Supplementary Information

# A unified benchmark of deep learning methods for RNA secondary structure prediction

Xiaoya Fan[1], Jingjing Chen[2], Tingting Li[3], Yubang Dong[2], Tingwei Chen[3*], Zheng Zhao[4*]

[1]School of Software Technology, Dalian University of Technology, Dalian, China.
[2]College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, 110016, China.
[3]Faculty of Information, Liaoning University, Shenyang, 110036, Liaoning, China.
[4]College of Artificial Intelligence, Dalian Maritime University, Dalian, China.

*Corresponding author(s). E-mail(s): twchen@lnu.edu.cn; zhaozheng@dlmu.edu.cn;

## S1 Supplementary Methods

### S1.1 Detailed dataset construction

#### S1.1.1 Statistics of benchmark datasets

The sample sizes of all benchmarks and their corresponding subsets are summarized in Supplementary Table S1. The construction rules for the intra-dataset and cross-dataset benchmarks have been described in the main text; this section further provides the detailed construction procedures for the cross-family, length-generalization, and pseudoknot benchmarks.

**Table S1** Benchmark datasets.

| Benchmark | Dataset | Train | Valid | Test |
|---|---|---|---|---|
| intra-dataset evaluation | Rivas | 2941 | 547 | 414 |
| | bpRNA-1m | 9920 | 1240 | 1241 |
| Cross-dataset evaluation | RNAStrAlign | 2475 | 618 | - |
| | ArchiveII | - | - | 422 |
| | bpRNA-1m | 9920 | 1240 | - |
| | bpRNA-new | - | - | 5055 |
| Easy cross-family evaluation | 16S rRNA | 5126 | 1282 | 464 |
| | 5S rRNA | 3897 | 975 | 2000 |
| | Group II intron | 4368 | 1093 | 1411 |
| | RNase P | 5192 | 1298 | 382 |
| | SRP RNA | 5005 | 1252 | 615 |
| | tRNA | 3897 | 975 | 2000 |
| Complex cross-family evaluation | 16S rRNA | 7956 | 1989 | 2000 |
| | 5S rRNA | 7956 | 1989 | 2000 |
| | Group I intron | 8236 | 2059 | 1650 |
| | Group II intron | 8392 | 2099 | 1454 |
| | RNase P | 8417 | 2105 | 1423 |
| | SRP RNA | 9057 | 2265 | 623 |
| | Telomerase RNA | 9502 | 2376 | 67 |
| | tmRNA | 8973 | 2244 | 728 |
| | tRNA | 7956 | 1989 | 2000 |
| Length generalization evaluation | Standard-length | 16239 | 2029 | Short 1819<br>Mid-length 213 |
| | Extended-length | 16363 | 2045 | Short 1810<br>Mid-length 200<br>Long 37 |
| | Extreme-length | 16482 | 2058 | Short 1791<br>Mid-length 183<br>Long 21<br>Extra-long 70 |
| Pseudoknot evaluation | Pseudoknot | 1124 | 140 | 142 |
| | Non-pseudoknot | 9920 | 1240 | - |

Note: Standard-length, extended-length, and extreme-length datasets denote progressively expanded training sets used for length-generalization evaluation.

### S1.1.2 Cross-family benchmark

To evaluate model generalization to unseen RNA families, we designed two cross-family benchmarks: an easy cross-family benchmark and a complex cross-family benchmark. Both settings used a leave-one-family-out strategy, in which one RNA family was held out as the test set in each run and the remaining families were used to construct the training and validation sets. For each family, if more than 2,000 sequences remained after redundancy removal, 2,000 sequences were randomly sampled; otherwise, all remaining sequences were retained. In the easy cross-family benchmark, we constructed a relatively controlled cross-family evaluation setting by removing sequences with pseudoknotted structures and restricting the sequence length to 511 nt. This setting retained six RNA families, including 16S rRNA, 5S rRNA, Group II intron, RNase P, SRP RNA, and tRNA. In each run, one family was used as the test set, and the remaining families were split into training and validation sets at a ratio of 80% and 20%. In contrast, the complex cross-family benchmark retained long sequences and pseudoknot-containing structures, thereby introducing greater variation in sequence length and higher structural complexity. In this setting, one family was likewise held out as the test set in each run, and the remaining families were used for model training and validation. This design avoids information leakage from the same RNA family and provides a more stringent evaluation of model generalization to unseen RNA families and complex structural patterns.

### S1.1.3 Length-generalization benchmark

The length-generalization benchmark was constructed based on the four sequence-length intervals defined in the main text: short sequences (1-255 nt), mid-length sequences (256-511 nt), long sequences (512-1023 nt), and extra-long sequences (1024+ nt). We integrated RNA sequences from Rivas, ArchiveII, RNAStrAlign, bpRNA-new, bpRNA-1m, CRW2, and RNASTRAND, and stratified the test sequences according to these length intervals to evaluate model performance across different sequence-length ranges. The sample sizes of the training, validation, and length-stratified test subsets are summarized in Supplementary Table S1.

### S1.1.4 Pseudoknot benchmark

The pseudoknot benchmark was further constructed based on bpRNA-1m. Specifically, we extracted pseudoknot-containing RNA sequences no longer than 511 nt from bpRNA-1m. After redundancy removal using CD-HIT, these sequences were split into training, validation, and test sets at a ratio of 8:1:1, resulting in 1,124, 140, and 142 sequences, respectively. The pseudoknot-containing training and validation sets were then merged into the training and validation sets of the bpRNA-1m benchmark, respectively, whereas the 142 pseudoknot-containing test sequences were kept separate for dedicated evaluation of pseudoknot prediction performance.

## S1.2 Detailed evaluation metrics

We used Precision, Recall, F1-score (F1), and Matthews correlation coefficient (MCC) as general evaluation metrics to assess overall base-pair prediction performance. Here, TP denotes a base pair present in both the predicted and reference structures, FP denotes a base pair present only in the predicted structure, FN denotes a base pair present in the reference structure but missed by the model, and TN denotes a candidate base pair absent from both the predicted and reference structures. The formulas for these metrics are provided in Supplementary Table S2. All metrics were computed for each RNA and then averaged across RNAs within the corresponding test set.

**Table S2** Evaluation metrics used in this study.

| **Metric** | **Definition** |
|---|---|
| Precision | $\frac{TP}{TP+FP}$ |
| Recall | $\frac{TP}{TP+FN}$ |
| F1 | $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$ |
| MCC | $\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$ |

For pseudoknot-specific evaluation, we computed PK-Precision, PK-Recall, and PK-F1 only for pseudoknotted base pairs. Here, $TP_{\mathrm{PK}}$ denotes correctly predicted pseudoknotted base pairs, $FP_{\mathrm{PK}}$ denotes predicted pairs that are absent from the reference set of pseudoknotted base pairs, and $FN_{\mathrm{PK}}$ denotes pseudoknotted base pairs in the reference structure that were missed by the model. PK-Precision, PK-Recall, and PK-F1 were computed in the same way as Precision, Recall, and F1, respectively. Because pseudoknotted base pairs account for only a very small fraction of the full candidate base-pair space, the definition of true negatives can be sensitive to the choice of candidate pairing space and background set. We therefore did not report pseudoknot-specific MCC.

## S1.3 Detailed model information

We compared 16 RNA secondary structure prediction models in this study. The main architecture, training strategy, maximum input length, and support for pseudoknot prediction of each model are summarized in Supplementary Table S3.
It should be noted that the maximum input lengths reported in Supplementary Table S3 refer to the practical settings used under our unified experimental environment, rather than necessarily representing the theoretical limits of the official implementations. Although the official RNA-FM implementation can support longer input sequences, we set its maximum input length to 600 nt due to GPU memory

constraints. Moreover, for RNADiffFold, we followed its implementation when setting the input length and initialized its UFold component with UFold parameters trained on the corresponding benchmark task; therefore, the maximum input length of RNADiffFold was kept consistent with that of UFold.

**Table S3** Summary of evaluated RNA secondary structure prediction models.

| Model | Model Architecture | Training Strategy | Maximum input (nt) | Pseudoknot support |
|---|---|---|---|---|
| MXfold2 [1] | Thermodynamics + Deep Neural Network | Train from scratch | - | × |
| RNAErnie [2] | Pretrained foundation model + MXfold2 prediction head | Freeze backbone, fine-tune prediction head | 512 | - |
| RNADiffFold [3] | Pretrained foundation model + Diffusion prediction head | Freeze backbone, fine-tune prediction head | 600 | ✓ |
| UFold [4] | CNN + U-Net architecture | Train from scratch | 600 | ✓ |
| REDfold [5] | Residual Encoder–Decoder CNN | Train from scratch | 720 | ✓ |
| SPOT-RNA [6] | ResNets + 2D-BLSTM | Train from scratch | 1024 | ✓ |
| Rfold [7] | Attention neural network + K-rook matching | Train from scratch | 1005 | ✓ |
| RNA-FM [8] | Pretrained foundation model | Freeze backbone, fine-tune prediction head | 600 | - |
| Neuralfold [9] | FNN + MEA decoding | Train from scratch | 512 | ✓ |
| GCNfold [10] | Graph Convolutional Network | Train from scratch | 600 | ✓ |
| CNNfold [11] | Convolutional Neural Network | Train from scratch | 512 | ✓ |
| LTPConstraint [12] | Deep learning model + hard constraints | Train from scratch | 512 | ✓ |
| CDPfold [13] | CNN + Dynamic Programming | Train from scratch | 512 | × |
| DMfold [14] | Deep learning + IBPMP | Train from scratch | 512 | ✓ |
| E2Efold [15] | Transformer + Unrolled algorithm | Train from scratch | 512 | × |
| MSFF-CDCGAN [16] | CDCGAN | Train from scratch | 512 | ✓ |

"-" in the Max Input column indicates that the method does not have a predefined architecture-level maximum input length under the setting used in this study. In the Pseudoknot column, "-" for pretrained foundation models indicates that pseudoknot prediction support is determined by the downstream prediction head or decoder rather than by the pretrained backbone itself.

## S1.4 Detailed training and inference settings

All training and inference experiments were conducted on four NVIDIA GeForce RTX 3080 GPUs, each equipped with 10 GB of GDDR6X memory. Unless otherwise specified, all models were trained, fine-tuned, or evaluated based on their official implementations and default hyperparameter settings. For the pretrained foundation models RNAErnie and RNA-FM, we used the official pretrained weights and retrained only the downstream structure prediction heads. For RNADiffFold, we used the official RNA-FM pretrained model and initialized its UFold component with UFold parameters trained on the corresponding benchmark task. All models were compared under the same data splits and evaluation protocol.

# S2 Supplementary Results

## S2.1 Detailed results on intra-dataset and cross-dataset benchmarks

Supplementary Figure S1 reports the rankings of the 16 models on each intra-dataset and cross-dataset benchmark, together with their overall average rankings across these benchmarks.

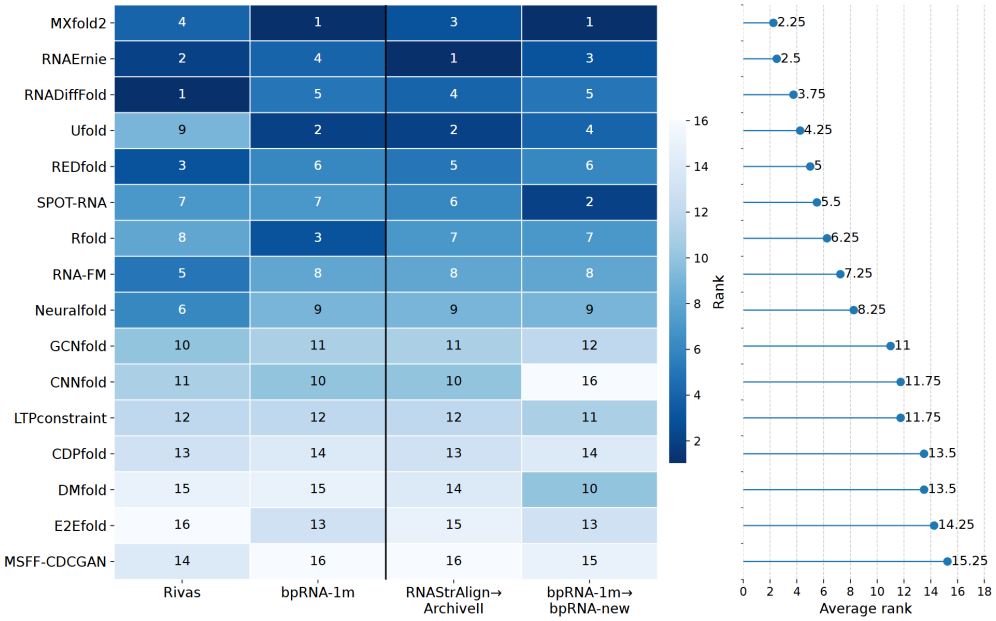


**Fig. S1** Ranking of the 16 models on the intra-dataset and cross-dataset benchmarks.

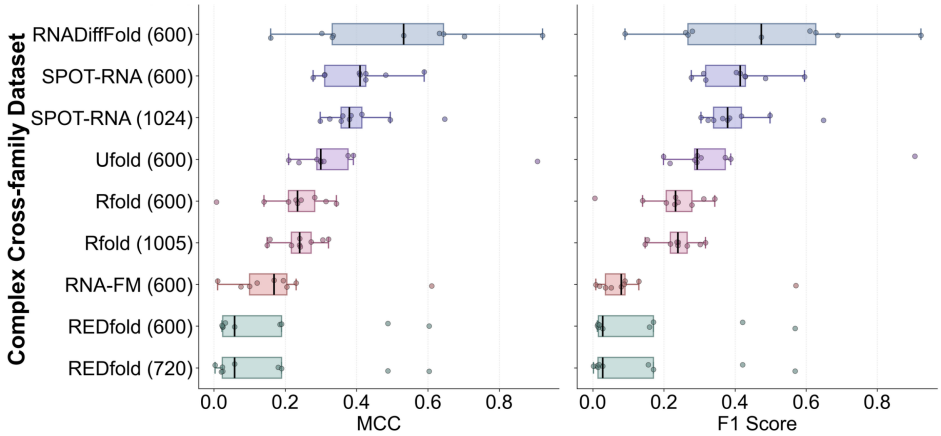


**Fig. S2** Complex cross-family evaluation results. (a) and (b) show the MCC and F1 distributions under the complex cross-family setting, respectively. The number in parentheses after each model name indicates the input-length setting used for both training and evaluation.

## S2.2 Detailed results on cross-family benchmarks

Supplementary Tables S4 and S5 report the detailed performance of representative models under the easy and complex cross-family benchmarks, respectively.

In the more challenging complex cross-family benchmark, we performed leave-one-family-out testing across nine RNA families, including 16S rRNA, 5S rRNA, Group I intron, Group II intron, RNase P, SRP RNA, Telomerase RNA, tRNA, and tmRNA. This setting retained longer sequences and pseudoknotted structures, thereby introducing greater family-level distribution shift, length variation, and structural complexity than the easy cross-family benchmark. RNAErnie was not included in this comparison because its evaluated implementation was limited to sequences no longer than 512 nt. MXfold2 was also excluded because it does not support pseudoknotted structures. To facilitate a fair comparison across models while retaining a substantial portion of the long-sequence data, we adopted 600 nt as the primary evaluation limit, as this length could be supported by most evaluated methods without introducing excessive model-specific exclusions. For models capable of processing longer sequences, additional evaluations were conducted using their respective maximum supported input lengths to assess performance under less restrictive settings.

The detailed results are shown in Fig. S2 and Supplementary Table S5. Among the evaluated settings, RNADiffFold(600) achieved the highest average MCC (0.4613) and Precision (0.7598), but its Recall was relatively low (0.3458), indicating a conservative prediction pattern. SPOT-RNA showed similar performance under the 600-nt and 1024-nt settings, with MCC values of 0.4042 and 0.4067, respectively, suggesting relatively stable behavior when evaluated with a larger input-length limit. UFold(600) achieved an average MCC of 0.3685, whereas Rfold and REDfold showed lower average MCC values across their evaluated input-length settings.

**Table S4** Average performance on the easy cross-family benchmark.

| Rank | Models | Precision | Recall | F1 | MCC |
|---|---|---|---|---|---|
| 1 | RNADiffFold | **0.7778** | 0.4113 | **0.4814** | **0.5139** |
| 2 | MXfold2 | 0.4835 | **0.4684** | 0.4723 | 0.4722 |
| 3 | UFold | 0.4618 | 0.3535 | 0.3905 | 0.3952 |
| 4 | SPOT-RNA | 0.3507 | 0.4361 | 0.3767 | 0.3816 |
| 5 | RNAErnie | 0.4043 | 0.3585 | 0.3770 | 0.3774 |
| 6 | REDfold | 0.4456 | 0.1829 | 0.2325 | 0.2634 |
| 7 | Rfold | 0.2227 | 0.1582 | 0.1748 | 0.1814 |
| 8 | RNA-FM | 0.7160 | 0.0456 | 0.0727 | 0.1058 |

**Table S5** Average performance on the complex cross-family benchmark.

| Rank | Models | Precision | Recall | F1 | MCC |
|---|---|---|---|---|---|
| 1 | RNADiffFold(600) | **0.7598** | 0.3458 | **0.4210** | **0.4613** |
| 2 | SPOT-RNA(600) | 0.3876 | **0.4422** | 0.4070 | 0.4042 |
| 2 | SPOT-RNA(1024) | 0.3938 | 0.4391 | 0.4069 | 0.4067 |
| 3 | UFold(600) | 0.4270 | 0.3337 | 0.3621 | 0.3685 |
| 4 | Rfold(600) | 0.2554 | 0.2007 | 0.2210 | 0.2228 |
| 4 | Rfold(1005) | 0.2790 | 0.2109 | 0.2354 | 0.2384 |
| 5 | RNA-FM(600) | 0.6387 | 0.0764 | 0.1193 | 0.1906 |
| 6 | REDfold(600) | 0.3279 | 0.1223 | 0.1566 | 0.1808 |
| 6 | REDfold(720) | 0.3126 | 0.1210 | 0.1544 | 0.1769 |

Note: The number in parentheses after each model name indicates the input-length setting used for both training and evaluation.

## S2.3 Detailed results on length-generalization benchmark

Supplementary Table S6 reports model performance across the four sequence-length intervals defined in the main text.

## S2.4 Detailed results on pseudoknot-containing RNAs

**Table S6** Model performance on RNAs of different lengths.

| Rank | Model | Precision | Recall | F1 | MCC |
|---|---|---|---|---|---|
| ***Short Sequences (1-255 nt)*** | | | | | |
| 1 | UFold | 0.6306 | **0.7157** | **0.6589** | **0.6643** |
| 2 | RNAErnie | 0.6001 | 0.6833 | 0.6283 | 0.6330 |
| 3 | Rfold | 0.5658 | 0.7096 | 0.6165 | 0.6249 |
| 4 | MXfold2 | 0.5797 | 0.6290 | 0.5909 | 0.5956 |
| 5 | SPOT-RNA | 0.5513 | 0.6262 | 0.5864 | 0.5731 |
| 6 | RNADiffFold | 0.6758 | 0.4979 | 0.5310 | 0.5433 |
| 7 | RNA-FM | **0.6988** | 0.4656 | 0.4621 | 0.4870 |
| 8 | REDfold | 0.4287 | 0.2301 | 0.2706 | 0.2941 |
| ***Mid-length Sequences (256-511 nt)*** | | | | | |
| 1 | RNAErnie | 0.5355 | **0.5673** | **0.5409** | **0.5451** |
| 2 | Rfold | 0.5350 | 0.5277 | 0.5238 | 0.5268 |
| 3 | UFold | 0.4951 | 0.5489 | 0.5128 | 0.5165 |
| 4 | MXfold2 | 0.5179 | 0.4947 | 0.5006 | 0.5027 |
| 5 | RNADiffFold | 0.5308 | 0.4746 | 0.4889 | 0.4953 |
| 6 | SPOT-RNA | 0.5024 | 0.4088 | 0.4508 | 0.4473 |
| 7 | RNA-FM | **0.6308** | 0.2007 | 0.2128 | 0.2407 |
| 8 | REDfold | 0.0959 | 0.0097 | 0.0171 | 0.0293 |
| ***Long Sequences (512-1023 nt)*** | | | | | |
| 1 | SPOT-RNA | **0.4506** | 0.3803 | **0.4125** | **0.4082** |
| 2 | MXfold2 | 0.3908 | **0.4180** | 0.3935 | 0.3983 |
| ***Extra-long Sequences (1024+ nt)*** | | | | | |
| 1 | MXfold2 | **0.35216** | **0.37493** | **0.35119** | **0.35673** |

Supplementary Table S7 reports model performance on pseudoknot-containing RNAs, with separate evaluations for pseudoknot base pairs, nested base pairs, and all base pairs.

**Table S7** Performance on pseudoknot-containing RNAs.

| Rank | Model | Precision | Recall | F1 | MCC |
|---|---|---|---|---|---|
| ***Pseudoknot base pairs*** | | | | | |
| 1 | RNADiffFold | **0.9930** | **0.6369** | **0.7529** | - |
| 2 | Rfold | 0.6831 | 0.4626 | 0.5348 | - |
| 3 | UFold | 0.6620 | 0.4649 | 0.5250 | - |
| 4 | RNA-FM | 0.3099 | 0.1811 | 0.2152 | - |
| 5 | REDfold | 0.1197 | 0.0901 | 0.0983 | - |
| 6 | SPOT-RNA | 0.2113 | 0.0309 | 0.0516 | - |
| ***Nested base pairs*** | | | | | |
| 1 | UFold | 0.5703 | 0.6213 | **0.5887** | **0.5913** |
| 2 | Rfold | 0.5781 | 0.5970 | 0.5827 | 0.5843 |
| 3 | RNA-FM | **0.6251** | 0.3487 | 0.4209 | 0.4477 |
| 4 | SPOT-RNA | 0.3542 | 0.5840 | 0.4287 | 0.4465 |
| 5 | REDfold | 0.2378 | 0.0909 | 0.1203 | 0.0991 |
| 6 | RNADiffFold | 0.0029 | **0.6400** | 0.0057 | 0.0333 |
| ***All base pairs*** | | | | | |
| 1 | UFold | 0.5820 | 0.6094 | **0.5897** | **0.5918** |
| 2 | Rfold | 0.5770 | 0.5831 | 0.5761 | 0.5771 |
| 3 | RNA-FM | **0.7486** | 0.3360 | 0.4110 | 0.4399 |
| 4 | SPOT-RNA | 0.3356 | 0.4887 | 0.3979 | 0.3941 |
| 5 | REDfold | 0.2305 | 0.1019 | 0.1183 | 0.0825 |
| 6 | RNADiffFold | 0.0032 | **0.6401** | 0.0064 | 0.0351 |

"-" indicates that MCC is not reported for pseudoknot base pairs.

# References


[1] Sato, K., Akiyama, M. & Sakakibara, Y. RNA secondary structure prediction using deep learning with thermodynamic integration. *Nat. Commun.* **12**, 941 (2021).

[2] Wang, N. *et al.* Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning. *Nat. Mach. Intell.* **6**, 548–557 (2024).

[3] Wang, Z. *et al.* RNADiffFold: generative RNA secondary structure prediction using discrete diffusion models. *Brief. Bioinform.* **26**, bbae618 (2025).

[4] Fu, L. *et al.* UFold: fast and accurate RNA secondary structure prediction with deep learning. *Nucleic Acids Res.* **50**, e14–e14 (2022).

[5] Chen, C.-C. & Chan, Y.-M. REDfold: accurate RNA secondary structure prediction using residual encoder-decoder network. *BMC Bioinformatics* **24**, 122 (2023).

[6] Singh, J., Hanson, J., Paliwal, K. & Zhou, Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nat. Commun.* **10**, 5407 (2019).

[7] Tan, C. *et al.* Deciphering RNA secondary structure prediction: a probabilistic K-rook matching perspective. In *Proceedings of the 41st International Conference on Machine Learning*, **235**, 47564–47578 (PMLR, 2024).

[8] Chen, J. *et al.* Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. Preprint at https://arxiv.org/abs/2204.00300 (2022).

[9] Akiyama, M., Sakakibara, Y. & Sato, K. Direct inference of base-pairing probabilities with neural networks improves prediction of RNA secondary structures with pseudoknots. *Genes* **13**, 2155 (2022).

[10] Yang, E. *et al.* GCNfold: a novel lightweight model with valid extractors for RNA secondary structure prediction. *Comput. Biol. Med.* **164**, 107246 (2023).

[11] Saman Booy, M., Ilin, A. & Orponen, P. RNA secondary structure prediction with convolutional neural networks. *BMC Bioinformatics* **23**, 58 (2022).

[12] Fei, Y., Zhang, H., Wang, Y., Liu, Z. & Liu, Y. LTPConstraint: a transfer learning-based end-to-end method for RNA secondary structure prediction. *BMC Bioinformatics* **23**, 354 (2022).

[13] Zhang, H. *et al.* A new method of RNA secondary structure prediction based on convolutional neural network and dynamic programming. *Front. Genet.* **10**, 467 (2019).

[14] Wang, L. *et al.* DMfold: a novel method to predict RNA secondary structure with pseudoknots based on deep learning and improved base pair maximization principle. *Front. Genet.* **10**, 143 (2019).

[15] Chen, X., Li, Y., Umarov, R., Gao, X. & Song, L. RNA secondary structure prediction by learning unrolled algorithms. Preprint at https://arxiv.org/abs/2002.05810 (2020).

[16] Yuan, S. *et al.* MSFF-CDCGAN: a novel method to predict RNA secondary structure based on generative adversarial network. *Methods* **204**, 368–375 (2022).