PriFold: Biological Priors Improve RNA Secondary Structure Predictions

Chenchen Yang[1,2*], **Hao Wu**[1,2*], **Tao Shen**[3], **Kai Zou**[4], **Siqi Sun**[1,2†]

[1]Research Institute of Intelligent Complex Systems, Fudan University
[2]Shanghai Artificial Intelligence Laboratory
[3]Zelixir Biotech
[4]NetMind.AI
{ccyang21, hwu24}@m.fudan.edu.cn, siqisun@fudan.edu.cn

Abstract

Predicting RNA secondary structures is crucial for understanding RNA function, designing RNA-based therapeutics, and studying molecular interactions within cells. Existing deep-learning-based methods for RNA secondary structure prediction have mainly focused on local structural properties, often overlooking the global characteristics and evolutionary features of RNA sequences. Guided by biological priors, we propose PriFold, incorporating two key innovations: 1) improving attention mechanism with pairing probabilities to utilize global pairing characteristics, and 2) implementing data augmentation based on RNA covariation to leverage evolutionary information. Our structured enhanced pretraining and finetuning strategy significantly optimizes model performance. Extensive experiments demonstrate that PriFold achieves state-of-the-art (SOTA) results in RNA secondary structure prediction on benchmark datasets such as bpRNA, RNAStrAlign and ArchiveII. These results not only validate our prediction approach but also highlight the potential of integrating biological priors, such as global characteristics and evolutionary information, into RNA structure prediction tasks, opening new avenues for research in RNA biology and bioinformatics.

Code — https://github.com/BEAM-Labs/PriFold

Introduction

Ribonucleic acid (RNA) is a crucial molecule involved in many biological functions like protein synthesis and controlling gene expression. RNA is made up of nucleotides with four different types of nitrogenous bases: adenine (A), cytosine (C), guanine (G), and uracil (U) (Cleaves 2011). These bases are the fundamental building blocks that form the RNA sequence, also known as primary structure. The bases are capable of forming bonds with each other, creating specific base pairs that define the RNA's secondary structure. The secondary structure of RNA based on these pairings can be depicted as a binary contact matrix, where a value of 1 indicates the presence of a base pair between nucleotides, and a value of 0 indicates no pairing. In secondary structure studies, we normally consider canonical pairings: Watson-Crick (WC) and Wobble. WC pairings, involving A pairing with U and G pairing with C, form the backbone of RNA structures through hydrogen bonds. On the other hand, Wobble pairings, between G and U, introduce a more flexible pairing mechanism. Such flexibility is crucial in genetic translation processes (Varani and McClain 2000). These pairings underpin the complex folding and functional capabilities of RNA, making their study essential for understanding and manipulating RNA behavior (Reuter and Mathews 2010).

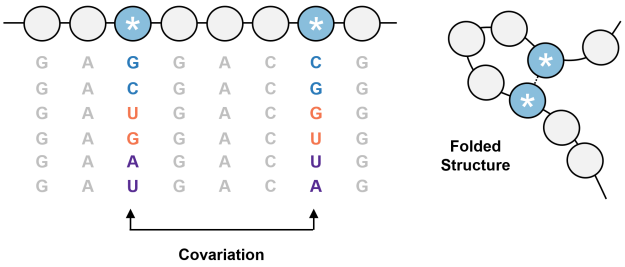

Figure 1: Visualization of RNA covariation. The left panel displays six aligned RNA sequences with coordinated nucleotide changes (highlighted in color) that preserve base pairing. The right panel shows the folded RNA secondary structure, demonstrating how covariation maintains structural integrity despite sequence changes.

Experimental methods for determining RNA secondary structures, such as nuclear magnetic resonance (NMR) spectroscopy (Fürtig et al. 2003), enzymatic probing (Cheong et al. 2004), and cryo-electron microscopy (Fica 2020), are valuable but often suffer from low throughput and high costs, rendering them impractical for large-scale studies. As a result, computational prediction methods have gained prominence due to their efficiency and scalability.

Traditional computational approaches, like dynamic programming (DP) algorithms, are rooted in energy minimization principles, where the energy function is derived from thermodynamic experiments or learned from empirical data (Waterman and Smith 1978; Reuter and Mathews 2010; Janssen and Giegerich 2015; Lorenz et al. 2011). These algorithms aim to compute the most stable RNA secondary structure from an enormous space of possibilities, a task known to be NP-complete (McDonald et al. 2005). How-

*These authors contributed equally to this work.
†Corresponding author.

ever, DP algorithms typically assume that RNA structures form nested configurations, where base pairs are constrained to stack without crossing, thereby overlooking complex but biologically significant structures like pseudoknots. A pseudoknot is an RNA structure that is minimally composed of two helical segments connected by single-stranded regions or loops, creating sophisticated folding patterns essential for RNA function and regulation (Staple and Butcher 2005).

Recently, deep-learning-based prediction methods have emerged as a promising approach to address the limitations of traditional DP algorithms. SPOT-RNA (Singh et al. 2019) uses ResNet (He et al. 2015) and LSTM (Hochreiter and Schmidhuber 1997) networks for end-to-end RNA secondary structure prediction, while E2Efold (Chen et al. 2020) and RFold (Tan, Gao, and Li 2022) integrate several hard constraints of secondary structure directly into the prediction process using an unrolled algorithm. Ufold (Fu et al. 2022) separates these hard constraints from the prediction module, applying them in post-processing to allow for a more flexible model design with U-Net (Ronneberger, Fischer, and Brox 2015). Large-scale models including RNAErnie (Wang et al. 2024), RiNALMo (Penić et al. 2024), RNA-FM (Chen et al. 2022), and ERNIE-RNA (Yin et al. 2024) leverage extensive RNA sequence databases to learn intrinsic patterns and structural characteristics through self-supervised pretraining.

While these deep learning approaches have achieved remarkable gains in prediction accuracy over traditional DP methods, they have primarily focused on local structural properties by incorporating constraints, such as WC and Wobble pairing rules, non-crossing base pairs, and maximum diagonal distance, to ensure biologically plausible RNA secondary structures (Chen et al. 2020), but often neglect other important biological properties. They have not taken into account the global structural information or evolutionary information. For example, global structural information includes that WC pairings have significantly higher formation probabilities than Wobble pairings. Evolutionary information like covariation, as shown in Figure 1, indicates that if both bases in a base pair simultaneously mutate to other bases that can still pair, the secondary structure remains unchanged (Parsch, Braverman, and Stephan 2000). These elements are crucial for a comprehensive understanding and accurate prediction of RNA structures. Therefore, we integrate such biological priors into our method.

In this paper, we present PriFold, a novel RNA secondary structure prediction method optimized by incorporating biological priors, including global pairing characteristics such as pairing frequencies and evolutionary features like RNA covariation. We summarize the main contributions of this work as follows:

- PriFold enhances the prediction pipeline by incorporating pairing probabilities into the attention mechanism with a positional factor, aligning predictions more closely with biological realities.
- PriFold adopts a novel data augmentation strategy based on RNA covariation, which enriches the training dataset by capturing conserved features throughout evolution.
- We incorporate several biological task-related modules and demonstrate that PriFold achieves SOTA performance on multiple datasets consistently.

Related Work

Energy-Based Dynamic Programming Methods

Traditional RNA secondary structure prediction algorithms, such as RNAfold (Lorenz et al. 2011), Mfold (Zuker 2003), RNAstructure (Reuter and Mathews 2010), and CONTRAfold (Do, Woods, and Batzoglou 2006), use DP techniques to minimize energy based on thermodynamic data acquired from experiments. However, these parameters are limited by the conditions under which the experiments are conducted, which may not capture all the complexities of RNA interactions under different biological conditions (Shi et al. 2014).

A more advanced approach, MXfold2 (Sato, Akiyama, and Sakakibara 2021), tries to improve these energy parameters using deep learning techniques. Then these learned parameters are integrated back into the DP framework. Despite these improvements, all these energy-based DP methods share a major limitation: they struggle to accurately predict non-nested secondary structures. This limitation reduces their effectiveness, especially when dealing with complex RNA structures that do not fit in the traditional nested patterns (Chen et al. 2020).

Deep Learning-Based Methods

In response to the limitations of DP methods, recent advancements have seen the development of deep-learning-based approaches to avoid using traditional DP frameworks. SPOT-RNA (Singh et al. 2019), for example, integrates LSTM networks with ResNet architectures, exhibiting a significant improvement over prior DP-based methods. It employs a transfer learning strategy, initially utilizing the extensive bpRNA (Danaee et al. 2018) dataset before refining with the smaller, yet more precise, PDB (Bank 1971) dataset. RNAformer (Franke, Runge, and Hutter 2023) uses axial attention and recycling to construct a scalable model.

Following these footsteps, E2Efold (Chen et al. 2020) introduces a post-processing technique that enforces hard constraints on the predicted RNA structures, ensuring that the predicted configurations adhere more closely to biologically plausible structures. Such technique has been influential, with subsequent methods like UFold (Fu et al. 2022) and RFold (Tan, Gao, and Li 2022) adopting similar strategies. UFold further innovates by incorporating a U-Net architecture, known for its efficiency in handling spatial hierarchies in data.

Pretrained Language Models

Pretrained language models (LMs) have revolutionized natural language processing (NLP) by dramatically improving machines' ability to understand and generate human language. Prominent models such as BERT (Devlin et al. 2018) and GPT (Floridi and Chiriatti 2020) leverage vast amounts of text data to learn complex linguistic patterns. These models are based on the transformer architecture, which employs

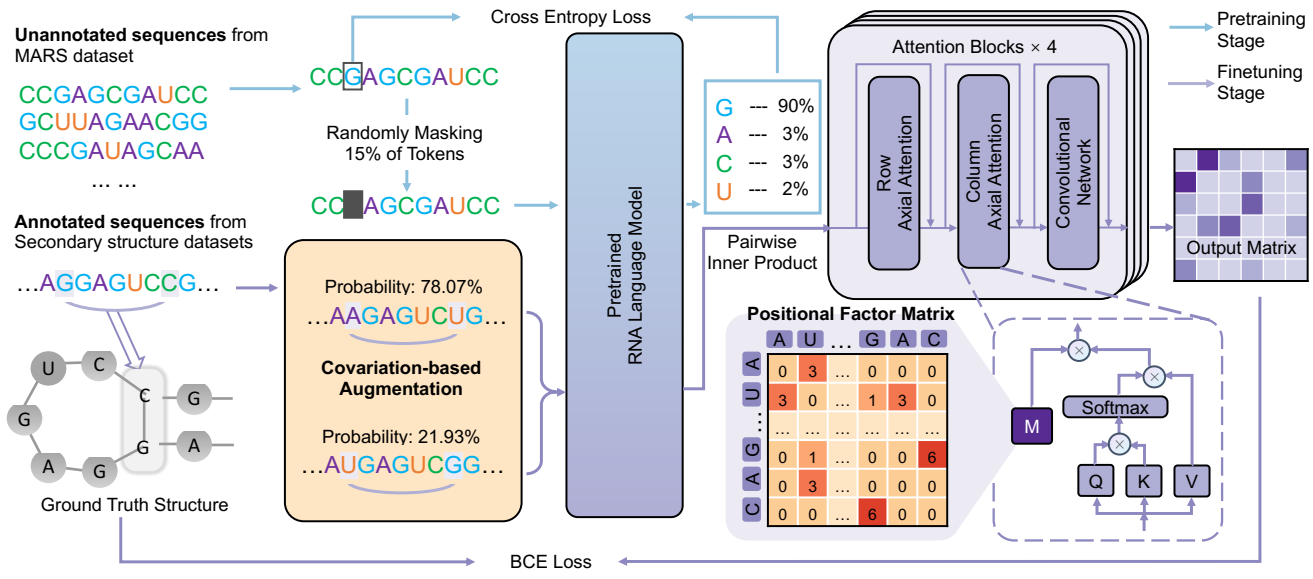

Figure 2: Model architecture of PriFold. The training process is divided into two stages: pretraining (blue arrows) and finetuning (purple arrows). The model is enhanced by two key innovations: (1) A data augmentation strategy based on RNA covariation, illustrated in the bottom-left of the figure, which generates sequence variants while maintaining structural integrity. (2) A positional factor matrix (bottom-right) that encodes pairing characteristics between nucleotides in the axial attention mechanism.

mechanisms like attention (Vaswani et al. 2017) to capture contextual relationships within the text. As a result, they provide a robust foundation for a variety of tasks, including text classification, question answering, and sentiment analysis, significantly outperforming earlier methods that relied on hand-engineered features or shallower learning architectures.

Recently, pretrained language models have been increasingly applied to RNA structure prediction, showcasing their potential in this domain. Models such as RNAErnie (Wang et al. 2024), RiNALMo (Penić et al. 2024), RNA-FM (Chen et al. 2022), and ERNIE-RNA (Yin et al. 2024) utilize large-scale RNA sequence datasets for pretraining, enabling them to capture RNA-specific patterns and structural features. By leveraging these pretrained representations, these models have achieved remarkable success in tasks like RNA secondary structure prediction.

Methods

This paper presents a novel RNA secondary structure prediction framework, PriFold, which aims to utilize biological priors. The training process is divided into two main stages: pretraining and finetuning. The overall framework of PriFold is illustrated in Figure 2.

Pretraining Stage

In the pretraining stage, we utilize unannotated sequences for the initial training of the model. Sequences are fed through an input embedding layer into the encoding module. The encoding module follows the Llama2-encoder style. Specifically, it is a model with 150 million parameters that consists of 12 Transformer blocks, each including RMS normalization, multi-head attention (MHA), and feedforward neural networks. The model is trained on MARS (Chen et al. 2023), a massive dataset containing 1.1 billion RNA sequences. The objective of the pretraining stage is Masked Language Modeling (MLM), which aims to minimize prediction error through a cross-entropy loss function, thereby enabling the model to learn latent representations of the sequences. The specific formula is shown as follows:

$$\mathcal{L} = \sum_{i \in \mathcal{M}} -\log P(x_i | X_{\setminus \mathcal{M}}) \tag{1}$$

where $\mathcal{M}$ is the index set of all masked tokens. The specific hyperparameters of the pretraining stage are detailed in the supplementary materials.

Finetuning Stage

In the finetuning stage, sequences annotated with secondary structures are introduced for further training. First, we utilize covariation-based data augmentation to enrich training samples while maintaining structural integrity. Then, we feed the augmented sequences through a deep network consisting of four axial attention blocks, proposed by Franke's (Franke, Runge, and Hutter 2023). These attention blocks encompass row axial attention, column axial attention, and convolutional networks. We utilize a positional factor matrix to enhance the model's understanding of pairing rules. The following paragraphs detail our finetuning approach.

Covariation-Based Data Augmentation First, the annotated sequences go through an augmentation process guided

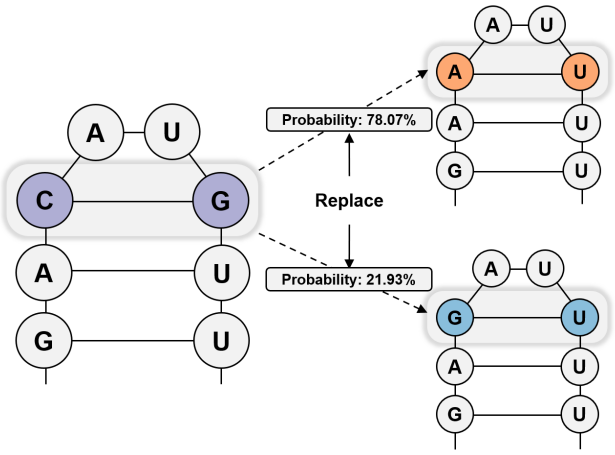

Figure 3: Data augmentation technique based on RNA covariation. The image shows how a single RNA structure can be augmented to create two variant sequences while maintaining the overall structural integrity. On the left is the original RNA structure with a C-G base pair. This base pair is then replaced with two possible alternatives: an A-U pair (with a 78.07% probability) and a G-U pair (with a 21.93% probability). The probabilities are derived from (3)

by RNA covariation. RNA covariation refers to the ability of RNA structures to adapt and persist throughout evolution by compensatory base pair changes that occur between covarying nucleotides within paired regions. This phenomenon involves the simultaneous mutation of paired nucleotides without altering the structure (Figure 1). For example, an A-U pair may be replaced by a G-C pair as a result of covariation. RNA covariation is an important biological phenomenon, emphasizing the flexibility of RNA structure and its ability to retain functionally essential regions (Rivas 2023). By taking advantage of covariation properties, RNA secondary structure datasets can be expanded without losing accuracy. Such expansion can be done by replacing paired nucleotides simultaneously, simulating the covariation process.

To accurately simulate the covariations between paired nucleotides in RNA secondary structures, we have proposed a concise substitution process. Our approach begins by randomly selecting 10% of sequences from each minibatch within the given dataset. Focusing on these selected sequences, we then randomly choose a certain proportion of paired nucleotides for substitution, determined by a replacement ratio α. The specific implementations are as follows.

Given an RNA sequence $X = (x_1, ..., x_L)$, where $x_i \in \{A, U, C, G\}$, its secondary structure is a contact map represented as a matrix $M \in \{0, 1\}^{L \times L}$. We define $\mathcal{B} := \{\text{AU, UA}\} \cup \{\text{GC, CG}\} \cup \{\text{GU, UG}\}$ to denote all possible canonical pairings (including WC and Wobble pairings). For any base pair $x_i x_j$ where $x_i x_j \notin \mathcal{B}$, $M_{ij} = 0$.

The substitution process is meticulously structured around the pairing frequencies, detailed in Table 1. Since Table 1 contains all pairings, we first normalize the probabilities to $\mathcal{B}$ only. Given RNA pairing frequency listed in Table 1, we obtain the pairing probability of y, where $y \in \mathcal{B}$, *e.g.*, $P_{GU} = 7.24\%$. We can then normalize these probabilities as follows:

	A	C	G	U
A	3.31	4.64	9.17	**25.77**
C	–	0.29	**46.3**	0.45
G	–	–	1.45	**7.24**
U	–	–	–	1.37

Table 1: Pairing frequency of nucleotide bases (%). Bold values indicate the three types of base pairs considered: WC pairs (A-U and G-C) and Wobble pair (G-U). The table is diagonally symmetric, so only the upper triangle is shown.

$$p'_y = \frac{p_y}{\sum_{z \in \mathcal{B}} p_z} \tag{2}$$

where p_y represents the pairing frequencies for y. p'_y represents the normalized probability for the specific base pair y, where the prime symbol ($'$) indicates normalization. Then we can design a transition probability formula given the normalized probabilities:

$$p'(x_i x_j \to y) = \frac{p'_y}{1 - p'_{x_i x_j}}, y \in \mathcal{B},\ y \neq x_i x_j \tag{3}$$

Here, x_i and x_j represent the original paired nucleotides, and y denotes a possible nucleotide pair from the set $\mathcal{B}$, excluding the original pair. This formula ensures that the substitutions are proportional to the realistic probabilities of nucleotide pairing, thus maintaining biological fidelity. The final replacement strategy is visualized in Figure 3.

Axial Attention Following data augmentation and processing by the pretrained language model, the input goes through the axial attention module. Given an input matrix $x \in \mathbb{R}^{h \times w \times d}$, where h is the number of rows, w is the number of columns, and d is the feature dimension, axial attention is divided into row axial attention and column axial attention steps, where we first compute the query (q), key (k), and value (v) matrices, and then calculate the attention weights and the output as follows:

$$o_{\{\text{row, col}\}} = \text{softmax}\left(\frac{q_{\{\text{row, col}\}} k^T_{\{\text{row, col}\}}}{\sqrt{d}}\right) v_{\{\text{row, col}\}} \tag{4}$$

By decomposing global 2D attention into two 1D attention operations, axial attention significantly reduces computational complexity (Ho et al. 2019).

Positional Factor The axial attention is enhanced by a positional factor guided by pairing frequency directly. Table 1 shows the frequency of various base pairs in the PDB (Lawson et al. 2024). According to (2), we can obtain the normalized probability p'_y.

To integrate these pairing statistics into the training process, we construct a positional factor matrix M', which directly modifies the attention score calculated in the axial attention. The elements of the matrix M' are computed for

each pair of positions within the sequences as follows:

$$M'_{ij} = \begin{cases} \lambda p'_{x_i x_j} + 1 & \text{if } x_i x_j \in \mathcal{B} \\ 1 & \text{if } x_i x_j \notin \mathcal{B} \end{cases}, \tag{5}$$

where x_i and x_j are the nucleotides at positions i and j of the sequence, respectively. λ is a hyperparameter that scales the probability, allowing adjustments to the influence of the probability term on the final matrix value.

After performing QK in the attention mechanism, we multiply the positional factor matrix M' to the attention weight, thereby computing the final attention output o' by:

$$o'_{\{\text{row, col}\}} = o_{\{\text{row, col}\}} \cdot M' \tag{6}$$

Experiments

We carry out a series of experiments to evaluate our proposed PriFold model against the leading and widely used approaches in RNA secondary structure prediction. Our evaluation encompasses various experimental setups, such as standard RNA secondary structure prediction task, prediction with pseudoknots, assessments of generalization capability, and large-scale benchmark experiments. In addition, we present ablation studies to analyze the contributions of different components of our model. Links to the datasets used in our experiments are provided in the supplementary materials.

Datasets Our study utilizes three benchmark datasets to evaluate the performance of our model.

To establish a standard benchmark for comparison, we use RNAStrAlign (Tan et al. 2017) as our primary dataset. It is one of the most comprehensive collections of RNA structures, comprising 37,149 structures from 8 RNA types. We split RNAStrAlign into training set and test set following UFold, E2Efold, and MXFold2.

To assess our model's generalization capabilities, we include ArchiveII (Mathews 2019), a widely used testing dataset in classical RNA folding methods, containing 3,975 RNA structures from 10 RNA types. This dataset serves as an additional independent test set to evaluate our model's performance across a broad range of RNA types.

Lastly, to leverage a large-scale dataset for comprehensive training and evaluation, we incorporate bpRNA (Danaee et al. 2018) dataset, an extensive benchmark containing 13,419 RNA sequences after excluding those with $> 80\%$ sequence identities. We divided bpRNA into training, validation and test set following SPOT-RNA.

Standard RNA Secondary Structure Prediction

First, we perform standard RNA secondary structure prediction on the RNAStrAlign dataset, training on its training set and evaluating on its test set. The results presented in Table 2 demonstrate the superior performance of our proposed method PriFold compared to existing approaches in RNA secondary structure prediction. Our model achieves an impressive F1 score of 0.988, surpassing all other methods in the comparison.

Method	Precision	Recall	F1
MFold	0.450	0.398	0.420
CONTRAfold	0.608	0.663	0.633
RNAstructure	0.537	0.568	0.550
RNAfold	0.516	0.568	0.540
LinearFold	0.620	0.606	0.609
E2Efold	0.866	0.788	0.821
UFold	0.905	0.927	0.915
RFold	<u>0.981</u>	<u>0.973</u>	<u>0.977</u>
PriFold	**0.988**	**0.989**	**0.988**

Table 2: Performances on the RNAStrAlign dataset. The highest score in each column is shown in bold, while the second highest is underlined. This applies to the following tables.

Prediction with Pseudoknots

We evaluate the model's capability to handle RNA sequences containing pseudoknots by extracting sequences with pseudoknots from the RNAStrAlign test set. We test the prediction result directly on the models trained on RNAStrAlign training set, and calculate the precision, recall, and F1 score on sequences containing pseudoknots.

Method	Precision	Recall	F1
UFold	0.453	0.509	0.478
SPOT-RNA	0.893	<u>0.910</u>	0.902
RFold	<u>0.932</u>	0.907	<u>0.918</u>
PriFold	**0.965**	**0.928**	**0.944**

Table 3: Performances on pseudoknot prediction in RNAStrAlign test set.

As shown in Table 3, PriFold demonstrates superior performance across all metrics, achieving a precision of 0.965, recall of 0.928, and F1 score of 0.944. This consistent improvement over previous methods validates the effectiveness of our method in capturing complex pseudoknot structures.

Generalization Evaluation

In order to examine the model's generalization abilities, we conduct experiments on testing dataset ArchiveII without any further adjustment, using the model trained with RNAStrAlign.

The results(Table 4) demonstrate that PriFold not only outperforms traditional thermodynamic models but also achieves higher scores than other advanced machine-learning-based approaches. Specifically, PriFold achieved an F1 score of 0.952, which is a significant improvement over other methods. The competitive gain in F1 score on the ArchiveII dataset, following training on a completely different set (RNAStrAlign), indicates that PriFold is not only precise in its predictions but also versatile across different RNA datasets.

Method	Precision	Recall	F1
MFold	0.668	0.590	0.621
RNAsoft	0.665	0.594	0.622
CONTRAfold	0.695	0.651	0.665
RNAstructure	0.664	0.606	0.628
Contextfold	0.873	0.821	0.842
RNAfold	0.663	0.613	0.631
SPOT-RNA	0.743	0.726	0.711
LinearFold	0.724	0.605	0.647
E2Efold	0.734	0.660	0.686
MXfold2	0.788	0.760	0.768
Eternafold	0.667	0.622	0.636
UFold	0.887	0.928	0.905
RFold	0.938	0.910	0.921
PriFold	**0.962**	**0.947**	**0.952**

Table 4: Performances on the ArchiveII dataset.

Large-Scale Benchmark Evaluation

To further validate our model's performance, we conduct experiments on the bpRNA dataset. We train the model on bpRNA-TR0 and evaluate the model on bpRNA-TS0. We compare PriFold with both traditional RNA structure prediction methods and recent deep learning approaches, including those based on RNA language models. The experimental results are reported in Table 5. Our proposed method, PriFold, outperforms all existing approaches across all metrics, achieving the highest precision (0.802), recall (0.756), and F1 score (0.770).

Method	Precision	Recall	F1
CONTRAfold	0.528	0.655	0.567
RNAstructure	0.494	0.622	0.533
SPOT-RNA	0.594	0.693	0.619
LinearFold	0.561	0.581	0.550
E2Efold	0.140	0.129	0.130
MXfold2	0.519	0.646	0.558
Eternafold	0.516	0.666	0.563
UFold	0.521	0.588	0.553
RFold	0.692	0.635	0.644
RNAErnie	0.575	0.678	0.622
RNA-FM	0.718	0.713	0.704
RiNALMo	0.784	0.730	0.747
ERNIE-RNA	0.780	0.735	0.748
PriFold	**0.802**	**0.756**	**0.770**

Table 5: Performances on the bpRNA-TS0 dataset.

Ablation Study

Overall Evaluation First, we explore the impact of omitting the covariation-based data augmentation. We train PriFold without data augmentation. Other training and testing settings remain unchanged. As shown in Table 6, the results indicate that PriFold secured a much higher F1 score compared to PriFold (w/o aug.), highlighting the importance of our covariation-based augmentation method.

To validate the effectiveness of our proposed positional factor, we also conducted experiments without both the positional factor and data augmentation. All other aspects of our training and testing settings remain the same. From Table 6, we find that excluding positional factors in addition to removing data augmentation (PriFold (w/o factor & aug.)) can result in poorer prediction performance. This contrast highlights the utility of the positional factor.

Method	Precision	Recall	F1
PriFold	0.802	**0.756**	**0.770**
- w/o aug.	**0.810**	0.710	0.737
- w/o factor & aug.	0.807	0.697	0.730

Table 6: Ablation study on the bpRNA-TS0 dataset.

Effectiveness of the Prediction Head To further validate the effectiveness of our prediction head architecture, we conducted additional experiments by applying our prediction head to RNA-FM's language model. Table 7 demonstrates the impact of our architectural choices.

Model	PriFold head	Data aug.	F1
RNA-FM	✘	✘	0.680
RNA-FM	✔	✘	0.702
RNA-FM	✔	✔	**0.744**

Table 7: Ablation study on PriFold head and data augmentation.

The results demonstrate that the two methods we proposed are effective for other secondary structure prediction methods based on language models.

Determining λ in Positional Factor In order to determine the best λ in the positional factor matrix M, we conducted experiments with different values of λ. Table 8 summarizes the results of these experiments on the bpRNA-TS0 dataset.

λ	Precision	Recall	F1
0 (w/o factor)	0.807	0.697	0.728±0.18
0.1	0.809	0.704	0.733±0.33
0.01	**0.810**	**0.710**	**0.737±0.23**
0.001	0.807	0.705	0.733±0.17

Table 8: Performances of different λ values on the bpRNA-TS0 dataset.

Including the positional factor ($\lambda > 0$) consistently improves the model's performance compared to the baseline without the positional factor ($\lambda = 0$). The model achieves the highest F1 score of 0.737 with $\lambda = 0.01$, providing the best balance between precision and recall. Decreasing λ beyond 0.01 results in diminishing returns and decreased

performance. Based on these observations, we selected $\lambda = 0.01$ for all subsequent experiments.

Effectiveness of Different Augmentation Strategies To provide a more comprehensive analysis of our data augmentation approach, we conducted an extensive comparative study examining three alternative data augmentation strategies: exclusive replacement with adenine-uracil (A-U) base pairings, exclusive replacement with cytosine-guanine (C-G) base pairings, and random replacement with arbitrary base combinations. These strategies were compared against our proposed method and a baseline without augmentation. The experimental results are presented in Table 9:

Augmentation Strategy	Precision	Recall	F1
w/o aug.	**0.810**	0.710	0.737
Random replacement	0.805	0.720	0.745
A-U exclusive	0.804	0.698	0.730
C-G exclusive	0.804	0.730	0.753
PriFold	0.802	**0.756**	**0.770**

Table 9: Performance comparison of different data augmentation strategies on the bpRNA dataset.

The detailed results demonstrate that our targeted, proportional replacement strategy based on RNA covariation significantly outperforms alternative approaches.

Hyperparameter of Covariation To explore the best settings of our proposed data augmentation technique, we conduct a series of experiments. As mentioned above, the replacement ratio α represents the proportion of paired nucleotides within selected sequences that undergo substitution. We systematically vary α from 10% to 100% in 10% increments and evaluate the model's performance of each setting.

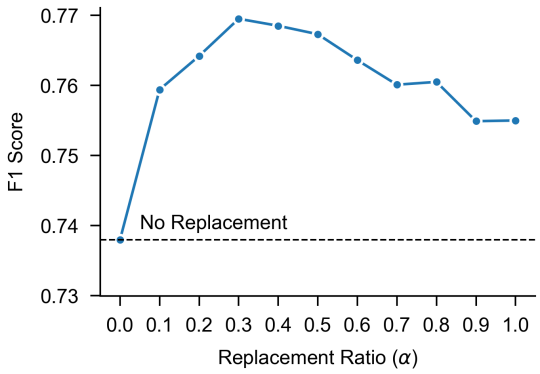

Figure 4: Ablation study on data augmentation(bpRNA-TS0 dataset).

The results of these experiments are visually represented in Figure 4, which clearly illustrates the overall improvement in model performance. Notably, the most significant gains are observed with a replacement ratio of 30%, beyond which the results diminish slightly. This suggests an optimal balance point, beyond which additional augmentation yields diminishing returns, possibly due to overfitting or the introduction of noise.

Visualization

After the quantitative evaluation of PriFold, we visualize two examples predicted by PriFold and RFold in bpRNA-TS0 to demonstrate their prediction abilities, represented as two rows in Figure 5. The two leftmost graphs show the ground truth structures, with F1 scores for each prediction labeled below the corresponding secondary structure graph. The first row shows a simple example of a nested structure, where PriFold achieves a perfect match score of 1.00. In contrast, the RFold model's prediction has some differences from the ground truth structure, with a score of 0.90. The second row illustrates a more complicated RNA secondary structure. PriFold's prediction closely resembles the ground truth, achieving a score of 0.98, while RFold's prediction deviates significantly, with a score of 0.80.

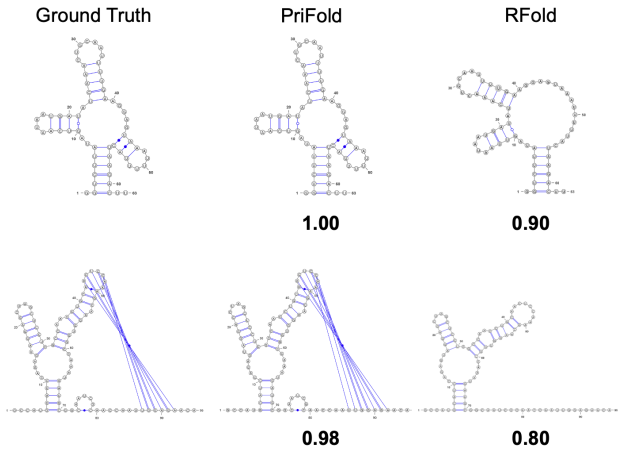

Figure 5: Visualization of the true and predicted structures.

Conclusion

In this study, we introduce PriFold, an innovative method for predicting RNA secondary structures that effectively integrates biological priors into the training process. In the light of pairing probabilities and RNA covariation, we have developed a positional factor to enhance the predicting pipeline and a novel data augmentation technique simulating RNA covariation. Comprehensive experiments demonstrate that PriFold outperforms existing RNA secondary structure prediction methods. We hope that PriFold can provide valuable insights for future developments in RNA secondary structure prediction algorithms.

Acknowledgments

This project was partially supported by Shanghai Artificial Intelligence Laboratory (S.S.).

References

Bank, P. D. 1971. Protein data bank. *Nature New Biol*, 233(223): 10–1038.

Chen, J.; Hu, Z.; Sun, S.; Tan, Q.; Wang, Y.; Yu, Q.; Zong, L.; Hong, L.; Xiao, J.; Shen, T.; King, I.; and Li, Y. 2022. Interpretable RNA Foundation Model from Unannotated Data for Highly Accurate RNA Structure and Function Predictions. ArXiv:2204.00300 [q-bio].

Chen, K.; Litfin, T.; Singh, J.; Zhan, J.; and Zhou, Y. 2023. The master database of all possible RNA Sequences and its integration with RNAcmap for RNA Homology Search. *bioRxiv*, 2023–02.

Chen, X.; Li, Y.; Umarov, R.; Gao, X.; and Song, L. 2020. RNA secondary structure prediction by learning unrolled algorithms. *arXiv preprint arXiv:2002.05810*.

Cheong, H.-K.; Hwang, E.; Lee, C.; Choi, B.-S.; and Cheong, C. 2004. Rapid preparation of RNA samples for NMR spectroscopy and X-ray crystallography. *Nucleic Acids Research*, 32(10): e84.

Cleaves, H. J. 2011. *Watson–Crick Pairing*, 1775–1776. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-11274-4.

Danaee, P.; Rouches, M.; Wiley, M.; Deng, D.; Huang, L.; and Hendrix, D. 2018. bpRNA: large-scale automated annotation and analysis of RNA secondary structure. *Nucleic acids research*, 46(11): 5381–5394.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Do, C. B.; Woods, D. A.; and Batzoglou, S. 2006. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14): e90–e98.

Fica, S. M. 2020. Cryo-EM snapshots of the human spliceosome reveal structural adaptions for splicing regulation. *Current opinion in structural biology*, 65: 139–148.

Floridi, L.; and Chiriatti, M. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30: 681–694.

Franke, J. K.; Runge, F.; and Hutter, F. 2023. Scalable deep learning for RNA secondary structure prediction. *arXiv preprint arXiv:2307.10073*.

Fu, L.; Cao, Y.; Wu, J.; Peng, Q.; Nie, Q.; and Xie, X. 2022. UFold: fast and accurate RNA secondary structure prediction with deep learning. *Nucleic Acids Research*, 50(3): e14.

Fürtig, B.; Richter, C.; Wöhnert, J.; and Schwalbe, H. 2003. NMR spectroscopy of RNA. *Chembiochem: A European Journal of Chemical Biology*, 4(10): 936–962.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. ArXiv:1512.03385 [cs].

Ho, J.; Kalchbrenner, N.; Weissenborn, D.; and Salimans, T. 2019. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Janssen, S.; and Giegerich, R. 2015. The RNA shapes studio. *Bioinformatics*, 31(3): 423–425.

Lawson, C. L.; Berman, H. M.; Chen, L.; Vallat, B.; and Zirbel, C. L. 2024. The Nucleic Acid Knowledgebase: a new portal for 3D structural information about nucleic acids. *Nucleic Acids Research*, 52(D1): D245–D254.

Lorenz, R.; Bernhart, S. H.; Höner zu Siederdissen, C.; Tafer, H.; Flamm, C.; Stadler, P. F.; and Hofacker, I. L. 2011. ViennaRNA Package 2.0. *Algorithms for molecular biology*, 6: 1–14.

Mathews, D. H. 2019. How to benchmark RNA secondary structure prediction accuracy. *Methods*, 162: 60–67.

McDonald, R.; Pereira, F.; Ribarov, K.; and Hajic, J. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, 523–530.

Parsch, J.; Braverman, J. M.; and Stephan, W. 2000. Comparative sequence analysis and patterns of covariation in RNA secondary structures. *Genetics*, 154(2): 909–921.

Penić, R. J.; Vlašić, T.; Huber, R. G.; Wan, Y.; and Šikić, M. 2024. Rinalmo: General-purpose rna language models can generalize well on structure prediction tasks. *arXiv preprint arXiv:2403.00043*.

Reuter, J. S.; and Mathews, D. H. 2010. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC bioinformatics*, 11: 1–9.

Rivas, E. 2023. RNA covariation at helix-level resolution for the identification of evolutionarily conserved RNA structure. *PLOS Computational Biology*, 19(7): e1011262.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Sato, K.; Akiyama, M.; and Sakakibara, Y. 2021. RNA secondary structure prediction using deep learning with thermodynamic integration. *Nature communications*, 12(1): 941.

Shi, Y.-Z.; Wu, Y.-Y.; Wang, F.-H.; and Tan, Z.-J. 2014. RNA structure prediction: progress and perspective. *Chinese Physics B*, 23(7): 078701.

Singh, J.; Hanson, J.; Paliwal, K.; and Zhou, Y. 2019. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature communications*, 10(1): 5407.

Staple, D. W.; and Butcher, S. E. 2005. Pseudoknots: RNA structures with diverse functions. *PLoS biology*, 3(6): e213.

Tan, C.; Gao, Z.; and Li, S. Z. 2022. RFold: RNA Secondary Structure Prediction with Decoupled Optimization.

Tan, Z.; Fu, Y.; Sharma, G.; and Mathews, D. H. 2017. TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs. *Nucleic acids research*, 45(20): 11570–11581.

Varani, G.; and McClain, W. H. 2000. The G·U wobble base pair. *EMBO Reports*, 1(1): 18–23.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, N.; Bian, J.; Li, Y.; Li, X.; Mumtaz, S.; Kong, L.; and Xiong, H. 2024. Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning. *Nature Machine Intelligence*, 1–10.

Waterman, M. S.; and Smith, T. F. 1978. RNA secondary structure: a complete mathematical analysis. *Mathematical Biosciences*, 42(3): 257–266.

Yin, W.; Zhang, Z.; He, L.; Jiang, R.; Zhang, S.; Liu, G.; Zhang, X.; Qin, T.; and Xie, Z. 2024. ERNIE-RNA: An RNA Language Model with Structure-enhanced Representations. *bioRxiv*, 2024–03.

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