Supplementary Materials for UFold: Fast and Accurate RNA Secondary Structure Prediction with Deep Learning

Laiyi Fu[1,2⋆], Yingxin Cao[2,5,6⋆], Jie Wu[3], Qinke Peng[1], Qing Nie[4,5,6], and Xiaohui Xie[2 ⋆⋆]

[1] Systems Engineering Institute, School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shannxi, 710049, China
[2] Department of Computer Science, University of California, Irvine, CA, 92697, USA
[3] Department of Biological Chemistry, University of California, Irvine, CA, 92697, USA
[4] Department of Mathematics, University of California, Irvine, CA, 92697, USA
[5] Center for Complex Biological Systems, University of California, Irvine, CA, 92697, USA
[6] NSF-Simons Center for Multiscale Cell Fate Research, University of California, Irvine, CA, 92697, USA

⋆ Equal contributions
⋆⋆ To whom correspondence should be addressed

Table of Contents

List of Tables

List of Figures

1 Feature Map for Pairing Probability

For sequences of length L, We adapted this matrix representation $W \in \mathbb{R}_{\geq 0}^{L \times L}$ of RNA secondary structure pairing from CDPfold [1]. The algorithm is illustrated in figure S1.

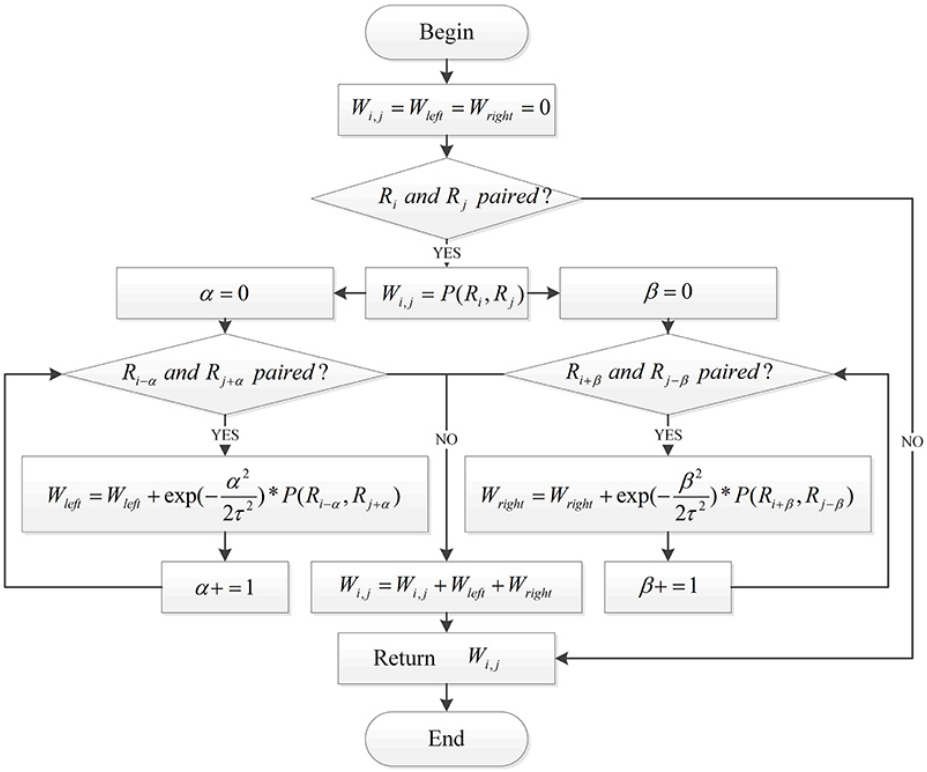

Figure S1: Algorithm for generating matrix representation of RNA secondary structure pairing probability [1].

Where each entry of the matrix W_{ij} represents probability of pairing at that position. The matrix is first initialized as zeros. $P(R_i, R_j)$ is determined by the type of base-pairing ($P(R_i, R_j) = 2$ for A-U paring, 3 for G-C pairing and $x \in (0, 2)$ for G-U pairing). The effect of neighborhood base-pairings are modeled as Gaussian functions with radius tracked by α and β.

2 Model Implementation

Input of our model is generated by taking the outer product of all combinations of four base channels of onehot encoding. This yields the first 16 channels as shown in Figure 1a. After that, an additional channel representing the paring probability (described in previous section) is then concatenated with the 16-channel sequence representation and together server as the input of our model.

Our model (shown in Figure 1b) is a variant of U-Net[2], which takes the 17-channel tensor as input and transform the data with consecutive convolution and max pooling operations. Following each convolutional layer, we added a batch normalization with running statistics turned off and a ReLU activation. There are in total 4 downsampling operations with 2 by 2 max pooling in encoding pathway, and 4 symmetric up convolution blocks each consists of an upsampling with scale 2 and a 2d convolution in decoding pathway. Each up convolution block is followed by a double convolution block same as encoding pathway. The input of the double convolution block consists of the tensor from previous level of decoding pathway as well as a copied tensor from encoding pathway at the same level. Two tensors are concatenated and then feed into the double convolution block. We use convolutional kernels with size 3, and number of channels for each operation are denoted in Figure 1b.

The general U-net architecture is an encoder-decoder framework, which consists of two main paths. The first path is the contraction path or so called encoder path, which is designed for extracting high level features from the input data. The encoder contains several layers of conventional convolutional and max pooling layers, which result in gradually shrunken feature maps as it is shown in Figure 1. The second path is the dedicated designed symmetric expanding path, which is also known as decoder path, it utilized transposed convolution kernels (i.e. up convolutional kernel) applied on the feature maps to enable precise localization. What is more, the feature maps of the same size with the decoder output features also concatenate together as a residual connection to further reserve the original features, so as to alleviate the gradient vanishing when doing back-propagation. This end-to-end fully convolutional network (FCN) naturally owns an advantages of the ability to receiving input data of any size because of the purely usage of convolutional layers instead of dense layer, which is suitable for the implementation of RNA secondary structure prediction because the sequence length of different RNAs can be various.

At the final out put, we used a 1 by 1 convolution to take the 32-dimension feature map to a 1-dimension score map. The map is multiplied with its transpose to yield a symmetric contact score map and then subject to weighted binary cross entropy loss with ground truth contact map for end to end training.

3 Supplementary Tables

Table S1: Training and Testing dataset numbers used in the study. "Before" and "After" denote the number of sequences before and after removing redundant sequences

Dataset	Training		Testing
	Before	After	
RNAStralign	37149	30451	2826
ArchiveII	/	/	3966
bpRNA-1m	102318	10814	1305
mutate sequence	216040	2768	/
bpRNA-new	/	/	5401
PDB	/	669	114

Table S2: pseudoknots, non-canonical sequence, and unstacking pairs occurrence percentage in different datasets.

Dataset	Non-canonical	Pseudoknot	unstacking
RNAStralign	100%	45%	62.2%
ArchiveII	0%	26.2%	58.2%
TS0*	0%	0%	54.6%
bpRNA-new*	0%	0%	36.4%
PDB	86.3%	55.6%	77.6%

* Dataset adapted from MXFold2 [3]

Table S3: Number of parameters in different deep learning models.

Method	Num of Parameters
UFold	8641377
MXfold2	47346
e2efold	718863
SPOT-RNA	7759445

Table S4: Benchmark results on the ArchiveII dataset.

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

Table S5: Benchmark results on the TS0 dataset.

Method	Prec	Rec	F1
UFold	**0.607**	**0.741**	**0.654**
SPOT-RNA	0.594	0.693	0.619
MXfold2	0.519	0.646	0.558
E2Efold	0.140	0.129	0.130
Mfold	0.501	0.627	0.538
Linearfold	0.561	0.581	0.550
Contrafold	0.528	0.655	0.567
Eternafold	0.516	0.666	0.563
ContextFold	0.529	0.607	0.546
RNAfold	0.494	0.631	0.536
RNAsoft	0.497	0.626	0.535
RNAStructure	0.494	0.622	0.533

Table S6: Benchmark results on the TS0 dataset of long range pairing.

Method	Prec	Rec	F1
UFold	**0.687**	**0.808**	**0.675**
Contrafold	0.306	0.439	0.349
MXfold2	0.318	0.450	0.360
SPOT-RNA	0.361	0.492	0.403
RNAsoft	0.310	0.448	0.353
Mfold	0.315	0.450	0.356
Eternafold	0.308	0.458	0.355
Linearfold	0.281	0.355	0.305
ContextFold	0.332	0.432	0.363
RNAfold	0.304	0.448	0.350
RNAStructure	0.299	0.428	0.339

Table S7: Number of pseudoknot pairs in RNAStralign dataset.

Type	Number
H_type	203
Kissing_hairpin	1069
Three knots	0
peusdoknot pairs	19277

Table S8: Benchmark results on the bpRNA-new dataset.

Method	Prec	Rec	F1
UFold	0.570	**0.742**	0.636
UFold(w/o data augmentation)	0.500	0.736	0.583
Contrafold	0.620	0.736	**0.661**
MXfold2	0.599	0.715	0.641
Eternafold	0.598	0.732	0.647
SPOT-RNA	0.635	0.641	0.620
E2Efold	0.047	0.031	0.036
RNAsoft	0.580	0.692	0.620
Mfold	0.584	0.692	0.623
Linearfold	**0.658**	0.645	0.633
ContextFold	0.596	0.636	0.604
RNAfold	0.593	0.720	0.640
RNAStructure	0.586	0.704	0.629

Table S9: Benchmark results on the TS1 dataset.

Method	Prec	Rec	F1
UFold-PDBfinetune	0.781	0.664	0.712
Contrafold	0.826	0.603	0.688
MXfold2	0.823	0.604	0.686
SPOT-RNA	**0.882**	**0.677**	**0.751**
Eternafold	0.827	0.634	0.709
E2Efold	0.243	0.214	0.218
RNAsoft	0.796	0.549	0.634
Mfold	0.787	0.546	0.630
Linearfold	0.826	0.545	0.644
ContextFold	0.853	0.516	0.621
RNAfold	0.801	0.588	0.666
RNAStructure	0.786	0.570	0.649

Table S10: Benchmark results on the TS2 dataset.

Method	Prec	Rec	F1
UFold-PDBfinetune	0.943	**0.848**	**0.891**
Contrafold	0.923	0.744	0.818
MXfold2	0.947	0.744	0.828
SPOT-RNA	0.922	0.790	0.843
E2Efold	0.247	0.259	0.239
RNAsoft	**0.965**	0.749	0.836
Mfold	0.961	0.740	0.829
Eternafold	0.894	0.721	0.793
Linearfold	0.918	0.696	0.780
ContextFold	0.920	0.683	0.777
RNAfold	0.947	0.751	0.831
RNAStructure	0.940	0.742	0.824

Table S11: Benchmark results on the TS3 dataset.

Method	Prec	Rec	F1
UFold-PDBfinetune	0.849	**0.647**	**0.731**
Contrafold	0.912	0.578	0.692
MXfold2	0.906	0.586	0.700
SPOT-RNA	0.931	0.599	0.717
E2Efold	0.173	0.117	0.132
RNAsoft	0.891	0.610	0.712
Mfold	0.901	0.610	0.714
Eternafold	0.848	0.549	0.654
Linearfold	**0.943**	0.515	0.642
ContextFold	0.853	0.516	0.621
RNAfold	0.850	0.548	0.677
RNAStructure	0.845	0.562	0.663

Table S12: non-canonical prediction F1 value of UFold compared with SPOT-RNA on PDB datasets.

	TS1	TS2	TS3
UFold	**0.413**	**0.637**	**0.355**
SPOT-RNA	0.278	0.282	0.212

Table S13: p-value of UFold compared with other methods on four datasets.

	ArchiveII	TS0	bpnew	PDB
Contextfold	2.902e-62	5.381e-22	3.183e-10	9.261e-03
MXfold2	8.475e-200	3.539e-17	2.287e-01	9.937e-02
SPOT-RNA	0.000e+00	1.309e-03	1.134e-03	9.160e-01
Contrafold	0.000e+00	1.023e-14	1.122e-07	7.258e-02
Linearfold	0.000e+00	2.877e-18	7.071e-01	8.449e-04
Eternafold	0.000e+00	4.078e-16	1.942e-02	7.232e-02
RNAfold	0.000e+00	7.005e-24	3.351e-01	3.560e-02
RNAStructure	0.000e+00	3.647e-24	1.983e-01	1.107e-02
RNAsoft	0.000e+00	1.231e-25	4.077e-03	1.021e-02
Mfold	0.000e+00	2.050e-24	1.880e-02	7.318e-03
e2efold	0.000e+00	3.923e-27	0.000e+00	1.154e-42

Table S14: 95% confidence intervals obtained by the bootstrap percentile method for PDB dataset with various bootstrap steps.

Bootstrap steps	TS1	TS2	TS3
20	(0.698, 0.726)	(0.880, 0.907)	(0.704, 0.730)
50	(0.704, 0.721)	(0.890, 0.906)	(0.719, 0.735)
100	(0.706, 0.719)	(0.890, 0.902)	(0.725, 0.737)
200	(0.710, 0.718)	(0.889, 0.897)	(0.728, 0.736)
500	(0.707, 0.712)	(0.888, 0.893)	(0.730, 0.736)
1000	(0.710, 0.714)	(0.889, 0.892)	(0.729, 0.732)
2000	(0.709, 0.712)	(0.890, 0.892)	(0.730, 0.732)
5000	(0.711, 0.712)	(0.890, 0.892)	(0.730, 0.732)

4 Supplementary Figure

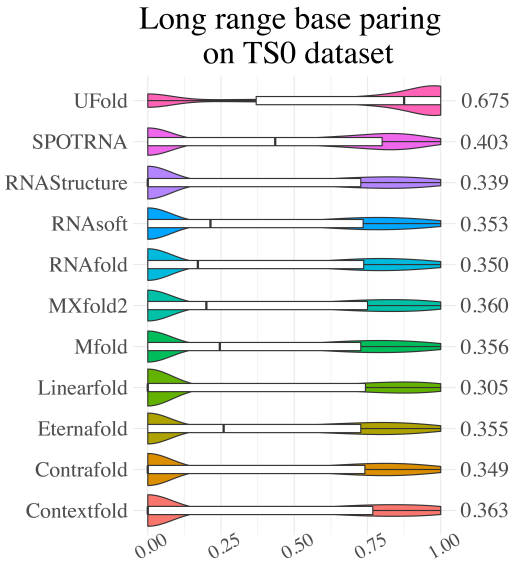

Figure S2: long range pairing benchmark result on TS0 dataset. Violin plot on the TS0 dataset. Visualization of F1 value of long-range base pair prediction of UFold versus other 10 RNA secondary structure predictions methods on TS0 dataset.

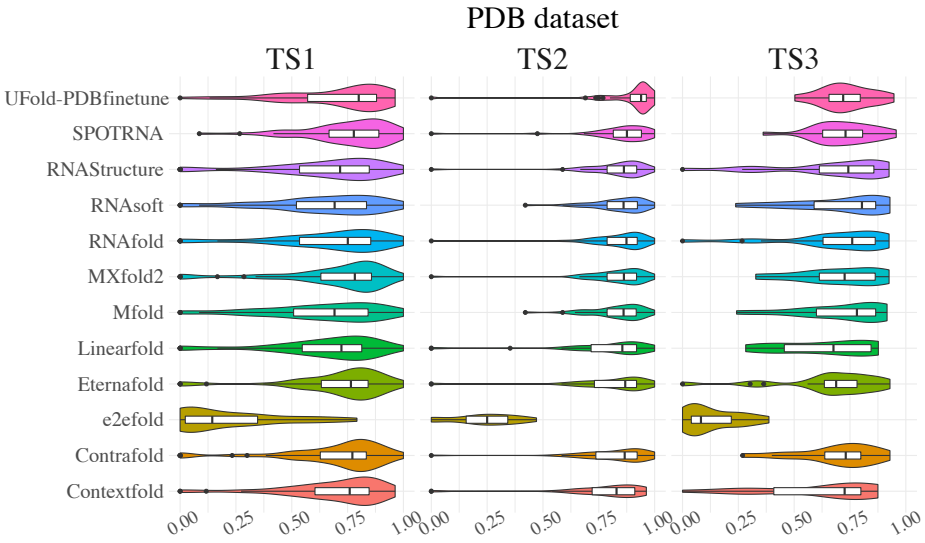

Figure S3: Violin plot on the PDB dataset. Visualization of F1 value of UFold versus other 11 RNA secondary structure predictions methods on three datasets(TS1,TS2,TS3) separately.

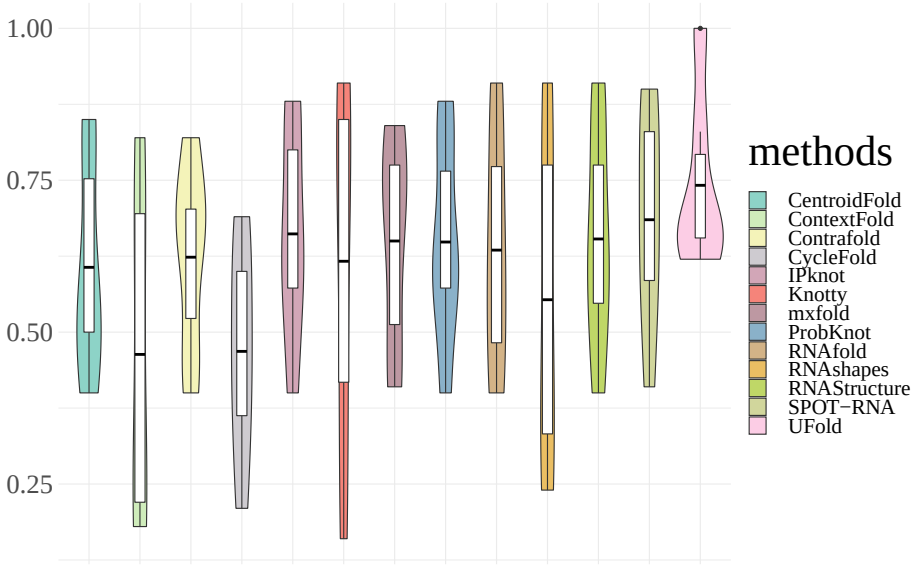

Figure S4: Violin plot on the 6 RNAs in PDB dataset. Visualization of F1 value of UFold versus other 12 RNA secondary structure predictions methods on 6 RNAs. The results of competitors are retrieved from SPOT-RNA paper[4].

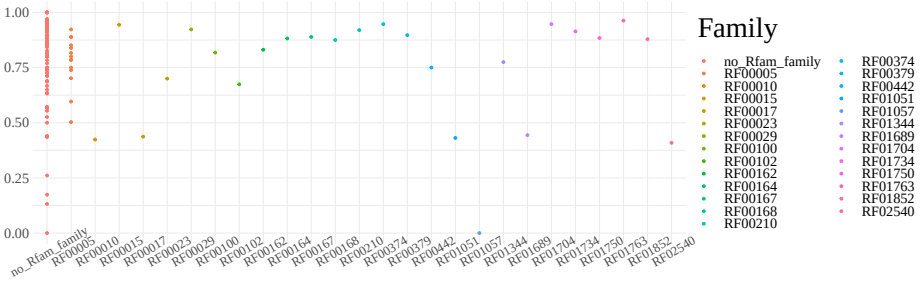

Figure S5: Performance comparison on 114 RNAs from PDB dataset(TS1+TS2+TS3) by mapping to RFAM families. Each dot represents the F1 value of each sequence prediction.

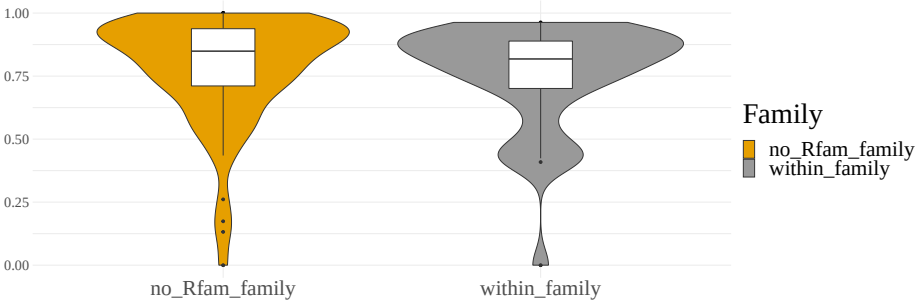

Figure S6: Violin plot of performance comparison between RNA sequences that are mapped to known Rfam families and those sequences that are not existed in any Rfam families.

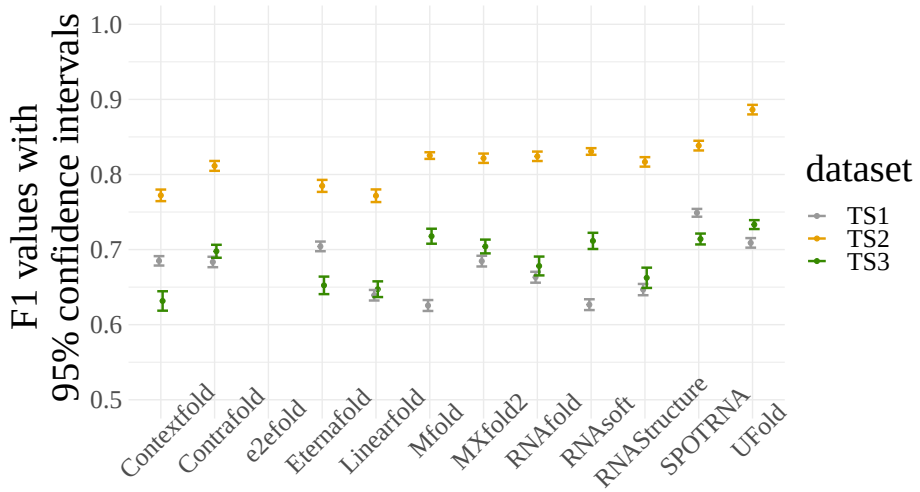

Figure S7: 95% bootstrap percentile confidence intervals for the F1 value of all the compared methods on PDB dataset.

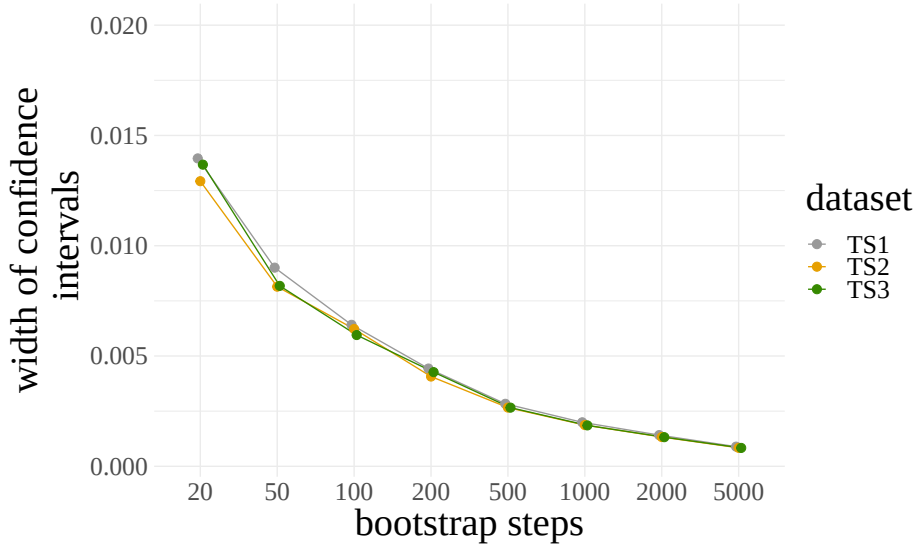

Figure S8: Confidence interval width versus bootstrap step size in the PDB dataset. Dot line plot of 95% confidence interval width of F1 values as the bootstrap size increase.

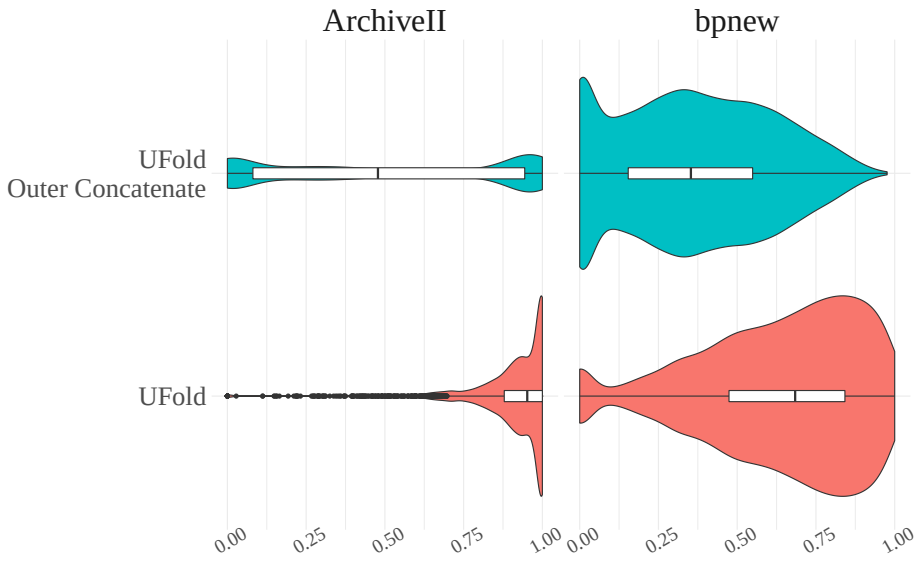

Figure S9: Performance benchmark result on ArchiveII and bpnew dataset using outer concatenation or Kronecker product input. Violin plot of F1 value prediction on the ArchiveII and bpnew dataset of UFold(Kronecker product input) vs. outer concatenation input.

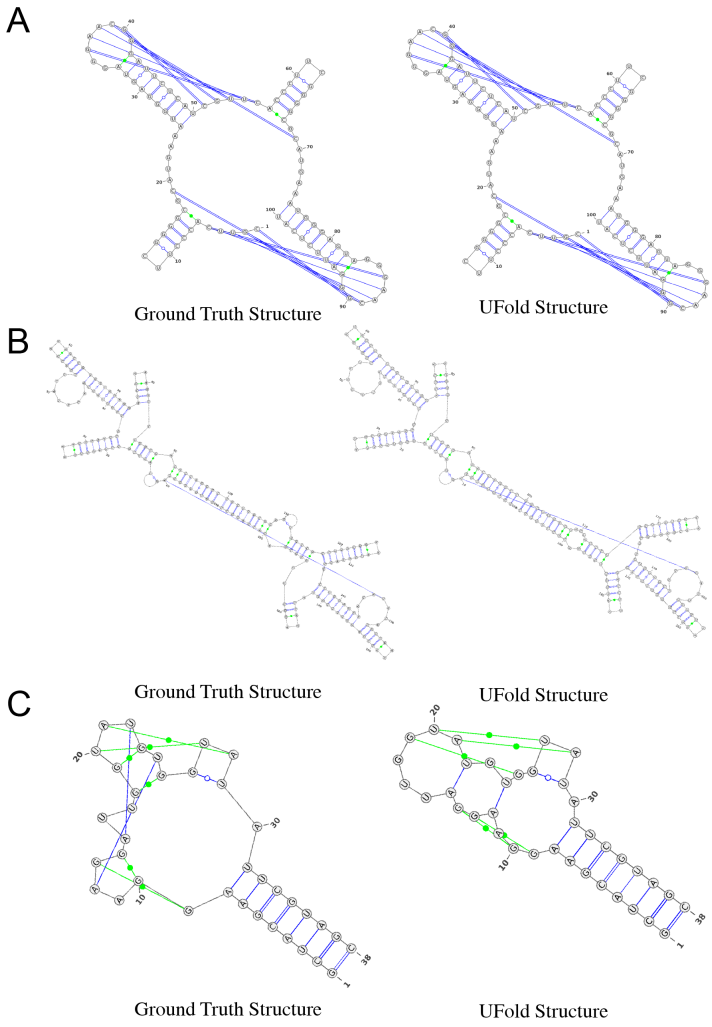

Figure S10: Comparison of UFold prediction with the ground truth structure on three recent released crystal structures. The secondary structure of a synthetic construct RNA with PDB ID A)6QN3 (Glutamine II Riboswitch RNA) B)6N2V (Manganese riboswitch from Xanthmonas oryzae bound to Mn(II)) C)6E8S (iMango-III aptamer bound to TO1-Biotin). Non-canonical base pairs are colored in light green.

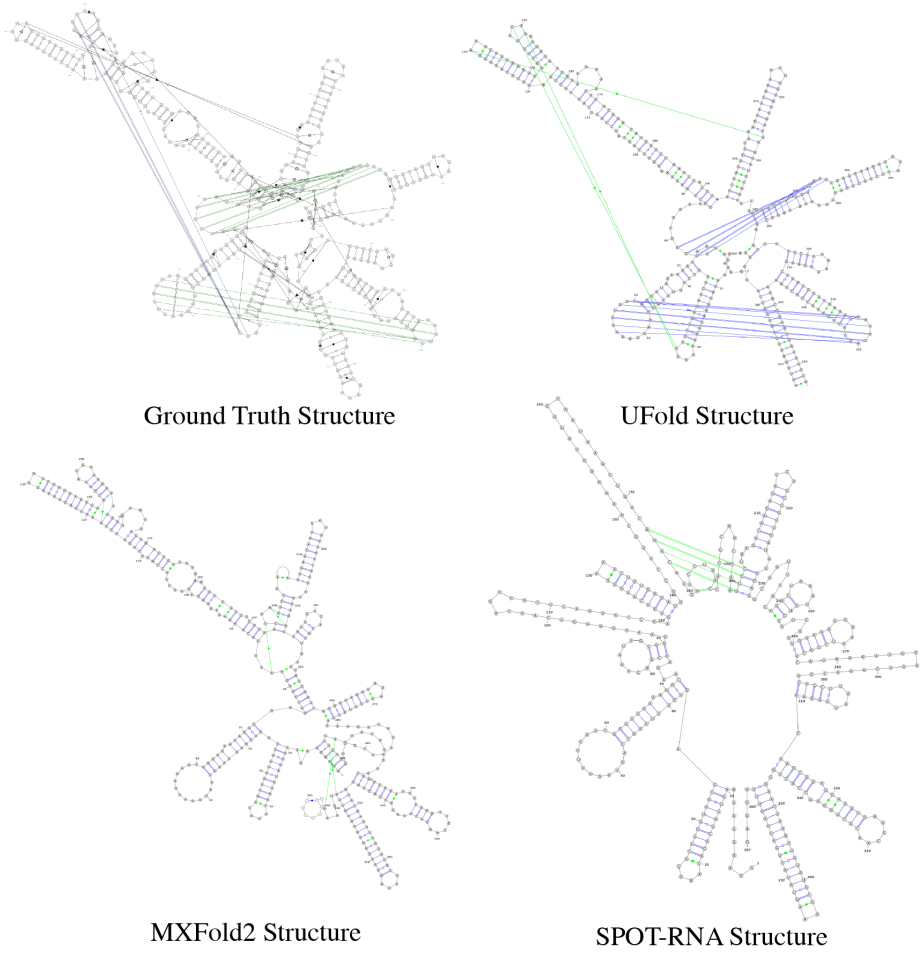

Figure S11: Comparison of UFold prediction with MXfold2 and SPOT-RNA predictions as well as ground truth structure on PDB database with RNA ID 7EZ0 (Apo L-21 ScaI Tetrahymena ribozyme) released in 2021. The predicted F1 values of UFold, MXFold2 and SPOT-RNA on the RNA 7EZ0 are 0.803,0.744 and 0.668 respectively.

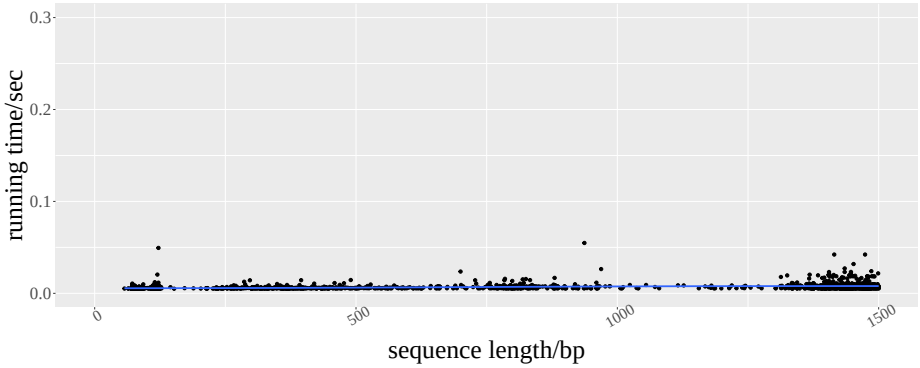

Figure S12: Running time vs. sequence length. Dot plots of running time on GPU against sequence length on RNAStralign dataset.

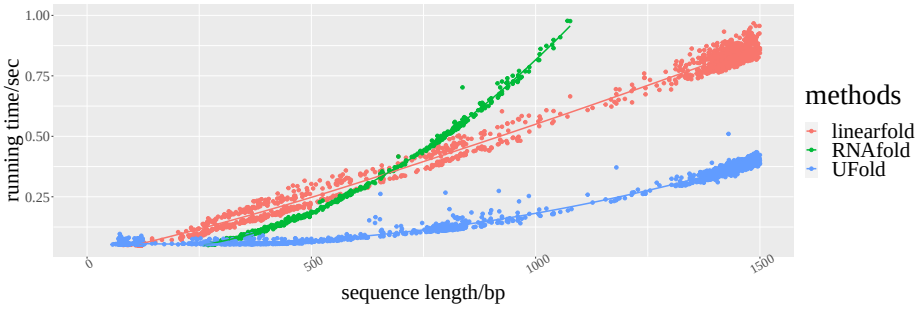

Figure S13: Running time vs. sequence length results of three compared methods. Dot plots of running time against sequence length on RNAStralign dataset on UFold, RNAfold and Linearfold.

References

1. Hao Zhang, Chunhe Zhang, Zhi Li, Cong Li, Xu Wei, Borui Zhang, and Yuanning Liu. A new method of rna secondary structure prediction based on convolutional neural network and dynamic programming. *Frontiers in genetics*, 10:467, 2019.
2. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
3. Kengo Sato, Manato Akiyama, and Yasubumi Sakakibara. Rna secondary structure prediction using deep learning with thermodynamic integration. *bioRxiv*, 2020.
4. Jaswinder Singh, Jack Hanson, Kuldip Paliwal, and Yaoqi Zhou. Rna secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature communications*, 10(1):1–13, 2019.