# ETERNAITY: RNA SECONDARY STRUCTURE PREDICTION USING TRANSFORMER-BASED MODELS

**Biswajit Banerjee & Asma Khimani**
Bioinformatics
Georgia Institute of Technology
{bbanerjee32,akhimani30}@gatech.edu

## 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:

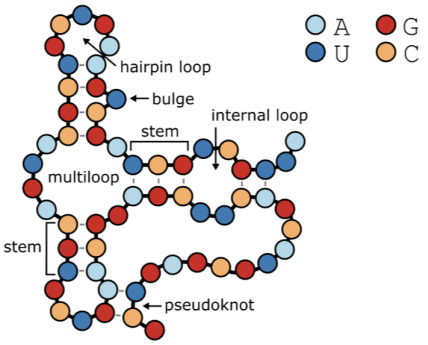
```
Sequence:  GGAAACUUCGGAACC
Structure: (((...)))...)))
```

As shown in Figure 1, 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.

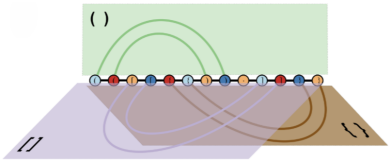
The formation of these structural elements is governed by thermodynamic principles, with the molecule typically adopting its minimum free energy configuration. However, the actual biological structure may sometimes differ due to factors such as protein interactions or kinetic effects during folding.

### 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 nu-




(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.

cleotides (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.

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.

## 2 Related Work

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. However, machine learning approaches can learn rich parameterizations but risk over-fitting due to limited training data. Thus, deep learning approaches have been explored to avoid overfitting and to avoid overfitting and ensure biologically meaningful predictions.

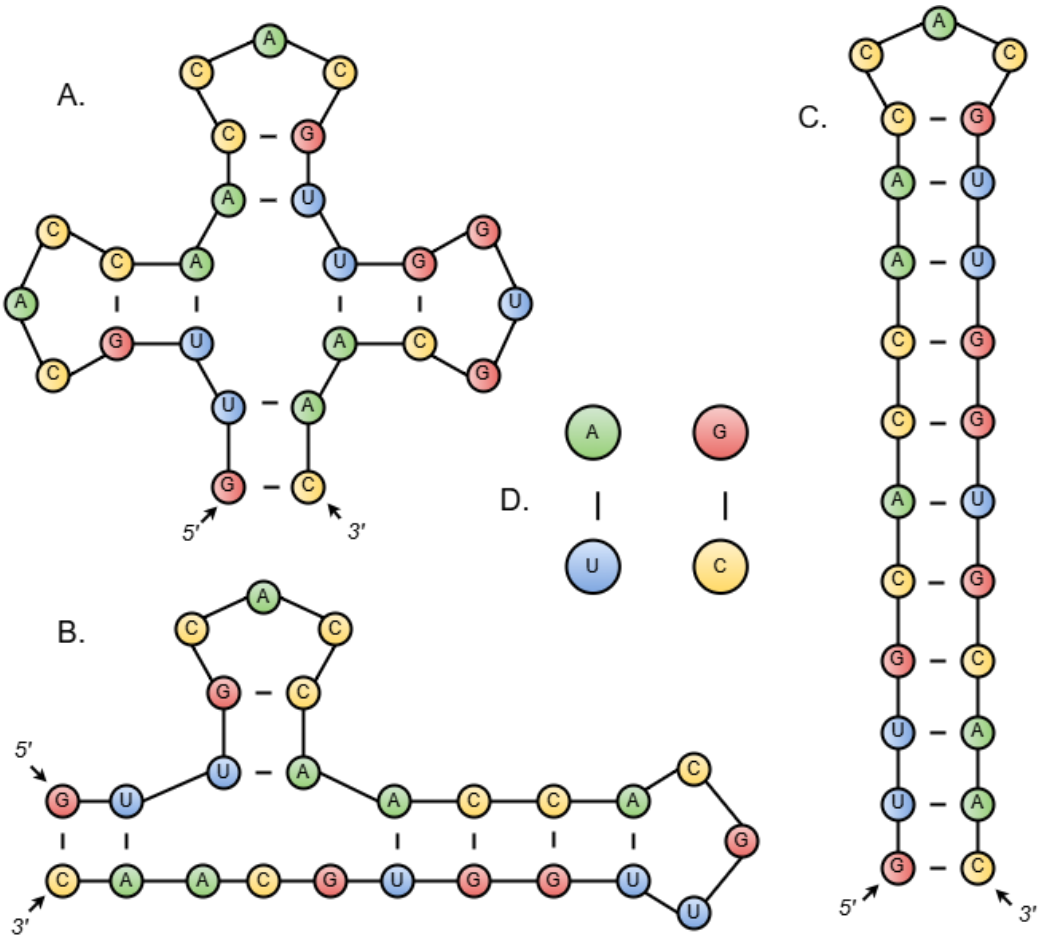



Figure 2: Different possible RNA secondary structures, for the same sequence: (A) A symmetrical RNA structure with a multi loop junction and three hairpin loops. (B) An alternative folding pattern for the same sequence showing different base-pairing arrangements consisting of a multi loop junction and 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.

### 2.1 Thermodynamic-Based Approaches

The field of prediction of RNA structures has evolved significantly since the Nussinov algorithm (Nussinov et al., 1978) (Nussinov & Jacobson, 1980). Approaches such as Mfold (Zuker & Stiegler, 1981) rely on minimum free energy calculations of RNA folding, using well-defined thermodynamic parameters to predict the most stable structure. Mfold has expanded on this by allowing users to explore suboptimal structures within a specified free energy range, visualized through energy dot plots (Zuker, 2003). Similarly, RNAfold (Hofacker et al., 1994) (Hofacker, 2003) can predict RNA secondary structures based on thermodynamics while integrating partition function calculations to estimate base-pair probabilities. However, the computational complexity of partition function calculations can be a challenge for larger RNA sequences. Furthermore, these approaches that rely on thermodynamic parameters can lead to reduced accuracy when parameters are unavailable or incomplete for various RNA molecules. Additionally, these methods assume that RNA has a single, most stable structure. This does not account for the alternative conformations that might also exist in real biological environments.

### 2.2 MACHINE LEARNING-BASED APPROACHES

To overcome the limitations of thermodynamics-based approaches, especially when experimental data from wet-laboratory experiments are unavailable, machine learning-based methods have been widely used. These methods learn parameters directly from training data, which consists of RNA sequences and their secondary structures, and can lead to developing more accurate predictive models with an increased number of parameters (Sato & Hamada, 2023). For instance, while CONTRAfold Do et al., 2006 uses around 300 parameters, ContextFold (Zakov et al., 2011) used a much larger parameter set, which enabled it to achieve highly accurate RNA secondary structure predictions. However, subsequent studies have shown that the rich parameterization of ContextFold makes it prone to overfitting, leading to reduced prediction accuracy when applied to new data (Rivas et al., 2012) (Rivas, 2013).

### 2.3 DEEP LEARNING-BASED APPROACHES

Recent advancements in deep learning have addressed many of the limitations faced by earlier machine learning methods by leveraging both thermodynamic and machine learning-based approaches. For example, MXfold2 (Sato et al., 2021) uses deep learning to calculate scores for loop substructures combined with Turner's energy parameters (thermodynamic-based method) resulting in highly accurate secondary structure prediction with a reduced risk of overfitting. Furthermore, Linearfold (Huang et al., 2019) uses a dynamic programming algorithm for RNA folding that speeds up runtime and interestingly has led to more accurate predictions on the long sequence families, such as 16S and 23S Ribosomal RNAs, as well as improved accuracies for long-range base pairs, such as those that are 500+ nucleotides apart.

Building on these advancements, our project aims to develop a deep learning model for RNA secondary structure prediction by leveraging insights from previous methods and incorporating domain-specific knowledge. Our approach attempts to improve prediction accuracy while mitigating issues like overfitting.

## 3 PRELIMINARY WORK: FINE-TUNING BERT

In our initial approach, we leveraged BERT, a transformer-based model, for RNA secondary structure prediction. RNA often forms motifs, which are smaller regions that fold locally within themselves, and these motifs determine secondary structure (Hendrix et al., 2005). Transformers, like BERT, are well-suited for local context learning, making them a promising choice for RNA secondary structure prediction. We treated the problem as a sequence-to-sequence task, where RNA sequences served as the input and their dot-bracket representations as the output.

### 3.1 METHODOLOGY

#### 3.1.1 DATA PREPROCESSING

RNA nucleotides and its corresponding dot-bracket notations were mapped to unique numerical indices. Also, additional mappings were made for characters, such as opening and closing bracket pairs, to capture more relationships in the data. We accounted for unknown tokens by assigning specific indices. Then, RNA sequences and their dot-bracket notations were tokenized and truncated to a maximum sequence length of 512 tokens due to the input constraints of BERT. We applied padding to make sequences one, uniform length, and attention masks were generated to distinguish between meaningful tokens and padding.

#### 3.1.2 MODEL AND TRAINING SETUP

The fine-tuned model was based on BertForTokenClassification, which is a pre-trained BERT model used for token-level predictions. The tokenizer was augmented with RNA-specific tokens, and we resized the model's embeddings to account for the expanded vocabulary. To address the issue of class imbalance, a focal loss function was implemented instead of cross-entropy loss because it focuses on samples that are harder to classify by penalizing errors more heavily for underrepresented classes.

We trained the model for three epochs using the AdamW optimizer with a learning rate of $5 \times 10^{-5}$. The input sequences, attention masks, and corresponding labels were passed through the model in mini-batches of size 32. For each batch, the model computed logits, which were then compared to the ground truth labels using the focal loss function. Also, the gradients were calculated through backpropagation, and the model parameters were updated iteratively.

### 3.2 EVALUATION

We performed the evaluation of the model on a small subset of the dataset. We evaluated the model using accuracy, calculated as the proportion of correctly predicted labels. Also, we printed out a classification report, which showed how the model performs across all structural classes to help assess its effectiveness in predicting RNA secondary structures.

### 3.3 RESULTS

Two variations of the model, one with cross-entropy loss and one with focal loss, were evaluated on the dataset. They achieved accuracies of 54.42% and 65.33%, respectively.

### 3.4 LIMITATIONS

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

Despite these challenges, these experiments provided valuable insights into the complexities of the task, which led us to develop a custom model.

## 4 METHODOLOGY

### 4.1 CUSTOM 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 designed for processing the RNA sequence.

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.

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.

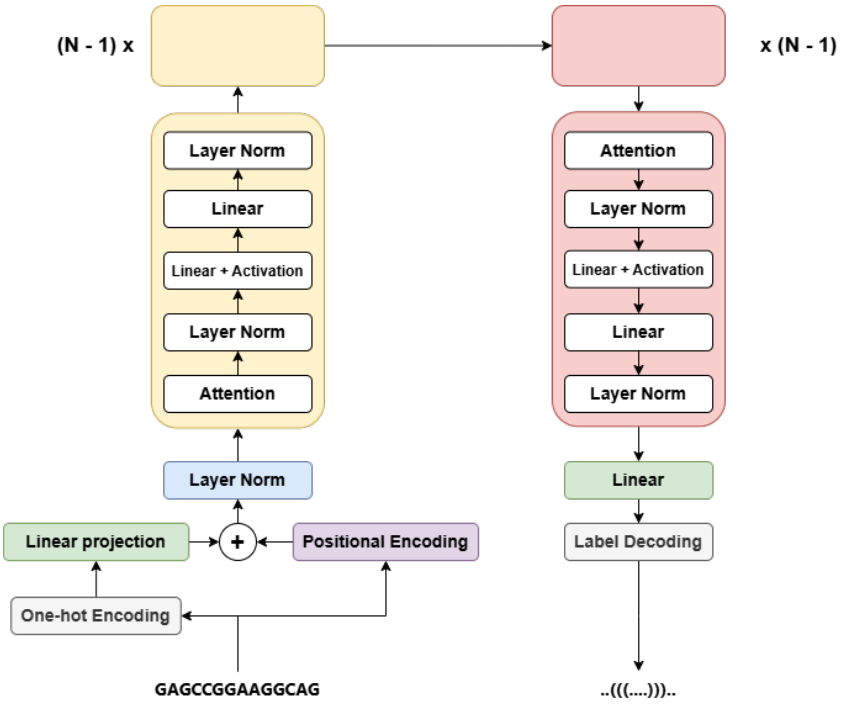



Figure 3: FoldFormer architecture diagram showing the complete encoder-decoder pipeline. The encoder (left, yellow) processes the input RNA sequence through multiple transformation layers, while the decoder (right, pink) generates the structural prediction. Key components include: attention mechanisms 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 (.(((...))).))

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.

### 4.2 Objective Function

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

$$\mathcal{L} = -\frac{1}{T}\sum_{t=1}^{T}\begin{cases} p_t \log(\hat{p}_t), & \text{if } p_t \text{ is not base-paired,} \\ p_{t_i}\log(\hat{p}_{t_i}) + p_{t_j}\log(\hat{p}_{t_j}), & \text{if } p_{t_i} \text{ is base-paired with } p_{t_j}. \end{cases} \tag{1}$$

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.

# 5 EXPERIMENTS AND RESULTS

## 5.1 DATASET

We utilized a dataset of 100K samples from various RNA families (Danaee et al., 2018). The data includes diverse structural motifs and varying sequence lengths. Processing 10% of the dataset requires approximately one hour of computation time.

## 5.2 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.

## 5.3 RESULTS

We used the results reported in Szikszai et al. (2022) as a benchmark for comparison. This paper provided F1 scores across several RNA families, such as 5S rRNA, SRP RNA, and tRNA. 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 can indeed generalize across RNA families, rejecting the need for family-specific considerations in model design.

| Family | RNAstructure | ReLU | MXfold2 | GeLU | UFold |
|---|---|---|---|---|---|
| 5S_rRNA | **0.63** | 0.50 | 0.54 | 0.48 | 0.53 |
| SRP | **0.64** | 0.49 | 0.5 | 0.46 | 0.26 |
| tRNA | **0.8** | 0.72 | 0.64 | 0.52 | 0.26 |
| tmRNA | 0.43 | 0.48 | 0.46 | **0.49** | 0.4 |
| RNaseP | **0.55** | 0.46 | 0.51 | 0.48 | 0.41 |
| group_I_intron | **0.53** | 0.47 | 0.45 | 0.48 | 0.45 |
| 16S_rRNA | **0.58** | 0.48 | 0.55 | 0.49 | 0.41 |
| telomerase | 0.5 | 0.45 | 0.34 | 0.47 | **0.8** |
| 23S_rRNA | **0.73** | 0.46 | 0.64 | 0.47 | 0.45 |
| Overall | **0.6** | 0.52 | 0.51 | 0.48 | 0.44 |

Figure 4: Comparative Evaluation of EteRNAity and State-of-the-Art Models

# 6 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.

# 7 CONTRIBUTIONS

This project was a collaborative effort. Asma's contributions included identifying and sourcing the dataset used for evaluating our custom model, conducting the related works review, and fine-tuning

BERT models for RNA secondary structure prediction. These efforts provided critical insights into the challenges of the task and laid the groundwork for the development of our custom model. Biswajit's contributions included designing, implementing, and evaluating the custom model. This involved creating an architecture to address the limitations of the BERT models, training the model, and evaluating the model against the state-of-the-art methods, as described in the Related Works section. Together, our contributions were complementary. This collaboration allowed us to iteratively improve our approach and achieve the project goals.

## 8 SOURCE CODE

Our code can be found on this github repository: https://github.com/Biswajit-Banerjee/Eternal

## REFERENCES

Padideh Danaee, Mason Rouches, Michelle Wiley, Dezhong Deng, Liang Huang, and David Hendrix. bprna: large-scale automated annotation and analysis of rna secondary structure. *Nucleic acids research*, 46(11):5381–5394, 2018.

Chuong B Do, Daniel A Woods, and Serafim Batzoglou. Contrafold: Rna secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):e90–e98, 2006.

Donna K Hendrix, Steven E Brenner, and Stephen R Holbrook. Rna structural motifs: building blocks of a modular biomolecule. *Quarterly reviews of biophysics*, 38(3):221–243, 2005.

Ivo L Hofacker. Vienna rna secondary structure server. *Nucleic acids research*, 31(13):3429–3431, 2003.

Ivo L Hofacker, Walter Fontana, Peter F Stadler, L Sebastian Bonhoeffer, Manfred Tacker, Peter Schuster, et al. Fast folding and comparison of rna secondary structures. *Monatshefte fur chemie*, 125:167–167, 1994.

Liang Huang, He Zhang, Dezhong Deng, Kai Zhao, Kaibo Liu, David A Hendrix, and David H Mathews. Linearfold: linear-time approximate rna folding by 5'-to-3'dynamic programming and beam search. *Bioinformatics*, 35(14):i295–i304, 2019.

Ruth Nussinov and Ann B Jacobson. Fast algorithm for predicting the secondary structure of single-stranded rna. *Proceedings of the National Academy of Sciences*, 77(11):6309–6313, 1980.

Ruth Nussinov, George Pieczenik, Jerrold R Griggs, and Daniel J Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied mathematics*, 35(1):68–82, 1978.

Elena Rivas. The four ingredients of single-sequence rna secondary structure prediction. a unifying perspective. *RNA biology*, 10(7):1185–1196, 2013.

Elena Rivas, Raymond Lang, and Sean R Eddy. A range of complex probabilistic models for rna secondary structure prediction that includes the nearest-neighbor model and more. *RNA*, 18(2): 193–212, 2012.

Kengo Sato and Michiaki Hamada. Recent trends in rna informatics: a review of machine learning and deep learning for rna secondary structure prediction and rna drug discovery. *Briefings in Bioinformatics*, 24(4):bbad186, 2023.

Kengo Sato, Manato Akiyama, and Yasubumi Sakakibara. Rna secondary structure prediction using deep learning with thermodynamic integration. *Nature communications*, 12(1):941, 2021.

Marcell Szikszai, Michael Wise, Amitava Datta, Max Ward, and David H Mathews. Deep learning models for rna secondary structure prediction (probably) do not generalize across families. *Bioinformatics*, 38(16):3892–3899, 2022.

Shay Zakov, Yoav Goldberg, Michael Elhadad, and Michal Ziv-Ukelson. Rich parameterization improves rna structure prediction. *Journal of Computational Biology*, 18(11):1525–1542, 2011.

Michael Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic acids research*, 31(13):3406–3415, 2003.

Michael Zuker and Patrick Stiegler. Optimal computer folding of large rna sequences using thermodynamics and auxiliary information. *Nucleic acids research*, 9(1):133–148, 1981.