ETERNAITY: RNA SECONDARY STRUCTURE PREDICTION USING TRANSFORMER-BASED MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

RNA secondary structure prediction is crucial to understanding biological functions and disease mechanisms. Traditional approaches rely on thermodynamic models or machine learning methods, each with their own limitations. We present EteRNAity, a transformer-based approach for RNA secondary structure prediction that treats the problem as a sequence-to-sequence translation task. Our model employs a simple encoder-decoder architecture with chunked linear attention and achieves competitive results compared to existing methods while using significantly fewer parameters (55K). We evaluated our approach across different RNA families and demonstrated the effectiveness of our novel objective function that explicitly accounts for base-pair relationships. We further show that language modeling approach can indeed generalize across various RNA families.

1 INTRODUCTION

1.1 RNA SECONDARY STRUCTURE

RNA secondary structure refers to the two-dimensional layout of RNA molecules formed through base-pairing interactions. While the primary structure is simply the linear sequence of nucleotides, the secondary structure emerges when these nucleotides form hydrogen bonds with each other, creating various structural elements: **Stem (or Helix):** Formed by consecutive base pairs, creating a double-stranded region. **Hairpin Loop:** A single-stranded loop that connects the two ends of a stem. **Bulge:** An unpaired region on one strand of a stem that causes a "bulge" in the structure. **Internal Loop:** Unpaired nucleotides on both strands of a stem. **Multiloop (Junction):** A junction where three or more stems meet. **Pseudoknot:** A structure formed when nucleotides in a loop pair with nucleotides outside the stem

The prediction task involves converting a linear sequence into a dot-bracket notation that captures these structural elements. For example:

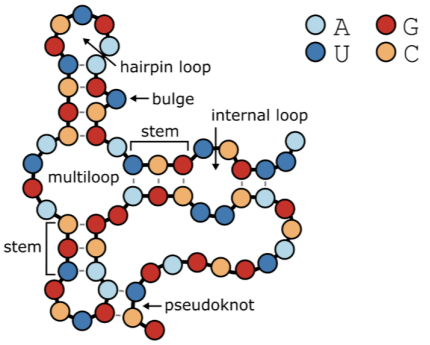
```
Sequence:  GGAAACUUCGGAACC
Structure: (((...)))...)))
```

As shown in Figure 2, this notation can be visualized as an arc diagram where each pair of brackets represents a base-pairing interaction. The prediction challenge lies in correctly identifying which nucleotides will form these structural elements, considering both local sequence patterns and long-range interactions.

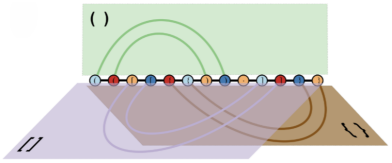
The formation of these structural elements is governed by thermodynamic principles, with the molecule typically adopting its minimum free energy configuration. However, the actual biological structure may sometimes differ due to factors such as protein interactions or kinetic effects during folding.

1.2 PROBLEM STATEMENT

RNA secondary structure prediction is fundamental to understanding the function of RNA and has applications in drug discovery and disease treatment. Given an RNA sequence composed of nucleotides (A, C, G, U), the goal is to predict its secondary structure represented in the dot-bracket notation, where dots represent unpaired bases, and matching brackets represent paired bases.



(a) Common RNA structural motifs showing different elements: stems (double-stranded regions), hairpin loops (end loops), bulges (single-strand protrusions), internal loops, multiloops (junctions), and pseudoknots.



(b) Arc representation of RNA structure showing base-pair interactions. Each arc connects paired nucleotides, with the sequence displayed linearly at the bottom. The diagram illustrates how dot-bracket notation maps to physical base-pairing relationships.

Figure 1: Visualization of RNA secondary structure elements and their representations. (a) Physical structure showing various RNA motifs and their spatial arrangement. (b) Abstract representation using arc diagram, demonstrating how base-pairs form long-range interactions in the sequence.

The challenge lies in the complex nature of RNA folding. A single RNA sequence can potentially fold into multiple different conformations (see Figure 1, structures A, B, and C), making the prediction task non-trivial. While thermodynamic principles suggest that RNA molecules typically adopt their minimum free energy (MFE) structure in nature, predicting this structure remains challenging due to several factors:

- The folding patterns are highly dependent on the RNA family, with different families exhibiting distinct structural motifs
- Only canonical base-pairs (A-U, G-C) are typically considered in prediction models, though non-canonical pairs exist in nature
- The set of known RNA families is incomplete and continually expanding, making it difficult to develop comprehensive prediction models
- The assumption that the MFE structure is always the correct biological conformation may not hold in all cases

These challenges are further complicated by the hierarchical nature of RNA folding, where local structure formation can influence global folding patterns. The prediction task must therefore consider both local sequence patterns and long-range interactions that can span significant distances in the primary sequence.

1.3 Exsiting Approaches

Traditional approaches to this problem fall into two categories: thermodynamic-based methods and machine learning-based methods. Thermodynamic methods rely on predefined energy parameters derived from experimental data but may not capture all sequence variations and typically only predict the minimum free energy structure. Machine learning approaches can learn rich parameterizations but risk overfitting due to limited training data.

2 Related Work

The field of prediction of RNA structures has evolved significantly since the 1978 Nussinov algorithm. Early approaches like Mfold (1981) and RNAfold (1994) relied primarily on thermodynamic principles. The 2000s saw the emergence of machine learning methods with tools like CONTRAfold (2006) and ContextFold (2011). Recent deep learning approaches include MXfold2 (2021) and LinearFold (2019).

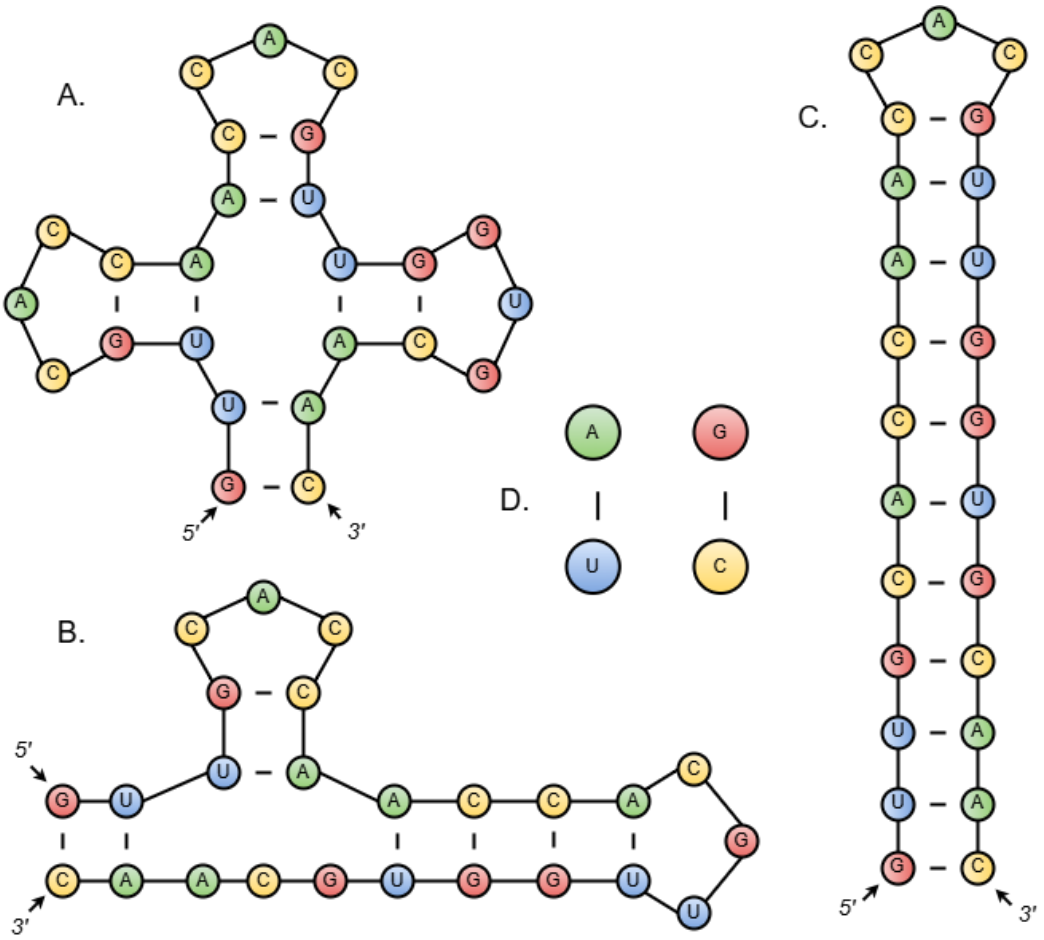


Figure 2: Different possible RNA secondary structures, for the same sequence: (A) A symmetrical RNA structure with a multi loop junction and three hairpin loops. (B) An alternative folding pattern for the same sequence showing different base-pairing arrangements consisting of a multi loop junction and two hairpin loops. (C) A linear stem structure demonstrating maximum base-pairing with only one hairpin. (D) The four nucleotides (A, U, G, C) that form canonical base pairs in RNA structures. Only canonical base pairs A-U and G-C are typically considered in structure prediction. Arrows indicate 5' to 3' directionality.

Our initial experiments with fine-tuning BERT revealed several limitations:

- Token length restrictions (512 tokens) proved problematic for longer RNA sequences
- The encoder-only architecture was suboptimal for this sequence-to-sequence task
- Despite achieving 65.33% accuracy with focal loss, the model struggled with long-range dependencies

3 METHOD

3.1 MODEL ARCHITECTURE

Our FoldFormer architecture adopts a transformer-based approach while maintaining a lightweight design of approximately 55,000 parameters. Figure 3 illustrates the complete model architecture, which consists of an encoder and decoder structure specifically tailored for RNA sequence processing.

The encoder pathway begins with a one-hot encoding of the input RNA sequence, representing each nucleotide (A, C, G, U) as a distinct vector. This encoded sequence is then combined with positional encoding through a linear projection and addition operation. The resulting representation passes through a layer normalization step before entering the main encoder block.

The encoder block, which repeats N-1 times, contains a sophisticated sequence of operations. Each block starts with an attention mechanism, followed by layer normalization. The normalized output then passes through a linear transformation coupled with an activation function, another layer normalization step, and finally a linear projection.

On the decoder side, we employ a similar but distinct structure that processes the encoded information to generate the structural prediction. The decoder also repeats N-1 times and includes attention mechanisms, layer normalizations, and linear transformations. The final output passes through a label decoding step that produces the dot-bracket notation representing the predicted RNA structure.

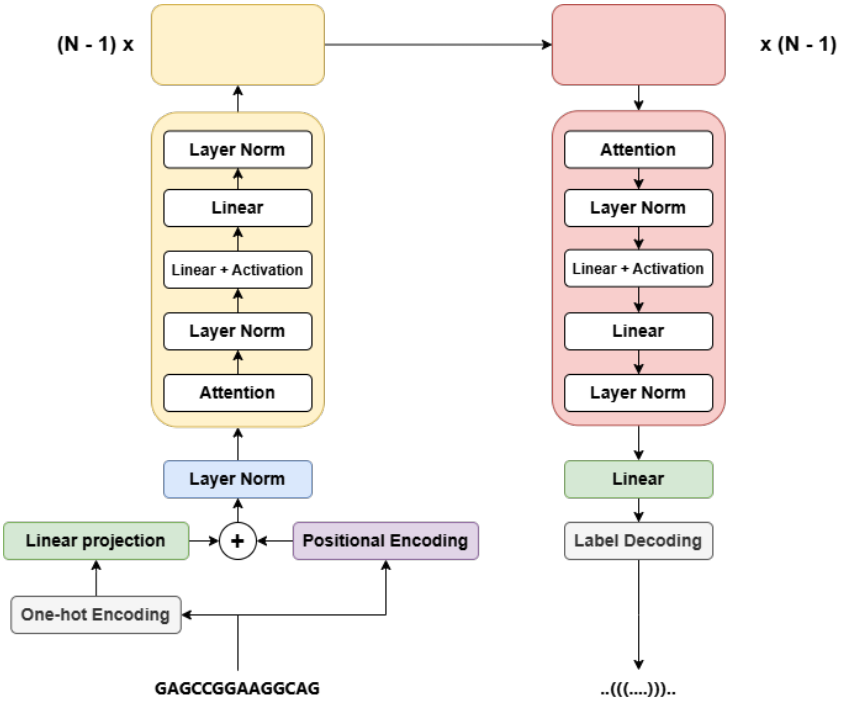


Figure 3: FoldFormer architecture diagram showing the complete encoder-decoder pipeline. The encoder (left, yellow) processes the input RNA sequence through multiple transformation layers, while the decoder (right, pink) generates the structural prediction. Key components include: attention mechanisms for capturing sequence relationships, layer normalization for training stability, and linear transformations with activation functions for feature extraction. The model employs chunked linear attention to efficiently handle longer sequences. Input example shows a sample RNA sequence (GAGCCGGAAGGCAG) being transformed into its corresponding dot-bracket notation (.(((...))).))

A notable feature of our architecture is the use of chunked linear attention, which allows efficient processing of longer RNA sequences while maintaining computational feasibility. This approach divides the input sequence into manageable chunks while preserving the ability to capture long-range dependencies critical for accurate structure prediction.

By deliberately omitting cross-attention mechanisms (due to implementation constraints), we maintain model simplicity while still achieving competitive performance. The entire architecture processes the input RNA sequence one-directionally, transforming the nucleotide sequence into a structural prediction represented in standard dot-bracket notation.

3.2 OBJECTIVE FUNCTION

We developed a specialized objective function that extends traditional cross-entropy loss to account for the paired nature of RNA structure:

$$\mathcal{L} = -\frac{1}{T}\sum_{t=1}^{T}\{\, p_{\,t}\log(\hat{p}_t), if p_t = Not - paired, p_{t_i}\log(\hat{p}_{t_i}) + p_{t_j}\log(\hat{p}_{t_j}), if p_t = Paired. \quad (1)$$

This formulation explicitly considers both opening and closing brackets when computing the loss for paired bases, improving the model's ability to learn structural patterns.

4 EXPERIMENTS AND RESULTS

4.1 DATASET

We utilized a dataset of 100K samples from various RNA families. The data includes diverse structural motifs and varying sequence lengths. Processing 10% of the dataset requires approximately one hour of computation time.

4.2 TRAINING

We compared ReLU and GeLU activation functions, finding that GeLU generally provided more stable training and better convergence. Training metrics showed steady improvement over time, with GeLU achieving more consistent performance across batches.

4.3 RESULTS

Our model's performance was evaluated against several state-of-the-art methods across different RNA families:

- Strong performance on tRNA (72% accuracy with ReLU)
- Competitive results on 5S_rRNA and SRP families
- Overall F1-score of 0.52 with ReLU activation

Notably, our model achieves these results with significantly fewer parameters than existing approaches. The results demonstrate that deep learning models may not generalize uniformly across RNA families, suggesting the need for family-specific considerations in model design.

5 CONCLUSION

We presented EteRNAity, a transformer-based approach to RNA secondary structure prediction that achieves competitive results while maintaining a lightweight architecture. Our work highlights the potential of specialized neural architectures and loss functions for biological sequence problems. Future work could explore cross-attention mechanisms and family-specific adaptations to improve performance across different RNA types.

REFERENCES