this field, owing to their self-attention mechanisms. Unlike RNNs and CNNs, transformers can analyze an entire sequence in parallel, regardless of the distance between sequence elements. This capability is crucial for RNA-protein interaction

predictions, enabling a deeper understanding of binding dynamics throughout the sequence. Models like AlphaFold have already validated the effectiveness of similar transformer-based architectures in related tasks like protein structure prediction [5].

By overcoming the limitations of earlier methods, transformers provide a more accurate and efficient framework for designing RNA sequences that specifically bind to target proteins.

2.2 Multimodal Approaches

Multimodal transformer models are designed to process text and images within a single framework. They learn patterns from both data types and produce outputs that address tasks such as image captioning, image-text matching, and classification. In this document, we discuss three well-known examples: CLIP, Flamingo, and GIT.

CLIP [6] aligns text and images by training a text encoder and an image encoder in a paired manner. It uses a large dataset of image-text pairs to learn rich associations between words and visual elements. After training, CLIP can match text to images, rank images based on text queries, and perform zero-shot classification.

Flamingo [7] extends the vision-language approach to various downstream tasks. It takes advantage of transformer-based blocks that fuse image and text signals, which helps it adapt to different scenarios, such as caption generation and visual question answering. It uses frozen a contrastive image encoder and a text decoder. It reshapes the image embeddings using perceiver resampler [8]. The image embeddings are fed to cross-attention blocks.

GIT [9] attempts to scale up multimodal learning by increasing the size of models and datasets. It employs attention mechanisms to combine textual and visual data, allowing it to tackle tasks that rely on understanding both images and text. Its image encoder is pretrained, but its text decoder is started with random weights. Unlike Flamingo, image embeddings are concatenated at the beginning of the text input as image tokens. Thus, the model does not uses any cross-attention layers.

2.3 Biological Multimodal Approaches

In our case, we have different modalities in biology that must orchestrate together to generate meaningful sequences. Therefore, it is crucial to understand how biological data types can be embedded into the same embedding spaces using different modalities.

Recently, multimodal approaches in computational biology have emerged as a transformative paradigm for understanding complex biological processes by integrating diverse data types. These approaches address challenges in modeling interactions between proteins, RNA, DNA, and small molecules by effectively merging them into a unified embedding space through the use of transformer models.

One notable study, ProSmith, introduced a framework that employs a multimodal Transformer Network to process protein amino acid sequences and small molecule representations simultaneously. This design enables effective information exchange between the two modalities, allowing the model to consider structural and functional interactions comprehensively. ProSmith integrates gradient boosting and deep learning predictions to achieve state-of-the-art results in predicting key biological metrics, such as Michaelis constants, enzyme substrate interactions, and protein-drug affinities. This model demonstrates the potential of multimodal approaches in enhancing predictive accuracy for protein-small molecule interactions [10].

Another significant contribution is IsoFormer, which addresses the complex task of modeling the relationships between DNA, RNA, and proteins. By leveraging pre-trained modality-specific encoders, IsoFormer connects these biological sequences to predict differential transcript expression across human tissues accurately. This approach not only outperforms existing methods but also showcases the power of transferring knowledge between modalities. IsoFormer's ability to model multiple RNA transcript isoforms originating from the same gene underlines its utility in understanding gene regulation and expression mechanisms, thereby providing a robust framework for tackling multimodal challenges in genomics [11].

In the context of single-cell biology, scMoFormer has emerged as a pivotal model for analyzing multimodal single-cell data. Traditional methods, which rely on static interaction graphs, often fail

to capture the dynamic interplay between cellular modalities. In contrast, scMoFormer employs transformers to model these interactions in an end-to-end manner, dynamically integrating domain knowledge and downstream task information. This innovative approach has demonstrated superior performance on benchmark datasets and provides critical insights into cellular states and dynamics. [12].

These works are pretty new and collectively highlight the growing significance of multimodal approaches in computational biology. However, there is no ongoing framework that works for sequence generation of different data types.

2.4 Cross-Lingual Frameworks

Our problem can also be modeled as a cross-lingual approach. If we regard RNA, protein, and SMILES representations as different languages, the problem naturally extends to a multilingual paradigm. This perspective is particularly relevant given the advancements in language models and the increasing popularity of multilingual frameworks. However, a critical challenge in multilingual learning is the issue of data imbalance, which has a significant impact on model performance. In this context, several studies have explored the effects of imbalanced datasets and proposed innovative strategies to mitigate their adverse effects.

One notable study by the Idiap Research Institute investigates the effects of language-specific class imbalance in multilingual fine-tuning. The researchers focus on imbalances in label distribution across languages and their impact on performance and latent space representations. They demonstrate that standard fine-tuning approaches exacerbate language separation in latent spaces and encourage reliance on uninformative features. To address this, they propose a modified class weighting approach that calculates class weights separately for each language. This method significantly mitigates the negative effects of imbalance, improving performance and reducing the reliance on language-specific features [13].

Similarly, Google researchers have examined the optimization dynamics of multi-task learning under significant data imbalance in multilingual contexts. They propose a straightforward yet effective strategy: pre-training on high-resource tasks followed by fine-tuning on a mix of high- and low-resource tasks. This approach leverages the advantages of high-resource data to establish a robust model foundation, which is then adapted to low-resource tasks. Their empirical analysis demonstrates consistent improvements over standard static weighting techniques, especially in neural machine translation (NMT) and multilingual language modeling. This work highlights the importance of task order and pre-training strategies in managing dataset imbalances [14].

In our work, we aim to address similar challenges of imbalance within the context of multilingual biological modeling, where RNA, protein, and SMILES data exhibit varying levels of resource availability. Drawing from these studies, we propose tailored strategies to ensure equitable representation across modalities while optimizing the transfer of knowledge between them.

2.5 Proposed Contributions and Innovations

This study introduces several key contributions and innovative approaches that advance the field of biological sequence generation and interaction prediction. Our primary contributions are outlined below:

1. **Multimodal Integration Using a Multilingual T5 Model:** We present a novel framework that leverages the multilingual T5 transformer architecture to seamlessly integrate diverse biological data types—specifically, protein sequences, RNA sequences, and chemical compounds represented as SMILES strings. This integration facilitates the capture of complex dependencies and interactions across different modalities within a unified embedding space.
2. **Specialized Tokenization Strategies:** Recognizing the unique structural and semantic properties of each biological modality, we develop tailored tokenization approaches. For protein and RNA sequences, we employ Byte Pair Encoding (BPE) to efficiently capture subsequence patterns, while for SMILES strings, we utilize WordLevel tokenization to preserve chemical semantics. This specialization ensures accurate representation and effective learning across all data types.

3. **Addressing Data Imbalance through Oversampling and Fine-Tuning:** We tackle the significant data imbalance between the larger protein-RNA and the smaller SMILES-RNA datasets by implementing strategic oversampling techniques and a pretraining-fine-tuning pipeline. The oversampling approach ensures adequate representation of the underrepresented SMILES-RNA data during training, while the fine-tuning phase leverages transfer learning to adapt the model to the specific characteristics of SMILES-RNA interactions.
4. **Architectural Modifications to the T5 Model for Multimodal Learning:** To accommodate the distinct vocabularies and embedding requirements of different biological modalities, we modify the original T5 architecture by implementing separate word token embedding (WTE) tables for the encoder and decoder. This separation prevents semantic misalignment and enhances the model's capacity to handle multimodal data effectively. Additionally, we integrate a merged tokenizer that maintains consistency in shared tokens while preserving modality-specific distinctions.
5. **Open-Source Contributions and Future Extensions:** Our architectural adaptations and tokenization strategies are designed with scalability and extensibility in mind. We plan to release these modifications as open-source enhancements to the Hugging Face T5 framework, facilitating their adoption and further development by the research community. Future work will explore the integration of additional biological modalities and the application of knowledge distillation techniques to enhance model efficiency and generalization.

These contributions provide a powerful tool for RNA sequence generation and interaction prediction that is both versatile and scalable across multiple biological data types.

3 Method

3.1 Data Acquisition

We have two distinct datasets: one representing protein-RNA interactions and the other capturing compound-RNA interactions.

3.1.1 Protein-RNA Interaction

This research utilized two extensive datasets representing Protein-RNA interactions, drawn from both human and mouse datasets. The human dataset comprises a comprehensive collection of binding protein-RNA pairs, totaling 75 GB in size. Each entry in the dataset is formatted as a single line, where an amino acid sequence is followed by a dollar sign ('$'), serving as a delimiter, and then by the corresponding RNA sequence with the potential to bind to the specified protein sequence. Similarly, the mouse dataset, though smaller in size at 16 GB, adheres to the same structural format. Each entry contains an amino acid sequence followed by the delimiter ('$') and its corresponding RNA sequence.

Despite the richness of these datasets, they presented significant challenges in terms of processing and inherent biases. For instance, specific RNA sequences appear disproportionately often in the dataset, leading the model to overfit and frequently generate those RNAs. Similarly, the overrepresentation of certain proteins causes the model to create embeddings that favor those proteins, resulting in a lack of generalizability. To address these issues, we developed a compression strategy to balance the dataset. By equalizing the number of distinct proteins and RNA sequences, we reduced the dataset to approximately 1 GB while maintaining its diversity and preserving biologically meaningful relationships.

3.1.2 Compound-RNA Interaction

The second dataset utilized in this research represents compound-RNA interactions. Unlike the Protein-RNA dataset, this dataset is relatively small, containing approximately 10,000 entries. Each entry is formatted similarly, with a SMILES (Simplified Molecular Input Line Entry System) string representing the compound followed by a dollar sign ('$') delimiter and the corresponding RNA sequence.

The small size of this dataset posed a significant challenge for effective training. In particular, the limited amount of data led to imbalances that hindered the model's ability to generalize. Furthermore,

the lack of robust data augmentation techniques for SMILES compounds exacerbates this issue. While traditional sequence augmentation techniques, such as randomization or shuffling, are widely applicable to natural languages, they are less effective for biological data due to the unique syntactic and semantic constraints of the representation. To overcome data imbalance, we apployed training strategies that will be discussed later.

3.2 Tokenization Strategies

Tokenization is a critical preprocessing step that segments data into smaller, manageable units called tokens, enabling machine learning models, especially those based on transformer architectures, to process and understand complex biological sequences. In our study, we focus on three distinct biological modalities: protein, RNA, and SMILES, requiring separate tokenization strategies tailored to each data type. We employed two primary types of tokenizers: Byte Pair Encoding (BPE) for protein and RNA, and WordLevel tokenization for SMILES.

3.2.1 Protein Tokenization

For protein sequences, we utilized a BPE tokenizer trained with a vocabulary size of 1000, aligning with literature standards. Byte Pair Encoding works by iteratively merging the most frequent pairs of characters or character sequences in the corpus to create new tokens, until the desired vocabulary size is reached. The merge operation can be expressed as:

$$\text{BPE Merge: } t_{\text{new}} = \text{argmax}_{t_i, t_j} f(t_i, t_j)$$

where t_i and t_j are token pairs in the vocabulary, and $f(t_i, t_j)$ represents their frequency in the corpus.

This tokenizer includes special tokens such as the start token (`<protein_start>`), end token (`</s>`), padding token (`<pad>`), and unknown token (`<unk>`). Truncation and padding strategies were implemented to ensure uniform sequence lengths of 1024 tokens, with truncation applied to the right of the sequence and padding added to meet the fixed length. Additionally, a ByteLevel pre-tokenizer was applied, incorporating lowercase normalization and adding prefix spaces for better segmentation. Post-processing was performed to append start and end tokens to the sequences.

3.2.2 RNA Tokenization

RNA sequences were tokenized using a similar BPE strategy with a 1000-token vocabulary. As with protein tokenization, BPE iteratively merges frequent subword units to optimize the representation of sequences. The same truncation and padding strategies were employed to standardize input lengths at 1024 tokens, with padding tokens (`<pad>`), end tokens (`</s>`), and unknown tokens (`<unk>`) added for consistency. A ByteLevel pre-tokenizer and lowercase normalizer were used to ensure reliable segmentation. After tokenization, start tokens (`<rna_start>`) and end tokens (`</s>`) were appended to mark sequence boundaries effectively.

3.2.3 SMILES Tokenization

For SMILES representations, we trained a WordLevel tokenizer, as the structural constraints of SMILES strings make BPE less effective. SMILES strings represent chemical structures as linear sequences where individual characters or groups of characters denote specific atoms, bonds, or functional groups. Unlike natural language, where BPE can efficiently merge frequently occurring subword units to reduce vocabulary size while preserving meaning, SMILES strings require precise segmentation to retain the chemical semantics. Merging tokens in SMILES through BPE can result in invalid or uninterpretable chemical units, disrupting the structural integrity of the representation. Therefore, WordLevel tokenization, which treats each predefined chemical unit as a separate token, is more suitable for capturing the inherent semantics and maintaining the validity of SMILES strings. Unlike BPE, WordLevel tokenization maps entire words or predefined units directly to tokens. The WordLevel tokenization process can be expressed as:

$$\text{WordLevel Token: } t_i = \text{lookup}(w_i)$$

where t_i is the token ID and w_i is the corresponding word or unit in the predefined vocabulary.

In our approach, individual tokens in SMILES represent distinct molecular components. The tokenizer includes unique start (<smiles_start>), end (</s>), padding (<pad>), and unknown (<unk>) tokens. Truncation was applied to a maximum length of 1022 tokens, with padding added to maintain uniform input sizes. A pre-tokenizer split SMILES strings into meaningful chemical units using regex-based rules, isolating functional groups and molecular structures for accurate representation.

3.2.4 Merged Tokenizer Design

To integrate the three modalities into a unified framework, we developed a merged tokenizer capable of handling protein, RNA, and SMILES data. A critical consideration during this merging process was ensuring that both the encoder and decoder utilize consistent token representations. Shared tokens, such as padding (‘<pad>‘), unknown (‘<unk>‘), and end tokens (‘</s>‘), were assigned identical IDs across all tokenizers to ensure uniform learning. However, unique start tokens (‘<protein_start>‘, ‘<rna_start>‘, ‘<smiles_start>‘) were maintained for each modality to distinguish the data type explicitly.

Since Hugging Face tokenizers do not natively support merging BPE and WordLevel tokenizers, we manually manipulated the SMILES tokenizer by assigning its unique tokens ID numbers that continue from the protein tokenizer's last unique token. This strategy allowed seamless integration of SMILES and protein tokenizers without altering their internal structures. By carefully aligning token IDs and maintaining modality-specific distinctions, the merged tokenizer facilitates efficient learning and generalization across all three biological data types.

3.3 Model Architecture Design

Transformers have evolved into a powerful framework for sequence-to-sequence tasks, with diverse architectural variants utilizing either the encoder, decoder, or both. Prominent examples include BERT (encoder-only), GPT (decoder-only), and T5 (encoder-decoder). For this research, we selected T5 due to its ability to handle text-to-text tasks effectively, making it particularly suited for tasks requiring the generation of biologically relevant sequences from structured inputs.

The T5 model, introduced by Google Research, incorporates both encoder and decoder components of the original Transformer architecture. This comprehensive design allows T5 to reframe diverse natural language processing tasks, including translation, summarization, question answering, and text classification, as text-to-text problems, where both input and output are treated as text strings. Our approach leverages T5 to process biological inputs and generate meaningful outputs, aligning perfectly with the requirements of our task. However, to adapt T5 for multilingual biological data, significant modifications were necessary.

3.3.1 T5-Based Framework

The original Transformer architecture, proposed by Vaswani et al., was designed for natural language translation. Unlike traditional sequence-to-sequence models such as RNNs and LSTMs, which process data sequentially, the Transformer architecture exploits parallel computation while utilizing the self-attention mechanism to capture dependencies between tokens, even when they are far apart in a sequence [15]. The attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

where Q is the query matrix, K is the key matrix, V is the value matrix, and d_k is the dimension of the key vectors.

The Transformer consists of encoder and decoder blocks. The encoder maps an input sequence to a continuous representation, while the decoder generates the output sequence from this representation. Each encoder and decoder block contains multi-head self-attention mechanisms and position-wise feed-forward neural networks (FNN). Additionally, decoder blocks incorporate an encoder-decoder attention layer that integrates the encoder's representation with the self-attention output. To retain positional information, positional encodings (either sine-cosine or learnable) are added to the input embeddings. Multi-head attention allows the model to focus on various parts of the sequence simultaneously, capturing complex relationships between tokens.

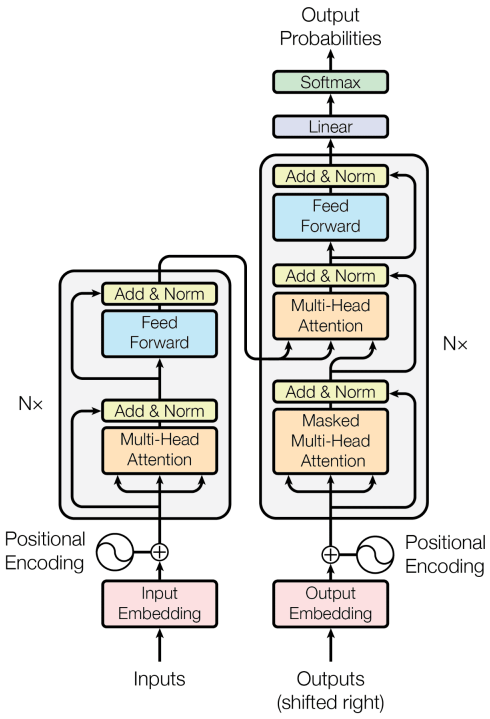


Figure 1: T5 Arcihtecture

Word Token Embedding Separation A significant challenge in adapting the T5 architecture for multilingual biological tasks lies in the shared word token embedding (WTE) table. In the original T5 implementation, the encoder and decoder utilize a single WTE table that is reference-bound, meaning the same token IDs are used across both components. This design assumes a shared semantic space between input and output tokens, which works well in traditional natural language processing tasks where the same tokenizer is used for both the input and output data. However, in our case, the token IDs correspond to distinct modalities—such as protein, RNA, and SMILES—and are not semantically aligned. For instance, a token ID representing a particular amino acid in the encoder could map to a completely unrelated chemical structure in the decoder. This misalignment introduces inconsistencies and hinders the model's ability to effectively process and generate biologically meaningful outputs.

This issue is less apparent in existing multilingual NLP approaches because those tasks typically involve a shared tokenizer that treats all languages as part of the same semantic space. In contrast, our work involves separate tokenizers for each biological modality, each tailored to the unique characteristics of its respective data type. Consequently, the shared WTE table in the original T5 architecture cannot accommodate the distinct vocabularies and embeddings required for our task.

To address this, we modified the T5 architecture to utilize separate WTE tables for the encoder and decoder. Specifically, we implemented unique WTE tables, each with approximately $5000 \times$ embed_size parameters, providing sufficient capacity to represent the vocabularies of all modalities. By severing the connection between the encoder and decoder WTE tables, we eliminated the interference caused by mismatched token IDs.

This architectural adjustment required significant manipulation of the T5 source code to ensure compatibility with our custom tokenizers. Additionally, the implementation maintains the flexibility to handle future expansions, such as adding new modalities or adjusting vocabulary sizes. These changes are part of our ongoing effort to contribute this adaptation as an open-source improvement to the Hugging Face T5 framework. By providing this enhancement, we aim to enable other researchers to utilize T5 for multimodal tasks across diverse domains effectively.

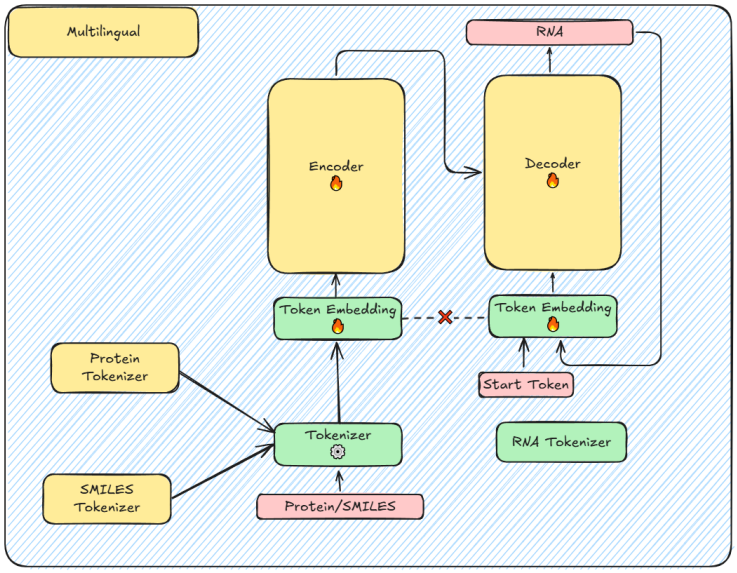


Figure 2: Proposed Changes

Tied Word Embedding Mechanisms In the original T5 architecture, the language modeling (LM) head, which is responsible for mapping the decoder's final hidden states to output probabilities, is inherently bound to the shared WTE table. This shared connection allows the LM head to leverage the same embedding space used by the WTE table, ensuring a consistent transformation between input tokens and output probabilities. However, when we introduced separate WTE tables for the encoder and decoder to accommodate different modalities, this connection was disrupted. As a result, the LM head could no longer rely on the shared WTE table for generating output probabilities.

To address this, we manipulated the architecture to re-establish a functional connection between the LM head and the WTE table, while maintaining the separation of WTE tables across the encoder and decoder. Specifically, we tied the LM head's weights to the newly separated WTE table in the decoder. This ensured that the LM head could still operate within the same embedding space as the decoder, preserving the model's ability to generate accurate output probabilities.

This modification required careful adjustments to the T5 source code to support the independent yet synchronized operation of the LM head and the WTE table. By tying their weights, we maintained consistency between the embeddings used for decoding and the probabilities generated by the LM head. Furthermore, this change optimized the model's computational efficiency by reducing the need for redundant parameter updates during training.

Re-establishing this connection also ensured that the semantic alignment between input tokens and output probabilities remained intact, even with the introduction of modality-specific WTE tables. This adjustment not only preserved the functionality of the LM head but also enhanced the flexibility of the T5 architecture, allowing it to handle diverse data types more effectively. Our modifications align with the broader goal of adapting T5 for multimodal tasks while maintaining its foundational strengths in sequence-to-sequence learning.

3.3.2 Imbalance Mitigation Strategies

Imbalance Mitigation Strategies are critical to address the inherent disparities in our datasets, specifically the protein-RNA dataset and the SMILES-RNA dataset. As described earlier, the protein-RNA dataset is significantly larger than the SMILES-RNA dataset, creating a substantial imbalance. To train a unified model effectively without favoring the larger dataset or neglecting the smaller one, it is essential to align these datasets and mitigate the imbalance.

We considered three strategies for imbalance mitigation and implemented two of them, achieving promising results. These strategies focus on addressing the imbalance either during training or through sequential pretraining and fine-tuning.

Unified Training Approach The unified training approach aims to train both datasets—protein-RNA and SMILES-RNA—together in a single framework, ensuring seamless integration while addressing the inherent imbalance. This strategy is particularly challenging due to the significant disparity in dataset sizes, which can cause the model to prioritize the larger dataset during training, potentially neglecting important features from the smaller dataset. To mitigate this, we implemented two techniques under the unified training approach: weighted loss and oversampling.

One method we employed was assigning a higher loss weight to the underrepresented SMILES-RNA dataset. By increasing the loss contribution of SMILES data during backpropagation, the model becomes more sensitive to this smaller dataset. This ensures that even though the SMILES dataset is smaller, its contribution to the overall optimization is amplified, forcing the model to learn its features more effectively. However, during experimentation, this approach presented challenges. We observed significant gradient norm fluctuations, particularly during backpropagation of SMILES data, which occasionally destabilized the training process. Therefore, we used a different strategy.

The main technique we explored was oversampling the SMILES-RNA dataset to increase its representation during training. This involved probabilistically sampling batches such that, for every 10 protein-RNA batches, one SMILES-RNA batch was included in the training loop. The 10:1 ratio was determined through extensive validation experiments. Ratios lower than 10:1 failed to adequately represent SMILES data, resulting in underfitting, while ratios higher than 10:1 caused the model to overfit to the SMILES dataset, negatively affecting generalization to protein-RNA tasks. This probabilistic sampling strategy effectively mimics the behavior of weighted loss but introduces the data more smoothly into the training process, avoiding sharp gradient changes. By ensuring regular exposure to SMILES data, this technique balanced the training process, allowing the model to capture important features from the smaller dataset without overfitting or destabilizing training.

Both techniques under the unified training approach aim to address the imbalance while maintaining a balance between training stability and performance. These methods highlight the need for careful parameter tuning and experimental validation to achieve optimal results.

Pretraining on Protein, Fine-Tuning on SMILES The second strategy, pretraining on the larger protein-RNA dataset followed by fine-tuning on the smaller SMILES-RNA dataset, leverages the hierarchical nature of transfer learning. This approach aligns with established methods described in the literature, such as Google's work on pretraining and fine-tuning for multilingual tasks. By pretraining on the protein-RNA dataset, the model learns robust, generalizable features from the extensive protein data, which serves as a strong foundation for subsequent adaptation.

During pretraining, the model captures high-level representations of biological sequences, including contextual relationships and patterns specific to the protein-RNA modality. These representations are stored in the model's parameters, creating a stable initialization point for the fine-tuning phase. Fine-tuning on the SMILES-RNA dataset refines these learned representations, adapting them to the unique characteristics of the SMILES data. This sequential approach minimizes the risk of overfitting to the smaller SMILES dataset while leveraging the knowledge gained from the protein-RNA dataset.

One significant advantage of this strategy is the smoother gradient transitions observed during training. The pretraining phase stabilizes the model's parameters, reducing the likelihood of abrupt gradient changes during fine-tuning. This stability is particularly beneficial when dealing with small datasets like SMILES-RNA, as it prevents catastrophic forgetting of the protein-RNA features while allowing the model to focus on the nuances of SMILES data. Furthermore, this method ensures that the model's embeddings remain coherent across modalities, improving its ability to generalize.

Multimodal Integration Framework The third strategy, which we did not implement, involves leveraging a true multimodal framework to integrate protein, RNA, and SMILES data simultaneously. While this approach has been successful in other domains, such as combining voice, image, and text data, it was not applicable to our datasets. In typical multimodal frameworks, data from different modalities corresponds directly to the same instances, allowing for cohesive integration. However, in our case, the RNA data corresponding to protein and SMILES modalities did not share a common source or structure. As a result, implementing a multimodal framework was not feasible within the constraints of our data.

By employing these strategies, we effectively addressed the dataset imbalance while ensuring the model could learn meaningful representations from both the larger protein-RNA dataset and the smaller SMILES-RNA dataset.

4 Experimental Setup

Training takes considerable time, making it unrealistic to optimize parameters through trial and error. Instead, widely used hyperparameter values from related studies were reviewed and adopted. The model was ultimately trained with a maximum step size of 1,000,000 and a learning rate of 2×10^{-5}. The batch size was set to 8, with gradient accumulation over 16 steps, and a weight decay of 0.01.

The training process took place on CicekLab's neo server, equipped with an NVIDIA TITAN RTX GPU. Huggingface's Transformers library and its `Trainer` class were used for model training. The model's state was saved every 100 steps, and gradient accumulation was employed to effectively simulate a larger batch size than the GPU's memory could handle. Additionally, the *wandb* platform was utilized for logging and visualizing the loss values, as well as for generating a loss graph.

For validation, metrics beyond loss values were considered. The generated RNA sequences were evaluated based on their binding affinities, GC content, and minimum free energy values.

5 Analysis and Discussion

In this section, we present an in-depth analysis and discussion of the results obtained from our experiments, focusing on the strategies employed to mitigate data imbalance and improve model performance. The analysis is structured to evaluate the impact of oversampled training, protein pretraining with SMILES fine-tuning, and their comparative effectiveness in addressing challenges associated with training a multimodal model on imbalanced datasets.

Key metrics for evaluation include validation loss, binding affinity scores, GC content analysis, and minimum free energy (MFE) insights. These metrics are particularly critical in assessing the model's ability to generate biologically meaningful sequences and predict accurate interactions between RNA, proteins, and small molecules. Additionally, we analyze the model's performance on specific test cases, such as the phenalenyl cation evaluation or RBM45 evaluation, to understand its capability in handling complex biological structures.

Although our results are much better than random, the time limitation we have does not allow us to train the model for a sufficient duration to achieve its potential. We expect performance to improve further as training progresses. By exploring these metrics and test cases across the employed strategies, this section aims to provide a comprehensive understanding of how the proposed methods contribute to the robustness and generalizability of the model.

5.1 Oversampled Training

Oversampling was a key strategy employed to mitigate data imbalance between the protein-RNA and SMILES-RNA datasets. The primary goal was to ensure that the underrepresented SMILES-RNA data was sufficiently incorporated into the training process, improving the model's ability to generalize across both modalities.

5.1.1 Validation Loss

The loss curve, shown in Figure 3, illustrates the convergence of the model during oversampled training. The loss starts at approximately 8 and exhibits a sharp decline in the initial steps, reaching around 3 by 5,000 steps. Beyond this point, the curve continues to decrease steadily, eventually stabilizing near a value of 2 at approximately 20,000 steps. This indicates that the model successfully learned from the training data and gradually reduced errors.

The oversampling strategy contributed significantly to this behavior. By probabilistically including one SMILES-RNA batch for every ten protein-RNA batches, we ensured consistent exposure to the smaller dataset without overwhelming the training process. This 10:1 ratio, determined through extensive validation, provided a balance between underfitting and overfitting to SMILES-RNA data.

The oversampling approach effectively smoothed the loss trajectory, preventing abrupt fluctuations that could destabilize training.

The steady decline in loss values reflects the model's ability to learn from both datasets simultaneously. Validation results confirm that the oversampled SMILES-RNA data positively influenced the model's performance, particularly in predicting accurate RNA sequences for underrepresented SMILES data. The absence of sharp spikes or plateaus in the loss curve further indicates that the oversampling strategy maintained a stable training process, allowing the model to generalize effectively across modalities.

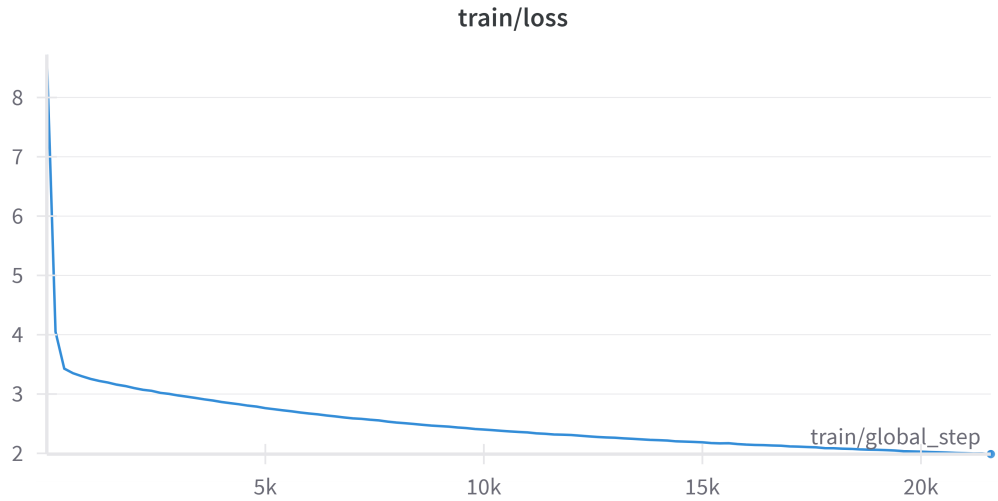


Figure 3: Validation Loss Curve during Oversampled Training. The sharp initial decline followed by gradual convergence indicates effective learning and stability.

The observed loss trajectory also highlights the importance of choosing the right sampling ratio. Lower ratios, such as 20:1, failed to provide sufficient representation for SMILES data, resulting in higher loss values and poorer generalization. Conversely, higher ratios, such as 5:1, caused overfitting to SMILES-RNA data, adversely impacting performance on protein-RNA tasks. The chosen 10:1 ratio represents an optimal trade-off, ensuring that both datasets contributed meaningfully to the learning process.

In summary, the loss analysis demonstrates that oversampling effectively mitigates data imbalance and allows the model to learn robust representations from both modalities. This strategy forms a critical component of our approach to addressing the challenges associated with imbalanced biological datasets.

To evaluate the binding affinity between proteins and RNA sequences, we employed the DeepClip model, a state-of-the-art deep learning framework specifically designed for RNA-protein interaction prediction. DeepClip leverages sequence features and structural patterns to provide accurate binding affinity scores, making it a reliable tool for such analyses [16].

For the binding affinity between SMILES representations of small molecules and RNA sequences, we utilized RSAPred. This method integrates chemical structure information with RNA sequence features to predict interactions, offering a robust approach for exploring small molecule-RNA binding mechanisms [17].

5.1.2 RBM45 Evaluation

RBM45 (RNA Binding Motif Protein 45) plays a crucial role in RNA binding and regulation, making it a key benchmark for evaluating the biological relevance of generated sequences. In this section, we evaluate three primary metrics—Binding Affinity Scores, GC Content Analysis, and Minimum Free Energy (MFE) Insights—by comparing the generated RNA sequences to random sequences.

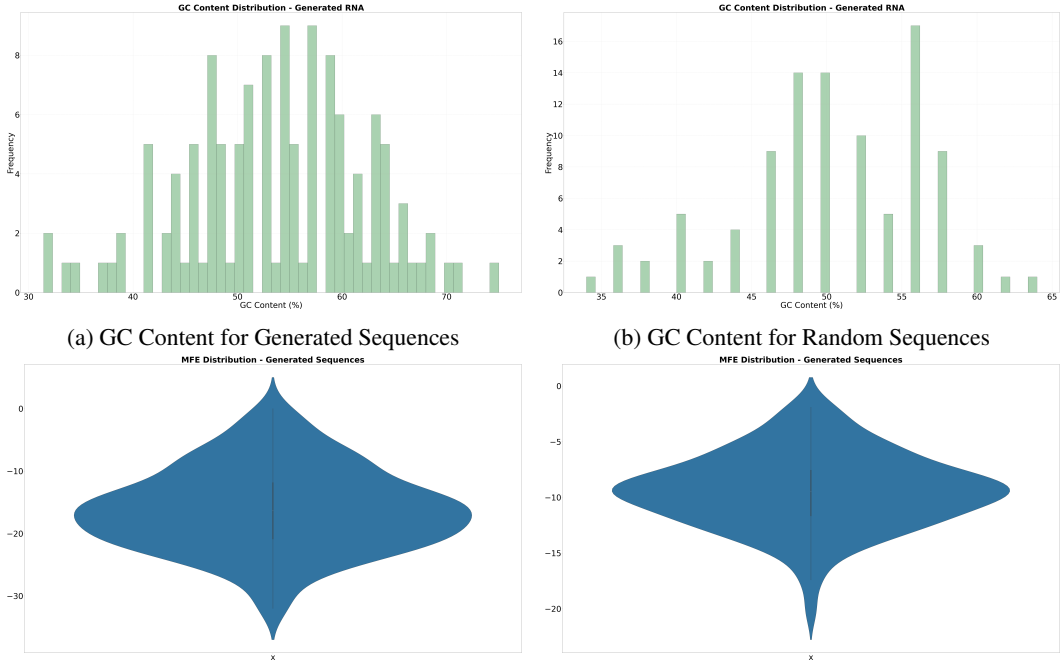
Binding Affinity Scores Binding affinity scores quantify the strength of interaction between RNA sequences and RBM45. Table 4 summarizes the mean and variance of binding affinity scores for generated and random sequences. The binding affinity scores are calculated from DeepClip's webserver.

Source	Mean	Variance
Generated	0.329641	0.114834
Random	0.096956	0.060459

Table 1: Binding Affinity Scores for Generated and Random Sequences. Generated sequences demonstrate significantly higher affinity, indicating strong interactions with RBM45.

The mean binding affinity for generated sequences is substantially higher (0.329641) than for random sequences (0.096956). This result highlights the model's capability to produce RNA sequences with biologically relevant interactions, emphasizing its ability to generalize well to RBM45-related tasks.

GC Content and Minimum Free Energy (MFE) Analysis GC content and MFE are critical metrics for assessing RNA sequences' biological relevance and thermodynamic stability. Figures 4 (a) and (b) show the GC content and MFE distributions for generated sequences, while Figures 4 (c) and (d) present the corresponding distributions for random sequences.



(a) GC Content for Generated Sequences

(b) GC Content for Random Sequences

(c) MFE for Generated Sequences

(d) MFE for Random Sequences

Figure 4: (a) and (b) compare the GC content distributions for generated for RBM45 and random RNA sequences, respectively. Generated sequences exhibit a biologically plausible GC content distribution. (c) and (d) present the MFE distributions, highlighting the greater thermodynamic stability of generated RNA sequences compared to random sequences.

The GC content distribution for generated sequences is centered around 40%-60%, reflecting biologically plausible nucleotide compositions. In contrast, random sequences exhibit a broader, less consistent range, indicating poorer biological plausibility.

Similarly, the MFE distributions reveal that generated sequences consistently achieve lower energy states, suggesting superior thermodynamic stability. Random sequences, however, display higher and more irregular MFE values, reinforcing the generated RNA's potential for reliable molecular interactions.

The evaluation of RBM45 interactions using Binding Affinity Scores, GC Content, and MFE metrics demonstrates the model's ability to generate biologically meaningful RNA sequences. The superior performance of generated sequences over random ones highlights the effectiveness of the training strategy. While the generated RNA sequences exhibit higher stability and plausible biological properties, further improvements can be explored through refined training techniques and additional comparative benchmarks.

5.1.3 Phenalenyl Cation Evaluation

Phenalenyl cation evaluation serves as a critical test case for assessing the model's ability to generate RNA sequences that bind effectively to small molecules. This subsection evaluates the model's performance by analyzing Binding Affinity Scores, GC Content, and Minimum Free Energy (MFE) of the generated sequences compared to random sequences.

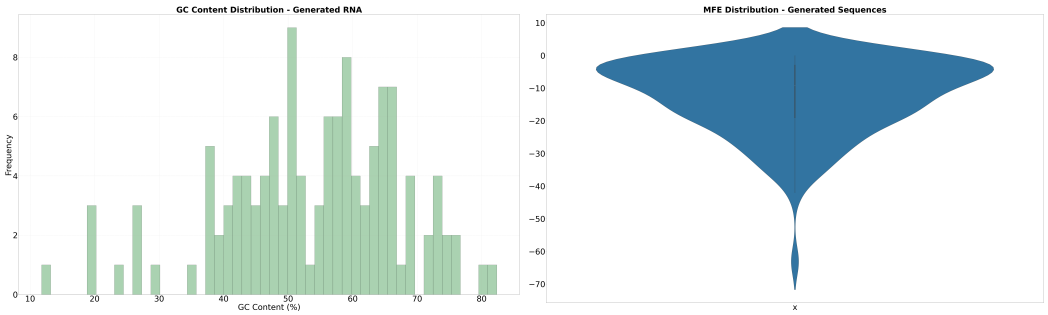
Binding Affinity Scores Binding affinity scores quantify the interaction strength between RNA sequences and phenalenyl cations. Table 2 summarizes the mean and variance of affinity scores for generated and random sequences.

Source	Mean Affinity Score	Variance Affinity Score
Generated	9.341915	8.110194
Random	7.674200	2.052518

Table 2: Binding Affinity Scores for Phenalenyl Cation. Generated sequences exhibit higher affinity and variance, suggesting a better capability to capture relevant molecular interactions.

The generated sequences achieve a higher mean affinity score (9.341915) compared to random sequences (7.674200). This indicates that the model effectively captures the key interaction patterns required for phenalenyl cation binding. The larger variance in generated scores reflects the diversity in the generated RNA sequences, which may indicate the model's ability to explore a wider range of molecular interactions.

GC Content and MFE Analysis GC content and MFE are critical metrics for understanding RNA stability and structural integrity as stated before. Figures 8 (a) and (b) compare the GC content and MFE distributions for generated and random sequences, respectively.



(a) GC Content Distribution (b) MFE Distribution

Figure 5: Comparison of GC Content and MFE Distributions for Generated RNA Sequences. (a) GC content distribution aligns well with biologically plausible ranges, while (b) MFE distribution reveals challenges in maintaining stable secondary structures for SMILES-RNA sequences.

For GC content, the generated sequences exhibit a narrower and more biologically relevant distribution compared to random sequences, with a clear concentration around 60%-70%, while random was around 50% with high variance. This suggests that the model effectively captures the nucleotide composition necessary for RNA stability.

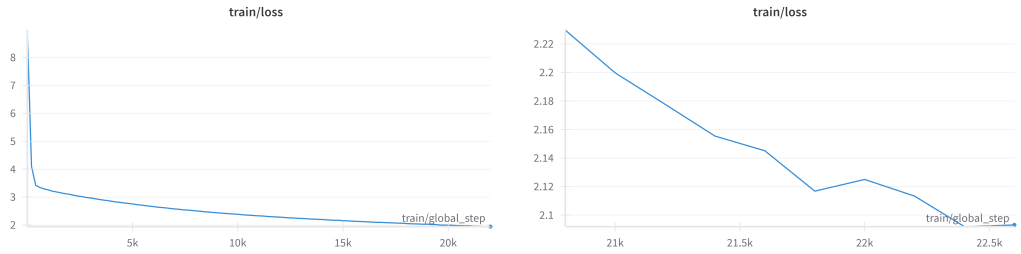
However, the MFE distribution for generated sequences presents a challenge. Contrary to expectations, the generated SMILES-RNA sequences exhibit higher MFE values, indicating less stable

secondary structures. This discrepancy may arise from the limited size and variability of the SMILES-RNA dataset, which could hinder the model's ability to generalize structural stability effectively. Additionally, the higher variance in MFE for SMILES-RNA sequences suggests inconsistencies in how the model captures thermodynamic properties during generation.

5.2 Protein Pretraining and SMILES Fine-Tuning

5.2.1 Validation Loss

Loss serves as a critical metric for understanding the model's learning progression and generalization capabilities. Figure 6 illustrates the training loss curves during both the pretraining phase on the Protein-RNA dataset and the fine-tuning phase on the SMILES-RNA dataset.



(a) Pretraining on Protein-RNA Dataset (b) Fine-Tuning on SMILES-RNA Dataset

Figure 6: Validation Loss Curves during Pretraining and Fine-Tuning Phases. The first graph shows steady convergence during pretraining, while the second highlights smoother loss reduction during fine-tuning.

During the pretraining phase (Figure 6a), the validation loss exhibits a steep decline initially, followed by a gradual reduction, indicating effective learning of general representations from the Protein-RNA dataset. This phase enables the model to capture essential features, such as protein-RNA interaction patterns, which serve as a foundation for fine-tuning.

In contrast, the fine-tuning phase (Figure 6b) on the SMILES-RNA dataset shows a smoother and slower reduction in validation loss. This behavior reflects the smaller dataset size and the transfer of knowledge from the pretraining phase. While the fine-tuning phase does not exhibit dramatic loss reductions, the steady trend indicates effective adaptation to SMILES-specific interactions without significant overfitting.

This combined strategy of pretraining and fine-tuning ensures that the model leverages the larger Protein-RNA dataset for robust feature extraction while adapting to the unique properties of the SMILES-RNA dataset.

5.2.2 RBM45 Evaluation

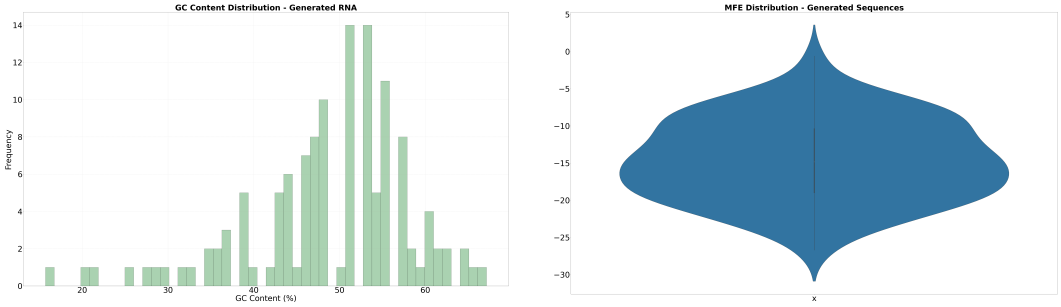
Binding Affinity Scores The binding affinity scores for the generated and random sequences were analyzed. The mean and variance of the scores are shown in Table 3. The generated sequences exhibit a significantly higher mean score (0.397) compared to the random sequences (0.097), indicating the model's ability to generate biologically relevant sequences. Additionally, the variance for the generated scores (0.009) is notably lower than that of the random scores (0.060), reflecting higher consistency among the generated sequences.

Source	Mean Affinity Score	Variance Affinity Score
Generated	0.397	0.009
Random	0.097	0.060

Table 3: Mean and variance of affinity scores for generated and random sequences.

GC Content and MFE Insights The GC content and minimum free energy (MFE) distributions for the generated RNA sequences are illustrated in Figure 7. The GC content demonstrates a peak around

50%, indicating biologically plausible sequences, while the MFE distribution reveals a clustering of sequences with stable negative energy values around -20 which was lower that the random MFE value These two metrics together provide insights into the thermodynamic stability and biological relevance of the generated RNA sequences.



(a) GC Content Distribution

(b) MFE Distribution

Figure 7: Comparison of GC content and MFE distribution for generated RNA sequences. The GC content around 50% correlates with thermodynamic stability indicated by the negative MFE values.

5.2.3 Phenalenyl Cation Evaluation

Phenalenyl cation evaluation is employed as a specialized case to assess the interaction of RNA sequences with small molecules as it is done in the first strategy. This evaluation provides critical insights into the structural and binding properties of the generated RNA sequences. We analyzed three aspects: binding affinity scores, GC content distribution, and minimum free energy (MFE) profiles, as detailed below.

Binding Affinity Scores The binding affinity scores between generated RNA sequences and the phenalenyl cation were compared against random RNA sequences. Table 4 summarizes the results:

Source	Mean Affinity Score	Variance Affinity Score
Generated	11.8950	14.235079
Random	7.6742	2.052518

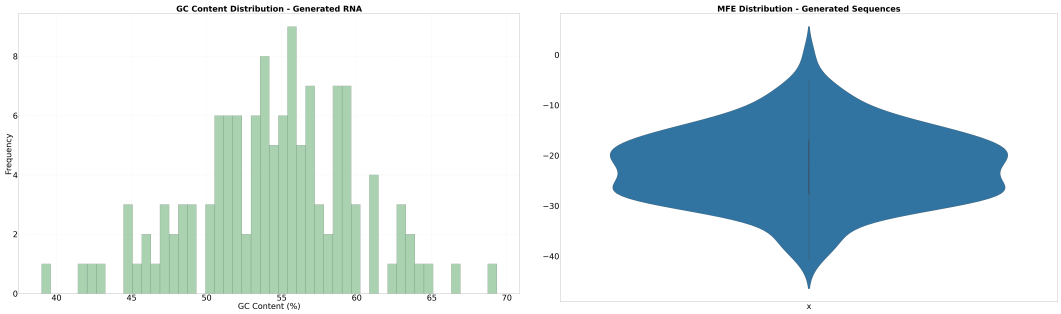
Table 4: Binding Affinity Scores for Phenalenyl Cation Evaluation. The generated sequences exhibit higher affinity and variance compared to random sequences, suggesting enhanced interaction capability.

The generated RNA sequences exhibit a significantly higher mean binding affinity score (11.8950) compared to random sequences (7.6742). This suggests that the generated sequences have a stronger propensity to form stable interactions with the phenalenyl cation. However, the higher variance (14.235079) indicates variability in the interaction strength, which might be influenced by the diversity in the generated sequences.

GC Content and MFE Insights The GC content and MFE distribution were analyzed to evaluate the nucleotide composition and thermodynamic stability of the generated RNA sequences. Figures 8 (a) and (b) present these distributions side by side for comprehensive comparison.

The GC content of the generated RNA sequences is concentrated within the range of 50%-60%, indicating a balanced nucleotide composition that is biologically plausible. Such consistency ensures structural integrity and functional relevance. In contrast, random sequences display more irregular and biologically less plausible GC content distributions.

The MFE distribution highlights that the generated sequences consistently achieve lower energy states compared to random sequences. The score is around -30, much lower than we get in random and in other cases. Lower MFE values indicate greater thermodynamic stability, which is crucial for maintaining the secondary structure of RNA in molecular interactions. The smooth MFE profile of the generated RNA supports its potential application in molecular biology tasks where stable RNA structures are essential.



(a) GC Content Distribution (b) MFE Distribution

Figure 8: (a) GC Content Distribution: The generated RNA sequences exhibit a concentrated GC content range of 50%-60%, which aligns with biologically plausible compositions. (b) MFE Distribution: The consistently lower MFE values of the generated RNA sequences suggest superior thermodynamic stability compared to random sequences.

5.3 Comparative Evaluation of Strategies

The comparative analysis between the oversampling approach and the pretraining-fine-tuning strategy highlights their respective strengths and limitations in addressing data imbalance and optimizing model performance. Both approaches demonstrated notable improvements in generating biologically relevant RNA sequences; however, their effectiveness varied depending on the specific evaluation metrics.

The oversampling strategy effectively addressed the underrepresentation of SMILES-RNA data by ensuring consistent exposure during training. This approach smoothed the validation loss curve and facilitated stable learning across modalities. Oversampling also yielded better binding affinity scores, particularly in the RBM45 evaluation, where generated sequences demonstrated significantly higher affinities compared to random sequences. The GC content and MFE analyses further reinforced the strategy's effectiveness, as generated sequences exhibited biologically plausible distributions and superior thermodynamic stability. However, oversampling introduced challenges such as increased variance in MFE values for SMILES-RNA sequences, indicating that the model sometimes struggled to generalize structural stability in this dataset.

On the other hand, the pretraining-fine-tuning strategy leveraged the large, balanced Protein-RNA dataset to establish robust foundational representations, which were then refined using the smaller SMILES-RNA dataset. This approach excelled in generating RNA sequences with stable secondary structures, as evidenced by consistently low MFE values and precise GC content distributions in both datasets. The smooth loss trajectory during fine-tuning highlighted the transfer of knowledge from the pretraining phase, ensuring better generalization. However, the reliance on sequential training phases made this strategy less adaptive to datasets with high variability, as seen in the phenalenyl cation evaluation, where binding affinity variance was higher than expected.

In summary, the oversampling strategy excelled in addressing immediate data imbalance and improving binding affinity metrics, while the pretraining-fine-tuning strategy provided a robust foundation for generating structurally stable sequences. A hybrid approach that combines the strengths of both strategies may offer a more comprehensive solution, balancing data representation and leveraging transfer learning to handle complex multimodal datasets effectively.

6 Future Directions

Due to time constraints, we were unable to train a Knowlege Distillation model but focused on designing and implementing it to address the problem more effectively. Future research direction includes training the below mentioned Knowledge Distillation model.

6.1 Leveraging Knowledge Distillation Techniques

Knowledge distillation (KD) is a powerful technique in deep learning that transfers knowledge from a larger, well-trained model (teacher) to a smaller, less complex model (student). In our context, KD can serve as an effective method for improving the generalization and efficiency of multimodal models, especially when dealing with diverse datasets such as protein-RNA and SMILES-RNA. By leveraging soft labels generated by the teacher model, the student model can learn richer representations, going beyond hard ground-truth labels. This framework holds promise for aligning diverse modalities while mitigating data imbalance and overfitting issues.

The proposed KD framework, illustrated in Figure 9, involves two primary stages: (1) knowledge transfer from the protein-RNA dataset using a frozen protein encoder and decoder, and (2) compression and adaptation to the SMILES-RNA dataset via a molecule encoder. Below, we elaborate on the key components of this framework:

Teacher Model and Soft Label Generation The teacher model is trained on the protein-RNA dataset, leveraging a frozen protein encoder to extract embeddings that effectively represent protein sequences. These embeddings are then passed to a frozen decoder, which generates token predictions. In addition to the original hard ground-truth labels, the teacher model produces soft labels—a probability distribution over possible tokens. These soft labels provide richer information about the model's confidence in its predictions, encapsulating inter-class relationships that are not captured by hard labels. The soft labels are generated as:

$$y_{\text{soft}} = \text{softmax}\left(\frac{z}{T}\right)$$

where z represents the logits, and T is the temperature parameter that controls the smoothness of the distribution.

The student model, designed for SMILES-RNA data, uses a molecule encoder to generate embeddings that align with the protein embeddings from the teacher model. The molecule encoder is trained to mimic the protein embeddings through KL divergence loss:

$$\mathcal{L}_{\text{mid}} = \text{KL}(p_{\text{teacher}} \parallel p_{\text{student}})$$

This mid-level KD ensures that the student model captures structural and functional relationships similar to the protein-RNA modality.

At the output level, the student model's predictions are aligned with the soft labels produced by the teacher model. The cross-entropy loss between the student model's token predictions and the soft labels ensures that the model captures the nuanced distribution learned by the teacher:

$$\mathcal{L}_{\text{out}} = -\sum_{i} y_{\text{soft},i} \log(y_{\text{student},i})$$

Figure 9 visually outlines the proposed KD framework, including the mid-level KD for embedding alignment and output-level KD for token prediction. The integration of these techniques into future work aims to enhance the robustness and scalability of multimodal models in computational biology.

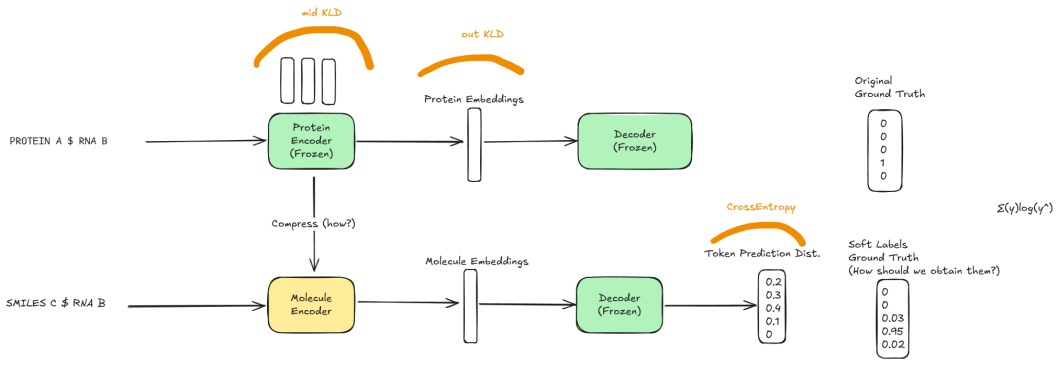


Figure 9: Knowledge Distillation Framework for Protein-RNA and SMILES-RNA Alignment.

7 Conclusion

In this study, we have successfully developed a multilingual T5-based framework for cross-modality sequence generation, targeting compound-RNA and protein-RNA interactions. By integrating diverse biological data types—proteins, RNA sequences, and chemical compounds represented as SMILES strings—into a unified model, we demonstrated the feasibility of generating RNA sequences with high binding affinity and thermodynamic stability. Our approach addressed critical challenges such as modality-specific tokenization and data imbalance through innovative strategies like oversampling and pretraining-fine-tuning pipelines.

The enhanced binding affinity scores in specific evaluations, such as RBM45 and phenalenyl cation interactions, underscore the model's capability to capture intricate biological interactions effectively for both compounds and proteins. The values were not as high as we expected. They could improve with additional training time.

Furthermore, the architectural modifications to the T5 model, including the separation of word token embeddings for different modalities and the incorporation of a merged tokenizer, have paved the way for more flexible and scalable multimodal biological modeling. These contributions collectively advance the application of transformer-based models in computational biology, offering a robust tool for RNA sequence generation and interaction prediction.

Looking ahead, future work will focus on expanding the model to incorporate additional biological modalities and further refining the knowledge distillation techniques to enhance model efficiency and generalization. Additionally, integrating more comprehensive datasets and exploring advanced training strategies could further improve the model's performance and applicability in diverse biological contexts. Overall, this research lays a strong foundation for leveraging advanced machine learning models in the intricate landscape of biological sequence interactions, with scope for further study exploration.

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] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[7] 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.

[8] A. Jaegle, F. Gimeno, A. Brock, A. Zisserman, O. Vinyals, and J. Carreira, "Perceiver: General perception with iterative attention," 2021.

[9] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang, "Git: A generative image-to-text transformer for vision and language," 2022.

[10] A. Kroll, S. Ranjan, and M. J. Lercher, "A multimodal transformer network for protein-small molecule interactions enhances drug-target affinity and enzyme-substrate predictions," *bioRxiv*, 2023. Accessed: 2024-12-24.

[11] J. J. Garau-Luis, P. Bordes, L. Gonzalez, M. Roller, B. P. de Almeida, L. Hexemer, C. Blum, S. Laurent, J. Grzegorzewski, M. Lang, T. Pierrot, and G. Richard, "Multi-modal transfer learning between biological foundation models," *arXiv preprint arXiv:2406.14150v1*, 2024. Accessed: 2024-12-24.

[12] W. Tang, H. Wen, R. Liu, J. Ding, W. Jin, Y. Xie, H. Liu, and J. Tang, "Single-cell multimodal prediction via transformers," *arXiv preprint arXiv:2303.00233*, 2023. Accessed: 2024-12-24.

[13] V. Jung and L. van der Plas, "Understanding the effects of language-specific class imbalance in multilingual fine-tuning," *arXiv preprint arXiv:2402.13016*, 2024. Accessed: 2024-12-24.

[14] D. Choi, D. Xin, H. Dadkhahi, J. Gilmer, A. Garg, O. Firat, C.-K. Yeh, A. M. Dai, and B. Ghorbani, "Order matters in the presence of dataset imbalance for multilingual learning," *arXiv preprint arXiv:2312.06134*, 2024. Accessed: 2024-12-24.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Łukasz Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017. Accessed: 2024-12-24.

[16] S. Brage and B. S. Andresen, "Deepclip: Predicting the effect of mutations on protein–rna binding with deep learning," *Nucleic Acids Research*, vol. 48, no. 13, pp. 7099–7118, 2020.

[17] S. R. Krishnan, A. Roy, and M. M. Gromiha, "Reliable method for predicting the binding affinity of rna-small molecule interactions using machine learning," *Briefings in Bioinformatics*, vol. 25, no. 2, p. bbae002, 2024.