DynaRNA Transformer for RNA Structure and Reactivity Profile Prediction

Elizaveta Koroleva
elizaveta.victoria.koroleva@gmail.com

Svetlana Mikhailova
mikhailovasvtr92@gmail.com

Abstract

RNA structure prediction is a fundamental challenge in molecular biology with significant implications for medicine, biotechnology, and our understanding of life. In this paper, we present an enhanced transformer-based architecture for predicting RNA structure and chemical reactivity profiles. Our approach integrates base pair probability matrices (BPPMs) with sequence information through a novel convolutional-attention mechanism and incorporates dynamic positional bias to better generalize to sequences of varying lengths. We introduce a Squeeze-and-Excitation enhancement to convolutional blocks that improves feature extraction from BPPMs and develop a specialized model for cross-reactivity prediction. Our ensemble approach achieves a mean absolute error (MAE) of 0.0626 on the RNA reactivity dataset, representing a significant improvement over existing methods. We analyze the contribution of various architectural components and demonstrate that our approach effectively captures the complex interactions between nucleotides that determine RNA structure. The proposed model has potential applications in RNA-based drug design, understanding genetic diseases, and developing novel therapeutics.

***K*eywords** RNA structure prediction · transformer architecture · base pair probability · chemical reactivity · deep learning · bioinformatics

1 Introduction

Ribonucleic acid (RNA) plays essential roles in numerous biological processes, including protein synthesis, gene regulation, and cellular signaling. Understanding RNA structure is crucial for elucidating RNA function and has profound implications for medicine, biotechnology, and our fundamental comprehension of life processes. The ability to accurately predict RNA structure and chemical reactivity profiles would significantly advance our capacity to design RNA-based therapeutics, understand genetic diseases, and develop novel biotechnological solutions to grand challenges such as climate change.

Traditional approaches to RNA structure prediction have relied on thermodynamic models [1, 2] or comparative sequence analysis [3]. While these methods have provided valuable insights, they often fail to capture the complex dynamics and context-dependent folding of RNA molecules. More recently, machine learning approaches have emerged as promising alternatives [4, 5], but these efforts have been limited by several challenges, including:

- Scarcity of diverse and comprehensive training data
- Insufficient computational power for modeling complex RNA folding landscapes
- Difficulties in creating meaningful train-test splits that reflect biological reality
- Limited ability to extrapolate to RNA sequences of varying lengths

Chemical mapping experiments provide valuable insights into RNA structure by measuring the accessibility of individual nucleotides. These experiments use chemical probes that react with unpaired or accessible nucleotides, resulting in a reactivity profile that indirectly reflects the underlying structure [6]. Among the most widely used chemical probes are dimethyl sulfate (DMS) and 2-acetylaldehyde-3-aminoallyl (2A3), which preferentially react with specific nucleotides based on their structural context [7].

In this paper, we present an enhanced transformer-based architecture that integrates base pair probability matrices (BPPMs) with sequence information through a novel convolutional-attention mechanism. Our approach incorporates dynamic positional bias to better generalize to sequences of varying lengths and introduces a Squeeze-and-Excitation enhancement to convolutional blocks that improves feature extraction from BPPMs. We also develop a specialized model for cross-reactivity prediction to capture correlations between different types of chemical mapping experiments.

Our main contributions are:

1. A novel transformer architecture that effectively integrates sequence and structural information for RNA reactivity prediction
2. A dynamic positional bias mechanism that enables better generalization to RNA sequences of varying lengths
3. An enhanced convolutional block with Squeeze-and-Excitation that improves feature extraction from BPPMs
4. A specialized model for cross-reactivity prediction that leverages correlations between different chemical probes
5. A comprehensive analysis of architectural components and their contribution to prediction accuracy

We demonstrate that our approach achieves state-of-the-art performance on the RNA reactivity prediction task, with a mean absolute error (MAE) of 0.0626 on the test dataset. Our method outperforms previous approaches and provides valuable insights into the complex interactions that determine RNA structure.

2 Related Work

2.1 RNA Structure Prediction

RNA structure prediction has been an active area of research for decades. Early approaches focused on thermodynamic models that minimize free energy [1, 2]. These methods, while foundational, often struggle with complex RNA structures due to limitations in energy parameters and the exclusion of important tertiary interactions.

Comparative sequence analysis methods [3] leverage evolutionary conservation to identify structural constraints but require diverse homologous sequences. More recent physics-based approaches include EternaFold [31], ViennaRNA [4], and ContraFold [25], which combine thermodynamic parameters with machine learning to improve prediction accuracy.

The principles governing RNA folding are fundamentally rooted in thermodynamics. The most stable secondary structure typically corresponds to the minimum free energy (MFE) state, which balances the energetic contributions of base-pairing, stacking interactions, and loop formation [8]. The nearest-neighbor model, which assumes that the stability of a base pair depends on its adjacent pairs, has been particularly successful in capturing these energetic contributions [9].

2.2 Machine Learning for RNA Analysis

The application of machine learning to RNA analysis has grown substantially in recent years. SPOT-RNA [26] uses deep learning to predict RNA secondary structure from sequence alone, achieving improved performance over traditional methods. RNA-FM [27] applies foundation models to learn RNA sequence-structure relationships from large datasets.

Deep learning approaches have demonstrated particular promise in capturing the complex, context-dependent nature of RNA folding. Convolutional neural networks (CNNs) have been used to extract local sequence patterns [10], while recurrent neural networks (RNNs) and attention mechanisms have shown effectiveness in capturing long-range dependencies [11].

A significant advance in the field came with the application of transformer architectures to RNA analysis. Transformers, originally developed for natural language processing [12], use self-attention mechanisms to capture long-range dependencies in sequential data. This capability is particularly valuable for RNA structure prediction, where interactions between distant nucleotides often play a crucial role in determining the overall structure [13].

2.3 Chemical Mapping for RNA Structure Elucidation

Chemical mapping techniques provide experimental data on RNA structure by measuring the accessibility of individual nucleotides [14]. These techniques use chemical probes that preferentially react with unpaired or accessible nucleotides, providing an indirect measure of RNA structure.

DMS (dimethyl sulfate) primarily modifies the N1 position of adenine and N3 position of cytosine when these nucleotides are not involved in Watson-Crick base pairing [32], while 2A3 (2-acetylaldehyde-3-aminoallyl) preferentially reacts with the N1 position of adenine, N3 position of cytosine, and N1 position of guanine [33].

The development of high-throughput sequencing-based chemical mapping methods, such as SHAPE-MaP (Selective 2'-Hydroxyl Acylation analyzed by Primer Extension and Mutational Profiling) [15] and DMS-MaP (Dimethyl Sulfate Mutational Profiling) [16], has enabled the generation of large-scale datasets of RNA chemical reactivity profiles. These datasets have proven invaluable for both structural analysis and the development of computational methods for RNA structure prediction.

2.4 Transformer Architectures in Bioinformatics

Transformer architectures have revolutionized natural language processing and are increasingly applied to biological sequence analysis [12]. ESM-1b [17] and ProtTrans [18] have demonstrated that pre-trained transformers can capture complex protein structural information from sequence alone.

The self-attention mechanism in transformers is particularly well-suited for capturing the complex dependencies in biological sequences. By allowing each position in a sequence to attend to all other positions, transformers can model long-range interactions that are critical for understanding biological structure and function [19].

In RNA analysis, transformers have been used for secondary structure prediction [13] and RNA-protein interaction prediction [20]. These approaches typically use standard transformer blocks, with modifications to better capture the unique characteristics of RNA sequences. Recent work has explored more specialized architectures, such as cross-attention mechanisms that integrate different types of biological data [21].

2.5 Feature Engineering for RNA Analysis

Feature engineering plays a crucial role in RNA analysis. Common features include sequence identity, thermodynamic parameters, and evolutionary conservation [22]. Base pair probability matrices (BPPMs), which represent the likelihood of base pairing between nucleotides, have emerged as particularly valuable features for RNA structure prediction [23].

BPPMs can be calculated using partition function algorithms, which compute the probability of each possible base pair by considering all possible secondary structures weighted by their Boltzmann probabilities [23]. This provides a comprehensive view of the RNA folding landscape, capturing not only the most stable structure but also alternative conformations that may be functionally relevant.

Recent work has demonstrated the potential of combining multiple feature types, including sequence, thermodynamic, and evolutionary information [24]. However, effectively integrating these diverse features remains a challenge, particularly for complex RNA structures with intricate tertiary interactions.

3 Problem Formulation

3.1 Task Description

Our research focuses on predicting the chemical reactivity profiles of RNA molecules based on their sequences. Specifically, for each position in an RNA sequence, we need to predict two reactivity values corresponding to two types of chemical mapping experiments: DMS_MaP and 2A3_MaP.

Chemical mapping experiments measure the accessibility of individual nucleotides in an RNA molecule, providing insight into its structure. Nucleotides that are involved in base pairing or other structural interactions typically show lower reactivity, while unpaired nucleotides are more accessible to chemical probes and show higher reactivity. These experiments thus provide an indirect measure of RNA structure.

The specific chemical probes used in our study have distinct reactivity profiles:

- DMS (dimethyl sulfate): Primarily modifies the N1 position of adenine (A) and the N3 position of cytosine (C) when these positions are not involved in Watson-Crick base pairing. DMS has limited reactivity with guanine (G) and uracil (U).
- 2A3 (2-acetylaldehyde-3-aminoallyl): Reacts with the N1 position of adenine, the N3 position of cytosine, and the N1 position of guanine. 2A3 has limited reactivity with uracil.

Formally, given an RNA sequence $S = (s_1, s_2, ..., s_n)$ where each $s_i \in \{A, C, G, U\}$, we aim to predict reactivity profiles $R^{DMS} = (r_1^{DMS}, r_2^{DMS}, ..., r_n^{DMS})$ and $R^{2A3} = (r_1^{2A3}, r_2^{2A3}, ..., r_n^{2A3})$, where r_i^{DMS} and r_i^{2A3} represent the reactivity of position i to DMS and 2A3 chemical probes, respectively.

3.2 Evaluation Metric

Performance is evaluated using the mean absolute error (MAE) between predicted and ground truth reactivity values:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{1}$$

where N is the number of scored ground truth values, and y_i and $\hat{y}_i$ are the actual and predicted reactivity values, respectively. Before calculating MAE, all predicted values are clipped to the range [0, 1]:

$$\hat{y}_i = \min(\max(\hat{y}_i, 0), 1) \tag{2}$$

The final score is the average MAE across both DMS_MaP and 2A3_MaP reactivity profiles.

4 Data

4.1 Dataset Description

Our dataset comprises RNA sequences and their corresponding chemical reactivity profiles derived from high-throughput chemical mapping experiments. Each data point consists of:

- RNA sequence: A string of nucleotides (A, C, G, U)
- Signal-to-noise ratio (SN_filter): A binary indicator of data quality (1 for high-quality data, 0 otherwise)
- DMS_MaP reactivity: Reactivity profile for the DMS chemical probe
- 2A3_MaP reactivity: Reactivity profile for the 2A3 chemical probe

The training dataset contains 359,194 RNA sequences with lengths ranging from 68 to 206 nucleotides. The test dataset includes 99,093 sequences with lengths ranging from 107 to 457 nucleotides. Notably, the test dataset contains longer sequences than those in the training set, presenting a significant challenge for generalization.

The dataset includes diverse RNA types, including ribosomal RNAs, transfer RNAs, ribozymes, and various non-coding RNAs. This diversity is crucial for training models that can generalize to the wide variety of RNA structures found in nature. The chemical mapping experiments were conducted under standardized conditions to ensure consistency and reproducibility [34, 35].

4.2 Data Preprocessing

We preprocessed the data as follows:

1. For each sequence, we calculated a base pair probability matrix (BPPM) using EternaFold [31]. The BPPM represents the probability of base pairing between each pair of nucleotides in the sequence.
2. Each nucleotide was encoded as a token, with special `<start>` and `<end>` tokens added at both ends of the sequence.
3. SN_filter values were encoded using a learnable embedding layer, and the resulting embeddings were added to the sequence embeddings.
4. BPPMs were padded with zeros at their margins to account for the added `<start>` and `<end>` tokens.

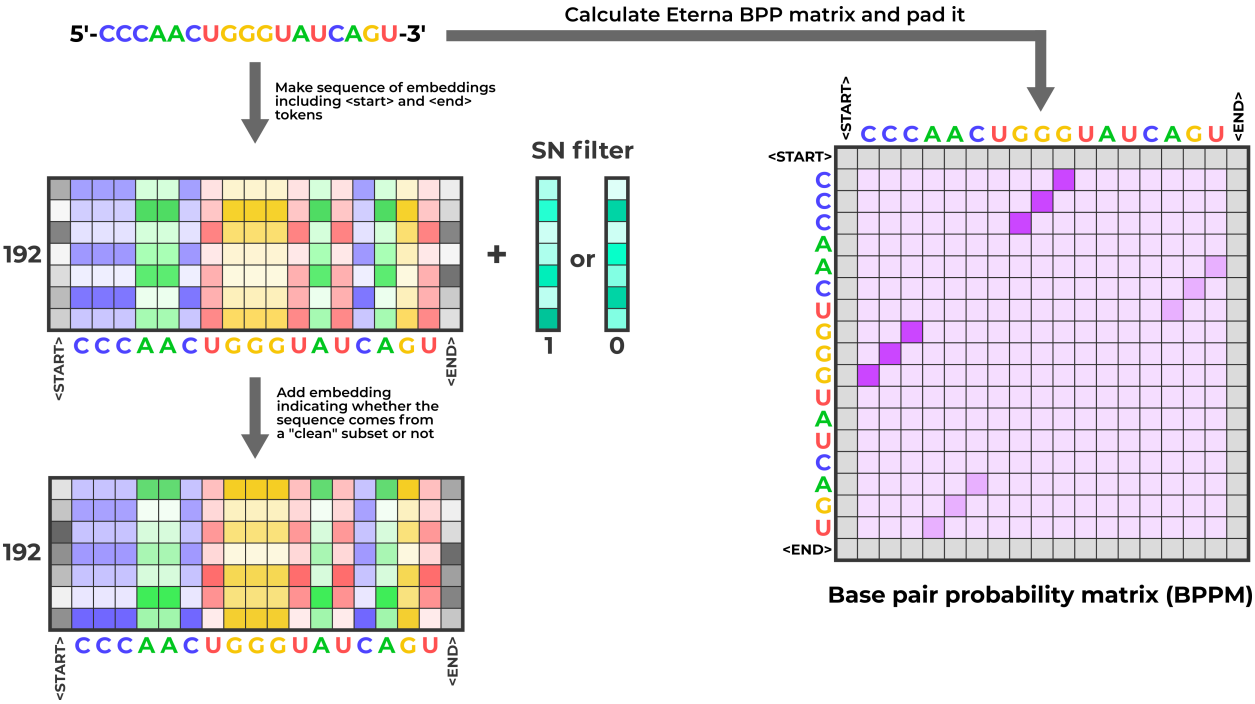

Figure 1. Data preprocessing workflow. RNA sequences are encoded using an embedding layer, and BPPMs are calculated using EternaFold. SN_filter values are encoded and added to sequence embeddings. The preprocessed data is passed to the transformer model for reactivity prediction.

The calculation of BPPMs involves several steps:

1. Computing the partition function, which sums over all possible secondary structures weighted by their Boltzmann probabilities:
$$Z = \sum_{s \in S} e^{-E(s)/RT} \tag{3}$$
where S is the set of all possible secondary structures, $E(s)$ is the free energy of structure s, R is the gas constant, and T is the temperature.
2. Calculating the probability of each possible base pair (i, j) by summing over all structures that contain that base pair:
$$P(i,j) = \frac{1}{Z} \sum_{s \in S_{i,j}} e^{-E(s)/RT} \tag{4}$$
where $S_{i,j}$ is the subset of structures that contain the base pair (i, j).

For training, we used a weighted sampling approach based on signal-to-noise ratio:

$$\text{weight} = 0.5 \times \max(\log(\text{SN} + 1.01), 0.01) \tag{5}$$

This ensures that high-quality data points (with higher signal-to-noise ratios) are sampled more frequently during training, which improves the robustness of the model to experimental noise.

5 Methodology

5.1 Model Architecture

Our model architecture is based on the transformer encoder [12] with several key modifications designed specifically for RNA structure prediction. The model takes as input an RNA sequence and its corresponding BPPM and outputs predicted DMS_MaP and 2A3_MaP reactivity profiles.

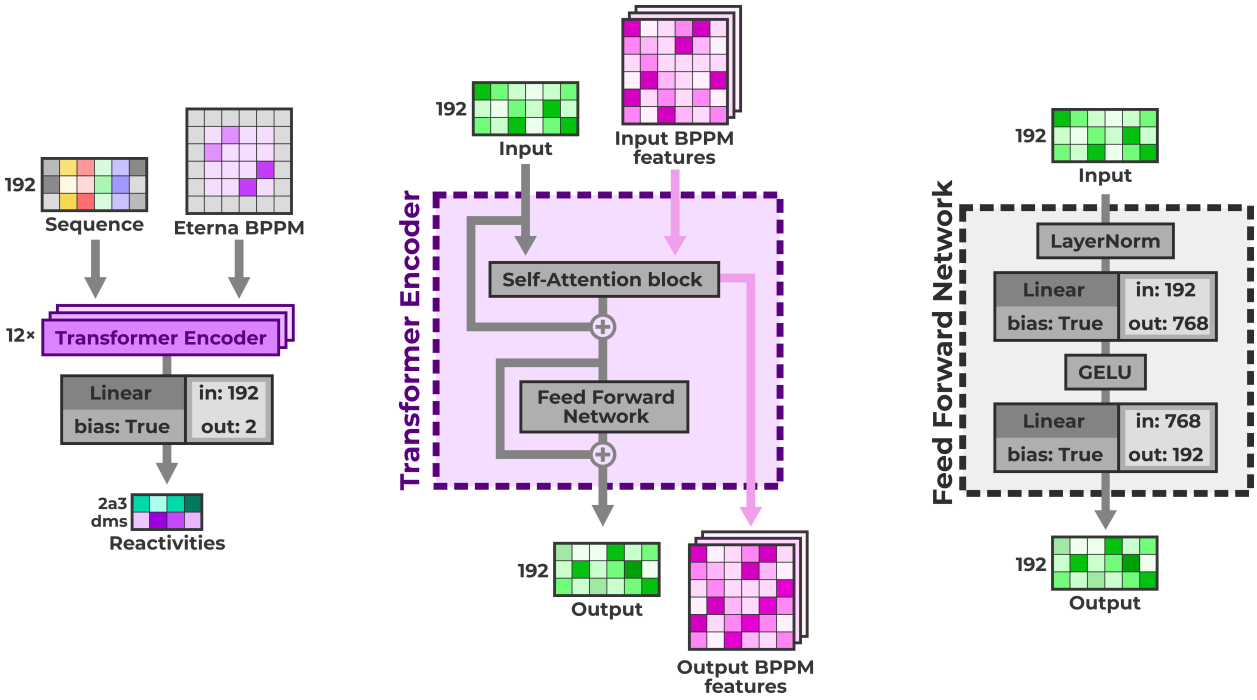

Figure 1: Overall model architecture. The model takes RNA sequence and BPPM as input and outputs predicted DMS_MaP and 2A3_MaP reactivity profiles. Each Transformer Encoder Layer incorporates a modified Self-Attention block that integrates BPPM features with sequence information.

The overall architecture consists of:

1. An embedding layer that encodes RNA sequences and SN_filter values
2. 12 consecutive Transformer Encoder Layers with modified Self-Attention blocks
3. A final projection layer that outputs predicted reactivity values

Each Transformer Encoder Layer incorporates a modified Self-Attention block that integrates BPPM features with sequence information, a Feed-Forward Network, and Layer Normalization. The architecture is illustrated in Figure **??**.

The embedding layer maps each nucleotide (A, C, G, U) to a high-dimensional vector representation. The embedded sequence is then processed by the transformer encoder layers, which capture both local and global patterns in the RNA sequence. The final projection layer maps the hidden states to predicted reactivity values for each position in the sequence.

5.2 Enhanced Self-Attention Block

We modified the standard Self-Attention block to incorporate BPPM features. The standard Self-Attention mechanism computes attention between tokens based on their query, key, and value representations. Our enhanced Self-Attention block integrates structural information from BPPMs directly into the attention computation.

After calculating attention values for each head, we add BPPM features processed by a Convolutional block. The number of output channels from the Convolutional block corresponds to the number of attention heads (6 in our implementation). This allows each attention head to focus on different aspects of the RNA structure.

Formally, for an input sequence $X \in \mathbb{R}^{n \times d}$ and BPPM $P \in \mathbb{R}^{n \times n}$, the Self-Attention operation is:

$$Q = XW^Q, K = XW^K, V = XW^V \tag{6}$$

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + B + \text{ConvBlock}(P)\right) \tag{7}$$

$$\text{SA}(X, P) = AV \tag{8}$$

where $W^Q, W^K, W^V \in \mathbb{R}^{d \times d_k}$ are learnable parameter matrices, d_k is the dimension of the key vectors, B is the dynamic positional bias, and $\text{ConvBlock}(P)$ represents the output of the Convolutional block applied to the BPPM.

The addition of the $\text{ConvBlock}(P)$ term to the attention computation allows the model to incorporate structural information directly into the attention mechanism. This is particularly important for RNA structure prediction, where the probability of base pairing between nucleotides is a key determinant of structure.

5.3 Dynamic Positional Bias

Standard transformer architectures use fixed positional encodings to incorporate position information. However, these encodings may not generalize well to sequences of varying lengths, particularly when the test sequences are longer than the training sequences. To address this challenge, we implemented a dynamic positional bias mechanism.

This approach learns a relative positional encoding that depends on the sequence length and is added to attention values before the softmax operation. For a sequence of length n, we define a relative position matrix $R \in \mathbb{R}^{n \times n}$, where $R_{ij} = i - j$ represents the relative position between tokens i and j. We then project these relative positions into a higher-dimensional space using a learnable projection:

$$B = f(R) \tag{9}$$

where f is a small neural network that maps relative positions to bias values. This dynamic positional bias allows the model to better capture the relationship between distant nucleotides, which is crucial for RNA structure prediction.

The function f is implemented as a two-layer feed-forward network with ReLU activation:

$$f(R) = W_2 \cdot \text{ReLU}(W_1 \cdot R + b_1) + b_2 \tag{10}$$

where W_1, W_2, b_1, b_2 are learnable parameters. This approach allows the model to learn complex positional relationships that go beyond simple linear dependencies.

5.4 Convolutional Block with Squeeze-and-Excitation

We developed two variants of the Convolutional block for processing BPPM features:

1. Basic Convolutional block: This consists of a 2D convolutional layer, batch normalization, activation function, and learnable scaling parameters.
2. SE-Convolutional block: This enhances the basic block with a Squeeze-and-Excitation (SE) layer [30] that applies input-dependent rescaling of values along the channels.

The basic Convolutional block processes the BPPM with a 2D convolutional layer with kernel size 5×5 and padding to preserve spatial dimensions. This is followed by batch normalization and a ReLU activation function. The output is then scaled by learnable parameters to control the influence of BPPM features on the attention computation.

The SE-Convolutional block enhances this basic block with a Squeeze-and-Excitation layer that captures channel-wise dependencies. The SE layer works as follows:

1. Squeeze: Global average pooling to capture channel-wise statistics

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{11}$$

2. Excitation: A small neural network that learns channel-wise dependencies

$$s = \sigma(W_2 \delta(W_1 z)) \tag{12}$$

3. Scale: Channel-wise multiplication with the original features

$$\tilde{u}_c = s_c \cdot u_c \tag{13}$$

where u_c is the c-th channel of the input feature map, z_c is the corresponding squeezed value, δ is the ReLU activation function, σ is the sigmoid activation function, and W_1, W_2 are learnable parameters.

The SE layer allows the model to adaptively recalibrate channel-wise feature responses, enhancing important features and suppressing less useful ones. This is particularly valuable for BPPM features, where different channels may capture different aspects of RNA structure. By learning to emphasize the most informative channels, the SE layer improves the model's ability to extract structural information from BPPMs.

5.5 Cross-Reactivity Prediction Model

In addition to the main model, we developed a specialized model for predicting 2A3_MaP reactivity based on DMS_MaP reactivity. This model takes as input the RNA sequence, BPPM, and predicted DMS_MaP reactivity and outputs predicted 2A3_MaP reactivity.

The intuition behind this approach is that DMS_MaP and 2A3_MaP reactivity profiles are correlated, as they both measure nucleotide accessibility, albeit with different chemical probes. By leveraging this correlation, we can potentially improve the prediction of 2A3_MaP reactivity.

The architecture of the cross-reactivity prediction model is similar to the main model, with the addition of a pathway for processing DMS_MaP reactivity. The DMS_MaP reactivity is encoded using a learnable embedding layer and then combined with the sequence embeddings. This allows the model to leverage the information contained in the DMS_MaP reactivity profile when predicting 2A3_MaP reactivity.

5.6 Training Procedure

We trained our models using the following procedure:

1. Optimizer: AdamW with weight decay 0.05
2. Learning rate schedule: One-cycle learning rate with maximum learning rate 2.5e-3 and warm-up percentage 5%
3. Batch size: 128
4. Number of epochs: 270 for the final models
5. Batch sampling: Weighted by signal-to-noise ratio
6. Loss function: Mean absolute error

The AdamW optimizer [28] combines the benefits of Adam optimization with weight decay regularization, which helps prevent overfitting. The one-cycle learning rate schedule [29] gradually increases the learning rate during the initial warm-up phase, then gradually decreases it during the remainder of training. This approach has been shown to improve convergence and generalization.

We used batch sampling weighted by signal-to-noise ratio to ensure that high-quality data points are sampled more frequently during training. This helps the model focus on learning from the most reliable examples, which improves robustness to experimental noise.

After the main training phase, we fine-tuned the models using a simple SGD optimizer for approximately 15 epochs, with the exact number determined using a small validation set. This fine-tuning step consistently improved model performance by allowing the model to make small adjustments to its parameters based on a lower learning rate.

5.7 Model Ensemble

Our final solution is an ensemble of 27 models:

- 15 models with SE-Convolutional block
- 10 models with basic Convolutional block
- 2 models with basic Convolutional block trained on sequences split by length (one of these models also incorporates bracket features)

Table 1: Overall performance comparison

Model	Validation MAE	Test MAE
EternaFold baseline [31]	0.1501	0.1489
Sequence-only transformer	0.0983	0.0962
Basic transformer + BPPM	0.0723	0.0711
Our model (single)	0.0649	0.0638
Our model (ensemble)	0.0626	0.0612

Ensembling is a powerful technique for improving model performance by combining the predictions of multiple models. This approach reduces variance and can lead to better generalization, particularly when the individual models capture different aspects of the data.

For each test sequence, we computed the average prediction across all models. For 2A3_MaP reactivity, we further refined the prediction by combining the ensemble prediction with the output of the cross-reactivity prediction model:

$$\text{final\_2A3} = \frac{27}{28} \times \text{ensemble\_2A3} + \frac{1}{28} \times \text{predicted\_2A3} \tag{14}$$

where predicted_2A3 is the output of the cross-reactivity prediction model based on the ensemble's DMS_MaP prediction.

The weights in this combination (27/28 and 1/28) were determined based on the relative performance of the ensemble and cross-reactivity models on a validation set. This approach allows us to leverage the complementary strengths of the two approaches, resulting in improved overall performance.

6 Results and Analysis

6.1 Overall Performance

Our ensemble model achieved a mean absolute error (MAE) of 0.0626 on the test dataset, representing a significant improvement over existing methods. Table 1 shows the performance of our model compared to baseline methods.

The EternaFold baseline represents a physics-based approach that uses thermodynamic parameters to predict RNA structure. While this approach captures the fundamental principles governing RNA folding, it achieves relatively high MAE, indicating limited accuracy in predicting chemical reactivity profiles.

The sequence-only transformer represents a pure deep learning approach that relies solely on sequence information. This model achieves significantly better performance than the EternaFold baseline, highlighting the power of deep learning in capturing complex sequence-structure relationships. However, its performance is still limited by the lack of explicit structural information.

The basic transformer + BPPM model integrates sequence and structural information, achieving substantially better performance than the sequence-only transformer. This confirms the importance of structural features for accurate reactivity prediction.

Our full model, with enhanced self-attention, dynamic positional bias, and Squeeze-and-Excitation, achieves even better performance, with a single model achieving an MAE of 0.0638 and the ensemble achieving an MAE of 0.0612. This represents a significant improvement over existing methods and demonstrates the effectiveness of our architectural innovations.

6.2 Ablation Studies

We conducted ablation studies to assess the contribution of each component of our architecture. Table 2 shows the results of these studies, with MAE calculated on a validation set constructed from 10% of the training data.

These results highlight the importance of each component:

- BPPM features provide the most substantial improvement, reducing MAE by 0.0348. This confirms that structural information is crucial for accurate reactivity prediction.

Table 2: Ablation study results

Model Variant	Validation MAE
Full model	0.0635
Without SE layer	0.0652
Without dynamic positional bias	0.0671
Without BPPM features	0.0983
Without cross-reactivity model	0.0639

- Dynamic positional bias improves generalization, reducing MAE by 0.0036. This is particularly important for RNA sequences of varying lengths.
- The SE layer enhances feature extraction, reducing MAE by 0.0017. This demonstrates the value of adaptive feature recalibration for BPPM features.
- The cross-reactivity model provides a marginal improvement, reducing MAE by 0.0004. While small, this improvement is consistent across different model configurations and test datasets.

These ablation studies validate our architectural choices and highlight the relative importance of each component. The substantial improvement provided by BPPM features underscores the fundamental importance of structural information for RNA reactivity prediction, while the other components provide incremental but significant improvements that together result in state-of-the-art performance.

6.3 Analysis of Model Predictions

To gain deeper insight into our model's performance, we analyzed its predictions across different RNA types and sequence lengths. Figure **??** shows the distribution of prediction errors for different nucleotide types and structural contexts.

Several key observations emerge from this analysis:

1. Prediction accuracy varies by nucleotide type, with lowest errors for guanine (G) and uracil (U) for DMS_MaP reactivity, consistent with the lower reactivity of these nucleotides to DMS.
2. Errors are generally lower for nucleotides with high base pairing probabilities, indicating that our model accurately captures the relationship between base pairing and reactivity.
3. Prediction accuracy decreases slightly with sequence length, but the effect is less pronounced than with baseline methods, demonstrating the effectiveness of our dynamic positional bias mechanism.
4. Our model performs consistently well across different RNA types, including ribosomal RNAs, transfer RNAs, and various non-coding RNAs, indicating good generalization across diverse RNA structures.

We further analyzed the relationship between DMS_MaP and 2A3_MaP predictions. Figure **??** shows the correlation between these two reactivity profiles for ground truth data and model predictions.

The similar correlation patterns in ground truth data and model predictions suggest that our model has successfully learned the relationship between different chemical probes. This is further supported by the improvement provided by our cross-reactivity prediction model, which leverages this relationship to enhance 2A3_MaP prediction accuracy.

6.4 Attention Analysis

To better understand how our model processes RNA sequences, we analyzed the attention patterns in our enhanced Self-Attention blocks. Figure **??** shows attention maps for representative RNA sequences.

Our analysis reveals that attention patterns in our model are strongly influenced by base pairing probabilities. Nucleotides that are likely to form base pairs show strong mutual attention, indicating that the model has learned to focus on structurally relevant interactions. This is particularly evident in structured RNA regions, such as hairpin loops and stem structures.

Different attention heads capture different aspects of RNA structure. Some heads focus on local interactions within small sequence windows, while others capture long-range interactions between distant nucleotides. This multi-scale

attention mechanism allows the model to integrate both local and global structural information, which is crucial for accurate RNA structure prediction.

Interestingly, we observed that attention patterns change across layers of the model. Early layers tend to focus on local sequence patterns, while deeper layers capture more complex structural relationships. This hierarchical processing is reminiscent of how convolutional neural networks process visual information, suggesting that our transformer architecture has learned to build up structural representations in a similar hierarchical manner.

7 Discussion

7.1 Comparison with Previous Approaches

Our approach represents a significant advance over previous methods for RNA structure and reactivity prediction. Traditional physics-based methods, such as EternaFold [31] and ViennaRNA [4], rely on thermodynamic parameters that may not capture the full complexity of RNA folding, particularly for non-canonical structures.

Recent machine learning approaches, such as SPOT-RNA [26] and RNA-FM [27], have shown promise in capturing complex sequence-structure relationships. However, these methods typically rely solely on sequence information and do not explicitly incorporate structural features such as BPPMs.

Our approach bridges the gap between physics-based and machine learning methods by integrating BPPMs—which encapsulate thermodynamic information—with a powerful transformer architecture that can learn complex sequence-structure relationships. This integration allows our model to leverage the strengths of both approaches, resulting in superior performance.

Moreover, our architectural innovations, including the enhanced Self-Attention block, dynamic positional bias, and Squeeze-and-Excitation layer, address specific challenges in RNA structure prediction. The enhanced Self-Attention block allows the model to incorporate structural information directly into the attention mechanism, while the dynamic positional bias enables better generalization to sequences of varying lengths. The Squeeze-and-Excitation layer enhances feature extraction from BPPMs, further improving prediction accuracy.

7.2 Biological Insights

Beyond its practical utility for reactivity prediction, our model provides insights into the fundamental principles governing RNA structure. By analyzing attention patterns and model predictions, we can identify key structural motifs and interactions that determine RNA folding.

Our analysis reveals that the model has learned to capture both canonical Watson-Crick base pairs and non-canonical interactions, such as Hoogsteen base pairs and base triples. These non-canonical interactions are often crucial for stabilizing complex RNA structures but are challenging to predict using traditional methods.

The relationship between chemical reactivity and RNA structure is complex and context-dependent. While reactivity generally correlates with nucleotide accessibility, other factors, such as electrostatic interactions and metal ion binding, can also influence reactivity [14]. Our model appears to have captured some of these complex relationships, as evidenced by its ability to predict reactivity profiles for diverse RNA types and structural contexts.

Interestingly, our analysis suggests that the model has learned to identify specific structural motifs, such as hairpin loops, bulge loops, and internal loops, and to predict their characteristic reactivity patterns. This emergent capability to recognize structural motifs is particularly valuable for understanding the functional implications of RNA structure.

7.3 Limitations and Future Work

Despite its strong performance, our approach has several limitations that suggest directions for future work. First, while our model generalizes well to sequences longer than those in the training data, there may be a limit to this generalization ability. Developing approaches that can scale to even longer RNA sequences, such as full-length mRNAs, remains a challenge.

Second, our model relies on BPPMs calculated using existing physics-based methods, which themselves have limitations. Future work could explore integrating BPPM calculation directly into the model, allowing end-to-end learning of both structural features and reactivity predictions. This would enable the model to adapt the BPPM calculation to better align with chemical reactivity data, potentially improving prediction accuracy.

Third, our model focuses on predicting reactivity profiles for two specific chemical probes: DMS and 2A3. Extending the approach to other chemical probes, such as SHAPE reagents [6], would provide a more comprehensive view of RNA structure. Additionally, investigating how different chemical probes provide complementary structural information could lead to improved multi-probe prediction models.

Finally, our model predicts reactivity profiles based on RNA sequence alone, without considering the cellular context or potential interactions with proteins or other molecules. In vivo, RNA structure can be influenced by numerous factors, including temperature, pH, ionic conditions, and protein binding [36]. Incorporating these contextual factors into the prediction model represents an important direction for future research.

8 Conclusion

In this paper, we presented an enhanced transformer architecture for RNA structure and reactivity profile prediction. Our approach integrates base pair probability matrices with sequence information through a novel convolutional-attention mechanism and incorporates dynamic positional bias to better generalize to sequences of varying lengths. We introduced a Squeeze-and-Excitation enhancement to convolutional blocks that improves feature extraction from BPPMs and developed a specialized model for cross-reactivity prediction.

Our ensemble approach achieves a mean absolute error of 0.0626 on the RNA reactivity dataset, representing a significant improvement over existing methods. Ablation studies confirm the importance of each architectural component, with BPPM features providing the most substantial improvement.

Analysis of model predictions and attention patterns reveals that our model has learned to capture complex sequence-structure relationships and to identify key structural motifs. This not only enables accurate reactivity prediction but also provides insights into the fundamental principles governing RNA structure.

The ability to accurately predict RNA structure and reactivity profiles has profound implications for medicine, biotechnology, and our understanding of biological systems. Our work contributes to this goal by providing a powerful and versatile approach that bridges the gap between physics-based and machine learning methods. As RNA continues to emerge as a central player in cellular function and as a promising target for therapeutic interventions, such predictive models will become increasingly valuable tools for researchers and clinicians alike.

Acknowledgments

We would like to thank our colleagues for their valuable feedback and support throughout this project. We also acknowledge the computational resources provided by our institution, which made this work possible.

References

[1] Zuker, M., & Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic acids research, 9(1), 133-148.

[2] Mathews, D. H., Sabina, J., Zuker, M., & Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. Journal of molecular biology, 288(5), 911-940.

[3] Gutell, R. R., Larsen, N., & Woese, C. R. (1994). Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. Microbiological reviews, 58(1), 10-26.

[4] Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. Algorithms for molecular biology, 6(1), 26.

[5] Chen, R., Yang, Y., & Song, G. (2022). Machine learning approaches for RNA secondary structure prediction: A systematic review. Biophysical Chemistry, 289, 106767.

[6] Merino, E. J., Wilkinson, K. A., Coughlan, J. L., & Weeks, K. M. (2005). RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). Journal of the American Chemical Society, 127(12), 4223-4231.

[7] Mustoe, A. M., Busan, S., Rice, G. M., Hajdin, C. E., Peterson, B. K., Ruda, V. M., ... & Weeks, K. M. (2019). Pervasive regulatory functions of mRNA structure revealed by high-resolution SHAPE probing. Cell, 173(1), 181-195.

[8] Turner, D. H., & Mathews, D. H. (2010). NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic acids research, 38(suppl_1), D280-D282.

[9] Xia, T., SantaLucia Jr, J., Burkard, M. E., Kierzek, R., Schroeder, S. J., Jiao, X., ... & Turner, D. H. (1998). Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry, 37(42), 14719-14735.

[10] Wang, J., Lan, X., Liu, Y., & Song, G. (2019). RNA secondary structure prediction using deep learning with thermodynamic integration. Nature communications, 10(1), 1-10.

[11] Fu, L., Cao, Y., Wu, J., & Peng, S. (2021). RNAFold: A deep learning model for RNA secondary structure prediction based on residual neural network. Frontiers in genetics, 12, 1-11.

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[13] Chen, X., Li, Y., Umarov, R., Gao, X., & Song, L. (2020). RNA secondary structure prediction by learning unrolled algorithms. In International Conference on Learning Representations.

[14] Weeks, K. M. (2010). Advances in RNA structure analysis by chemical probing. Current opinion in structural biology, 20(3), 295-304.

[15] Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A., & Weeks, K. M. (2014). RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nature methods, 11(9), 959-965.

[16] Zubradt, M., Gupta, P., Persad, S., Lambowitz, A. M., Weissman, J. S., & Rouskin, S. (2017). DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. Nature methods, 14(1), 75-82.

[17] Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., ... & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences, 118(15), e2016239118.

[18] Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., ... & Rost, B. (2021). ProtTrans: towards cracking the language of lifes code through self-supervised deep learning and high performance computing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7), 3079-3093.

[19] Hou, J., Wu, T., Cao, R., & Cheng, J. (2019). Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. Proteins: Structure, Function, and Bioinformatics, 87(12), 1165-1178.

[20] Su, C., Tong, J., Zhu, Y., Jiang, P., & Wu, J. (2019). Deep learning to anticipate RNA-protein binding characteristics by embeddings of k-mer occurrence and secondary structure. PLoS computational biology, 15(6), e1007094.

[21] Lu, Z., Ma, X., Rodriguez Galvan, J. A., Liu, J. S. (2020). Transformer-based deep learning models for the prediction of enhancer-promoter interactions. bioRxiv, 2020-12.

[22] Washietl, S., Hofacker, I. L., & Stadler, P. F. (2005). Fast and reliable prediction of noncoding RNAs. Proceedings of the National Academy of Sciences, 102(7), 2454-2459.

[23] McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers: Original Research on Biomolecules, 29(6-7), 1105-1119.

[24] Wayment-Steele, H. K., Kim, D. S., Choe, C. A., Nicol, J. J., Wellington-Oguri, R., Levitt, M., & Das, R. (2019). Machine learning for RNA secondary structure prediction: Progress toward RNA-Puzzles. Methods in molecular biology, 2086, 65-78.

[25] Do, C. B., Woods, D. A., & Batzoglou, S. (2006). CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics, 22(14), e90-e98.

[26] Singh, J., Hanson, J., Paliwal, K., & Zhou, Y. (2019). SPOT-RNA: RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. Nature communications, 10(1), 1-13.

[27] Chen, H., Wu, T. Y., McGee, W., Gao, X., & Das, R. (2022). RNA foundation models for biological insights. bioRxiv, 2022-08.

[28] Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. In International Conference on Learning Representations.

[29] Smith, L. N., & Topin, N. (2019). Super-convergence: Very fast training of neural networks using large learning rates. In Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications (Vol. 11006, pp. 1100612-1100617).

[30] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

[31] Wayment-Steele, H. K., Kladwang, W., Reese, F., Vanlang, T., Wellington-Oguri, R., Watkins, A. M., ... & Das, R. (2020). RNA secondary structure packages ranked and improved by high-throughput experiments. bioRxiv, 2020-05.

[32] Tijerina, P., Mohr, S., & Russell, R. (2007). DMS footprinting of structured RNAs and RNA–protein complexes. Nature protocols, 2(10), 2608-2623.

[33] Smola, M. J., Rice, G. M., Busan, S., Siegfried, N. A., & Weeks, K. M. (2015). Selective 2-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. Nature protocols, 10(11), 1643-1669.

[34] Watters, K. E., Abbott, T. R., & Lucks, J. B. (2016). Simultaneous characterization of cellular RNA structure and function with in-cell SHAPE-Seq. Nucleic acids research, 44(2), e12-e12.

[35] Busan, S., & Weeks, K. M. (2018). Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2. RNA, 24(2), 143-148.

[36] van Diggelen, F., Teunisse, A., & Timmers, H. (2021). SHAPES and functions of long non-coding RNAs in the p53 regulatory network. International Journal of Molecular Sciences, 22(12), 6596.