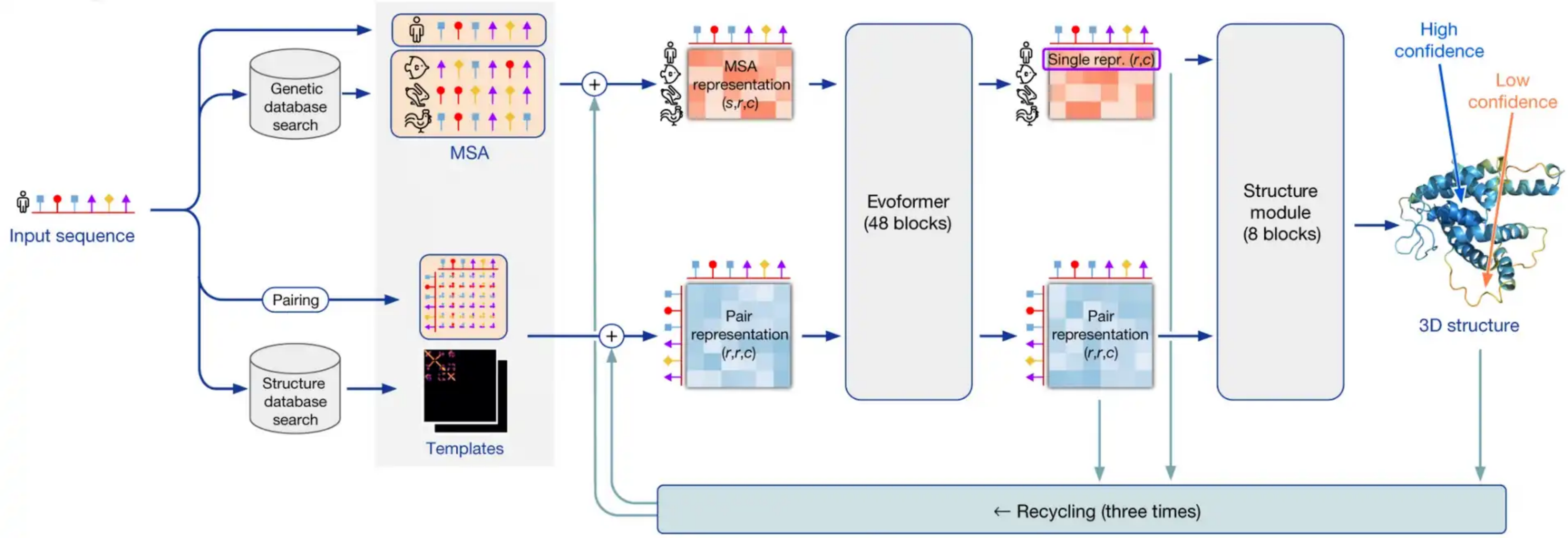
CS 6824:
RNA Structure Prediction in the Post-AlphaFold2 Era

Acknowledgement:
Many of the images in the slides are derived from images.google.com or other publicly available sources.

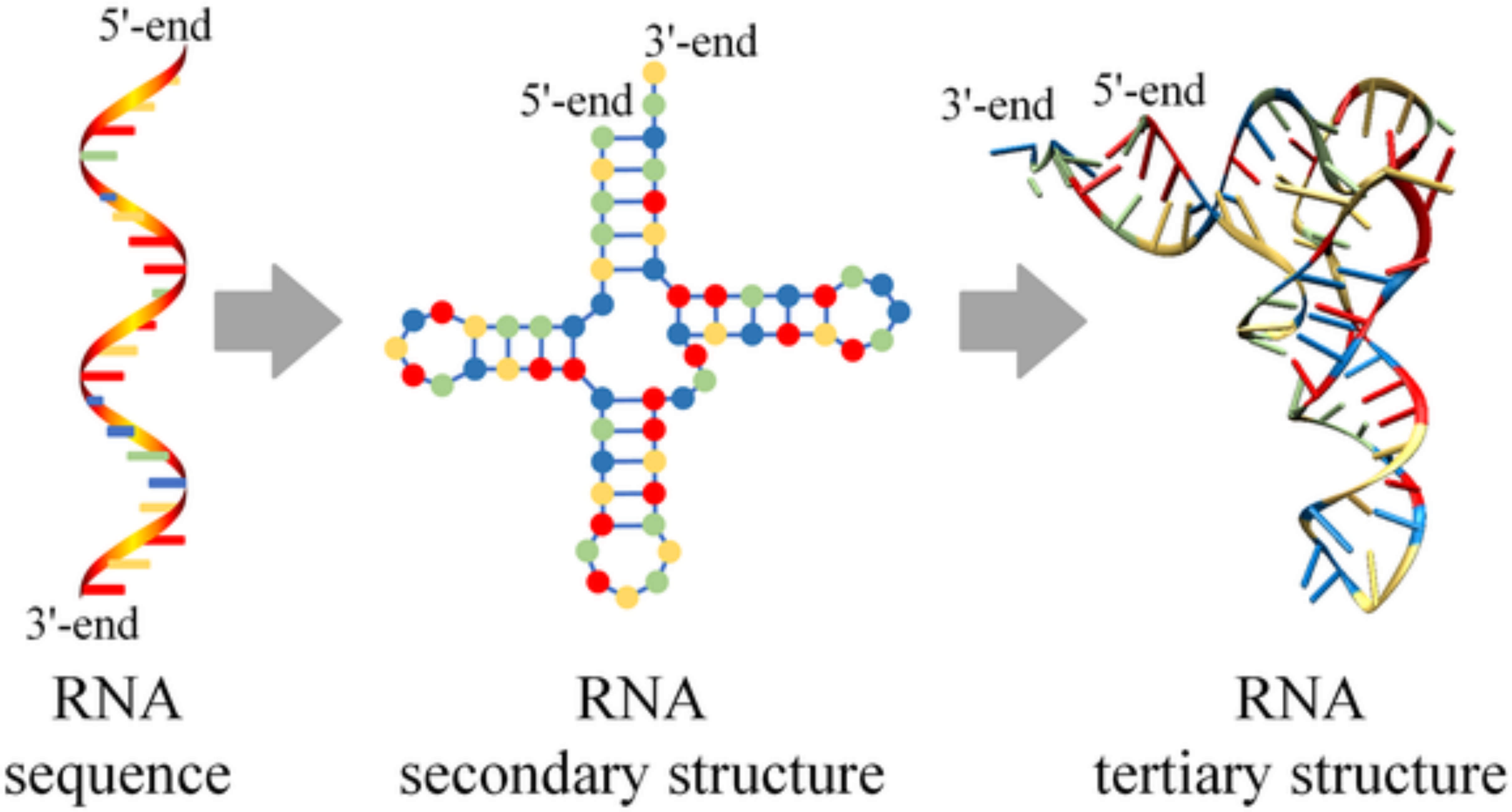
The AlphaFold 2 Era


Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

https://www.nature.com/articles/s41586-021-03819-2

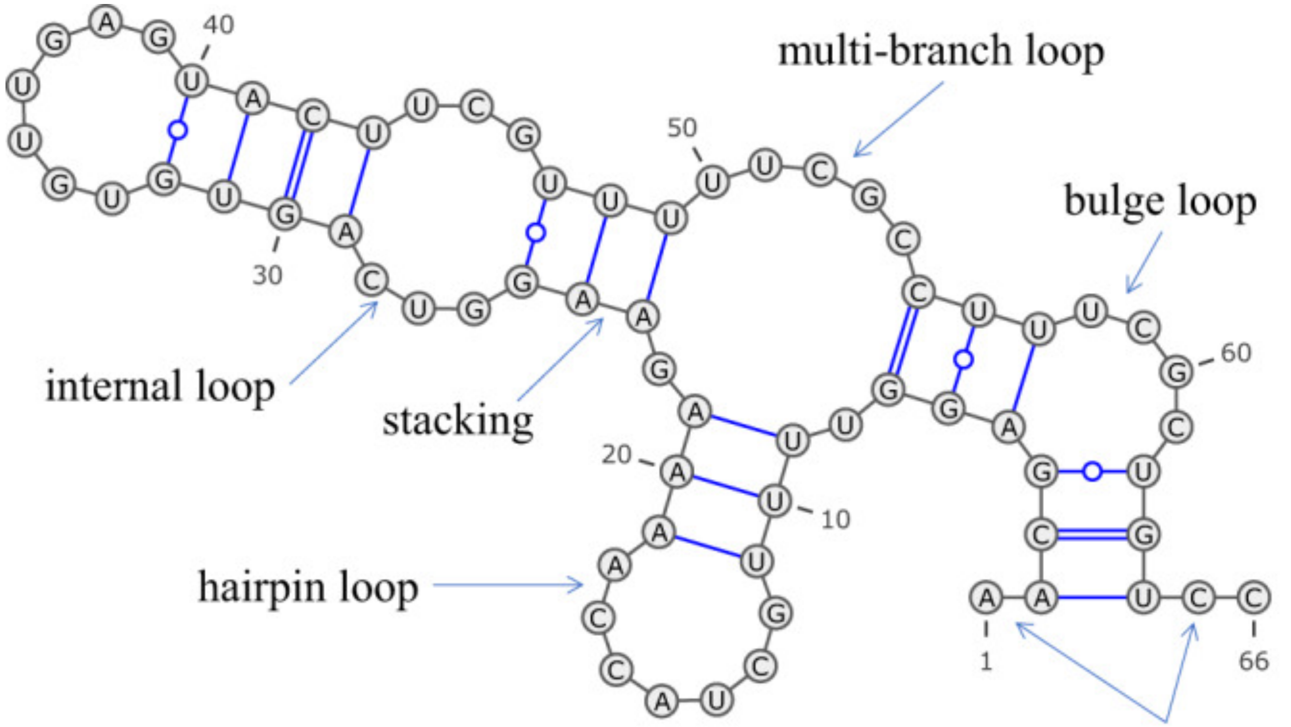
- Can AF2-like architectures generalize for other bimolecular modeling, such as RNA?

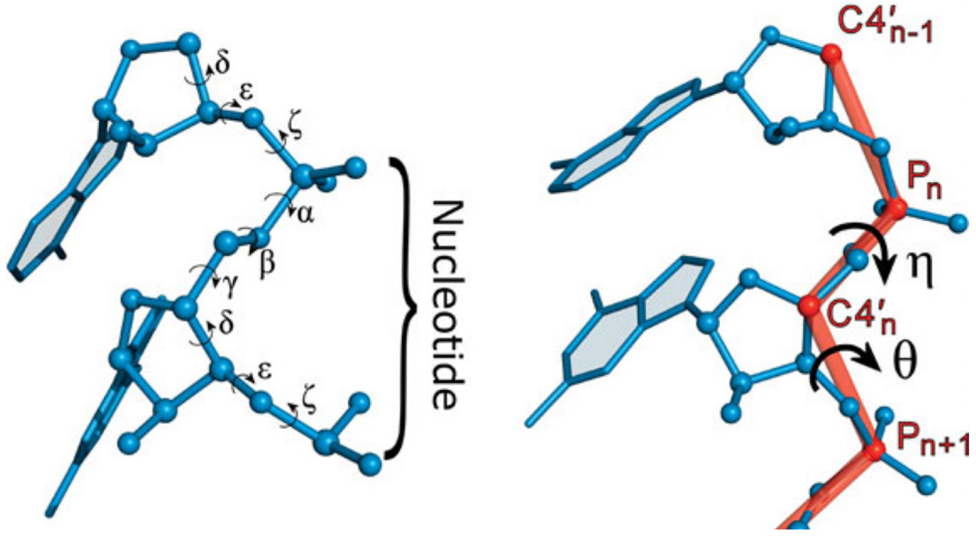
Hierarchical Conformations in RNA

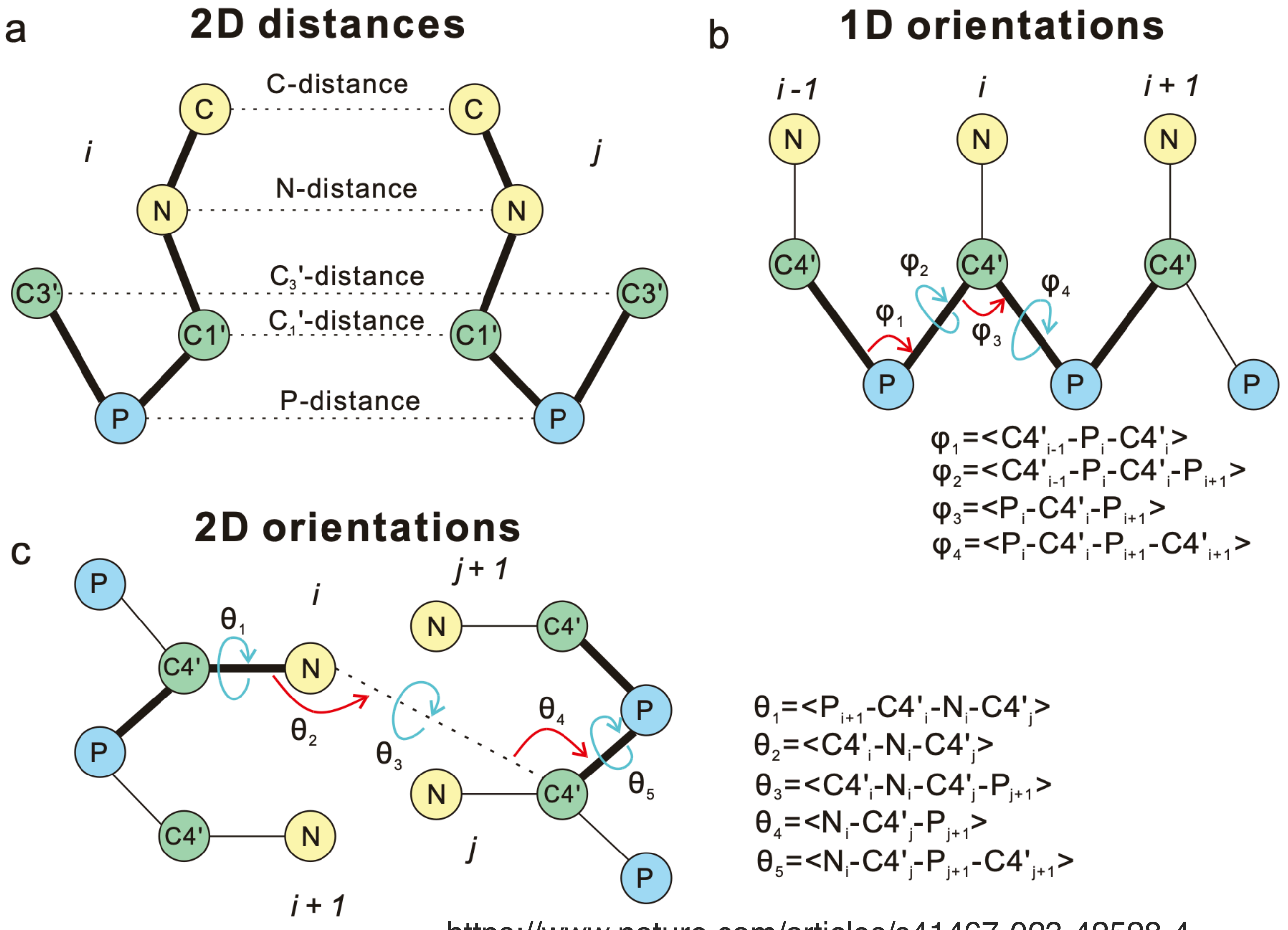


https://journals.plos.org/ploscompbiol/article/figures?id=10.1371/journal.pcbi.1009291

RNA Conformational Patterns

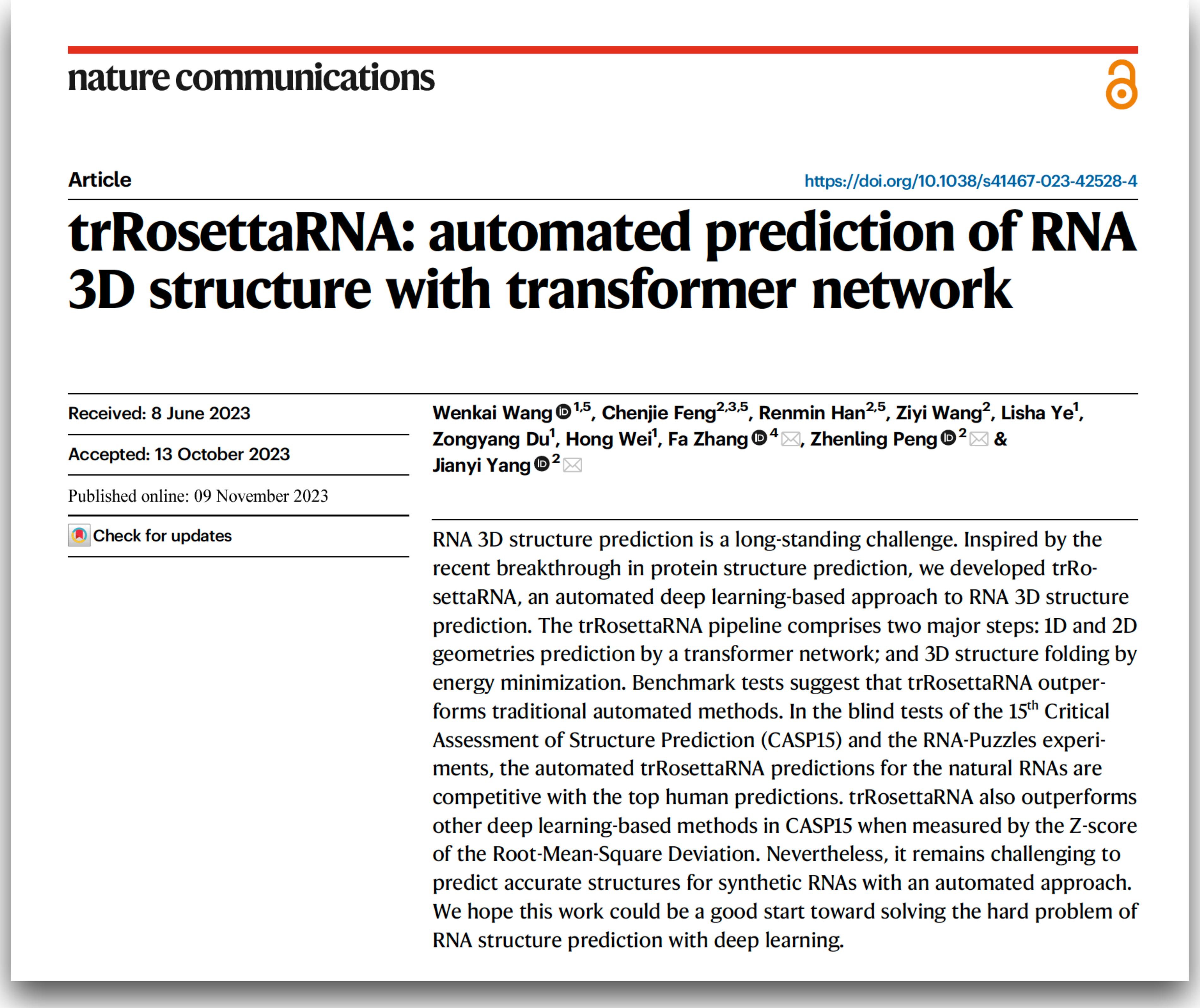





a
2D distances
C-distance
N-distance
C_3'-distance
C_1'-distance
P-distance
i
j
b
1D orientations
i - 1
i
i + 1
φ_1
φ_2
φ_3
φ_4
φ_1=<C4'$_{i-1}$-P$_i$-C4'$_i$>
φ_2=<C4'$_{i-1}$-P$_i$-C4'$_i$-P$_{i+1}$>
φ_3=<P$_i$-C4'$_i$-P$_{i+1}$>
φ_4=<P$_i$-C4'$_i$-P$_{i+1}$-C4'$_{i+1}$>
c
2D orientations
i
j + 1
j
i + 1
θ_1
θ_2
θ_3
θ_4
θ_5
θ_1=<P$_{i+1}$-C4'$_i$-N$_i$-C4'$_j$>
θ_2=<C4'$_i$-N$_i$-C4'$_j$>
θ_3=<C4'$_i$-N$_i$-C4'$_j$-P$_{j+1}$>
θ_4=<N$_i$-C4'$_j$-P$_{j+1}$>
θ_5=<N$_i$-C4'$_j$-P$_{j+1}$-C4'$_{j+1}$>

https://www.nature.com/articles/s41467-023-42528-4

trRosettaRNA



nature communications

Article

https://doi.org/10.1038/s41467-023-42528-4

trRosettaRNA: automated prediction of RNA 3D structure with transformer network

Received: 8 June 2023

Accepted: 13 October 2023

Published online: 09 November 2023

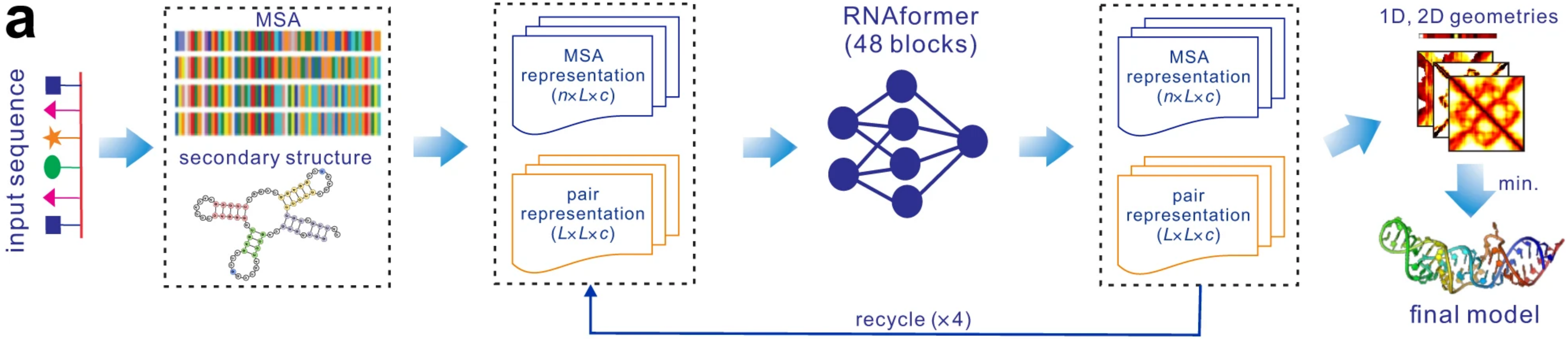
Check for updates

Wenkai Wang[1,5], Chenjie Feng[2,3,5], Renmin Han[2,5], Ziyi Wang[2], Lisha Ye[1], Zongyang Du[1], Hong Wei[1], Fa Zhang[4] ✉, Zhenling Peng[2] ✉ & Jianyi Yang[2] ✉

RNA 3D structure prediction is a long-standing challenge. Inspired by the recent breakthrough in protein structure prediction, we developed trRosettaRNA, an automated deep learning-based approach to RNA 3D structure prediction. The trRosettaRNA pipeline comprises two major steps: 1D and 2D geometries prediction by a transformer network; and 3D structure folding by energy minimization. Benchmark tests suggest that trRosettaRNA outperforms traditional automated methods. In the blind tests of the 15th Critical Assessment of Structure Prediction (CASP15) and the RNA-Puzzles experiments, the automated trRosettaRNA predictions for the natural RNAs are competitive with the top human predictions. trRosettaRNA also outperforms other deep learning-based methods in CASP15 when measured by the Z-score of the Root-Mean-Square Deviation. Nevertheless, it remains challenging to predict accurate structures for synthetic RNAs with an automated approach. We hope this work could be a good start toward solving the hard problem of RNA structure prediction with deep learning.

https://www.nature.com/articles/s41467-023-42528-4

The trRosettaRNA System


a
input sequence
MSA
secondary structure
MSA representation (n×L×c)
pair representation (L×L×c)
RNAformer (48 blocks)
MSA representation (n×L×c)
pair representation (L×L×c)
1D, 2D geometries
min.
final model
recycle (×4)

For a given query RNA, the first step of trRosettaRNA is to prepare an MSA and a secondary structure.

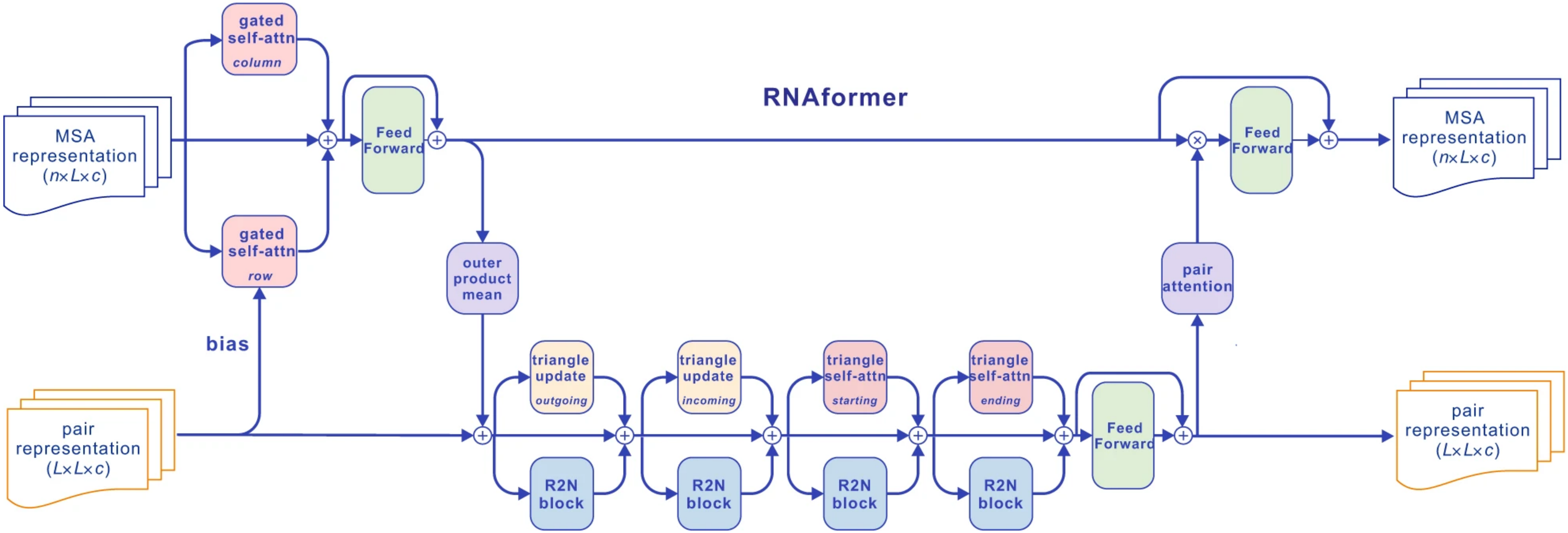
MSA is generated by using the program rMSA against multiple sequence databases (NCBI's nt, Rfam, and RNAcentral). The secondary structure is predicted by SPOT-RNA from the query sequence. Here we use the predicted probability matrix as the input, which contains more information than the dot-bracket representation.

The second step of trRosettaRNA is to predict the 1D and 2D geometries by deep learning.

Similar to trRosetta, trRosettaRNA generates full-atom structure models by energy minimization with deep learning potentials and physics-based energy terms in Rosetta.

https://www.nature.com/articles/s41467-023-42528-4

trRosettaRNA - RNAformer



1. **MSA to MSA**. To update the MSA representation by itself, we perform row- and column-wise gated self-attention operations and combine the corresponding results.

2. **MSA to pair**. We perform an outer product operation on the self-updated MSA representation to transform it into the pair format.

3. **Pair to pair**. After the above step, they perform the triangle updates, followed by a feed-forward layer. For each triangle update layer, use a Res2Net to enhance the ability to model the local details.

4. **Pair to MSA**. The updated pair representation is then linearly projected to the pair-wise attention maps, which are then multiplied on the MSA representation, followed by a feed-forward layer.

https://www.nature.com/articles/s41467-023-42528-4

trRosettaRNA - Structure Generation

$$E = w_1 E_{dist} + w_2 E_{ori} + w_3 E_{cont} + w_4 E_{ros}$$

$$E_{ori} = E_{ori,2D} + \frac{L}{2} E_{ori,1D}$$

where E_{dist}, E_{ori}, and E_{cont} represent the distance-, orientation-, contact-based restraints and Rosetta's internal energy terms, respectively; $E_{ori,2D}$ and $E_{ori,1D}$ represent the restraints from 2D and 1D orientations, respectively; L is the length of the sequence. The weights ($w_1 = 1.03$, $w_2 = 1.0$, $w_3 = 1.05$, $w_4 = 0.05$) are decided on hundreds of RNAs randomly selected from the training set to minimize the average RMSD.

The folding procedure is implemented with pyRosetta. From each RNA, 20 full-atom starting structures are first generated using the RNA_HelixAssembler protocol in pyRosetta. L-BFGS is then applied to refine these structures by minimizing the total energy, resulting in 20 refined full-atom structure models.

Finally, the model with the lowest total energy is selected as the final prediction.

https://www.nature.com/articles/s41467-023-42528-4

trRosettaRNA - Training

All the RNA chains released before 2022-01 in PDB are used as training. In total, they obtained 8849 samples. Then they tried to generate MSA for each query sequence and removed the sequences without sequence homologs. Finally, 3633 RNA chains were retained for training the network models of trRosettaRNA.

Self-distillation training set from bpRNA database with experimental secondary structures, consisting of 13,202 RNA chains.

In the first step, they trained an un-distilled model using the PDB set by 15 epochs. This model was then used to generate the labels for RNAs in the self-distillation set. In the second step, the un-distilled model was further trained on the combination of the PDB set and the self-distillation set with another 15 epochs. In the third step, they fine-tuned the models on the long sequences (>100 nucleotides) selected from the PDB set.

They used the Adam optimizer to minimize the loss function (see below) with different learning rates (0.0001 for the first two steps, 0.00005 for the third step).

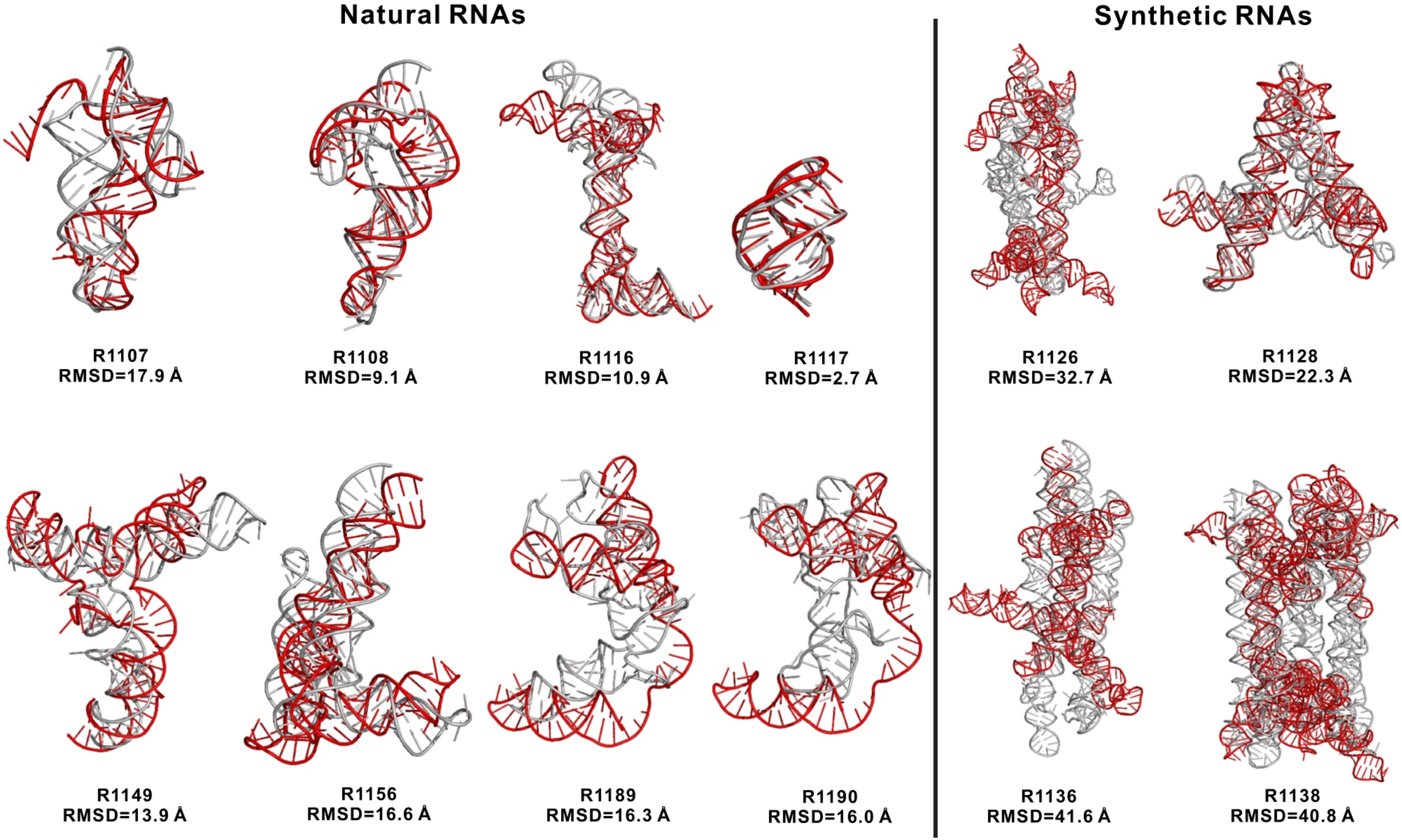
For all training steps, the loss function is defined as the cross entropy between the predicted distributions and the real or generated labels. In total, the loss function can be written as:

$$Loss = L_{2D} + L_{1D} + 5L_{cont}$$

where L_{2D}, L_{1D}, and L_{cont} are the loss for the 2D distances and orientations, 1D orientations, and 2D contacts.

https://www.nature.com/articles/s41467-023-42528-4

trRosettaRNA - CASP15 Results



https://www.nature.com/articles/s41467-023-42528-4

trRosettaRNA - CASP15 Results

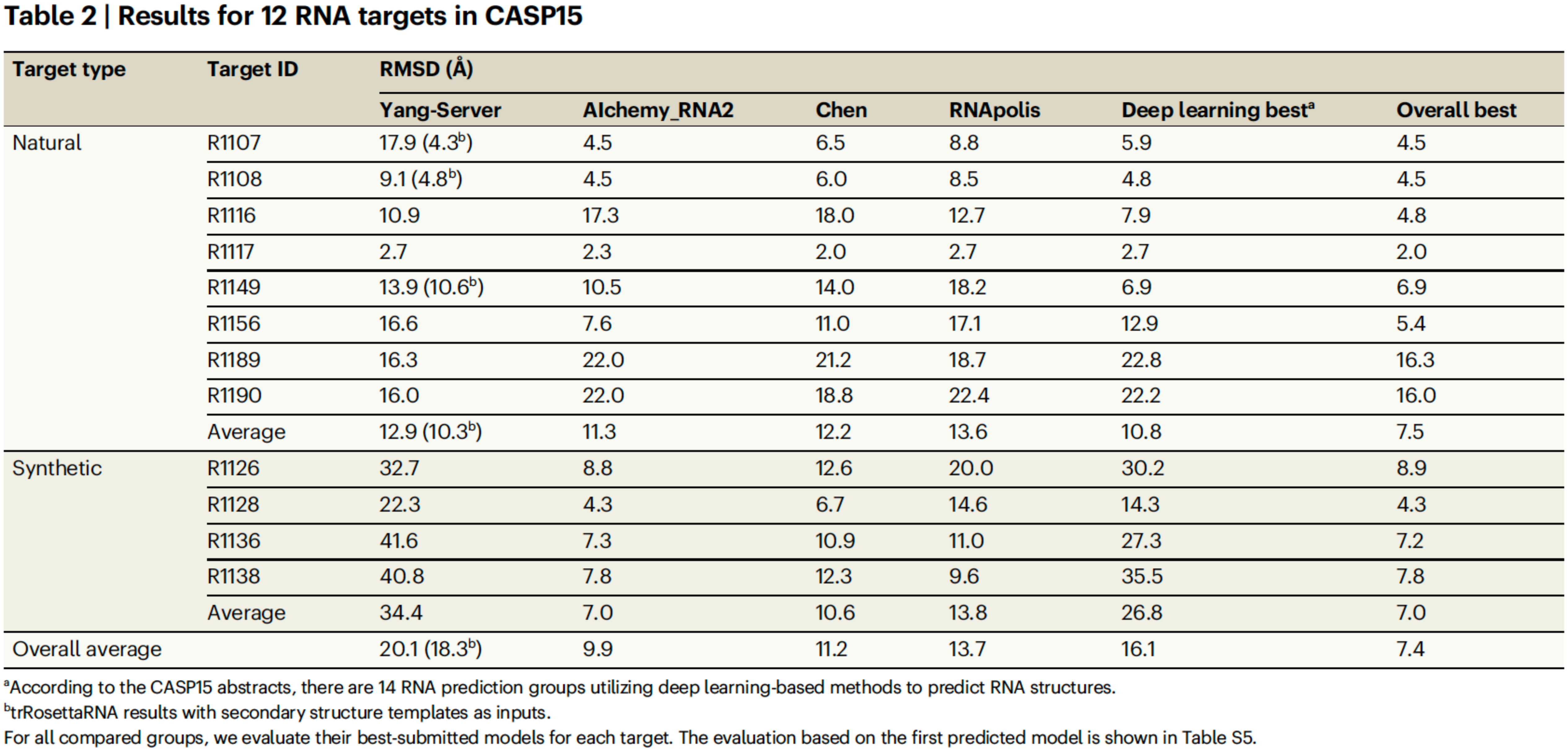


Table 2 | Results for 12 RNA targets in CASP15

Target type	Target ID	RMSD (Å)					
		Yang-Server	Alchemy_RNA2	Chen	RNApolis	Deep learning best[a]	Overall best
Natural	R1107	17.9 (4.3[b])	4.5	6.5	8.8	5.9	4.5
	R1108	9.1 (4.8[b])	4.5	6.0	8.5	4.8	4.5
	R1116	10.9	17.3	18.0	12.7	7.9	4.8
	R1117	2.7	2.3	2.0	2.7	2.7	2.0
	R1149	13.9 (10.6[b])	10.5	14.0	18.2	6.9	6.9
	R1156	16.6	7.6	11.0	17.1	12.9	5.4
	R1189	16.3	22.0	21.2	18.7	22.8	16.3
	R1190	16.0	22.0	18.8	22.4	22.2	16.0
	Average	12.9 (10.3[b])	11.3	12.2	13.6	10.8	7.5
Synthetic	R1126	32.7	8.8	12.6	20.0	30.2	8.9
	R1128	22.3	4.3	6.7	14.6	14.3	4.3
	R1136	41.6	7.3	10.9	11.0	27.3	7.2
	R1138	40.8	7.8	12.3	9.6	35.5	7.8
	Average	34.4	7.0	10.6	13.8	26.8	7.0
Overall average		20.1 (18.3[b])	9.9	11.2	13.7	16.1	7.4

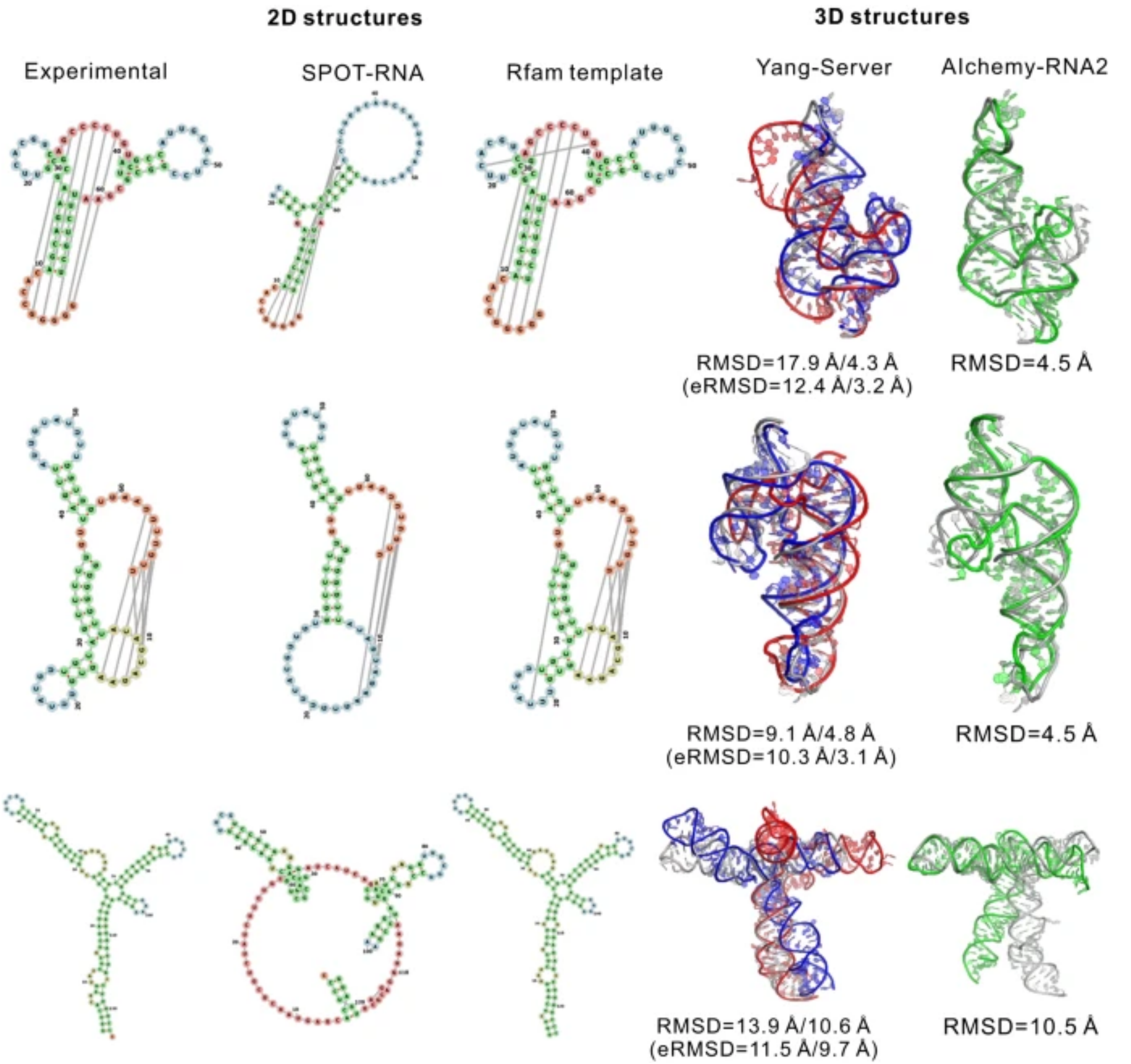
[a] According to the CASP15 abstracts, there are 14 RNA prediction groups utilizing deep learning-based methods to predict RNA structures.

[b] trRosettaRNA results with secondary structure templates as inputs.

For all compared groups, we evaluate their best-submitted models for each target. The evaluation based on the first predicted model is shown in Table S5.

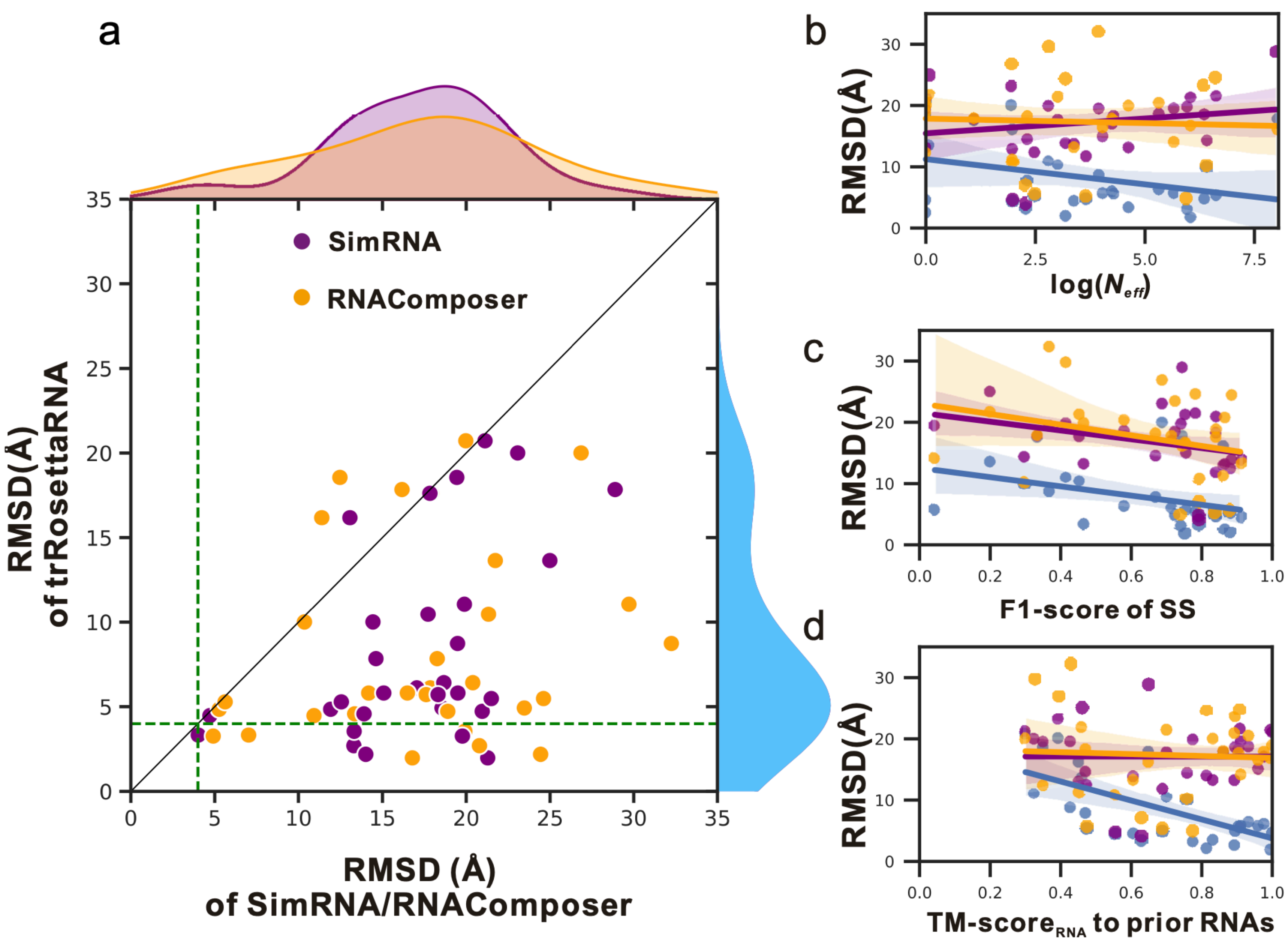
https://www.nature.com/articles/s41467-023-42528-4

trRosettaRNA - CASP15 Results



https://www.nature.com/articles/s41467-023-42528-4

trRosettaRNA vs. Automated Non-DL methods



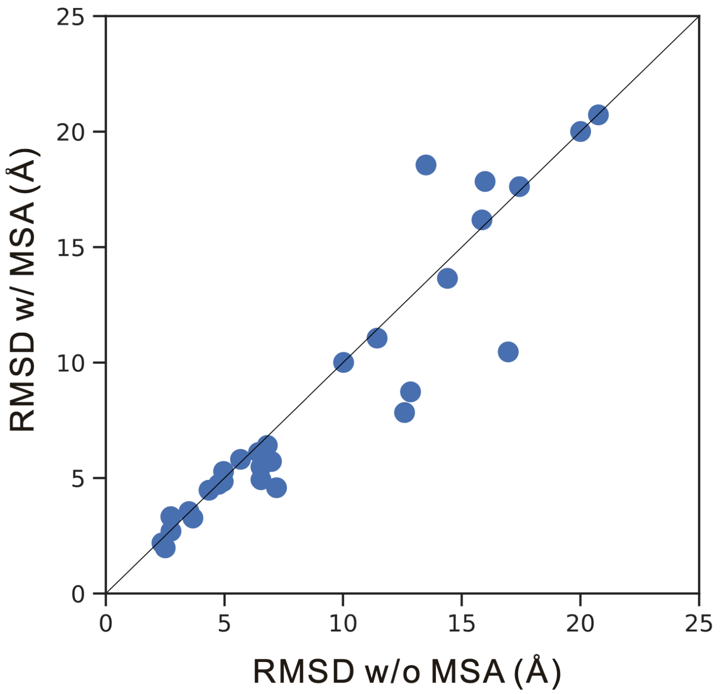


Figure S2. Performance on 30 independent RNAs. (a) head-to-head comparison between trRosettaRNA and two representative methods, SimRNA and RNAComposer (n=30 RNAs). The dashed horizontal and vertical lines correspond to an RMSD of 4 Å. The bar plots show the RMSD distributions. (b) the RMSD as a function of the logarithm of the MSA depth (N_{eff}). (c) RMSD as a function of the F1-score of the predicted secondary structure (denoted by SS). (d) RMSD as a function of the maximum TM-score$_{RNA}$ to prior RNAs. The gray and black dash lines in (d) refer to the TM-score$_{RNA}$ thresholds of 0.45 and 0.6 (homology match and very good homology match) respectively. The blue, purple, and orange dots in B-D refer to trRosettaRNA, SimRNA, and RNAComposer, respectively. Source data are provided as a Source Data file.

https://www.nature.com/articles/s41467-023-42528-4

trRosettaRNA vs. Automated Non-DL methods

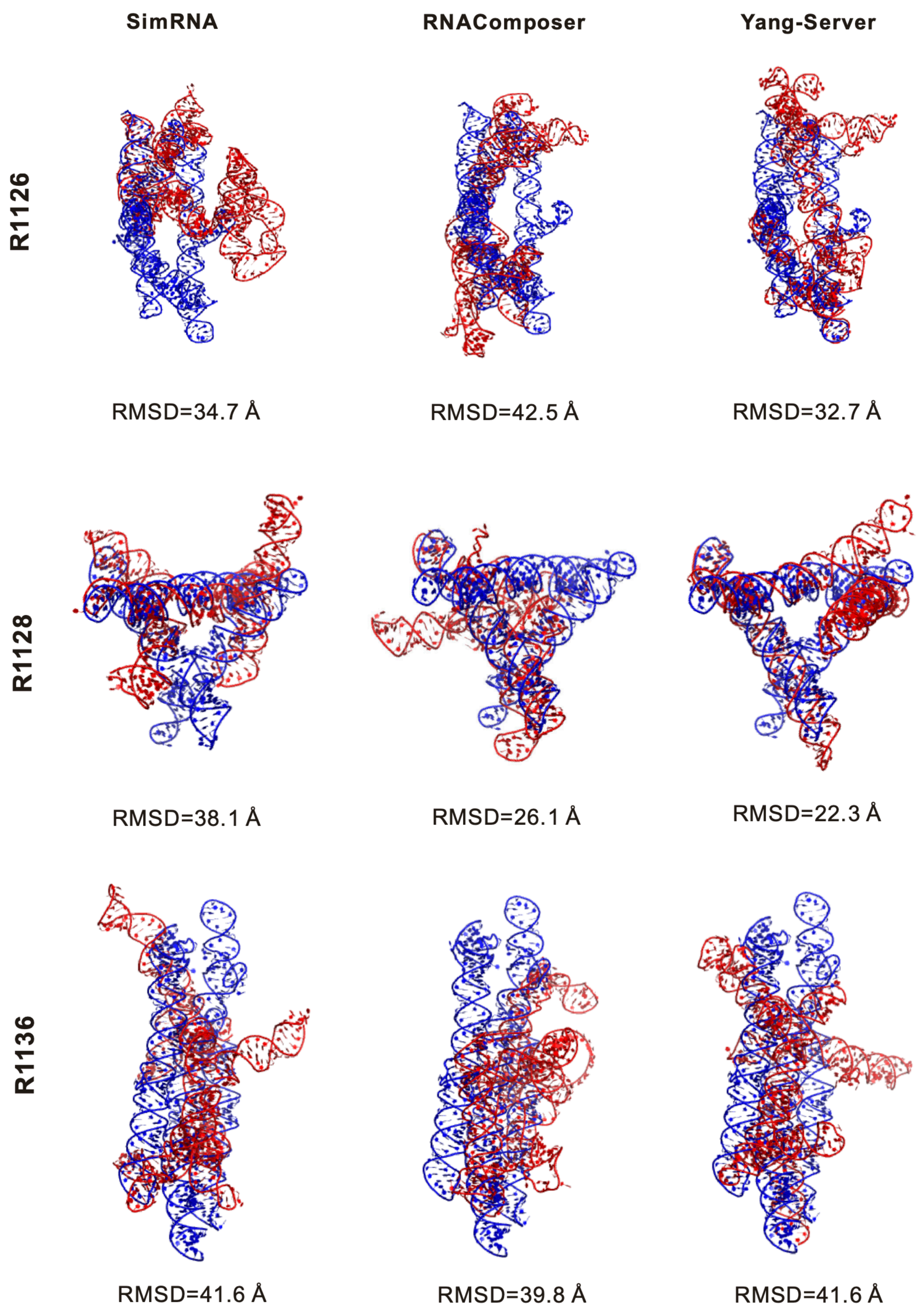
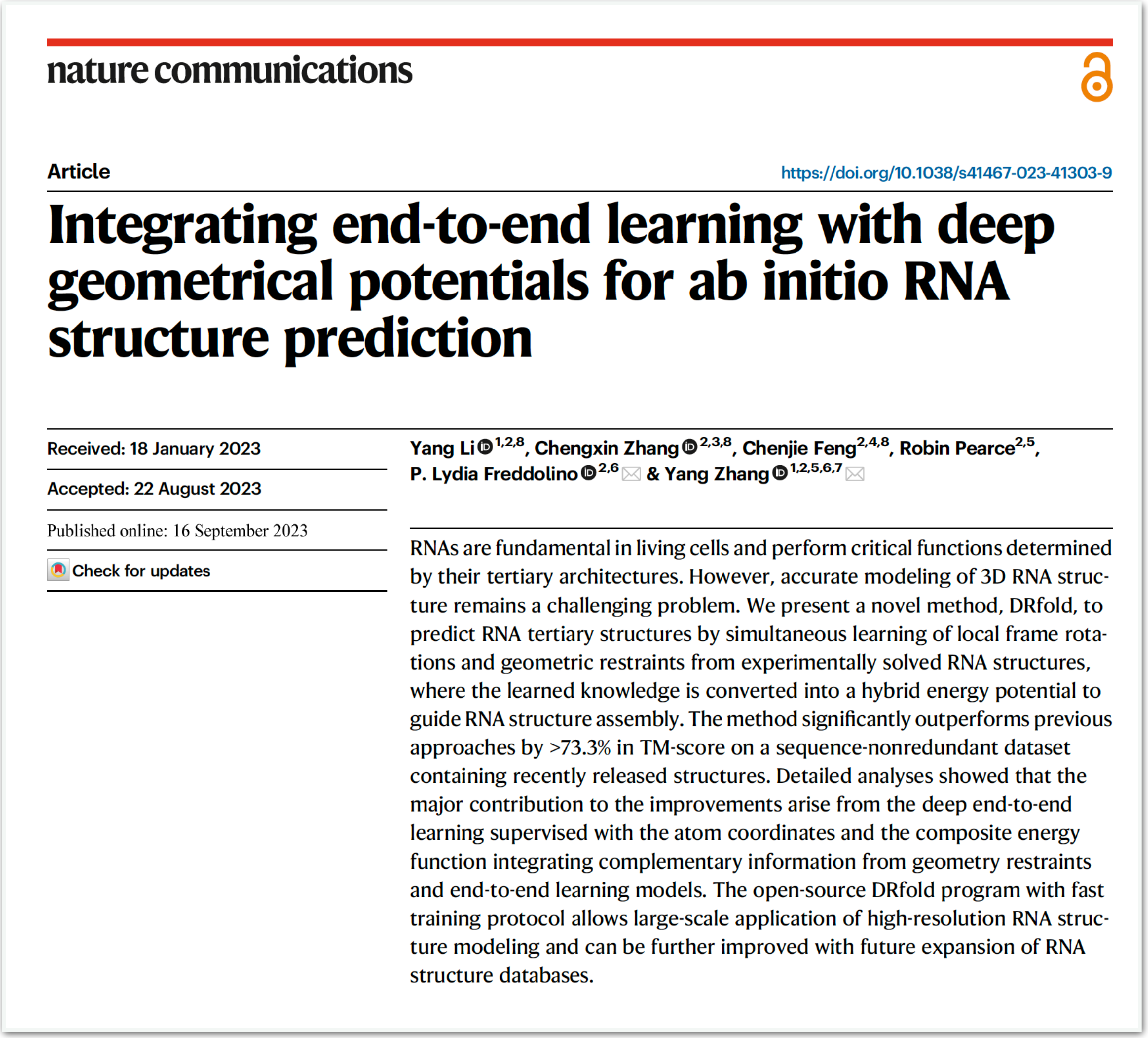


Figure S7. Comparison of 3D modelling results for synthetic RNAs in CASP15 between Yang-Server and representative automated methods. Both predicted 3D structures (in the red cartoon) are superimposed onto the experimental structures (in the blue cartoon).

https://www.nature.com/articles/s41467-023-42528-4

DRfold



nature communications

Article

https://doi.org/10.1038/s41467-023-41303-9

Integrating end-to-end learning with deep geometrical potentials for ab initio RNA structure prediction

Received: 18 January 2023

Accepted: 22 August 2023

Published online: 16 September 2023

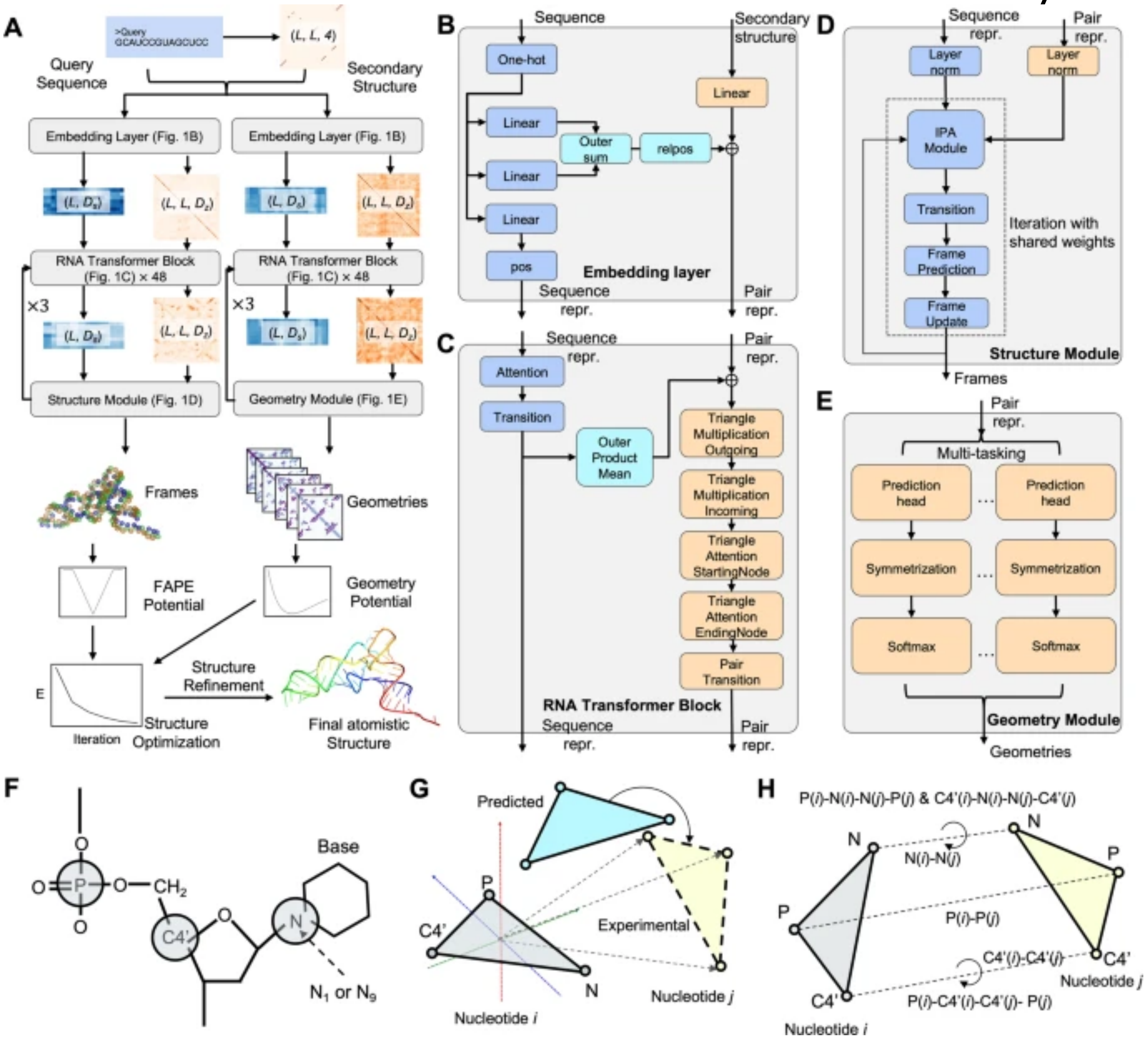
Check for updates

Yang Li[1,2,8], Chengxin Zhang[2,3,8], Chenjie Feng[2,4,8], Robin Pearce[2,5], P. Lydia Freddolino[2,6] & Yang Zhang[1,2,5,6,7]

RNAs are fundamental in living cells and perform critical functions determined by their tertiary architectures. However, accurate modeling of 3D RNA structure remains a challenging problem. We present a novel method, DRfold, to predict RNA tertiary structures by simultaneous learning of local frame rotations and geometric restraints from experimentally solved RNA structures, where the learned knowledge is converted into a hybrid energy potential to guide RNA structure assembly. The method significantly outperforms previous approaches by >73.3% in TM-score on a sequence-nonredundant dataset containing recently released structures. Detailed analyses showed that the major contribution to the improvements arise from the deep end-to-end learning supervised with the atom coordinates and the composite energy function integrating complementary information from geometry restraints and end-to-end learning models. The open-source DRfold program with fast training protocol allows large-scale application of high-resolution RNA structure modeling and can be further improved with future expansion of RNA structure databases.

https://www.nature.com/articles/s41467-023-41303-9

The DRfold System



A DRfold pipeline for sequence-based RNA structure prediction.

B–E Details of embedding layer, RNA transformer block, and structural and geometry modules, respectively.

F Reduced representation of nucleotide residues by a 3-bead model (C4′, P, glycosidic N) in DRfold. **G** Illustration of the frame aligned point error (FAPE).

H Prediction terms of inter-nucleotide geometry.

https://www.nature.com/articles/s41467-023-41303-9

DRfold - Training

For training the end-to-end models, two types of loss functions, including the FAPE loss and the inter-N atom distance loss, are used, i.e.,

$$L_{e2e} = 1.5L_{FAPE} + 0.6L_{dist}$$

The FAPE loss is adapted from AF2. The distance loss function takes the cross-entropy form.

For the geometry model, the Euclidean distance between the P, C4', and glycosidic N atoms are calculated, where the distance values for the inter-P atoms, inter-C4' atoms, and inter-N atoms are discretized into 56, 44, and 32 bins in the ranges of [2, 30 Å], [2, 24 Å], and [2, 18 Å]. The dihedral angle values are discretized into 36 bins. The loss function of the geometry models is the cross-entropy loss of the distance and dihedral angle terms.

Adam optimizer was used with an initial learning rate of 1e−3 for 100 epochs. The whole end-to-end model was trained on a single Nvidia A40 GPU with 32GB of memory, where 6 end-to-end models and 3 geometry models with different random parameter initializations were trained, and training each of them took 2 weeks. For the 3 geometry models, it took around 50 epochs of training for 5 days each.

https://www.nature.com/articles/s41467-023-41303-9

DRfold - Inference

Following the end-to-end and geometry modeling, a combination of two deep-learning energy terms is used to guide the next step of RNA structure optimization as follows:

$$E_{DL} = E_{e2e} + E_{geo}.$$

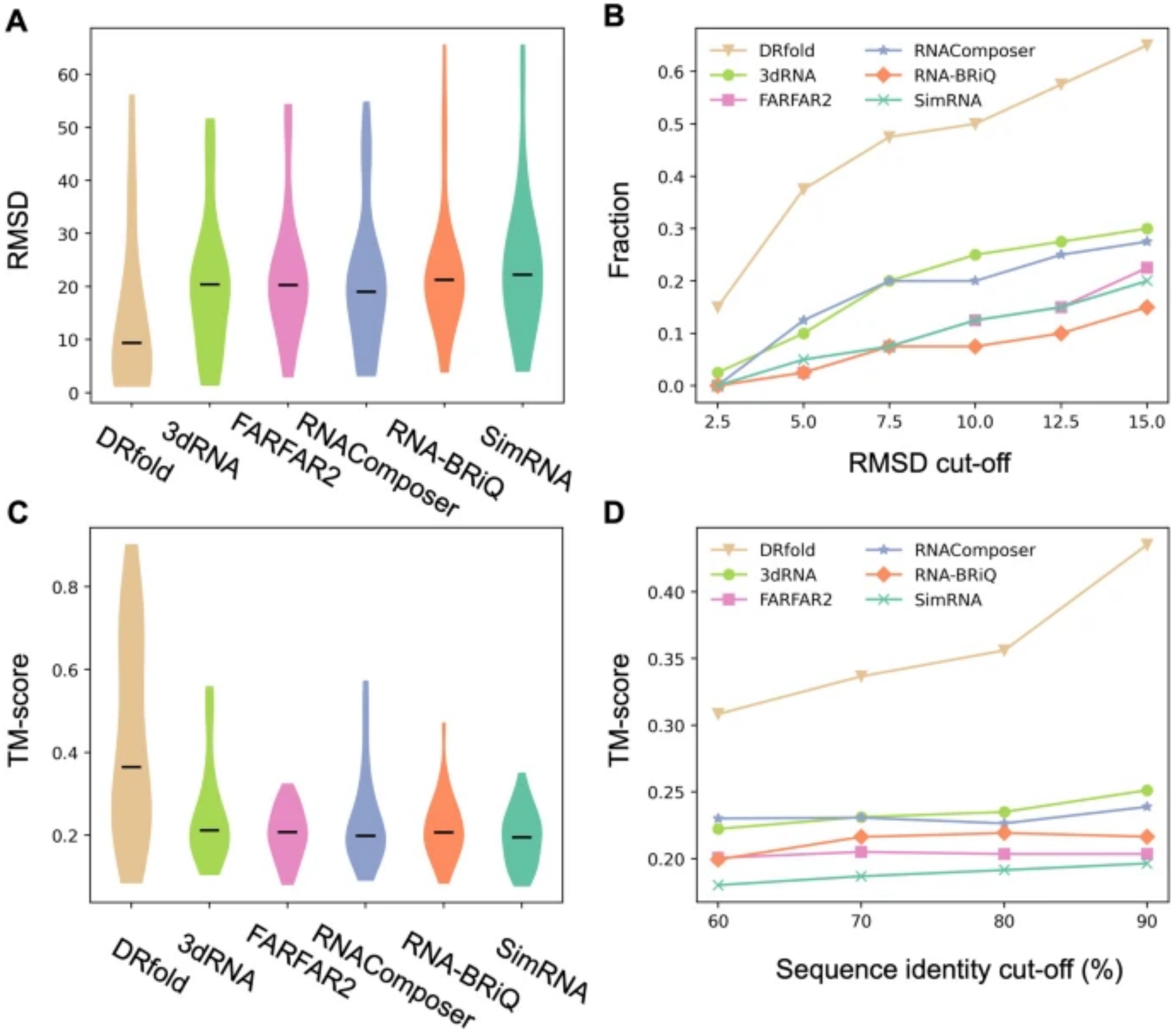
6 models predicted by 6 independent end-to-end models are predicted also used as initial structures for the optimization system to run L-BFGS algorithm to iteratively update the parameters of the system which determines the 3D conformations of the RNA models. The conformation with the lowest energy is considered as the final predicted structure among the 6 different L-BFGS trajectories.

During the first step, they use Arena to construct the standard conformations of the full-atomic structure.

Finally, a full-atom MD minimization is performed using OpenMM to further refine the local structure geometry, including steric clash and bond-length/angle violation removal.

https://www.nature.com/articles/s41467-023-41303-9

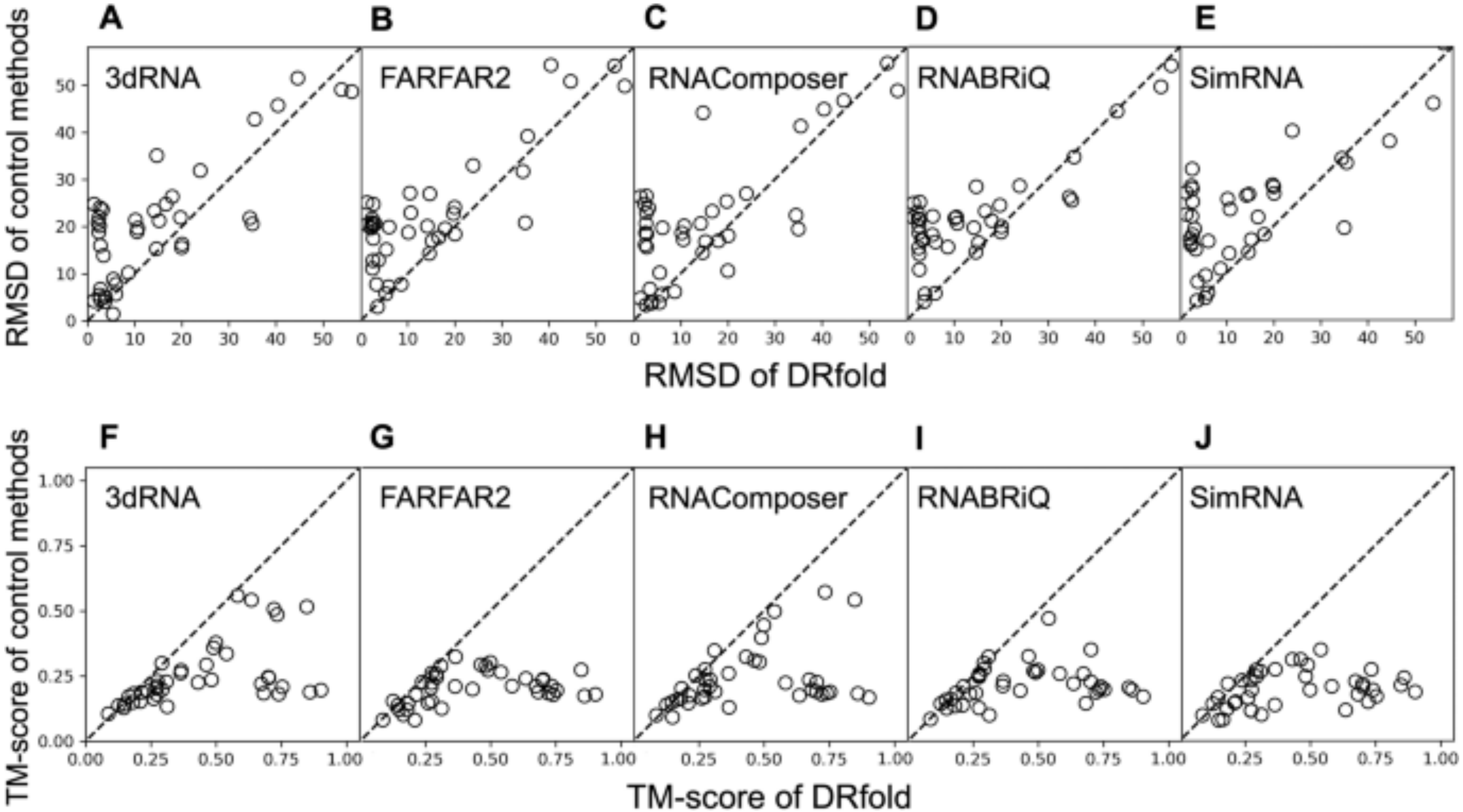
DRfold - Results



A Distribution of RMSD (Å) of the predicted models to the target structure. The central mark indicates the median. **B** Fraction of the test RNAs achieving successful structure prediction at different RMSD cut-offs. **C** TM-score distribution of different methods. The central mark indicates the median. **D** The average TM-score by different methods versus the sequence identity cut-off between the test and DRfold training datasets.

https://www.nature.com/articles/s41467-023-41303-9

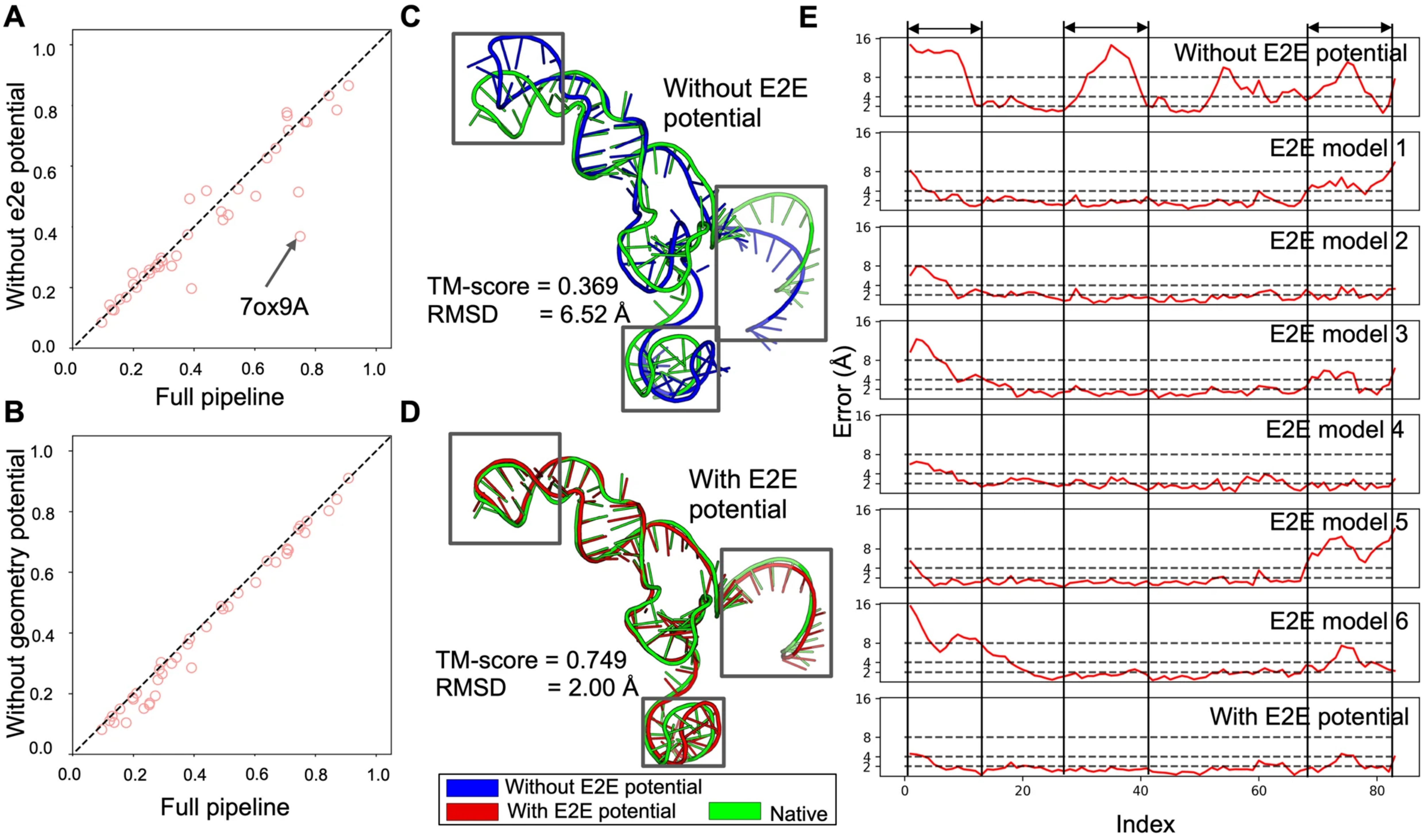
DRfold - Results



Head-to-head comparisons between DRfold and the control methods on the 40 test RNA structures.

https://www.nature.com/articles/s41467-023-41303-9

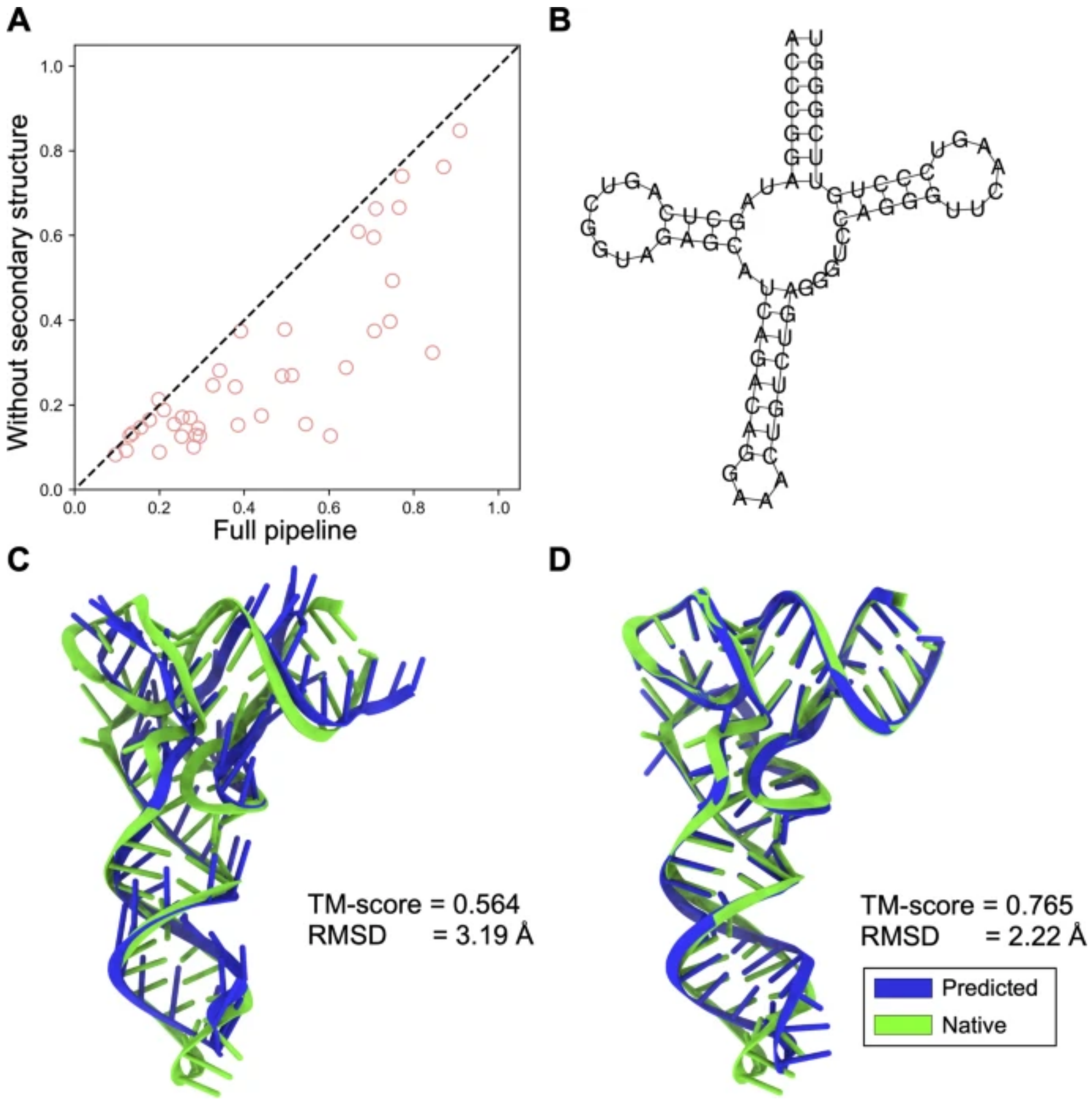
DRfold - Results



end-to-end models improve performance

https://www.nature.com/articles/s41467-023-41303-9

DRfold - Results



Secondary structure feature improves performance

https://www.nature.com/articles/s41467-023-41303-9

Discussions

Protein Structure Prediction via AF2

1.

2.

3.

4.

RNA Structure Prediction in the post-AF2 Era

1.

2.

3.

4.