Deep learning models of RNA base-pairing structures generalize to unseen folds and make accurate zero-shot predictions of base-base interactions of RNA complexes

Mei Lang[1], Thomas Litfin[2], Ke Chen[1], Jian Zhan[1,3 a], and Yaoqi Zhou[1,2a]

[1]Institute of Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen, 518107, China
[2]Institute for Glycomics, Griffith University, Parklands Dr, Southport, QLD 4222, Australia
[3]Ribopeutic Inc, Guangzhou International Bio Island, Guangdong, 510320, China

[a]For correspondence, please email to Dr. Yaoqi Zhou (zhouyq@szbl.ac.cn) or Dr. Jian Zhan (zhanjian@szbl.ac.cn).

ABSTRACT

The intricate network of RNA-RNA interactions, crucial for orchestrating essential cellular processes like transcriptional and translational regulation, has been unveiling through high-throughput techniques and computational predictions. With the emergence of deep learning methodologies, the question arises: how do these cutting-edge techniques for base-pairing prediction compare to traditional free-energy-based approaches, particularly when applied to the challenging domain of interaction prediction via chain concatenation? In this study, we employ base pairs derived from three-dimensional RNA complex structures as the gold standard benchmark to assess the performance of 22 different methods, including recently developed deep learning models. Our results demonstrate that the deep-learning-based methods, SPOT-RNA and coevolution-information-powered SPOT-RNA2, can be generalized to previously unseen RNA structures and are capable of making accurate zero-shot predictions of RNA-RNA interactions. The finding underscores the potential of deep learning as a robust tool for advancing our understanding of these complex molecular interactions.

Introduction

Recent advancements in high-throughput techniques have unveiled a complex network of RNA-RNA interactions (RRIs) critical for governing transcriptional and translational processes. These interactions are pivotal in the biogenesis of various RNA molecules, including mRNAs, rRNA, tRNA, microRNAs, and circRNAs [1–3]. Large-scale detection of RRIs has been achieved through innovative approaches that combine cross-linking techniques with high-throughput sequencing. Notably, techniques such as PARIS [4], SPLASH[5], LIGR-seq[6], and COMRADES[7] have employed psoralen or its derivatives, as well as formaldehyde in the case of RIC-Seq [8], for cross-linking. While these methods hold promise, they are not without limitations stemming from probe biases and ligation efficiencies [1–3]. Furthermore, many of these high-throughput techniques have yet to achieve the single nucleotide resolution.

Attaining the nucleotide-level resolution in RNA structures has historically relied on

traditional structure-determination methods such as X-ray crystallography, Nuclear Magnetic Resonance (NMR), and Cryo-electron microscopy. Yet, compared to proteins, determining RNA structures presents formidable challenges due to the unique physiochemical properties of nucleotides and the inherent fragility of RNA structures [9]. This is reflected from the fact that only a meagre 3% of structures in the Protein Data Bank contain RNAs, with even fewer dedicated to RNA-RNA complexes (681 as of March 16, 2023, after redundancy removal)[10]. This stark contrast becomes even more pronounced when considering the extensive collection of more than 31 million noncoding RNA sequences catalogued in the RNAcentral database [11]. Given the cost and challenges associated with experimental approaches, there is an imperative need for development of complementary computational prediction techniques.

The existing methods for predicting RNA-RNA interactions (RRIs) can be broadly classified into alignment-based, free-energy-based, and homology modeling approaches [12,13]. Alignment-based techniques, such as GUUGle [14] and RIsearch [15], focus on inter-RNA base pairs while overlooking potential intra-RNA interactions. Free-energy-based methods can be categorized into those considering only intermolecular interactions for expediency (such as RNAhybrid[16], RNAduplex[17], RNAplex-c[18], and DuplexFold[19]), those factoring in intramolecular interactions based on solvent accessibility (such as RNAup[20], IntaRNA[21], RNAplex-a[22], and AccessFold[23]), and those accommodating both intra- and inter-molecular base pairs through sequence concatenation (such as PAIRFOLD[24], RNAcofold[25] and biFold[19]) or without restrictions (such as RactIP[26]). Homology-based techniques, exemplified by TargetRNA2 [27], CopraRNA [28], RNAaliduplex [17] and PETcofold[29], utilize evolutionary information to infer binding.

Presently, 'de novo' RRI prediction methods predominantly rely on free-energy-based approaches, limited by their approximate energy or scoring functions, akin to the challenges faced in RNA secondary structure prediction[30]. Recent advancements have seen the emergence of deep learning-based methods, starting with SPOT-RNA[31], which achieved the first end-to-end prediction of intra-RNA base pairs. Subsequent developments include mxfold2[32], UFold[33], and 2dRNA[34]. To further enhance prediction accuracy, SPOT-RNA2[35] was developed to integrate evolutionary profiles and mutational coupling data generated by RNAcmap [36]

In this study, we conducted a comprehensive benchmark of various methods for predicting intra- and inter-RNA interactions. Our evaluation encompassed traditional energy-based techniques and newly developed deep learning models based on simple sequence concatenation. To ensure a rigorous assessment, we employed base pairs derived from experimentally-determined RNA-RNA complex structures and eliminated monomer structures employed for training of SPOT-RNA and SPOT-RNA2 through a strict structural similarity cutoff (TM-score [37] <0.3). This challenging set of the complex of unseen structures revealed significant improvements of SPOT-RNA's performance over the other 21 methods evaluated, underscoring the transferability of deep learning from intra-RNA to inter-RNA interaction prediction.

Results

Method Comparison on Inter-RNA Base Pair Prediction

In this study, we compare 22 different RRI predictors on a benchmark set of 64 RNA-RNA pairs after excluding all monomer structures remotely similar to the RNA structures employed in the training and validation sets for SPOT-RNA and SPOT-RNA2. We evaluate their performance in inter-RNA base pair prediction through precision/recall curves and F1 score distributions, as shown in Figure 1. The performance metrics, including overall F1-score, Mathews correlation coefficient (MCC) values, and the median F1-score and standard deviation of individual RNA pairs, are also summarized in Table 1. Predictors with probabilistic outputs are represented by precision-recall (PR) curves, while others are represented as single points. For all methods with chain concatenation for predicting RNA-RNA interactions, a low case "c" is appended to the method name. They are RNAfoldc, UFoldc, MXfold2c, SPOT-RNAc, and SPOT-RNA2c. No linker was employed because adding a 3-nucleotide link did not lead to performance improvement (See Methods).

As shown in Figure 1A, SPOT-RNA2c is the most accurate predictor at low sensitivity (<0.2). However, the overall PR curve given by SPOT-RNAc has the best performance. We can also measure the performance by the overall F1-Score for all RNA pairs and the median F1-Score for individual RNA pairs. The thresholds for determining F1-scores of SPOT-RNAc and SPOT-RNA2c in the test set were set according to the thresholds for producing the highest F1-scores in the validation dataset for SPOT-RNAc and SPOT-RNA2c, respectively. Table 1 confirmed the result from the PR curve that SPOT-RNAc achieved the best overall performance with an overall F1-Score of 0.583, outperforming SPOT-RNA2c (the overall F1-Score of 0.561) and RNAcoFold (F1-Score of 0.494). SPOT-RNAc improves over SPOT-RNA2c by more than 4% and outperforms other methods by over 18%, a pattern similarly observed in MCC values (Table 1).

Figure 1B presents the distribution of F1 scores for individual RNA pairs, including median, 25th, and 75th percentiles. SPOT-RNAc continues to achieve the best performance with the highest median F1 score of 0.582, outperforming the next best SPOT-RNA2c (median F1-score of 0.533) with a 10% improvement. The improvement of SPOT-RNAc over all methods are statistically significant with a p-value of 0.012 when comparing to the second best method SPOT-RNA2c (Table 1). SPOT-RNAc also has the narrowest distribution among the top 5 predictors.

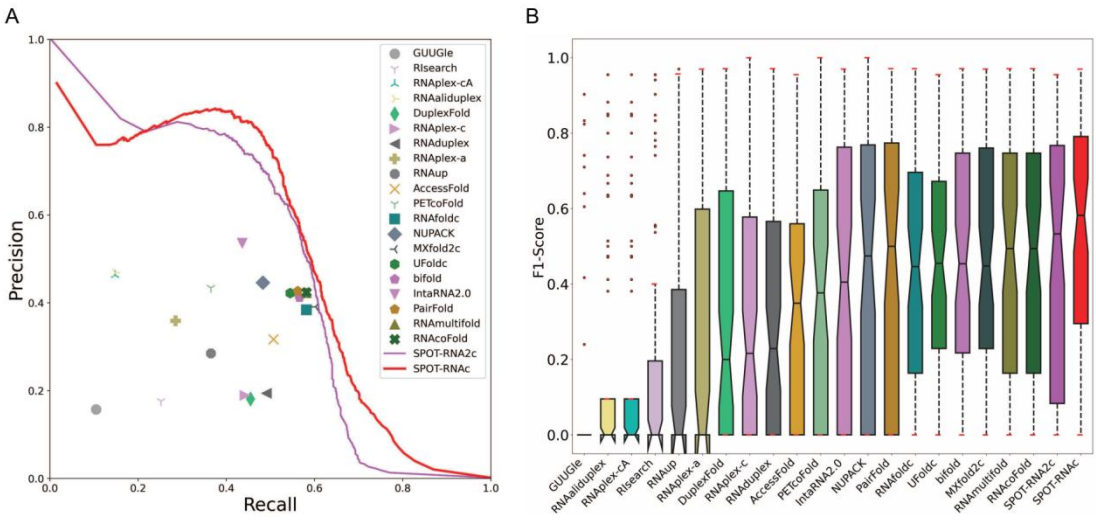

Figure 1: Performance Comparison of 22 Methods for Inter-RNA Base-Pair Prediction on the 64 Complexes of RNA Structures Unseen by SPOT-RNA and SPOT-RNA2. All structures in the test set have the structural similarity score TM-Score<0.3 compared to the monomeric structures used in training and validating SPOT-RNA and SPOT-RNA2 methods. (A) Precision-recall curves (for those methods with probabilistic outputs) or points given by 22 methods (B) Distribution of F1 scores for inter-RNA base pair prediction for individual RNA pairs by the same 22 methods. Each boxplot shows the median, 25th, and 75th percentiles, with outliers represented by "•". SPOT-RNAc exhibits the best performance for both overall and individual measures of F1-scores.

Table 1: Performance Comparison of 22 Predictors of Inter-RNA Base Pairs on 64 Complexes of RNA Structures Unseen by SPOT-RNA and SPOT-RNA2. All single-chain structures in the test set have the structural similarity score TM-Score<0.3 compared to the monomeric structures used in training and validating SPOT-RNA and SPOT-RNA2 methods.

Methods	Precision	Recall	Overall F1 Score	MCC	Median F1 Score ± Std	P-value
GUUGle	0.157	0.105	0.126	0.124	0.000 ± 0.241	$7.90e^{-15}$
RIsearch	0.176	0.252	0.207	0.205	0.000 ± 0.319	$2.03e^{-12}$
RNAplex-cA*	0.463	0.148	0.224	0.259	0.000 ± 0.296	$1.56e^{-11}$
RNAaliduplex*	0.469	0.148	0.225	0.261	0.000 ± 0.296	$1.56e^{-11}$
DuplexFold	0.181	0.455	0.259	0.281	0.200 ± 0.344	$6.12e^{-08}$
RNAplex-c	0.189	0.442	0.265	0.283	0.216 ± 0.343	$1.10e^{-08}$
RNAduplex	0.194	0.491	0.278	0.303	0.229 ± 0.329	$8.49e^{-09}$
RNAplex-a	0.359	0.285	0.318	0.316	0.000 ± 0.331	$1.43e^{-09}$
RNAup	0.285	0.365	0.32	0.318	0.000 ± 0.342	$4.13e^{-12}$
AccessFold	0.317	0.507	0.39	0.397	0.349 ± 0.315	$6.05e^{-06}$
PETcoFold*	0.434	0.365	0.397	0.395	0.376 ± 0.352	$9.24e^{-05}$
RNAfoldc	0.384	0.582	0.463	0.469	0.446 ± 0.318	0.007
NUPACK	0.446	0.483	0.464	0.461	0.474 ± 0.355	0.001
MXfold2c	0.391	0.603	0.474	0.482	0.448 ± 0.320	0.004
UFoldc	0.422	0.545	0.476	0.476	0.455 ± 0.302	0.002

bifold	0.412	0.566	0.477	0.48	0.454 ± 0.325	0.014
IntaRNA2.0	0.536	0.436	0.481	0.481	0.405 ± **0.377**	$9.75e^{-05}$
PairFold	0.427	0.562	0.485	0.486	0.500 ± 0.339	0.007
RNAmultifold	0.423	0.582	0.49	0.493	0.494 ± 0.325	0.015
RNAcoFold	0.424	**0.582**	0.491	0.494	0.494 ± 0.325	0.015
SPOT-RNA2c*	0.583	0.540	0.561	0.559	0.533 ± 0.331	0.012
SPOT-RNAc	**0.620**	0.551	**0.583**	**0.583**	**0.582** ± 0.326	

Note: The overall F1-score is harmonic mean of precision and recall for all RNA pairs. MCC denotes Matthew's correlation coefficient. The star * denotes the use of evolution information**.** Methods with an ending of "c" indicate the use of chain concatenation for RNA-RNA interaction prediction. Median F1 means the median value of single RNA. The P-value of a given method was computed by against the result from SPOT-RNAc.

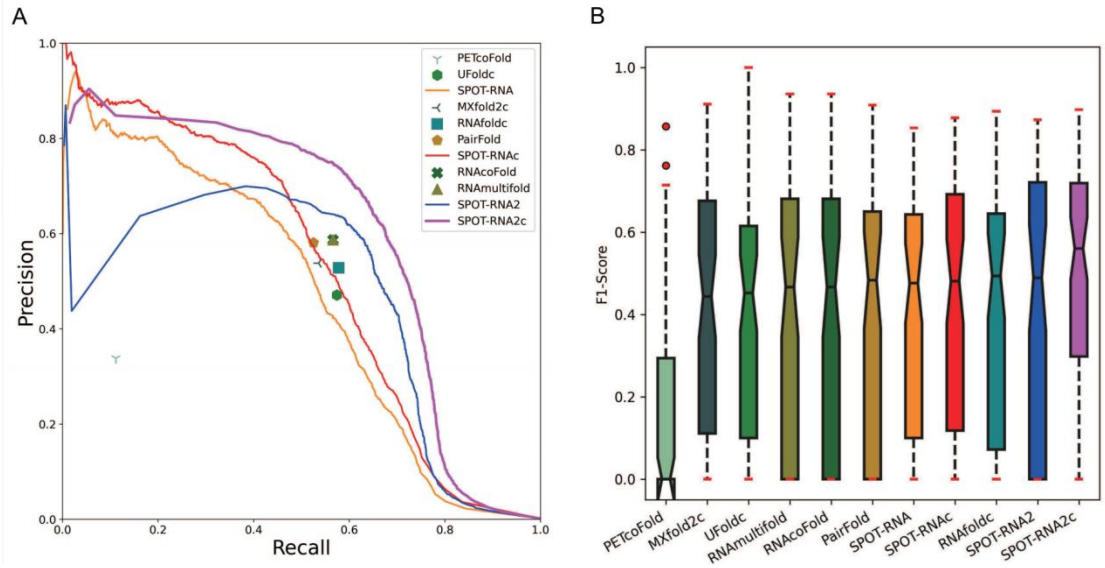

Figure 2: Performance Comparison of Intra-RNA Base Pair Prediction by 11 methods (A) Precision-recall curves and points illustrate the performance rankings. (B) A distribution of F1-scores for intra-RNA base pair prediction. SPOT-RNA and SPOT-RNAc represent intramolecular base pair prediction as single and concatenated chains, respectively. Evolution-based SPOT-RNA2c (or SPOT-RNA2) outperforms SPOT-RNAc (or SPOT-RNA) for intra-RNA base pairs.

It's important to assess how these methods perform on intra-molecular interactions, although not all RRI methods offer predictions for such interactions. We remove these RNAs without intra-base-pairing structures. This leads to 77 RNA chains. Figure 2A compares the PR curves or PR points given by 11 methods. Interestingly, PR curves indicate that SPOT-RNA2c now has the best performance. The overall performance according to overall F1-scores given by SPOT-RNA2c (Table 2) is the highest, surpassing the next best methods (SPOT-RNA2 without chain concatenation) by a 6% improvement in F1-score, and the third best (RNAmultifold or RNAcoFold) by 12%, with SPOT-RNAc ranking as the fourth best. This trend is consistent across MCC values.

Table 2: Performance Comparison of 11 Predictors for Predicting Intra-RNA Base Pairs on 77 unseen RNA chains that has intra secondary structure in 64 RNA-RNA Complex Structures.

Methods	Precision	Recall	Overall F1-Score	MCC	Individual F1-Score Median ±Std	p-value
PETcoFold*	0.338	0.112	0.168	0.194	0.000 ± 0.217	$1.62e^{-15}$
UFold	0.471	0.574	0.517	0.519	0.452 ± 0.278	$5.33e^{-07}$
SPOT-RNA	0.584	0.486	0.531	0.531	0.476 ± 0.271	$1.51e^{-06}$
MXfold2	0.538	0.536	0.537	0.536	0.444 ± 0.291	0.0001
RNAfoldc	0.528	0.578	0.552	0.551	0.494 ± 0.287	$1.06e^{-05}$
PairFold	0.581	0.525	0.552	0.551	0.483 ± 0.308	$8.15e^{-05}$
SPOT-RNAc	0.653	0.493	0.562	0.567	0.481 ± 0.291	0.0005
RNAcoFold	0.587	0.566	0.576	0.576	0.467 ± 0.305	0.001
RNAmultifold	0.587	0.566	0.576	0.576	0.467 ± 0.305	0.001
SPOT-RNA2*	0.589	0.629	0.608	0.607	0.489 ± 0.310	0.007
SPOT-RNA2c*	**0.626**	**0.665**	**0.645**	**0.644**	**0.560** ± 0.294	-

Note: The overall F1-score is harmonic mean of precision and recall for all RNA pairs. MCC denotes Matthew's correlation coefficient. The star * denotes the use of evolution information,Std means standard deviation . Median F1 means the median value of individual RNA chains. The P-value of a given method is compute by against the result of SPOT-RNAc.

Figure 2B further delves into performance by analyzing the distribution of F1-scores for individual RNAs. SPOT-RNA2c continues to lead with the highest median F1-score at 0.560, while SPOT-RNA2 follows closely with the second-best median F1-score of 0.608. The differences in F1-score distributions between SPOT-RNA2c and other methods are all statistically significant, with a p-value of 0.007 when comparing SPOT-RNA2c to the second best SPOT-RNA2 (Table 2).

Intuitively, a better intra-RNA base-pairing prediction should lead to a better inter-RNA base-pairing prediction. However, although SPOT-RNA2c has the best performance for intra-RNA interaction prediction, it is SPOT-RNAc with the best performance for inter-RNA interaction prediction. If we remove these RNA RRI pairs of which both chains do not have intra-base-pairing structures, this leads to 53 RRI pairs. Figure 3A compares intermolecular F1-scores for individual RNA pairs from SPOT-RNAc with the average intramolecular F1-scores. No correlation was found. Similar uncorrelated intra- and inter-RNA F1-scores are observed for SPOT-RNA2c (Figure 3B). This suggests that the evolution information contained in SPOT-RNA2c did contain the co-evolution information for predicting intra-RNA, but not inter-RNA interactions.

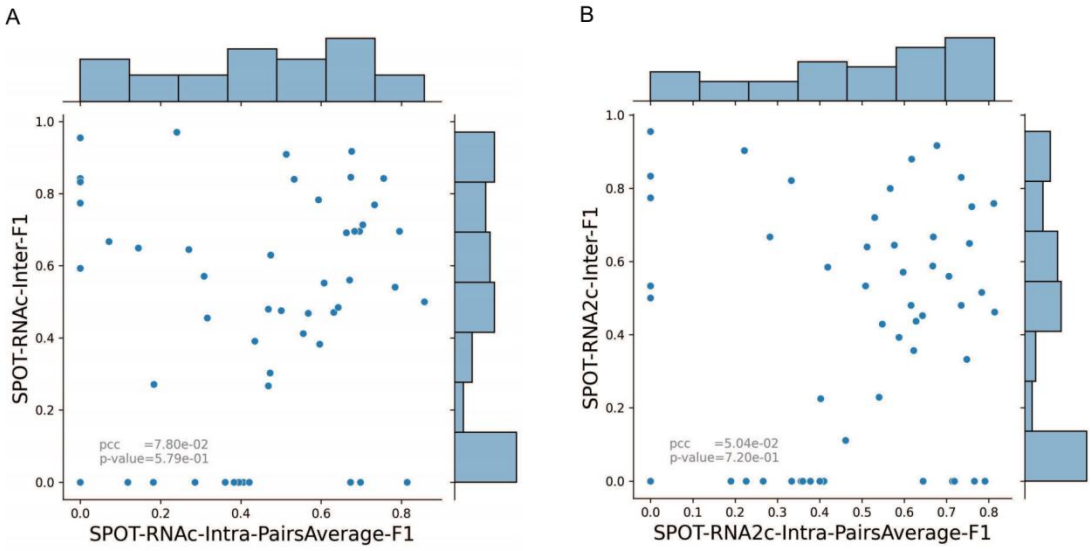

Figure 3 No Correlation between Inter-RNA F1-scores and Intra-RNA F1-scores. (A) Inter-RNA F1-scores versus the average Intra-RNA F1-scores of SPOT-RNAc for 53 RNA complex structures with intra-RNA base pairs for both chains. (B) Inter-RNA F1-scores versus Intra-RNA F1-scores of SPOT-RNA2c for 53 RNA complex structures.

We also assessed the impact of sequence concatenation on the intra-RNA base-pair prediction (i.e., SPOT-RNA versus SPOT-RNAc, SPOT-RNA2 versus SPOT-RNA2c). Table 2 shows that SPOT-RNAc/SPOT-RNA2c are better than SPOT-RNA/SPOT-RNA2 based on either overall F1-score or the median of individual F1-scores. In both cases, the difference is statistically significant with a p-value of 0.0007 between SPOT-RNA and SPOT-RNAc and a p-value of 0.007 between SPOT-RNA2 and SPOT-RNA2c. This indicates that knowing the binding partner improves the intra-RNA base pair prediction.

In Figures 3A and 3B, some inter-RNA (and intra-RNA) interactions were predicted with F1-scores of 0. To understand the reasons behind these poor predictions, we examined F1-scores given by SPOT-RNAc as a function of sequence length (L1+L2) in Figure 4A. No obvious correlation was found. However, when we plotted F1-scores against the number of true inter-RNA base pairs divided by the square root of (L1*L2), a clear and strong correlation emerged with a Pearson's correlation coefficient (PCC) of 0.464 (Figure 4B). Thus, poor predictions, including those with F1-scores of 0, can be attributed to the scarcity of inter-RNA contacts relative to the sequence lengths. This observation holds true for intra-RNA base pair prediction as well: intra-RNA interactions with F1-scores of 0 also involve very few intra-RNA base pairs (<5).

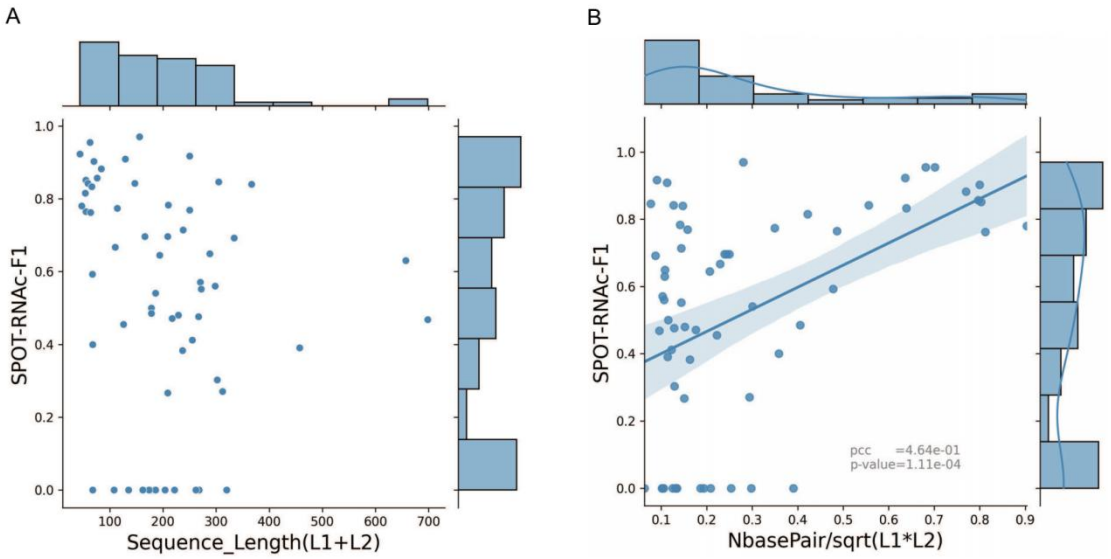

Figure 4: **Relationship Between Inter-RNA Base Pairing Prediction and Sequence/Interaction Characteristics.** (A) The inter-RNA F1 scores for individual RNA-RNA complexes from SPOT-RNAc plotted against the sum of sequence lengths (L1+L2). (B) The inter-RNA F1 scores for individual RNA-RNA complexes from SPOT-RNAc plotted against the normalized number of inter-RNA base pairs (the number of true inter-RNA base pairs divided by the square root of L1*L2). The performance does not correlate with sequence length but is related to the normalized number of inter-RNA base pairs.

We selected one example to illustrate SPOT-RNA's performance in RRI prediction. Figure 4A displays predicted and actual base-pairing maps for the subunits L2a rRNA complexed with L3b rRNA from Chlamydomonas reinhardtii mitoribosome (PDB ID 7PKT, chain ID 2 and chain ID 3). Figure 5A displays the intra- and inter-base-pairing maps of these two RNAs with Figure 5B for inter-base-pairing maps only. The F1 scores for intra-RNA base pairs are 0.32 for L2a rRNA and 0.30 for L3b rRNA, while the F1-score for inter-RNA base pairs is 0.571. Correctly predicted base pairs are highlighted with red dots in Figures 5B and 5C, and their 3D locations were shown as red-colored bases in Figure 5C. For this complex structure, precision for inter-RNA base pairs is 0.435, and sensitivity is relatively high (0.769).

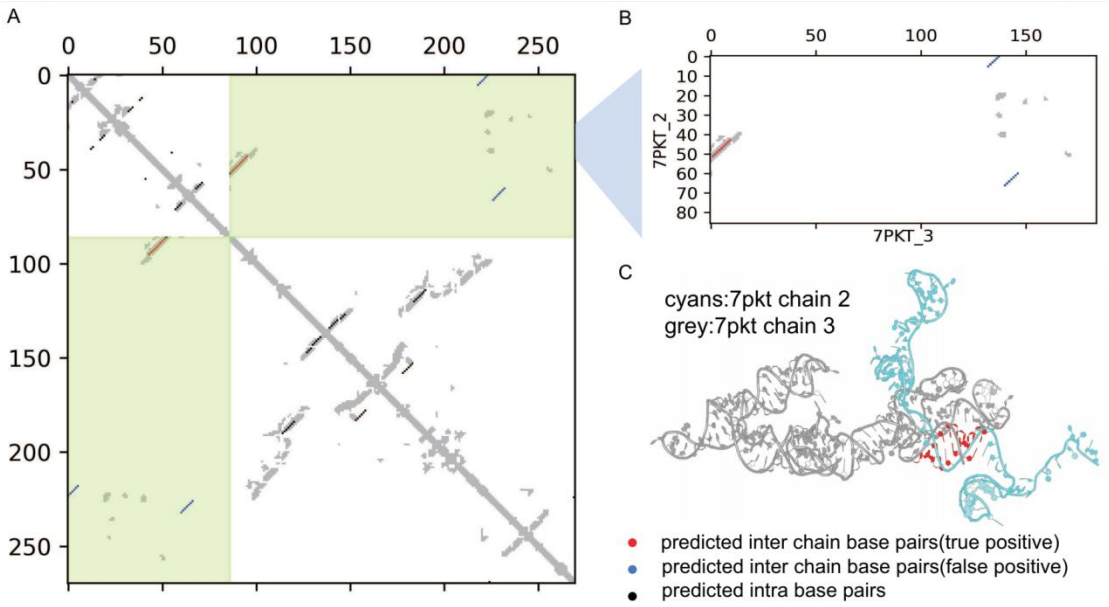

Figure 5: Accurate Prediction of Key Inter-RNA Base-pairing Contacts in RNA Complexes by SPOT-RNAc. Complex structure of large subunits of the Chlamydomonas reinhardtii mitoribosome (L2aRNA and L3bRNA in PDB ID 7PKT) with true distance-contact map and predicted intra and inter-RNA base pairs (A), inter-RNA base pairs only (B), and 3-D structure (C). Predicted intra-base pairs and inter-base pairs are denoted by black and red dots, respectively, in the base-pairing maps. In the 3-D structure (C), correctly predicted inter-RNA base pairs are highlighted in red.

Discussion

This study represents a comprehensive benchmark for assessing more than 20 methods in predicting RNA-RNA interactions. Previous benchmarks were constrained to interactions involving small RNAs without intra-RNA base pairs. For instance, Lai and Meyer [13] compared 14 RRI methods based on experimentally confirmed interactions in fungal snoRNA-rRNA and bacterial sRNA-mRNA pairs. Umu and Gardner (Umu and Gardner, 2017a) examined 15 RRI methods using a dataset focused on short linear base-pair matching. Antonov et al.[38] compared 13 RRI methods on mammalian lncRNAs with experimentally proven hybridizations. In contrast, our study presents the first comprehensive benchmark of 22 RRI prediction methods using known RRI interactions derived from 3D structures at the base pair level. Notably, most of these complexes (53/64) include RNAs with 3D structures and intra-RNA base pairs. Additionally, this study marks the first inclusion of deep-learning based methods for comparisons, utilizing chain concatenation.

In contrast to proteins, the available non-redundant data for RNA structures is limited. This raises concerns about the generalizability of deep learning models trained on such limited data. Previous studies by Szikszai et al [39] and Qiu [40] highlighted the challenges of deep-learning models when applied to unseen families not present in the training and validation sets. To evaluate the

adaptability of SPOT-RNA and SPOT-RNA2 beyond their training and validation data, we conducted a test by removing all test set structures with the structural similarity score (TM-score) ≥ 0.3 compared to those in the training and validation sets. SPOT-RNA and SPOT-RNA2 performed exceptionally well, outperforming most other methods for both intra- and inter-RNA base pair prediction. This demonstrates that deep learning, even with a limited number of 3D structures for training, can yield generalizable models for base pair prediction. This may be attributed to the fact that a large set of approximate secondary structures from bpRNA was used for training, followed by transfer learning with 3D structure-derived base pairs in SPOT-RNA and SPOT-RNA2.

SPOT-RNA2, which incorporates evolution and co-evolution information, outperforms SPOT-RNA for intra-RNA base pairs, aligning with previous findings (Table 2) [35]. However, SPOT-RNA2c underperforms SPOT-RNAc for inter-RNA base pairs. Notably, SPOT-RNA2 exhibits a positive correlation with the number of effective homologous sequences for intra-RNA base-pair prediction (PCC=0.5, p-value=2×10^{-6}) Supplementary Figure 1C), but this correlation nearly diminishes for SPOT-RNA2c for intra-RNA base pairs (PCC=0.27, p-value=0.05) and turned negative for inter-RNA base pair prediction (PCC=-0.3, p-value=0.02). This suggests that using linked sequences in SPOT-RNA2c for homology search may have provided harmful information for inferring inter-RNA interactions. In the future, it may be necessary to utilize sequences from the same species for homology searches, as co-evolution information can only be detected through inter-species comparisons via Multiple Sequence Alignment (MSA) pairing as has been done for proteins[41]. Such data for RNAs is few but could be explored in future studies.

For predicting RNA-RNA interactions, we concatenated two chains (A and B) as a single chain. To eliminate artificial sequence-order dependence, we predicted results for both AB and BA chains and then calculated the average. Interestingly, we found that in some cases, one sequence order (e.g., BA) outperformed the other (i.e., AB). Upon closer examination, we discovered for some RNA pairs that the order with a shorter separation in sequence positions for contacting base pairs tended to perform better. This observation is intuitive, as longer-range interactions are inherently more challenging to predict. However, there were some outerliers deserving further studies.

Methods

Benchmark dataset

We retrieved all RNA structures from the Protein Data Bank (PDB) in March 2023[10] and specifically selected structures featuring two RNA chains with a minimum of 5 inter-RNA base pairs. The identification of base pairs was carried out using DSSR[42]. To eliminate redundancy, we applied CD-HIT-EST[43], removing binding pairs with over 80% sequence identity between either chain. This initial step resulted in 155 unique RNA-RNA interaction (RRI) pairs.

To ensure stringency, we further filtered out RRI pairs that exhibited single-chain structural similarities with any RNAs in the SPOT-RNA training set, defined by TM-score ≥ 0.3 using RNA-align[44] with the length of the query sequence for normalization. This rigorous process yielded a final benchmark set comprising 64 unique RRI pairs. The PDB IDs of the benchmark set can be found in Supplementary Table 2.

Performance evaluation

We evaluated performance using common metrics: Recall (sensitivity), Precision, and the F1 score. Precision is TP/(TP+FP), Recall is TP/(TP+FN), and the F1 score is 2(Recall*Precision)/(Recall + Precision). Here, TP, FP, and FN are true positive, false positive and false negative, respectively. We also calculated Matthew's correlation coefficient (MCC) to provide a balanced measure as below:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}},$$

where TN denotes true negatives. MCC considers true negatives (TN) and measures the correlation between expected and observed classes. It ranges from 0 (no correlation) to 1 (highest correlation).

Methods for interaction prediction

We summarized the comparison methods in Table 3, with detailed settings provided in the Supplementary method description in the Supplementary Material. As per Lai and Meyer[13], we categorized the algorithms into four types: 1) 'Interaction only' methods predict intermolecular hybridization, ignoring RNA secondary structures; 2) 'Accessibility' methods consider RNA secondary structures using a partition function for unpaired probability; 3) 'Concatenation' algorithms treat two input sequences as a single chain and predict a joint secondary structure; and 4) 'Complex joint' methods also predict joint secondary structures but without concatenating the input sequences.

We used concatenation for comparing recent deep-learning methods with traditional free-energy-based methods. When dealing with concatenated chains, we made predictions for both sequence orders (AB and BA) and reported the average if probabilities were predicted. For those methods providing a two-state prediction, we considered the union of base pairs predicted for both sequence orders as a positive prediction (i.e. a positive prediction from either AB or BA concatenation will be considered as positive).

We experimented with and without a three-nucleotide linker (AAA, UUU, CCC, or GGG) and found that direct concatenation without any linker yielded slightly better results, although not statistically significant compared to AAA/CCC linkers (Supplementary Table 1). Therefore, we report results based on direct concatenation without any linkers.

Table 1 RRI interaction tools employed in this study for comparison, listed according to the year of publication along with their categories, the use of evolution information (MSA), the algorithm, and the capability of predicting intra-RNA base pairs.

Methods	Ref	Year	Category[a]	MSA[b]	Algorithms[c]	Intra-RNA base pairs[d]
RNAfoldc[e]	45	2004	concatenation	No	MFE	Yes
PairFold	24	2005	concatenation	No	MFE	Yes
RNAup	20	2006	accessibility	No	MFE	No
GUUGle	14	2006	interaction only	No	MFE	No
RNAcoFold	25	2006	concatenation	No	partition + MFE	Yes
NUPACK	46	2007	concatenation	No	MFE	No
RNAplex-c	18	2008	interaction only	No	MFE	No
RNAplex-a	18	2008	accessibility	No	MFE	No
RNAplex-cA	18	2008	interaction only	Yes	MFE	No
bifold	19	2010	complex joint	No	MFE	No
DuplexFold	47	2010	interaction only	No	MFE	No
PETcoFold	29	2011	complex joint	Yes	MFE	Yes
RNAduplex	17	2011	interaction only	No	MFE	No
RNAaliduplex	17	2011	interaction only	Yes	MFE	No
RNAmultifold	17	2011	concatenation	No	MFE	Yes
RIsearch	15	2012	interaction only	No	MFE	No
AccessFold	23	2016	concatenation	No	MFE	No
IntaRNA2.0	48	2017	accessibility	No	MFE + partition	No
SPOT-RNAc[e]	31	2019	concatenation	No	DL	Yes
SPOT-RNA2c[e]	35	2021	concatenation	Yes	DL	Yes
MXfold2c[e]	32	2021	concatenation	No	DL	Yes
UFoldc[e]	33	2022	concatenation	No	DL	Yes

[a] Category: the broad category of the method (see text for more details);
[b]MSA—indicate whether it takes multiple sequence alignment as input;
[c]Algorithm: MFE: minimum free energy, Partition; partition function, DL: deep learning;
[d]Intra-RNA Base Pairs —whether the output of a software also contains base pairing information for intramolecular interactions.
[e]c – indicates concatenation of two chains. For example SPOT-RNAc and SPOT-RNA2c denotes SPOT-RNA and SPOT-RNA2 with sequence concatenation, respectively, to distinguish from the methods dedicated to individual chains (SPOT-RNA and SPOT-RNA2, respectively)

Homology search

For these methods that do not predicted RRI by employing a linked chain such as RNAplex-cA, PETcoFold, RNAaliduplex and SPOT-RNA2, we searched the homologs using the single chain. For SPOT-RNA2c using a link chain, we searched the homologs with the linked chain.

Data and code availability

All data and codes are available at https://github.com/meilanglang/RNA-RNA-Interaction

Acknowledgement

We express our gratitude to Dr. Jaswinder Singh for valuable discussions and insights. This research received support from the National Natural Science Foundation of China (Grant # 22350710182), the Shenzhen Science and Technology Program (Grant No. KQTD20170330155106581), and benefited from access to the supercomputing resources at Shenzhen Bay Laboratory.

Conflict of Interest.

All authors declare no financial interest. Zhan and Zhou are the CEO and the chair of the scientific advisor board for Ribopeutic, respectively.

Reference

1. Dai, X., Zhang, S. & Zaleta-Rivera, K. RNA: interactions drive functionalities. *Mol Biol Rep* **47**, 1413–1434 (2020).
2. Singh, S., Shyamal, S. & Panda, A. C. Detecting RNA–RNA interactome. *WIREs RNA* **13**, e1715 (2022).
3. Gong, J., Ju, Y., Shao, D. & Zhang, Q. C. Advances and challenges towards the study of RNA-RNA interactions in a transcriptome-wide scale. *Quant Biol* **6**, 239–252 (2018).
4. Lu, Z. *et al.* RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. *Cell* **165**, 1267–1279 (2016).
5. Aw, J. G. A. *et al.* In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. *Mol Cell* **62**, 603–617 (2016).
6. Sharma, E., Sterne-Weiler, T., O'Hanlon, D. & Blencowe, B. J. Global Mapping of Human RNA-RNA Interactions. *Mol Cell* **62**, 618–626 (2016).
7. Ziv, O. *et al.* COMRADES determines in vivo RNA structures and interactions. *Nat Methods* **15**, 785–788 (2018).
8. Cai, Z. *et al.* RIC-seq for global in situ profiling of RNA–RNA spatial interactions. *Nature* **582**, 432–437 (2020).
9. Xu, B. *et al.* Recent advances in RNA structurome. *Sci China Life Sci* **65**, 1285–1324 (2022).
10. Burley, S. K. *et al.* RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res* **49**, D437–D451 (2021).
11. RNAcentral Consortium. RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res* **49**, D212–D220 (2021).
12. Umu, S. U. & Gardner, P. P. A comprehensive benchmark of RNA–RNA interaction prediction

tools for all domains of life. *Bioinformatics* **33**, 988–996 (2017).

13. Lai, D. & Meyer, I. M. A comprehensive comparison of general RNA-RNA interaction prediction methods. *Nucleic Acids Res* **44**, e61 (2016).

14. Gerlach, W. & Giegerich, R. GUUGle: a utility for fast exact matching under RNA complementary rules including G-U base pairing. *Bioinformatics* **22**, 762–764 (2006).

15. Wenzel, A., Akbasli, E. & Gorodkin, J. RIsearch: fast RNA-RNA interaction search using a simplified nearest-neighbor energy model. *Bioinformatics* **28**, 2738–2746 (2012).

16. Krüger, J. & Rehmsmeier, M. RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res* **34**, W451-454 (2006).

17. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**, 26 (2011).

18. Tafer, H. & Hofacker, I. L. RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics* **24**, 2657–2663 (2008).

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

20. Mückstein, U. *et al.* Thermodynamics of RNA-RNA binding. *Bioinformatics* **22**, 1177–1182 (2006).

21. Busch, A., Richter, A. S. & Backofen, R. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics* **24**, 2849–2856 (2008).

22. Tafer, H., Amman, F., Eggenhofer, F., Stadler, P. F. & Hofacker, I. L. Fast accessibility-based prediction of RNA-RNA interactions. *Bioinformatics* **27**, 1934–1940 (2011).

23. DiChiacchio, L., Sloma, M. F. & Mathews, D. H. AccessFold: predicting RNA-RNA interactions with consideration for competing self-structure. *Bioinformatics* **32**, 1033–1039 (2016).

24. Andronescu, M., Zhang, Z. C. & Condon, A. Secondary structure prediction of interacting RNA molecules. *J Mol Biol* **345**, 987–1001 (2005).

25. Bernhart, S. H. *et al.* Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol Biol* **1**, 3 (2006).

26. Kato, Y. *et al.* RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming. *Bioinformatics* **26**, i460-466 (2010).

27. Kery, M. B., Feldman, M., Livny, J. & Tjaden, B. TargetRNA2: identifying targets of small regulatory RNAs in bacteria. *Nucleic Acids Res* **42**, W124-129 (2014).

28. Wright, P. R. *et al.* Comparative genomics boosts target prediction for bacterial small RNAs. *Proc Natl Acad Sci U S A* **110**, E3487-3496 (2013).

29. Seemann, S. E., Richter, A. S., Gesell, T., Backofen, R. & Gorodkin, J. PETcofold: predicting conserved interactions and structures of two multiple alignments of RNA sequences. *Bioinformatics* **27**, 211–219 (2011).

30. Zhao, Y., Wang, J., Zeng, C. & Xiao, Y. Evaluation of RNA secondary structure prediction for both base-pairing and topology. *Biophys Rep* **4**, 123–132 (2018).

31. 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).

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

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

446 34. Mao, K., Wang, J. & Xiao, Y. Length-Dependent Deep Learning Model for RNA Secondary
447 Structure Prediction. *Molecules* **27**, 1030 (2022).
448 35. Singh, J. *et al.* Improved RNA Secondary Structure and Tertiary Base-pairing Prediction Using
449 Evolutionary Profile, Mutational Coupling and Two-dimensional Transfer Learning. *Bioinformatics*
450 btab165 (2021) doi:10.1093/bioinformatics/btab165.
451 36. Zhang, T. *et al.* RNAcmap: A Fully Automatic Pipeline for Predicting Contact Maps of RNAs by
452 Evolutionary Coupling Analysis. *Bioinformatics* btab391 (2021)
453 doi:10.1093/bioinformatics/btab391.
454 37. Gong, S., Zhang, C. & Zhang, Y. RNA-align: quick and accurate alignment of RNA 3D
455 structures based on size-independent TM-scoreRNA. *Bioinformatics* **35**, 4459–4461 (2019).
456 38. Antonov, I. V., Mazurov, E., Borodovsky, M. & Medvedeva, Y. A. Prediction of lncRNAs and
457 their interactions with nucleic acids: benchmarking bioinformatics tools. *Briefings in*
458 *Bioinformatics* **20**, 551–564 (2019).
459 39. Szikszai, M., Wise, M., Datta, A., Ward, M. & Mathews, D. H. Deep learning models for RNA
460 secondary structure prediction (probably) do not generalize across families. *Bioinformatics* **38**,
461 3892–3899 (2022).
462 40. Qiu, X. Sequence similarity governs generalizability of de novo deep learning models for RNA
463 secondary structure prediction. *PLOS Computational Biology* **19**, e1011047 (2023).
464 41. Improved prediction of protein-protein interactions using AlphaFold2 | Nature
465 Communications. https://www.nature.com/articles/s41467-022-28865-w.
466 42. Lu, X.-J., Bussemaker, H. J. & Olson, W. K. DSSR: an integrated software tool for dissecting the
467 spatial structure of RNA. *Nucleic Acids Res* **43**, e142 (2015).
468 43. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein
469 or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
470 44. Zheng, J., Xie, J., Hong, X. & Liu, S. RMalign: an RNA structural alignment tool based on a
471 novel scoring function RMscore. *BMC Genomics* **20**, 276 (2019).
472 45. Mathews, D. H. *et al.* Incorporating chemical modification constraints into a dynamic
473 programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci U S A* **101**,
474 7287–7292 (2004).
475 46. Dirks, R. M., Bois, J. S., Schaeffer, J. M., Winfree, E. & Pierce, N. A. Thermodynamic Analysis
476 of Interacting Nucleic Acid Strands. *SIAM Rev.* **49**, 65–88 (2007).
477 47. Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure
478 prediction and analysis. *BMC Bioinformatics* **11**, 129 (2010).
479 48. Mann, M., Wright, P. R. & Backofen, R. IntaRNA 2.0: enhanced and customizable prediction
480 of RNA–RNA interactions. *Nucleic Acids Research* **45**, W435–W439 (2017).
481