Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Xiaoya Fan
Dalian University of Technology
Jingjing Chen
Northeastern University
Tingting Li
Liaoning University
Yubang Dong
Northeastern University
Tingwei Chen
Liaoning University
Zheng Zhao
zhaozheng@dlmu.edu.cn
Dalian Maritime University

Article

Keywords: RNA secondary structure prediction, deep learning, benchmarking, generalization, pseudoknot prediction

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11520838/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

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;

Abstract

Accurate RNA secondary structure prediction is essential for understanding RNA function, molecular interactions, and regulatory mechanisms across diverse biological processes. Reported performance for RNA secondary structure predictors is difficult to compare because studies use different datasets, preprocessing pipelines, splits and metrics. We present a unified benchmark for RNA secondary structure prediction and evaluate 16 representative methods spanning end-to-end deep models, hybrid models with thermodynamic or structural constraints, and pretrained RNA foundation models. Across five settings, intra-dataset, cross-dataset, cross-family, length generalization and pseudoknot prediction, performance rankings vary markedly with evaluation protocol, and no single method dominates all settings. Methods incorporating biophysical priors or explicit structural constraints, together with some pretrained models, are generally more robust under distribution shift, whereas long RNAs and pseudoknots remain challenging for most predictors.

Keywords: RNA secondary structure prediction, deep learning, benchmarking, generalization, pseudoknot prediction

1 Introduction

The biological functions of RNA largely depend on its structural state, rather than being determined solely by its primary sequence. For many non-coding RNAs (ncRNAs) and cis-regulatory elements, secondary structure participates in catalysis, ligand recognition, translational regulation, stability control, and RNA-RNA or RNA-protein interactions. It therefore represents a key layer that links sequence information to three-dimensional conformation and is essential for understanding RNA structure-function relationships [1–5]. This functional relevance also causes secondary structure to often exhibit stronger evolutionary conservation than the sequence itself, making it a long-standing basis for ncRNA discovery, functional annotation, and conservation analysis [1, 6]. In addition, reliable secondary structure information provides an important foundation for three-dimensional structure modeling, RNA-targeted drug design, and synthetic biology design. Consequently, RNA secondary structure prediction has become a central problem in RNA biology and related application-oriented research [1, 3].

RNA secondary structures can be determined by experimental approaches such as chemical probing, enzymatic cleavage, nuclear magnetic resonance (NMR) spectroscopy, and crystallography [7–9]. However, these approaches are often limited by low throughput and relatively high cost, and RNA secondary structure analysis has therefore long relied heavily on computational prediction [10–13]. Classical computational methods are mainly built upon nearest-neighbor thermodynamic models and dynamic programming frameworks, with ViennaRNA and RNAstructure remaining important baseline tools for RNA secondary structure prediction [10, 11]. In recent years, advances in deep learning have been gradually reshaping RNA secondary structure prediction from a field dominated by classical algorithms toward a data-driven paradigm centered on deep learning. Existing methods can be broadly grouped into three categories. The first category comprises hybrid models that incorporate neural network-based scoring functions, energy terms, or structural constraints into traditional folding frameworks, such as MXfold2 and related constraint-based methods [14–17]. The second category includes end-to-end models that directly learn base-pairing patterns from sequences or contact maps, such as SPOT-RNA, UFold, and RNADiffFold [18–24]. The third category consists of foundation models that learn general-purpose RNA representations through large-scale unsupervised pretraining and transfer them to downstream structure prediction tasks, such as RNA-FM and RNAErnie [25, 26]. Although the number of methods has continued to increase and reported performance has steadily improved, practical applications require a shift in focus from identifying the best-performing model on a single dataset to understanding which types of methods are more robust and reliable in specific application scenarios.

However, a clear answer is still lacking regarding the relative reliability of different methods in practical application scenarios. Existing studies often report performance under different data sources, redundancy filtering criteria, training-validation-test splits, sequence length ranges, treatments of pseudoknots, and evaluation metrics, making published results difficult to compare directly and fairly

[12, 13, 27]. More importantly, when training and test sets overlap in family composition, homologous sequences, or structural patterns, model performance can be substantially overestimated. In contrast, challenges that are more relevant to real-world use, including cross-dataset transfer, cross-family generalization, length extrapolation, and pseudoknot prediction, are often not systematically evaluated within the same framework [12, 13, 28, 29]. Therefore, what is needed is not the continued accumulation of "state-of-the-art" results under mutually incomparable experimental settings, but a third-party, unified, and reproducible comparative benchmark that can identify the strengths, limitations, and applicability boundaries of different modeling paradigms.

Motivated by this need, we construct a unified benchmark for RNA secondary structure prediction (An overview of the benchmark is shown in Fig. 1). Under a unified protocol for data preprocessing, redundancy control, and evaluation metrics, we retrain or fine-tune 16 representative models on the same benchmark datasets based on their publicly available implementations, and conduct an independent systematic comparison. Specifically, we evaluate their performance across five scenarios: intra-dataset evaluation, cross-dataset evaluation, cross-family evaluation, length-generalization evaluation, and pseudoknot prediction. Unlike studies that focus on a single dataset or a single family of models, our work emphasizes model robustness and generalization under distribution shifts, increased sequence lengths, and greater structural complexity. Our goal is not to identify a context-independent "best" model, but to reveal the applicability boundaries of different modeling paradigms and provide a reusable framework for future method comparison and model selection.

2 Results

2.1 Performance on the intra-dataset benchmark

We first compared 16 RNA secondary structure prediction methods under the intra-dataset setting, where the training, validation, and test sets were derived from splits of the same dataset (Table 1). Considering the high sparsity of RNA base-pair contact maps, we ranked and compared the following results primarily by MCC, while also using Precision, Recall, and F1 to characterize the predictive behavior of different methods.

On the smaller Rivas dataset, RNADiffFold achieved the best overall performance, with an MCC of 0.559 and an F1 of 0.549. RNAErnie, REDfold [30], and MXfold2 followed closely, all achieving MCC values above 0.50. The performance gaps among the top four methods were relatively small, whereas the bottom four models still had MCC values close to zero, indicating a clear stratification in effective predictive ability even under in-distribution evaluation. Notably, although RNA-FM achieved the highest Precision (0.634), its Recall was only 0.310, and its overall MCC ranked fifth, suggesting a clear tendency toward conservative prediction.

On the larger and more compositionally complex bpRNA-1m dataset, overall model performance improved markedly, but the ranking of methods also changed substantially. MXfold2 achieved the highest MCC (0.711), with well-balanced Precision

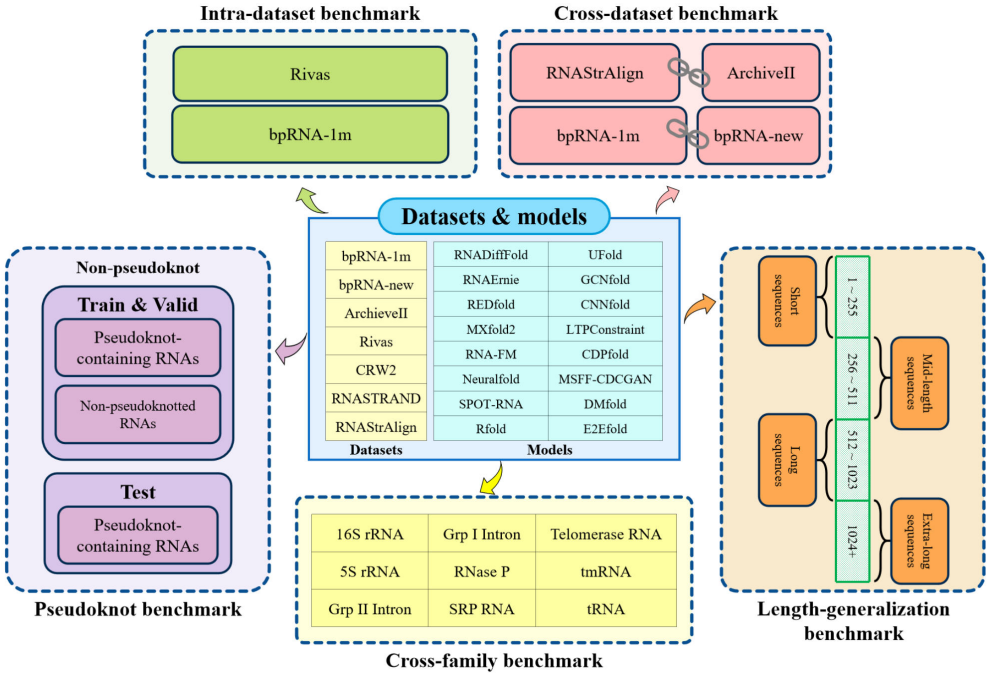

Fig. 1 Overview of the unified benchmark for RNA secondary structure prediction, integrating the datasets, prediction models and five evaluation scenarios used in this study.

and Recall values of 0.713 and 0.715, respectively. It outperformed the second-ranked UFold (0.612) by 0.099 in MCC, showing a clear leading advantage. UFold, Rfold [31], RNAErnie, and RNADiffFold ranked next, with MCC values concentrated in the range of 0.594-0.612. Compared with their performance on the Rivas dataset, UFold and Rfold rose from ninth and eighth to second and third, respectively, whereas the relative rankings of RNAErnie and RNA-FM decreased. These changes suggest that, with larger-scale supervised data, methods specifically designed for RNA secondary structure prediction become more competitive.

2.2 Performance on the cross-dataset benchmark

We further evaluated the generalization ability of different models under distribution shifts using the cross-dataset setting, with the results shown in Table 2. Here, RNAStrAlign→ArchiveII indicates that models were trained on RNAStrAlign and tested on ArchiveII, whereas bpRNA-1m→bpRNA-new indicates that models were trained on bpRNA-1m and tested on bpRNA-new.

In the RNAStrAlign→ArchiveII task, the top seven models all achieved MCC values above 0.7. RNAErnie obtained the best performance, with an MCC of 0.7929 and an F1 of 0.7924, followed by UFold and MXfold2, with MCC values of 0.7877 and 0.7640, respectively. Notably, the top three methods all explicitly incorporate structural or thermodynamic constraints: RNAErnie and MXfold2 integrate thermodynamic information, whereas UFold introduces base-pairing rules. RNADiffFold,

REDfold, SPOT-RNA, and Rfold followed next, with comparable F1 and MCC values, but their Precision-Recall balance was generally weaker than that of the top-performing methods. For example, SPOT-RNA achieved a high Recall of 0.8045 but a relatively low Precision of 0.6716, resulting in an overall ranking of sixth.
In the bpRNA-1m→bpRNA-new task, performance differences among models became more pronounced. MXfold2 again achieved the highest MCC of 0.5969 and clearly outperformed the other methods. SPOT-RNA showed a marked improvement in ranking compared with its intra-dataset performance, ranking second with an MCC of 0.5365 and an F1 of 0.5467. In contrast, although RNAErnie (MCC = 0.5203) and UFold (MCC = 0.5076) remained among the leading methods, their rankings decreased. Compared with the results on RNAStrAlign→ArchiveII, the relative advantages of different models were not fully consistent across cross-dataset transfer tasks, indicating substantial differences in their sensitivity to distribution shifts.
Consistent with the intra-dataset results, RNA-FM still exhibited a combination of high Precision and low Recall in the cross-dataset evaluation. In addition, several methods showed substantial performance degradation in at least one cross-dataset task, with F1 and MCC values close to zero.

2.3 Cross-family generalization

Based on the overall rankings in Supplementary Fig. S1, the following analyses of more complex scenarios focus on the 8 models with higher overall rankings and relatively stable performance.
Cross-family evaluation requires the training and test sets to come from different RNA families. It is therefore more stringent than simple cross-dataset transfer and better assesses whether a model has learned transferable folding principles rather than relying mainly on family-specific similar samples. Following the definition in the Benchmark design section, we evaluated model performance on both the easy and complex cross-family benchmarks.
On the easy cross-family benchmark, the average performance of different models is shown in Fig. 2 and Supplementary Table S4. RNADiffFold achieved the highest MCC (0.5139), F1 (0.4814), and Precision (0.7778), but its Recall was relatively low (0.4113), indicating a clear conservative prediction pattern. In contrast, MXfold2 achieved the highest Recall (0.4684) and the second-highest F1 (0.4723), showing a more balanced Precision-Recall trade-off. UFold and SPOT-RNA showed intermediate overall performance, with UFold having higher Precision but lower Recall, whereas SPOT-RNA showed the opposite pattern. RNA-FM also exhibited a conservative pattern with high Precision (0.7160) but extremely low Recall (0.0456), yielding the lowest F1 (0.0727) and MCC (0.1058) in this setting. Rfold also performed poorly (F1 = 0.1748, MCC = 0.1814). Overall, model performance generally declined when the training and test sets were separated by RNA family.
We further performed a more challenging complex cross-family evaluation that retained longer sequences and pseudoknotted structures; because this setting involved model-specific input-length constraints, the detailed results are reported in Supplementary Figure S2 and Supplementary Table S5. The additional analysis

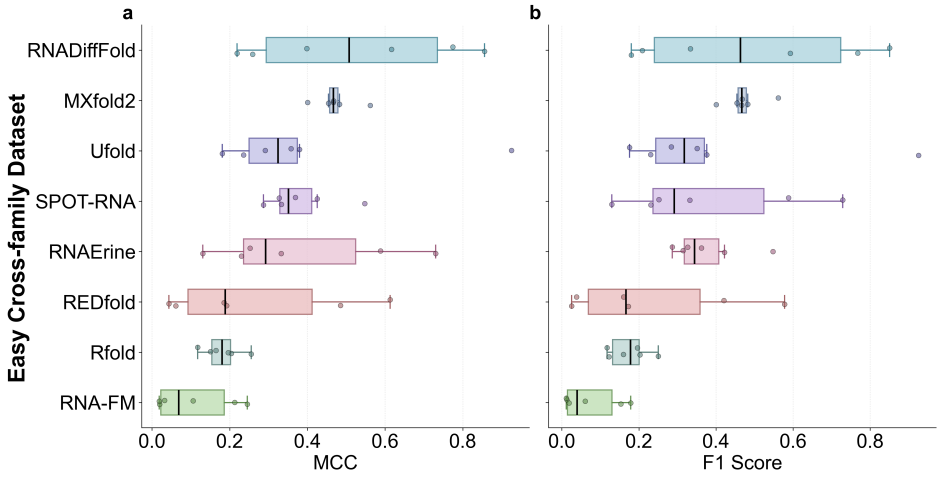

Fig. 2 Easy cross-family evaluation results. (a) and (b) show the MCC and F1 distributions under the easy cross-family setting, respectively. Each point represents the result for one RNA family.

showed that model rankings were sensitive to both family-level distribution shift and model-specific input-length constraints.

2.4 Length generalization

As described in the Benchmark design section, we stratified the test set by sequence length and compared model performance across different length intervals (Supplementary Table S6) to evaluate length generalization. Overall, model performance generally declined as sequence length increased, while the number of applicable methods decreased substantially. This suggests that increasing sequence length not only makes structure prediction more difficult, but also imposes stricter requirements on model scalability.

The best-performing model varied across sequence-length intervals (Fig. 3). For short sequences (1-255 nt), UFold achieved the highest mean F1 (0.6589) and mean MCC (0.6643). When the sequence length increased to mid-length (256-511 nt), RNAErnie replaced UFold as the best-performing method, with a mean F1 of 0.5409 and a mean MCC of 0.5451. This shift indicates that model advantages are strongly length-dependent, and that methods leading on short sequences do not necessarily maintain their advantage on longer sequences. Differences among models became more pronounced for longer RNAs. In the 512-1023 nt interval, only SPOT-RNA and MXfold2 were able to generate predictions, and both showed further performance declines compared with the mid-length interval. For extra-long sequences (1024+ nt), only MXfold2 remained applicable, achieving an F1 of 0.3512 and an MCC of 0.3567. These results show that the challenge introduced by increasing sequence length lies not only in reduced prediction accuracy, but also in the limited scalability of most methods to long RNAs. In contrast, MXfold2 showed the strongest applicability to extra-long sequences.

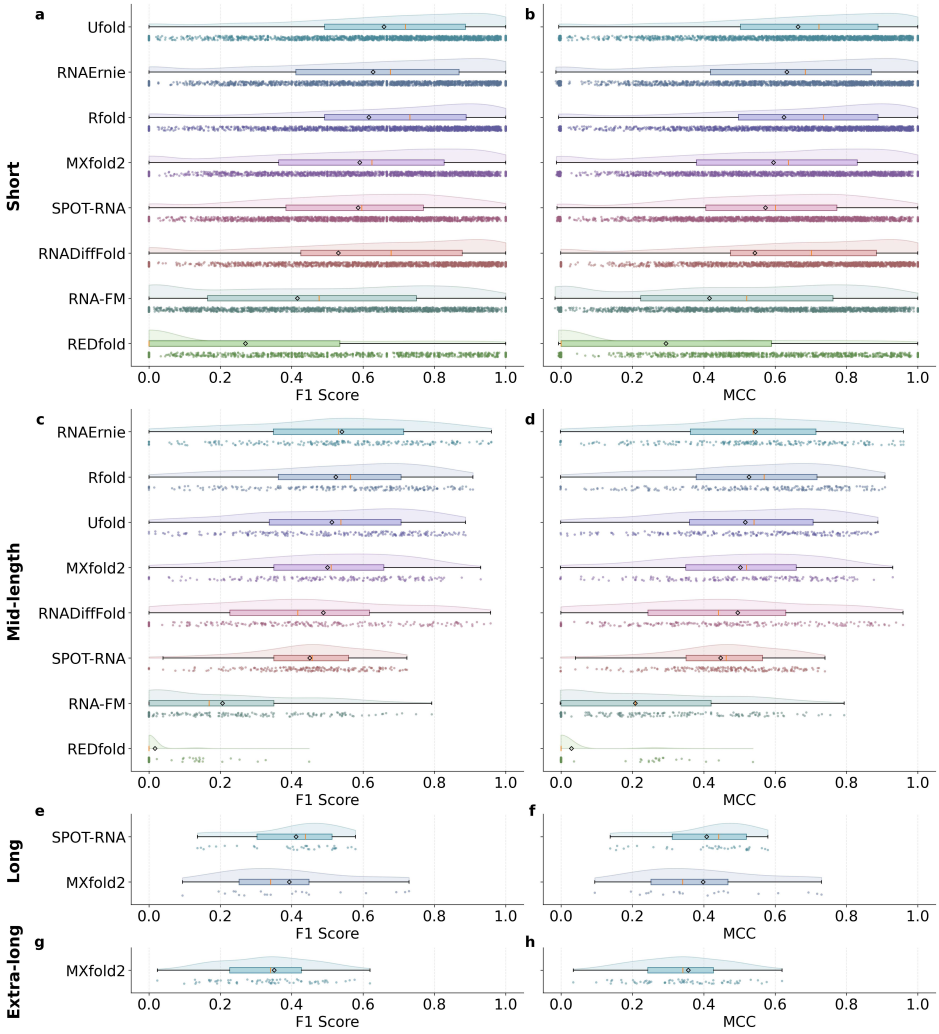

Fig. 3 Prediction performance distributions across different sequence-length intervals. (a), (c), (e), and (g) show the F1 distributions for short sequences (1-255 nt), mid-length sequences (256-511 nt), long sequences (512-1023 nt), and extra-long sequences (1024+ nt), respectively; (b), (d), (f), and (h) show the corresponding MCC distributions. Each point represents the result for one RNA in the test set. Black diamonds represent the mean value.

2.5 Pseudoknot prediction performance

Because MXfold2 and the evaluated RNAErnie implementation use dynamic programming-based decoders that generate only pseudoknot-free secondary structures, they were excluded from the pseudoknot prediction evaluation. The remaining evaluated models were retrained after incorporating pseudoknot-containing training samples and were evaluated on a test set consisting of 142 pseudoknot-containing

RNA sequences (see Methods; Supplementary Table S7). To distinguish the predictive ability of each model for different types of base pairs, we separately reported results for pseudoknot base pairs, nested base pairs, and all base pairs. For pseudoknot base pairs, only Precision, Recall, and F1 were reported, as defined in the Evaluation metrics section. The results are shown in Fig. 4.

For pseudoknot base pairs, RNADiffFold achieved the highest Precision (0.9930) and F1 (0.7529), substantially outperforming the other methods. However, this advantage did not translate into better overall structure prediction performance. For nested base pairs and all base pairs, its Precision values were only 0.0029 and 0.0032, respectively, and the corresponding F1 and MCC values were also close to zero. These results suggest that although RNADiffFold can identify a subset of pseudoknotted base pairs, it struggles to simultaneously maintain consistency in canonical nested pairs and the overall base-pairing pattern. By contrast, UFold and Rfold showed more balanced performance across the three evaluation categories. Both achieved relatively high F1 scores for pseudoknot base pairs (UFold: 0.5250; Rfold: 0.5348) and maintained the best overall performance for nested base pairs and all base pairs. UFold ranked first in both categories, with an F1 of 0.5887 and an MCC of 0.5913 for nested base pairs, and an F1 of 0.5897 and an MCC of 0.5918 for all base pairs, followed closely by Rfold. Although their overall performance was slightly lower than that on the bpRNA-1m dataset in the intra-dataset scenario, these results indicate that both methods can better preserve overall structure prediction ability in the presence of pseudoknots. RNA-FM continued to show a conservative prediction strategy: it achieved high Precision for all base pairs and nested base pairs (0.7486 and 0.6251, respectively), but its F1 for pseudoknot base pairs was only 0.2152. This indicates that RNA-FM can output a small number of high-confidence base pairs, but still has limited coverage of pseudoknotted topology. Notably, although SPOT-RNA retained reasonable predictive performance for nested base pairs (MCC = 0.4465), its performance on pseudoknot base pairs was close to failure (Precision = 0.2113, Recall = 0.0309, F1 = 0.0516), suggesting that this method may be more sensitive to changes in data distribution.

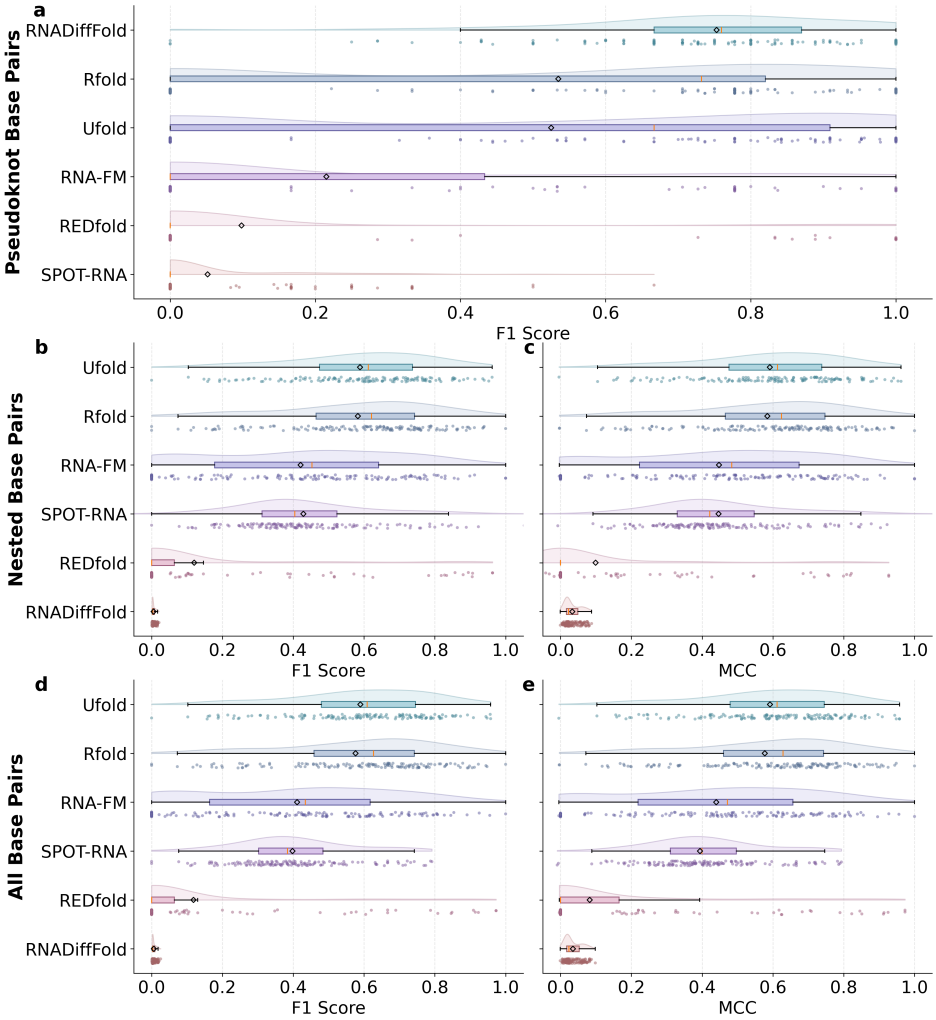

Fig. 4 Model performance on the pseudoknot-containing RNA test set. (a) F1 distribution for pseudoknot base pairs; (b,c) F1 and MCC distributions for nested base pairs; (d,e) F1 and MCC distributions for all base pairs. Each point represents one RNA in the test set. Black diamonds represent the mean value.

Table 1 Intra-dataset benchmark performance comparison.

Rank	Model	Precision	Recall	F1	MCC
Dataset A: Rivas					
1	RNADiffFold	0.61897	0.5239	**0.54945**	**0.55919**
2	RNAErnie	0.4962	0.5825	0.5302	0.5328
3	REDfold	0.60086	0.49474	0.52034	0.53161
4	MXfold2	0.4586	0.5745	0.5048	0.5085
5	RNA-FM	**0.6336**	0.3100	0.4163	0.4419
6	Neuralfold	0.31932	0.54138	0.40170	0.41574
7	SPOT-RNA	0.34492	0.51706	0.4138	0.41375
8	Rfold	0.4009	0.4292	0.4076	0.4094
9	UFold	0.42762	0.31208	0.34301	0.34892
10	GCNfold	0.2610	0.0040	0.0060	0.2450
11	CNNfold	0.24140	0.23530	0.23827	0.23139
12	LTPConstraint	0.02050	0.02819	0.02373	0.02381
13	CDPfold	0.1552	0.3377	0.2126	0.0025
14	MSFF-CDCGAN	0.33393	**0.72354**	0.45696	0.00035
15	DMfold	0.39680	0.22241	0.27717	-0.02177
16	E2Efold	0.00663	0.00173	0.00268	-0.04005
Dataset B: bpRNA-1m					
1	MXfold2	0.7127	**0.7151**	**0.7106**	**0.7111**
2	UFold	0.5800	0.6639	0.6054	0.6119
3	Rfold	0.5732	0.6433	0.5916	0.5982
4	RNAErnie	0.5989	0.6099	0.5918	0.5963
5	RNADiffFold	0.6373	0.5833	0.5847	0.5943
6	REDfold	0.5753	0.5185	0.5227	0.5325
7	SPOT-RNA	0.4345	0.6205	0.5111	0.5023
8	RNA-FM	**0.8028**	0.3046	0.4416	0.4937
9	Neuralfold	0.3406	0.5574	0.4228	0.4357
10	CNNfold	0.3232	0.3194	0.3213	0.3145
11	GCNfold	0.1330	0.0372	0.0507	0.0640
12	LTPConstraint	0.0352	0.0522	0.0420	0.0426
13	E2Efold	0.0000	0.0000	0.0000	0.0000
14	CDPfold	0.1645	0.3333	0.2203	-0.0033
15	DMfold	0.4872	0.3060	0.3382	-0.0035
16	MSFF-CDCGAN	0.3001	0.4124	0.3474	-0.0422

Models are sorted by MCC in descending order. For each metric, the highest value is indicated in bold.

Table 2 Cross-dataset benchmark performance comparison.

Rank	Model	Precision	Recall	F1	MCC
RNAStrAlign→ArchiveII					
1	RNAErnie	0.7817	**0.8082**	**0.7924**	**0.7929**
2	UFold	0.7919	0.7917	0.7860	0.7877
3	MXfold2	0.7528	0.7800	0.7635	0.7640
4	RNADiffFold	0.8379	0.6903	0.7305	0.7426
5	REDfold	0.7881	0.6926	0.7252	0.7316
6	SPOT-RNA	0.6716	0.8045	0.7320	0.7272
7	Rfold	0.7062	0.7509	0.7239	0.7254
8	RNA-FM	**0.8386**	0.4351	0.5729	0.6035
9	Neuralfold	0.4595	0.6372	0.5339	0.5411
10	CNNfold	0.3402	0.3322	0.3361	0.3300
11	GCNfold	0.3280	0.1380	0.1860	0.2080
12	LTPConstraint	0.0366	0.0622	0.0461	0.0474
13	CDPfold	0.1967	0.3571	0.2531	0.0291
14	DMfold	0.4813	0.2352	0.3077	-0.0171
15	E2Efold	0.0144	0.0094	0.0112	-0.0440
16	MSFF-CDCGAN	0.0956	0.6695	0.1673	-0.1438
bpRNA-1m→bpRNA-new					
1	MXfold2	0.5702	0.6384	**0.5932**	**0.5969**
2	SPOT-RNA	0.5018	0.6004	0.5467	0.5365
3	RNAErnie	0.5247	0.5319	0.5168	0.5203
4	UFold	0.4951	0.5368	0.5030	0.5076
5	RNADiffFold	0.5535	0.4380	0.4631	0.4751
6	REDfold	0.5662	0.3741	0.4143	0.4368
7	Rfold	0.4444	0.4401	0.4314	0.4349
8	RNA-FM	**0.7354**	0.1097	0.1909	0.2833
9	Neuralfold	0.0228	0.0527	0.0319	0.0347
10	DMfold	0.5050	0.2723	0.3380	0.0318
11	LTPConstraint	0.0157	0.0270	0.0198	0.0203
12	GCNfold	0.0387	0.0075	0.0117	0.0159
13	E2Efold	0.0000	0.0000	0.0000	0.0000
14	CDPfold	0.1642	0.3333	0.2200	-0.0038
15	MSFF-CDCGAN	0.3293	**0.9250**	0.4857	-0.0038
16	CNNfold	0.0054	0.0051	0.0053	-0.0057

3 Discussion

The core contribution of this study is an independent, third-party systematic comparison of 16 representative RNA secondary structure prediction methods under a unified benchmark. Using consistent protocols for data processing, data splitting, and evaluation, we retrained or fine-tuned these models based on their publicly available implementations. Our results show that RNA secondary structure predictors cannot be reliably assessed by a single in-distribution dataset or a single aggregate metric. Model rankings changed substantially across dataset shift, family shift, length extrapolation, and pseudoknot prediction settings, indicating that current performance claims are highly protocol-dependent. More challenging scenarios exposed clear limitations of existing predictors: performance declined under family-level distribution shifts, increasing sequence length reduced both accuracy and model applicability, and accurate recovery of pseudoknotted base pairs did not necessarily translate into strong whole-structure prediction. These findings support the need for reproducible, multi-scenario benchmarks that reflect biologically relevant distribution shifts and structural complexity.

From a methodological perspective, a relatively consistent observation is that methods explicitly incorporating structural rules or thermodynamic priors tend to show better robustness across multiple challenging scenarios. Methods represented by MXfold2 and RNAErnie showed strong competitiveness in cross-dataset transfer and some cross-family settings, suggesting that biophysical constraints remain valuable in current RNA secondary structure prediction. One possible reason is that high-quality labeled data available for supervised learning in RNA secondary structure prediction remain limited, and existing datasets show clear distributional biases in sequence length, structural complexity, and RNA type. Under such conditions, purely data-driven models may more easily learn statistical associations tied to the training distribution, leading to larger performance fluctuations when the distribution changes. In contrast, thermodynamic constraints or explicit structural rules can provide physical priors that are relatively independent of the training data, reducing outputs that substantially deviate from RNA folding principles and thereby improving model robustness and transferability. Nevertheless, this does not mean that data-driven methods lack advantages. Our results also show that end-to-end deep models can still achieve strong performance under in-distribution or short-sequence settings, indicating that task-specific deep architectures remain important for local pattern learning.

This study also suggests that integrating different modeling paradigms may be an important direction for future research. RNADiffFold combines one-hot sequence representations, UFold-derived structural information, and RNA-FM embeddings and attention features, thereby integrating end-to-end structure prediction, foundation model representations, and structural priors. Its strong performance on several benchmarks suggests that foundation model representations do not necessarily need to be used directly for final structure decoding, but can also serve as informative feature sources within specialized prediction frameworks. RNAErnie further combines pretrained representations with thermodynamic constraints and showed

relatively stable performance across multiple cross-distribution scenarios. By contrast, RNA-FM, as a general-purpose RNA foundation model, is not specifically optimized for secondary structure prediction and therefore exhibited a conservative pattern with high Precision and low Recall in this study. However, its representational information can still be exploited by specialized models such as RNADiffFold. Overall, future advances in high-performance RNA secondary structure prediction may depend not only on scaling up a single paradigm, but more importantly on effective synergy among data-driven learning, foundation model representations, and biophysical constraints.

Overall, the evaluation of RNA secondary structure prediction models should not rely on a single dataset or a single metric, but should instead consider model stability across cross-dataset, cross-family, variable-length, and complex-structure scenarios. Although integrating different paradigms may improve both accuracy and robustness, it may also introduce greater model complexity and computational cost. Therefore, achieving a better balance among prediction accuracy, generalization ability, and computational efficiency remains an important problem for future research. The unified benchmark established in this study provides a more transparent reference for future method comparison and may help promote RNA secondary structure prediction toward more robust, generalizable, and practically useful models.

4 Methods

4.1 Benchmark design

We designed five benchmark scenarios, including intra-dataset, cross-dataset, cross-family, length-generalization, and pseudoknot evaluation. The overall benchmark design is shown in Fig. 1, and the dataset sizes for each scenario are summarized in Supplementary Table S1. Considering that some methods impose constraints on input length, we limited RNA sequences in the intra-dataset and cross-dataset scenarios to no more than 512 nucleotides (nt), thereby enabling comparisons across different models under consistent conditions. All datasets were dereplicated using CD-HIT with a sequence identity threshold of 80%, to reduce the influence of data redundancy and potential homologous overlap on the evaluation results.

Intra-dataset scenario. We constructed two intra-dataset benchmarks based on the Rivas dataset [32] and the bpRNA-1m dataset [33], respectively (Supplementary Table S1). In this setting, the training, validation, and test sets were derived from splits of the same dataset, both datasets were dereplicated using CD-HIT-EST at an 80% sequence identity threshold. For the Rivas dataset, after redundancy removal, we followed its classical split and used TrainSetA, TestSetA, and TestSetB as the training, validation, and test sets, respectively. For the bpRNA-1m dataset, we retained only non-pseudoknotted sequences and, after redundancy removal, split them into training, validation, and test sets at a ratio of 8:1:1.

Cross-dataset scenario. The cross-dataset benchmark included two settings to evaluate the generalization ability of different models under changes in data source and family composition. In the first setting, the RNAStrAlign dataset [34] was split

into training and validation sets at a ratio of 8:2, and the models were tested on the ArchiveII dataset [35]. In the second setting, we directly used the models trained on bpRNA-1m in the intra-dataset scenario and evaluated them on bpRNA-new [14], thereby assessing their transfer performance under cross-dataset conditions.
Cross-family scenario. We collected RNA sequences with family annotations from RNAStrAlign, ArchiveII, CRW2 [36], bpRNA-1m, and RNASTRAND [37], and constructed an initial dataset covering nine RNA families: 16S rRNA, 5S rRNA, Group I intron, Group II intron, RNase P, SRP RNA, Telomerase RNA, tmRNA, and tRNA. Based on this dataset, we further designed two types of cross-family evaluation. The first was an easy cross-family scenario, in which pseudoknots were removed and sequence length was restricted. The second was a complex cross-family scenario, in which long sequences and pseudoknotted structures were retained to evaluate model generalization under more challenging cross-family conditions.
Length-generalization scenario. To evaluate length generalization, we first defined four sequence-length intervals: short sequences (1-255 nt), mid-length sequences (256-511 nt), long sequences (512-1023 nt), and extra-long sequences (1024+ nt). Based on these intervals, we further constructed a progressively expanded length-generalization scenario to examine how model performance changes in long-range dependency modeling and length extrapolation as the length ranges of the training and test sets increase. During training, we select the sub-dataset (standard length, extended length, extreme length) for each model based on its maximum output length. The corresponding dataset sizes are summarized in Supplementary Table S1.
Pseudoknot prediction scenario. To evaluate the ability of different models to capture pseudoknotted structures, we extracted pseudoknot-containing RNA sequences no longer than 511 nt from bpRNA-1m. After redundancy removal, these sequences were split into training, validation, and test sets at a ratio of 8:1:1. We then merged the pseudoknot-containing training and validation sets into the bpRNA-1m training and validation sets used in the intra-dataset scenario, while keeping the pseudoknot-containing test set separate for dedicated evaluation of pseudoknot prediction performance.

4.2 Evaluation metrics

We evaluated model performance using Precision, Recall, F1-score (F1) and Matthews correlation coefficient (MCC). Their definitions in terms of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) are summarized in Supplementary Table S2.
For pseudoknot-specific evaluation, we restricted the comparison to pseudoknotted base pairs only: TP_{PK} denotes correctly predicted pseudoknotted pairs, FP_{PK} denotes predicted pseudoknotted pairs absent from the reference, and FN_{PK} denotes reference pseudoknotted pairs missed by the prediction; PK-Precision, PK-Recall and PK-F1 were computed analogously. We did not report MCC for pseudoknot-specific evaluation because defining true negatives for this highly sparse subset is non-trivial and can introduce arbitrary choices that hinder comparability.

4.3 Models and implementation details

We compared 16 RNA secondary structure prediction models in this study, and their main architectures, training strategies, maximum input lengths, and support for pseudoknot prediction are summarized in Supplementary Table S3. All models were trained, fine-tuned, or evaluated on our benchmark datasets following their official implementations and default hyperparameter settings unless otherwise specified. For the pretrained foundation models RNAErnie and RNA-FM, we used the official pretrained weights and retrained only the downstream prediction heads. For RNADiffFold, we used the official RNA-FM pretrained model and initialized its UFold component with parameters trained on the corresponding benchmark task. All training and inference experiments were conducted on four NVIDIA GeForce RTX 3080 GPUs, each equipped with 10 GB of GDDR6X memory.

Author contributions

X.F. conceived and designed the benchmark, conducted the experiments and analyses, interpreted the results, and drafted the manuscript. J.C., T.L., and Y.D. contributed to data curation, model evaluation, and validation of the results. T.C. and Z.Z. contributed to study conceptualization, supervised the research, and critically revised the manuscript. All authors reviewed and approved the final manuscript.

Data availability

The data used in this study are available at https://doi.org/10.5281/zenodo.20792478.

Code availability

Code and evaluation scripts are available at https://github.com/XploreAI-Lab/A-unified-benchmark-of-deep-learning-methods-for-RNA-secondary-structure-prediction. Archived model files are available at https://doi.org/10.5281/zenodo.20775472.

Data availability

No new RNA nucleotide sequences or experimentally determined RNA structures were generated in this study. All RNA sequences and reference secondary-structure annotations were obtained from previously published and publicly available datasets, including Rivas, bpRNA-1m, RNAStrAlign, ArchiveII, bpRNA-new, CRW2, and RNA STRAND. The original accession numbers for individual RNA sequences are provided in the corresponding benchmark data files.
The curated benchmark datasets, including RNA sequences, secondary-structure annotations, original accession numbers, and training, validation, and test splits, are publicly available at Zenodo: https://doi.org/10.5281/zenodo.20792478.

Funding

This work was supported by the Liaoning Provincial Science and Technology Plan Joint Program Project (Natural Science Foundation General Program; Grant No. 2025-MSLH-112).

References

[1] Mortimer, S. A., Kidwell, M. A. & Doudna, J. A. Insights into RNA structure and function from genome-wide studies. *Nature Reviews Genetics* **15**, 469–479 (2014).

[2] Doudna, J. A. & Cech, T. R. The chemical repertoire of natural ribozymes. *Nature* **418**, 222–228 (2002).

[3] Tinoco, I. & Bustamante, C. How RNA folds. *Journal of Molecular Biology* **293**, 271–281 (1999).

[4] Serganov, A. & Nudler, E. A decade of riboswitches. *Cell* **152**, 17–24 (2013).

[5] Wu, L. & Belasco, J. G. Let me count the ways: mechanisms of gene regulation by miRNAs and siRNAs. *Molecular Cell* **29**, 1–7 (2008).

[6] Fu, Y., Xu, Z. Z., Lu, Z. J., Zhao, S. & Mathews, D. H. Discovery of novel ncRNA sequences in multiple genome alignments on the basis of conserved and stable secondary structures. *PLOS ONE* **10**, e0130200 (2015).

[7] Westhof, E. Twenty years of RNA crystallography. *RNA* **21**, 486–487 (2015).

[8] Fürtig, B., Richter, C., Wöhnert, J. & Schwalbe, H. NMR spectroscopy of RNA. *ChemBioChem* **4**, 936–962 (2003).

[9] Strobel, E. J., Yu, A. M. & Lucks, J. B. High-throughput determination of RNA structures. *Nature Reviews Genetics* **19**, 615–634 (2018).

[10] Lorenz, R. *et al.* ViennaRNA package 2.0. *Algorithms for Molecular Biology* **6**, 26 (2011).

[11] Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**, 129 (2010).

[12] Zhao, Q. *et al.* Review of machine learning methods for RNA secondary structure prediction. *PLOS Computational Biology* **17**, e1009291 (2021). URL https://doi.org/10.1371/journal.pcbi.1009291.

[13] Zhao, Q. *et al.* ML-based RNA secondary structure prediction methods: a survey. *AI Medicine* **1**, 6 (2024).

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

[15] 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).

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

[17] 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).

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

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

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

[21] Yang, E. *et al.* GCNfold: a novel lightweight model with valid extractors for RNA secondary structure prediction. *Computers in Biology and Medicine* **164**, 107246 (2023).

[22] Chen, X., Li, Y., Umarov, R., Gao, X. & Song, L. RNA secondary structure prediction by learning unrolled algorithms. *arXiv preprint arXiv:2002.05810* (2020).

[23] Wang, Z. *et al.* RNADiffFold: generative RNA secondary structure prediction using discrete diffusion models. *Briefings in Bioinformatics* **26**, bbae618 (2025).

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

[25] Chen, J. *et al.* Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *arXiv preprint arXiv:2204.00300* (2022).

[26] Wang, N. *et al.* Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning. *Nature Machine Intelligence* **6**, 548–557

(2024).

[27] Mathews, D. H. How to benchmark RNA secondary structure prediction accuracy. *Methods* **162–163**, 60–67 (2019).

[28] Szikszai, M., Wise, M., Datta, A., Ward, M. & Mathews, D. H. Deep learning models for RNA secondary structure prediction (probably) do not generalize across families. *Bioinformatics* **38**, 3892–3899 (2022). URL https://doi.org/10.1093/bioinformatics/btac415.

[29] Zablocki, L. I. *et al.* Comprehensive benchmarking of large language models for RNA secondary structure prediction. *Briefings in Bioinformatics* **26**, bbaf137 (2025). URL https://doi.org/10.1093/bib/bbaf137.

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

[31] Tan, C. *et al.* Salakhutdinov, R. *et al.* (eds) *Deciphering RNA secondary structure prediction: A probabilistic k-rook matching perspective.* (eds Salakhutdinov, R. *et al.*) *Proceedings of the 41st International Conference on Machine Learning*, Vol. 235 of *Proceedings of Machine Learning Research*, 47564–47578 (PMLR, 2024). URL https://proceedings.mlr.press/v235/tan24a.html.

[32] Rivas, E., Lang, R. & Eddy, S. R. A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more. *RNA* **18**, 193–212 (2012).

[33] Danaee, P. *et al.* bpRNA: large-scale automated annotation and analysis of RNA secondary structure. *Nucleic Acids Research* **46**, 5381–5394 (2018).

[34] Tan, Z., Fu, Y., Sharma, G. & Mathews, D. H. TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs. *Nucleic Acids Research* **45**, 11570–11581 (2017).

[35] Sloma, M. F. & Mathews, D. H. Exact calculation of loop formation probability identifies folding motifs in RNA secondary structures. *RNA* **22**, 1808–1818 (2016).

[36] Gutell, R. R. & Chan, C. X. CRW2: Comparative RNA web-2. https://crw2-comparative-rna-web.org/database/ (2025). URL https://crw2-comparative-rna-web.org/database/. Accessed: 2025-10-25.

[37] Andronescu, M., Bereg, V., Hoos, H. H. & Condon, A. RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics* **9**, 340 (2008).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- suppl.pdf