A Comparison of mainstream RNA secondary structure prediction methods

We compare our proposed method RFold with several other leading RNA secondary structure prediction methods and summarize the results in Table 8. RFold satisfies all three constraints (a)-(c) for valid RNA secondary structures, while the other methods do not fully meet some of the constraints. RFold utilizes a sequence-to-map attention mechanism to capture long-range dependencies, whereas SPOT-RNA simply concatenates pairwise sequence information and E2Efold/UFold uses hand-crafted features. In terms of prediction accuracy on the RNAStralign benchmark test set, RFold achieves the best F1 score of 0.977, outperforming SPOT-RNA, E2Efold and UFold by a large margin. Regarding inference time, RFold is much more efficient and requires only 0.02 seconds to fold the RNAStralign test sequences. In summary, RFold demonstrates superior performance over previous methods for RNA secondary structure prediction in both accuracy and speed.

Table 8: Comparison between RNA secondary structure prediction methods and RFold.

Method	SPOT-RNA	E2Efold	UFold	RFold
constraint (a)	×	✓	✓	✓
constraint (b)	×	✓	✓	✓
constraint (c)	×	×	×	✓
sequence-to-map	pairwise concat	pairwise concat	hand-crafted	seq2map attention
F1 on RNAStralign test	0.711	0.686	0.915	**0.977**
Inference time on RNAStralign test	77.80 s	0.40 s	0.16 s	**0.02** s

B Discussion on Abnormal Samples

Although we have illustrated three hard constraints in 3.2, there exist some abnormal samples that do not satisfy these constraints in practice. We have analyzed the datasets used in this paper and found that there are some abnormal samples in the testing set that do not meet these constraints. The ratio of valid samples in each dataset is summarized in the table below:

Table 9: The ratio of valid samples in the datasets.

Dataset	RNAStralign	ArchiveII	bpRNA
Validity	93.05%	96.03%	96.51%

As shown in Table 6, RFold forces the validity to be 100.00%, while other methods like E2Efold only achieve about 50.31%. RFold is more accurate than other methods in reflecting the real situation.

Nevertheless, we provide a soft version of RFold to relax the strict constraints. A possible solution to relax the rigid procedure is to add a checking mechanism before the Rol-Col Argmax function in the inference. Specifically, if the confidence given by the Rol-Col Softmax is low, we do not perform Rol-Col Argmax and assign more base pairs. It can be implemented as the following pseudo-code:

```
1  y_pred = row_col_softmax(y)
2  int_one = row_col_argmax(y_pred)
3
4  # get the confidence for each position
5  conf = y_pred * int_one
6  all_pos = conf > 0.0
7
8  # select reliable position
9  conf_pos = conf > thr1
10
11 # select unreliable position with the full row and column
12 uncf_pos = get_unreliable_pos(all_pos, conf_pos)
13
14 # assign "1" for the positions with the confidence higher than thr2
15 # note that thr2 < thr1
16 y_pred[uncf_pos] = (y_pred[uncf_pos] > thr2).float()
17 int_one[uncf_pos] = y_pred[uncf_pos]
```

543 We conduct experiments to compare the soft-RFold and the original version of RFold in the RNAS-
544 tralign dataset. The results are summarized in the Table 10. It can be seen that soft-RFold improves
545 the recall metric by a small margin. The minor improvement may be because the number of abnormal
546 samples is small.

Table 10: The results of soft-RFold and RFold on the RNAStralign.

Method	Precision	Recall	F1
RFold	0.981	0.973	0.977
soft-RFold	0.978	0.974	0.976

Table 11: The results of soft-RFold and RFold on the abnormal samples on the RNAStralign.

Method	Precision	Recall	F1
RFold	0.956	0.860	0.905
soft-RFold	0.949	0.889	0.918

547 We then select those samples that do not obey the three constraints to further analyse the performance.
548 The total number of such samples is 179. It can be seen that soft-RFold can deal with abnormal
549 samples well. The improvement of the recall metric is more obvious.

C Proofs of Theorems

C.1 Proof of Theorem 1

Theorem 1. Given a symmetric matrix $\widehat{\boldsymbol{H}} \in \mathbb{R}^{L \times L}$, the matrix Row-Col-Argmax($\widehat{\boldsymbol{H}}$) is also a symmetric matrix.

Proof: From Eq. 10 and Eq. 11, we can know that:

$$
\begin{aligned}
&\text{Row-Col-Argmax}(\widehat{\boldsymbol{H}}_{ij}) = 1, \\
&\text{if } \max\{\{\widehat{\boldsymbol{H}}_{ik}\}_{k=1}^{L} \cup \{\widehat{\boldsymbol{H}}_{kj}\}_{k=1}^{L}\} = \widehat{\boldsymbol{H}}_{ij},
\end{aligned}
\tag{20}
$$

Then, we can infer that:

$$
\begin{aligned}
&\text{Row-Col-Argmax}(\widehat{\boldsymbol{H}}_{ji}) = 1, \\
&\text{if } \max\{\{\widehat{\boldsymbol{H}}_{jk}\}_{k=1}^{L} \cup \{\widehat{\boldsymbol{H}}_{ki}\}_{k=1}^{L}\} = \widehat{\boldsymbol{H}}_{ji},
\end{aligned}
\tag{21}
$$

As $\widehat{\boldsymbol{H}}$ is a symmetric matrix, $\widehat{\boldsymbol{H}}_{jk} = \widehat{\boldsymbol{H}}_{kj}$ and $\widehat{\boldsymbol{H}}_{ki} = \widehat{\boldsymbol{H}}_{ik}$. Thus, Row-Col-Argmax($\widehat{\boldsymbol{H}}_{ji}$) can be rewritten as:

$$
\begin{aligned}
&\text{Row-Col-Argmax}(\widehat{\boldsymbol{H}}_{ji}) = 1, \\
&\text{if } \max\{\{\widehat{\boldsymbol{H}}_{kj}\}_{k=1}^{L} \cup \{\widehat{\boldsymbol{H}}_{ik}\}_{k=1}^{L}\} = \widehat{\boldsymbol{H}}_{ij},
\end{aligned}
\tag{22}
$$

It can be seen that only if $\max\{\{\widehat{\boldsymbol{H}}_{kj}\}_{k=1}^{L} \cup \{\widehat{\boldsymbol{H}}_{ik}\}_{k=1}^{L}\} = \widehat{\boldsymbol{H}}_{ij} = \widehat{\boldsymbol{H}}_{ji}$, then $\widehat{\boldsymbol{H}}_{ij} = \widehat{\boldsymbol{H}}_{ji} = 1$.

Thus, Row-Col-Argmax($\widehat{\boldsymbol{H}}$) is also a symmetric matrix.

C.2 Proof of Theorem 2

Theorem 2. Given a symmetric matrix $\widehat{\boldsymbol{H}} \in \mathbb{R}^{L \times L}$, the matrix Row-Col-Softmax($\widehat{\boldsymbol{H}}$) is also a symmetric matrix.

Proof: $\forall i, j \in \{1, ..., L\}$,

$$
\begin{aligned}
&\text{Row-Col-Softmax}(\widehat{\boldsymbol{H}}_{ji}) \\
&= \frac{1}{2}\left(\frac{\exp(\widehat{\boldsymbol{H}}_{ji})}{\sum_{k=1}^{L} \exp(\widehat{\boldsymbol{H}}_{jk})} + \frac{\exp(\widehat{\boldsymbol{H}}_{ji})}{\sum_{k=1}^{L} \exp(\widehat{\boldsymbol{H}}_{ki})} \right) \\
&= \frac{1}{2}\left(\frac{\exp(\widehat{\boldsymbol{H}}_{ij})}{\sum_{k=1}^{L} \exp(\widehat{\boldsymbol{H}}_{kj})} + \frac{\exp(\widehat{\boldsymbol{H}}_{ij})}{\sum_{k=1}^{L} \exp(\widehat{\boldsymbol{H}}_{ik})} \right) \\
&= \text{Row-Col-Softmax}(\widehat{\boldsymbol{H}}_{ij}).
\end{aligned}
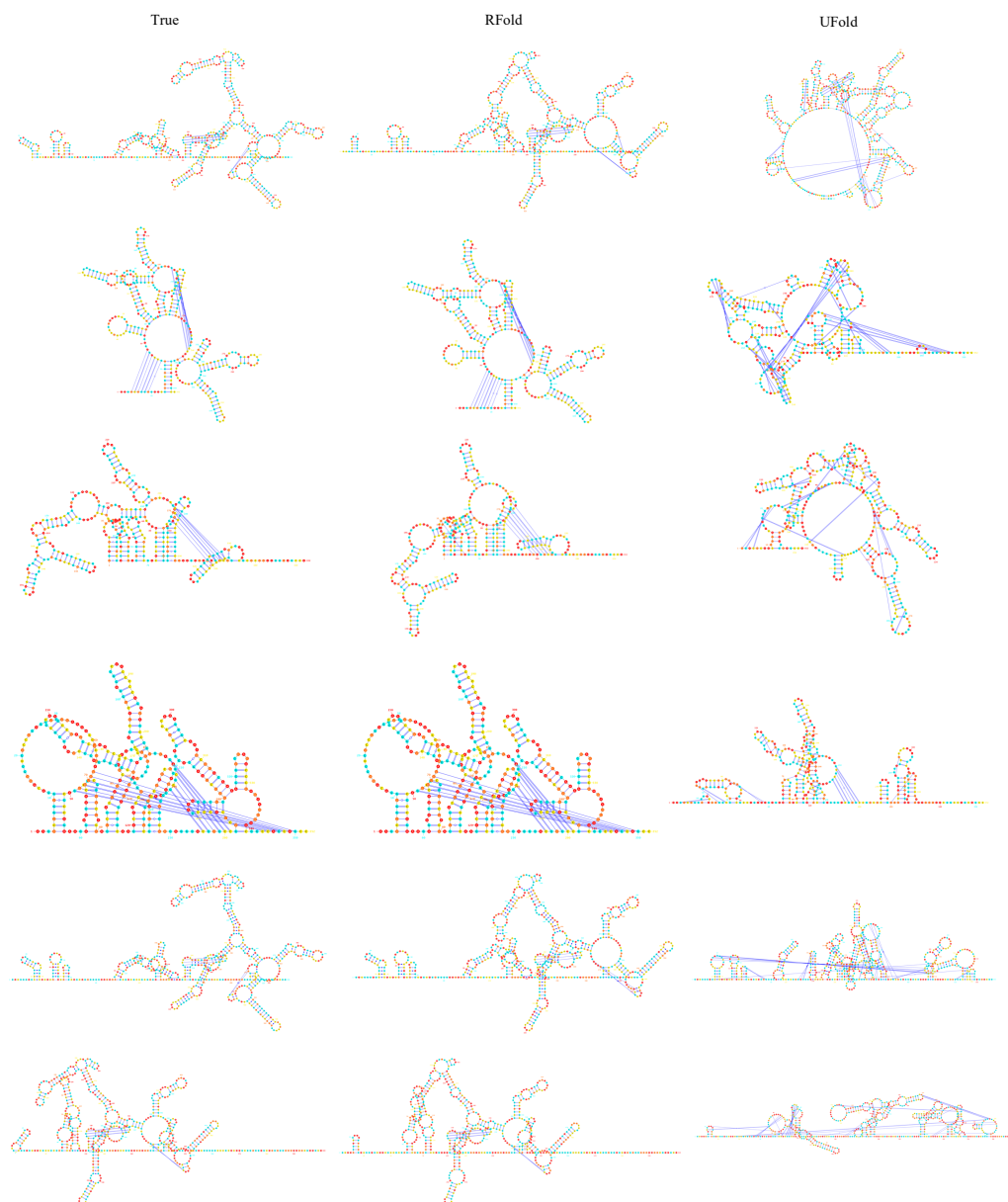\tag{23}
$$

15

D Visualization

True RFold UFold

Figure 7: Visualization of the true and predicted structures.