Supplementary Materials for

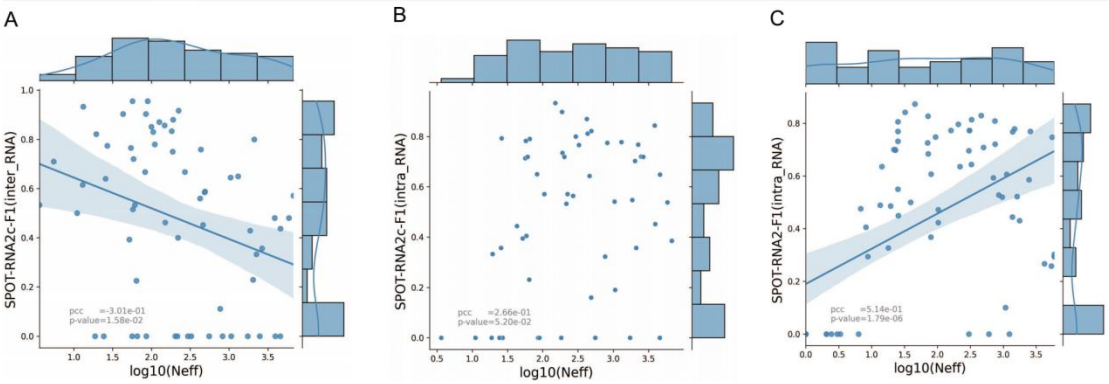
Deep learning models of RNA base-pairing structures generalize to unseen folds and make accurate zero-shot predictions of base-base interactions of RNA complexes

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)

Supplementary Figures and Tables



Supplementary Figure 1. (A) Inter-RNA F1-scores as function of the number of effective homologous sequences (Neff) given by SPOT-RNA2c. (B) Intra-RNA F1-scores (average F1-scores of two RNA complex) as a function of Neff given by SPOT-RNA2c. (C) intra-RNA F1-scores as a function of Neff given by SPOT-RNA2 with a positive correlation coefficient of 0.514.

Supplementary Table 1 The performance SPOT-RNAc concatenation with different linkers on the base pair level.

Methods	Precision	Recall	Over All F1-Score	MCC	P-value
SPOT-RNAc-AAA	**0.626**	**0.544**	0.582	0.581	0.086
SPOT-RNAc-UUU	0.608	0.543	0.574	0.572	0.038
SPOT-RNAc-CCC	0.602	0.544	0.572	0.570	0.079
SPOT-RNAc-GGG	0.604	0.542	0.571	0.570	0.031
SPOT-RNAc-NoLink	0.620	0.551	**0.583**	**0.583**	

Note:The P-value of a given method was computed by against the result from SPOT-RNAc-NoLink.

Supplementary Table 2: List of PDB IDs of 64 RNA-RNA complexes in the test set.

#PDB_ID chain1 chain2
7D4I_3A5A 7D5S_3A5A 7D5T_3A5A 7D63_3A5A 7MQ8_L0L2 7PKT_23 7PKT_36 7PKT_34 7PKT_45 7SBB_XZ 7SUK_L0L2 7V01_GU 7W59_GF 7W59_GH 7W59_FH 7W5A_FG 7WAH_RD 7XSS_AD 7XT4_ar 7Y83_BC 7YNA_BC 8A22_A2A3 8A22_A4A5 8A22_B1B2 1DZ5_CD 5GIO_GH 5GMK_LM 5LJ3_IZ 5LQW_29 5MQF_26 5NRL_26 5NRL_2I 5NRL_46 5WSG_EL 5WSG_LM 5Y88_DE 5Y88_EF 5ZWM_FH 5ZWM_FI 5ZWM_GH 6AH0_FH 6AH0_FI 6AHD_GH 6EXN_26 6EXN_2I 6FF4_2Z 6IFL_IJ 6J6G_BL 6J6N_BL 6NUD_HU 6O1O_MN 6O7H_GH 6QDV_26 6QDV_2I 6QX9_62 6QX9_64 6S91_UV 6ZYM_2Y 7ABI_62 7B9V_26 7DCO_FH 7DVQ_GH 7OQB_I2 3SIV_CF

Supplementary Method Description

1. Tools Settings

For some tools setting, we followed the excellent review by Lai and Meyer[1]. But due to the different sequence length of our benchmark data set, we changed some parameters.

1.1 Interaction only prediction tools

Inputs (target.fa and query.fa) are two standard FASTA files of two RNA sequences. Each file contains one header line followed by an RNA sequence.

1.1.1 RNAduplex

RNAduplex[2] designed for computing the structure upon hybridization of two RNA strands. Algorithmically, it is a modification of Zuker's classical RNA secondary structure algorithm, simplified to just consider intermolecular base pairs.

RNAduplex is a part of the ViennaRNA Package 2.5[2], we downloaded the package from https://www.tbi.univie.ac.at/RNA/#download.

cat query.fa target.fa | RNAduplex -s > output.txt

-s sort the printed output by free energy

1.1.2 RNAplex-c

RNAplex-c[3] was especially designed to quickly find possible hybridization sites for a query RNA in large RNA databases. It utilizes a slightly different energy model which reduces the computational time significantly, compared to RNAhybrid.

RNAplex-c is also a part of the ViennaRNA Package 2.5[2] and the package was obtained from https://www.tbi.univie.ac.at/RNA/#download.

RNAplex -q query.fa -t target.fa > output.txt

1.1.3 DuplexFold

DuplexFold[4] predicts the lowest free energy structure for two interacting sequences, not allowing intramolecular base pairs.

DuplexFold is a part of the RNAstructure Package 6.4[4] and we obtained the package from https://rna.urmc.rochester.edu/RNAstructure.html.

DuplexFold query.fa target.fa > output.txt

1.1.4 RIsearch

RIsearch[5] enables quick localization of potential near-complementary interactions between given query and target sequences. RIsearch uses a modified Smith-Waterman-Gotoh algorithm based on di-nucleotides to approximate nearest-neighbor energy parameters.

We obtained the tool from https://rth.dk/resources/risearch/.

RIsearch -q query.fa -t target.fa -p > output.txt

-p report predictions in detailed format

1.1.5 RNAplex-cA

An updated version of RNAplex[3] can additionally take alignments as input, and like RNAaliduplex, also incorporates a covariation score into the stability of the predicted duplex in the style of RNAalifold. We also refer to this as RNAplex-cA as the review paper did[1].

RNAplex-cA is also part of the ViennaRNA Package 2.5[2] and the package was obtained from https://www.tbi.univie.ac.at/RNA/#download.

RNAplex -q query_alignment.fa -t target_alignment.fa -A > output.txt

-A: Instructs the tool to compute interactions based on alignments

1.1.6 RNAaliduplex

RNAaliduplex predicts conserved RNA-RNA interactions between two alignments, also part of the same ViennaRNA Package 2.5[2]. The calculation takes only inter−molecular base pairs into account. The use of alignments allows one to focus on binding sites that are evolutionary conserved. The program must read two alignments of RNA sequences in CLUSTAL format and predicts optimal and suboptimal binding sites, hybridization energies, and the corresponding structures.

RNAaliduplex also is a part of ViennaRNA Package 2.5[2]. We downloaded the package from https://www.tbi.univie.ac.at/RNA/#download.

RNAaliduplex query_alignment.aln target_alignment.aln > output.txt

1.1.7 GUUGle

GUUGle[6]Version 1.2 obtained from https://bibiserv2.cebitec.uni-bielefeld.de/guugle?id=guugle_view_ download

guugle -d 5 target.fa query.fa > output.txt

-d Obligatory option, minimum interaction length to be output

1.2 Accessibility prediction tools

1.2.1 IntaRNA

IntaRNA2.0[7] is a program for fast and accurate prediction of interactions between two RNA molecules. It was designed to predict mRNA target sites for given non-coding RNAs (ncRNAs) like eukaryotic microRNAs (miRNAs) or bacterial small RNAs (sRNAs), but it can also be used for predicting other types of RNA-RNA interactions.

Version 2.3.1 download from http://www.bioinf.uni-freiburg.de/Software/#IntaRNA

IntaRNA -q query.fa -t target.fa --outMode=C --outOverlap=N -n 1 > output.txt

--outMode=C 'C' CSV output (see --outCsvCols)

--outOverlap=N suboptimal output : interactions can overlap, 'N' in none of the sequences

-n number of (sub)optimal interactions to report

1.2.2 RNAup

RNAup[8] calculates the thermodynamics of RNA−RNA interactions, by decomposing the binding into two stages: (1) the probability that a potential binding sites remains unpaired (equivalent to the free energy needed to open the site) is computed; (2) this accessibility is combined with the interaction energy to obtain the total binding energy. All calculations are done by computing partition functions over all possible conformations.

RNAup is also a part of the ViennaRNA Package 2.5[2] . The package was obtained from https://www.tbi.univie.ac.at/RNA/#download.

cat query.fa target.fa | RNAup -b --interaction_pairwise > output.txt

--interaction_pairwise: Activate the pairwise interaction mode.

1.2.3 RNAplex-a

The new version of RNAplex-c [3], supplies a new option -a, which can take in RNAplfold. And we also named it as RNAplex-a.

RNAplex-a and RNAplfold are a part of ViennaRNA Package 2.5[2]. We downloaded the package from https://www.tbi.univie.ac.at/RNA/#download.

cat query.fa target.fa | RNAplfold -O -u min[seq1,seq2] -W 10

RNAplex -q query.fa -t target.fa -a accessibility_dir -a -l min_seq_len > output.txt

-O Switch output from probabilities to their logarithms

-u Compute the mean probability that regions of length 1 to a given length are unpaired.
min_seq_len is minimum rounded known interaction length.

-W Average the pair probabilities over the window of a given size.

-l Maximal length of interaction. min_seq_len is minimum rounded known interaction length.

1.3 DeepLearning models based concatenation tools

Inputs target_seq and query_seq as a linked single RNA sequence.

1.3.1 SPOT-RNA

SPOT-RNA[9] is an end-to-end predictor of intra-RNA base pairs by deep learning. The method was initially trained from a large approximate database of RNA secondary structures, followed by transfer learning on more precise base pairs from RNA 3D structures.

We download SPOT-RNA from https://github.com/jaswindersingh2/SPOT-RNA/.

SPOT-RNA -i qurey_seq + target_seq

1.3.2 SPOT-RNA2

SPOT-RNA[10] was subsequently improved by SPOT-RNA2, which incorporated evolution profiles and mutational coupling automatically generated by RNAcmap.

We download SPOT-RNA2 from https://github.com/jaswindersingh2/SPOT-RNA2/.

SPOT-RNA 2 -i qurey_seq + target_seq

1.3.3 UFold

UFold[11] is a RNA secondary structure prediction tool by using deep learning.

We download UFold from https: //github.com/uci-cbcl/UFold.

UFold qurey_seq + target_seq

1.3.4 MXfold2

MXfold2[12] is a RNA secondary structure prediction tool by using deep learning.

We download it from https://github.com/keio-bioinformatics/

1.4 MFE/partition-based concatenation tools

Input target_seq and query_seq as a linked single RNA sequence.

1.4.1 RNAfold

RNAfold[13] calculates minimum free energy secondary structures and partition function of RNAs. Here, we also employed it to predict intermolecular base pairs by chain concatenation.

RNAfold is a part of ViennaRNA Package 2.5[2]. We downloaded the package from https://www.tbi.univie.ac.at/RNA/#download.

RNAfold --MAE qurey_seq + target_seq

1.4.2 RNAcoFold

RNAcofold[14] calculates secondary structures of two RNAs with dimerization. The program works much like RNAfold, but allows one to specify two RNA sequences which are then allowed to form a dimer structure.

RNAcoFold is part of ViennaRNA Package 2.5[2]. We downloaded the package from https://www.tbi.univie.ac.at/RNA/#download.

qurey_seq + "\'&\'" + target_seq | RNAcoFold > output.txt

1.4.3 RNAmultifold

RNAmultifold[2] computes secondary structures of multiple interacting RNAs. RNAmultifold is the successor of the RNA-RNA dimer interaction prediction tool RNAcofold and effectively lifts the restriction to just two interacting strands. It follows the same principle of concatenating the RNA strands that shall form a complex and then predicts MFE and partition function. Along with that, it can also compute equilibrium concentrations of the complexes formed.

RNAmultifold is part of ViennaRNA Package 2.5[2]. We downloaded the package from https://www.tbi.univie.ac.at/RNA/#download.

qurey_seq + "\'&\'" + target_seq | RNAmultifold > output.txt

1.4.4 NUPAC

NUPAC[15] algorithms operate over two fundamental ensembles:

Complex ensemble: the ensemble of all (unpseudoknotted connected) secondary structures for an arbitrary number of interacting RNA or DNA strands.

Test tube ensemble: the ensemble of a dilute solution containing an arbitrary number of RNA or DNA strand species (introduced at user-specified concentrations) interacting to form an arbitrary number of complex species.

The version 4.0 user guide was from https://docs.nupack.org/

1.4.5 PairFold

PairFold[13] predicts the minimum free energy of secondary structure formed by two input DNA or RNA molecules.

We downloaded the tools from http://www.rnasoft.ca/

PairFold qurey_seq target_seq > output.txt

1.4.6 AccessFold

AccessFold [16]predicts bimolecular base pairs while accounting for the preexisting structure. AccessFold employs Pseudo-energy minimization to minimize the sum of free energy change and a pseudo-free energy penalty for bimolecular pairing of nucleotides that are unlikely to be accessible for bimolecular structure.

AccessFold is part of the RNAstructure Package 6.4[4]. We obtained the package from https://rna.urmc.rochester.edu/RNAstructure.html.

AccessFold query.fa taget.fa > output.txt

1.5 complex joint

1.5.1 bifold

bifold[4] predicts the lowest free energy structure for two interacting sequences, allowing intramolecular base pairs.

bifold is part of the RNAstructure Package 6.4[4]. We obtained the package from https://rna.urmc.rochester.edu/RNAstructure.html.

bifold query.fa taget.fa > output.txt

1.5.2 PETcoFold

PETcoFold[17] predicts the joint secondary structure including RNA-RNA interactions from two RNA alignments.

Version 3.3 was obtained from http://rth.dk/resources/petcofold/download.php

PETcofold -f query_alignment.fa -f target_alignment.fa --intermol > output.txt

--intermol Structure output of intermolecular base pairs

Reference

1. Lai, D. & Meyer, I. M. A comprehensive comparison of general RNA–RNA interaction prediction methods. *Nucleic Acids Res* **44**, e61 (2016).
2. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**, 26 (2011).
3. Tafer, H. & Hofacker, I. L. RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics* **24**, 2657–2663 (2008).
4. Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**, 129 (2010).
5. Wenzel, A., Akbaşli, E. & Gorodkin, J. RIsearch: fast RNA–RNA interaction search using a simplified nearest-neighbor energy model. *Bioinformatics* **28**, 2738–2746 (2012).
6. 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).
7. Mann, M., Wright, P. R. & Backofen, R. IntaRNA 2.0: enhanced and customizable prediction of RNA–RNA interactions. *Nucleic Acids Research* **45**, W435–W439 (2017).
8. Mückstein, U. *et al.* Thermodynamics of RNA–RNA binding. *Bioinformatics* **22**, 1177–1182 (2006).

9. 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).
10. Singh, J. *et al.* Improved RNA secondary structure and tertiary base-pairing prediction using evolutionary profile, mutational coupling and two-dimensional transfer learning. *Bioinformatics* **37**, 2589–2600 (2021).
11. Fu, L. *et al.* UFold: fast and accurate RNA secondary structure prediction with deep learning. *Nucleic Acids Res* **50**, e14 (2022).
12. Sato, K., Akiyama, M. & Sakakibara, Y. RNA secondary structure prediction using deep learning with thermodynamic integration. *Nat Commun* **12**, 941 (2021).
13. Andronescu, M., Zhang, Z. C. & Condon, A. Secondary Structure Prediction of Interacting RNA Molecules. *Journal of Molecular Biology* **345**, 987–1001 (2005).
14. Bernhart, S. H. *et al.* Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol Biol* **1**, 3 (2006).
15. Dirks, R. M., Bois, J. S., Schaeffer, J. M., Winfree, E. & Pierce, N. A. Thermodynamic Analysis of Interacting Nucleic Acid Strands. *SIAM Rev.* **49**, 65–88 (2007).
16. DiChiacchio, L., Sloma, M. F. & Mathews, D. H. AccessFold: predicting RNA-RNA interactions with consideration for competing self-structure. *Bioinformatics* **32**, 1033–1039 (2016).
17. 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).