Agency for Science, Technology and Research
SINGAPORE

Šikić Lab
AI IN GENOMICS

NUS National University of Singapore
School of Computing

FER
UNIVERSITY OF ZAGREB
Faculty of Electrical Engineering and Computing

Benchmarking deep learning-based methods for RNA 3D structure prediction

Ivona Martinović

PhD student

Computational Approaches to RNA Structure and Function

Benasque Science Center, Jul 21 - Aug 03, 2024

CREATING GROWTH, ENHANCING LIVES

Outline

- Benchmarking **deep learning-based tools** for RNA 3D structure prediction
  - overview of the tools
  - datasets
  - results

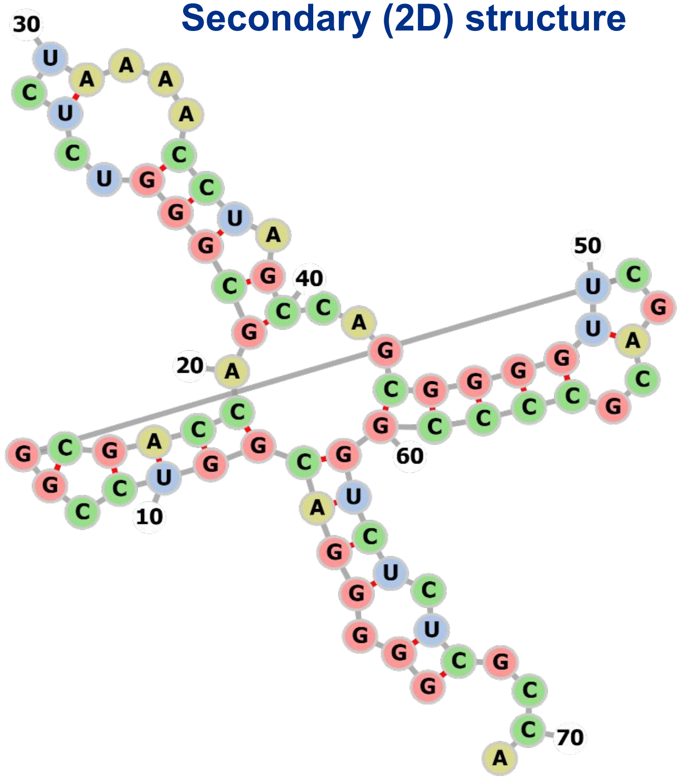
- Next steps towards our structure prediction model: our **RNA language model** - RiNALMo

RNA 3D structure prediction

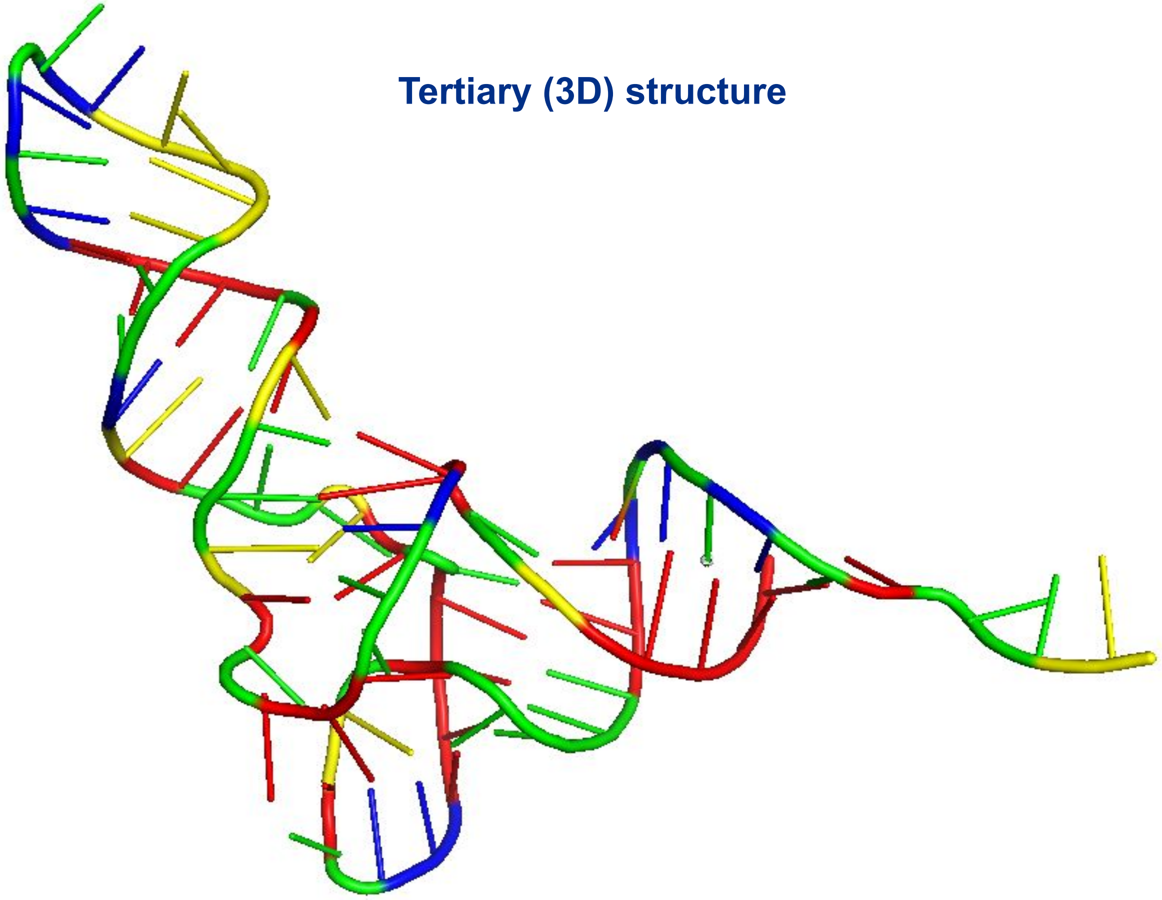
Primary structure / sequence

>8UPT_A
GGGGGACGGUCCGGCGACCAGCGGGUCUCUAAAACCUAGCCAGCGGGGUUCGACGCCCCGGUCUCUCGCCA

Secondary (2D) structure



Tertiary (3D) structure

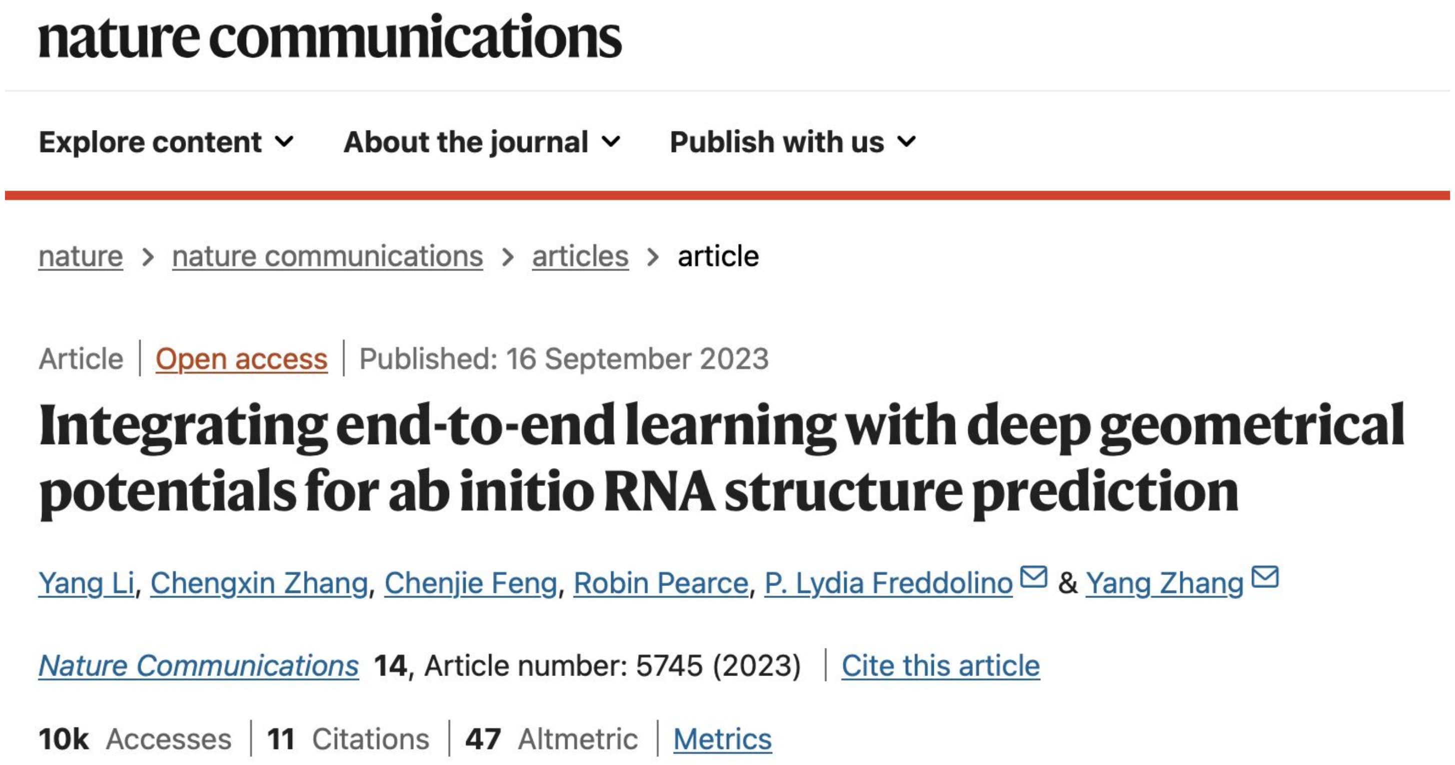


Computational Approaches to RNA Structure and Function, Benasque 2024

Deep learning based RNA 3D structure prediction tools

Deep learning based RNA 3D structure prediction tools

- DRfold



nature communications

nature > nature communications > articles > article

Article | Open access | Published: 16 September 2023

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

Yang Li, Chengxin Zhang, Chenjie Feng, Robin Pearce, P. Lydia Freddolino & Yang Zhang

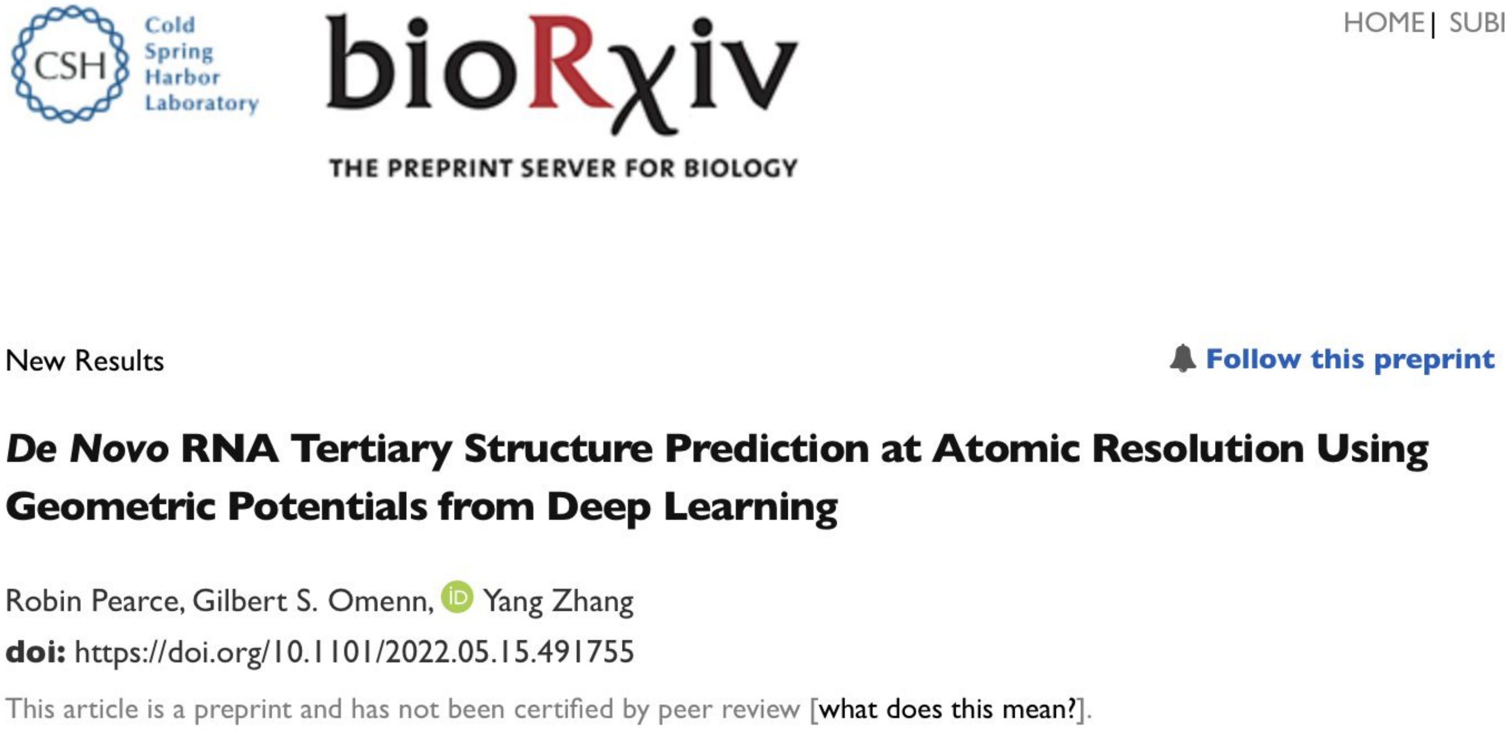
Nature Communications **14**, Article number: 5745 (2023) | Cite this article

10k Accesses | **11** Citations | **47** Altmetric | Metrics

CREATING GROWTH, ENHANCING LIVES

Deep learning based RNA 3D structure prediction tools

- DRfold
- DeepFoldRNA


CSH
Cold Spring Harbor Laboratory
bioRχiv
THE PREPRINT SERVER FOR BIOLOGY
HOME | SUBI
Follow this preprint

New Results

De Novo RNA Tertiary Structure Prediction at Atomic Resolution Using Geometric Potentials from Deep Learning

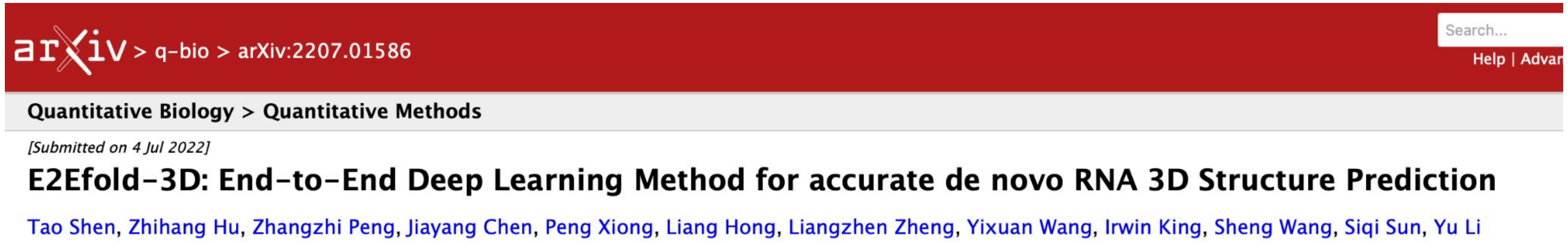
Robin Pearce, Gilbert S. Omenn, Yang Zhang

doi: https://doi.org/10.1101/2022.05.15.491755

This article is a preprint and has not been certified by peer review [what does this mean?].

Deep learning based RNA 3D structure prediction tools

- DRfold
- DeepFoldRNA
- RhoFold



Quantitative Biology > Quantitative Methods

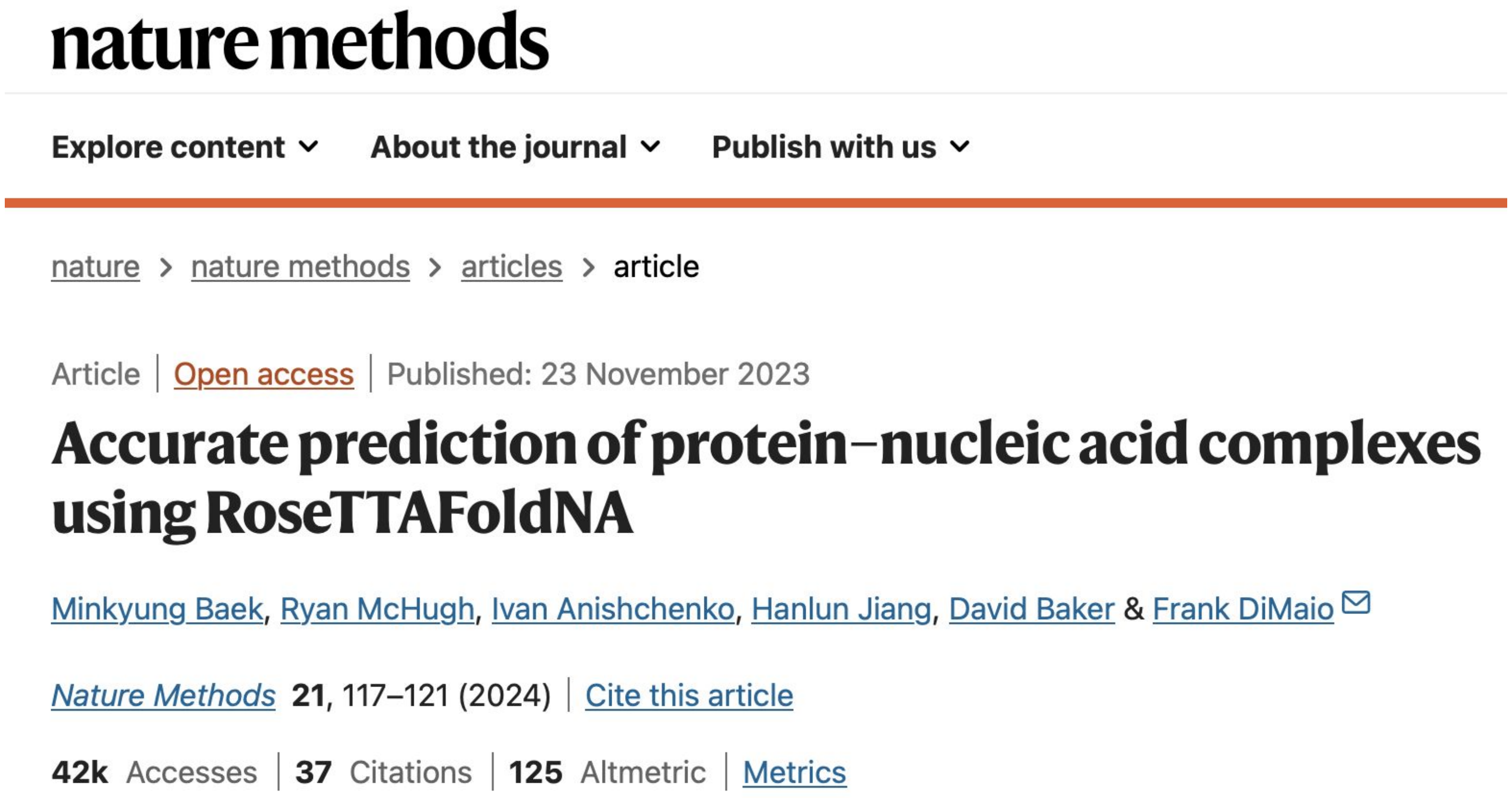
[Submitted on 4 Jul 2022]

E2Efold-3D: End-to-End Deep Learning Method for accurate de novo RNA 3D Structure Prediction

Tao Shen, Zhihang Hu, Zhangzhi Peng, Jiayang Chen, Peng Xiong, Liang Hong, Liangzhen Zheng, Yixuan Wang, Irwin King, Sheng Wang, Siqi Sun, Yu Li

Deep learning based RNA 3D structure prediction tools

- DRfold
- DeepFoldRNA
- RhoFold
- RoseTTAFoldNA



nature methods

nature › nature methods › articles › article

Article | Open access | Published: 23 November 2023

Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA

Minkyung Baek, Ryan McHugh, Ivan Anishchenko, Hanlun Jiang, David Baker & Frank DiMaio

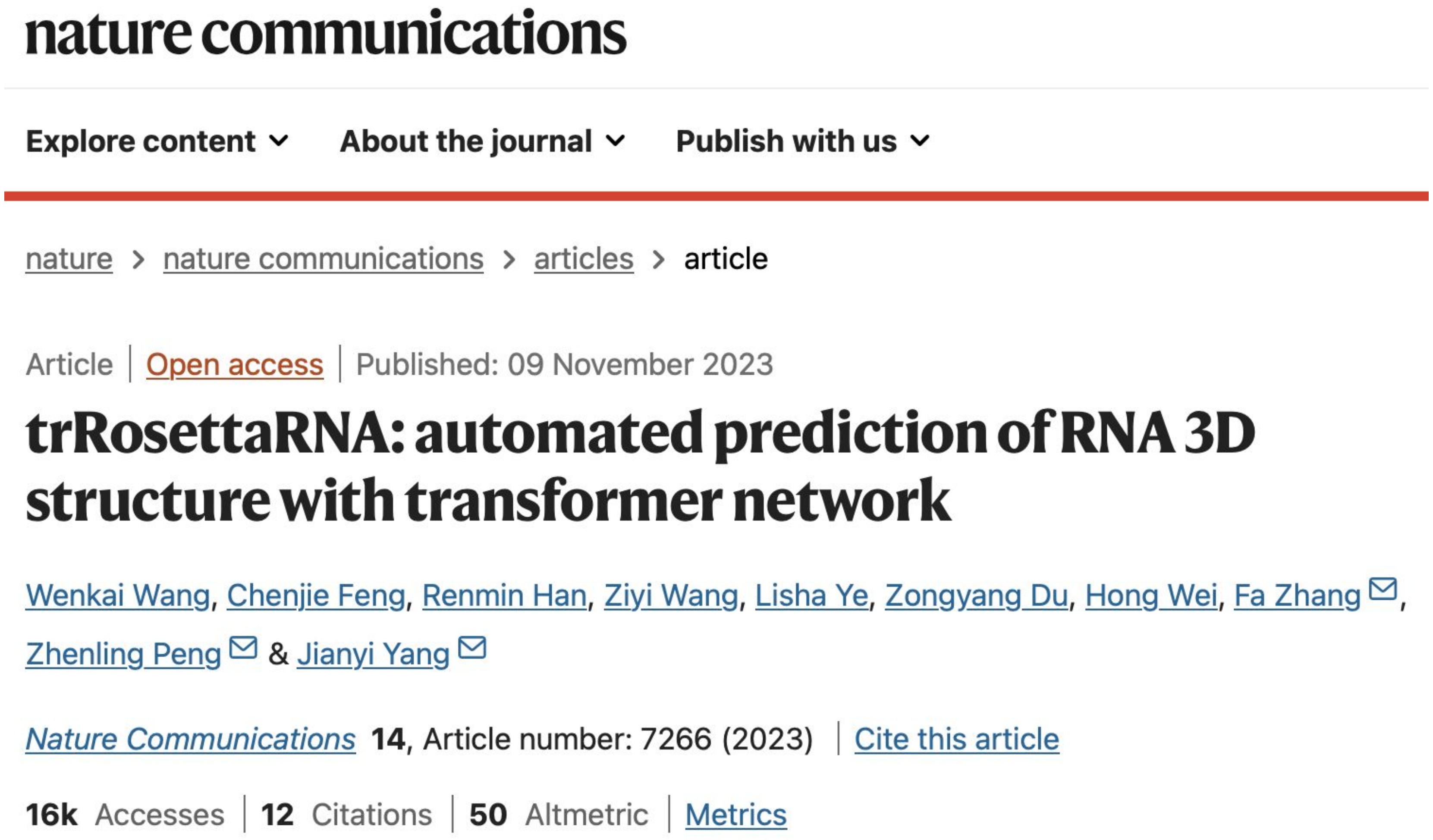
Nature Methods **21**, 117–121 (2024) | Cite this article

42k Accesses | **37** Citations | **125** Altmetric | Metrics

CREATING GROWTH, ENHANCING LIVES

Deep learning based RNA 3D structure prediction tools

- DRfold
- DeepFoldRNA
- RhoFold
- RoseTTAFoldNA
- trRosettaRNA



nature > nature communications > articles > article

Article | Open access | Published: 09 November 2023

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

Wenkai Wang, Chenjie Feng, Renmin Han, Ziyi Wang, Lisha Ye, Zongyang Du, Hong Wei, Fa Zhang, Zhenling Peng & Jianyi Yang

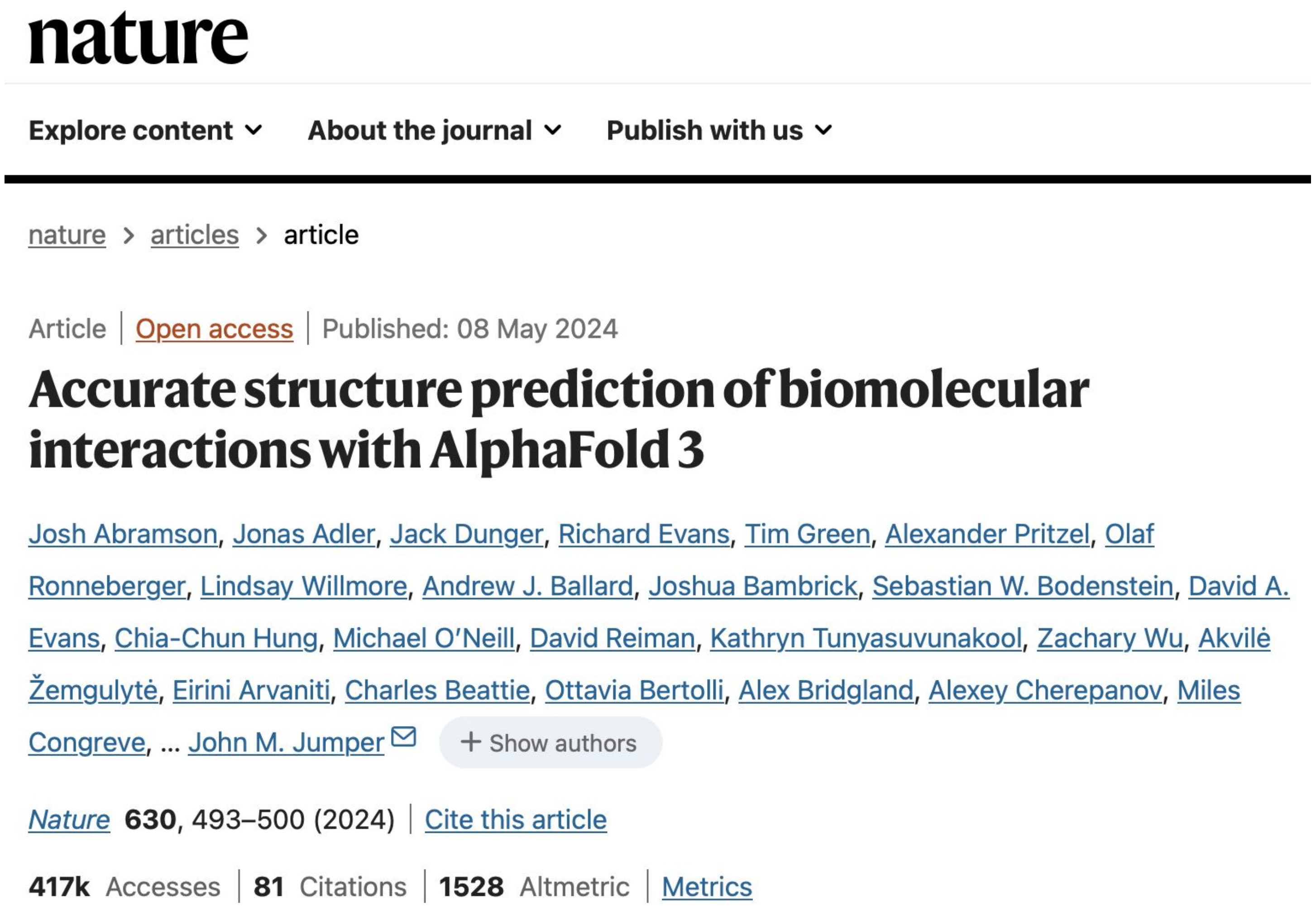
Nature Communications **14**, Article number: 7266 (2023) | Cite this article

16k Accesses | **12** Citations | **50** Altmetric | Metrics

Deep learning based RNA 3D structure prediction tools

- DRfold
- DeepFoldRNA
- RhoFold
- RoseTTAFoldNA
- trRosettaRNA

- AlphaFold3



Article | Open access | Published: 08 May 2024

Accurate structure prediction of biomolecular interactions with AlphaFold 3

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, ... John M. Jumper

Nature **630**, 493–500 (2024) | Cite this article

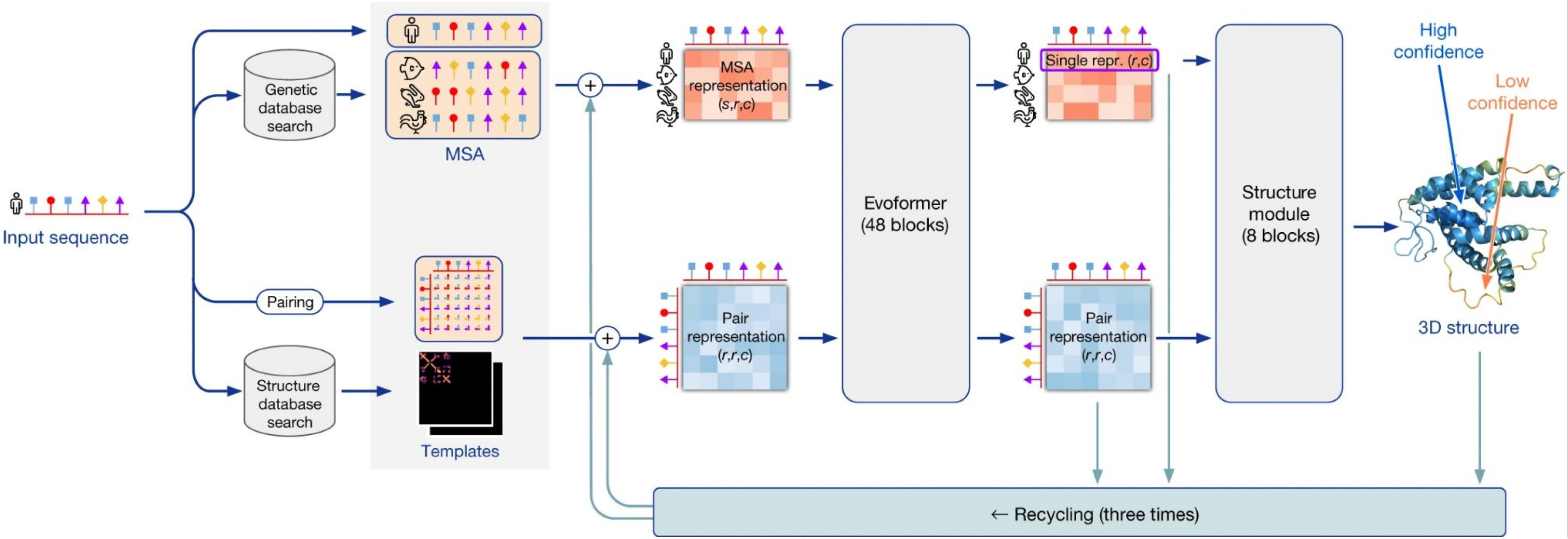
417k Accesses | **81** Citations | **1528** Altmetric | Metrics

CREATING GROWTH, ENHANCING LIVES

Research questions

- which of the tools performs best across different datasets and evaluation metrics?
- do certain design choices and methodologies impact accuracy?
- how well these tools generalize to RNA sequences different from those used in their training?
- can we choose the best predicted structure using ARES or Rosetta score?
- (with AF3) how much does having context (other chains from the complex) help in structure prediction?

AlphaFold2’s architecture



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)

*Figure adapted from [Jumper et al., 2021]

Data preprocessing

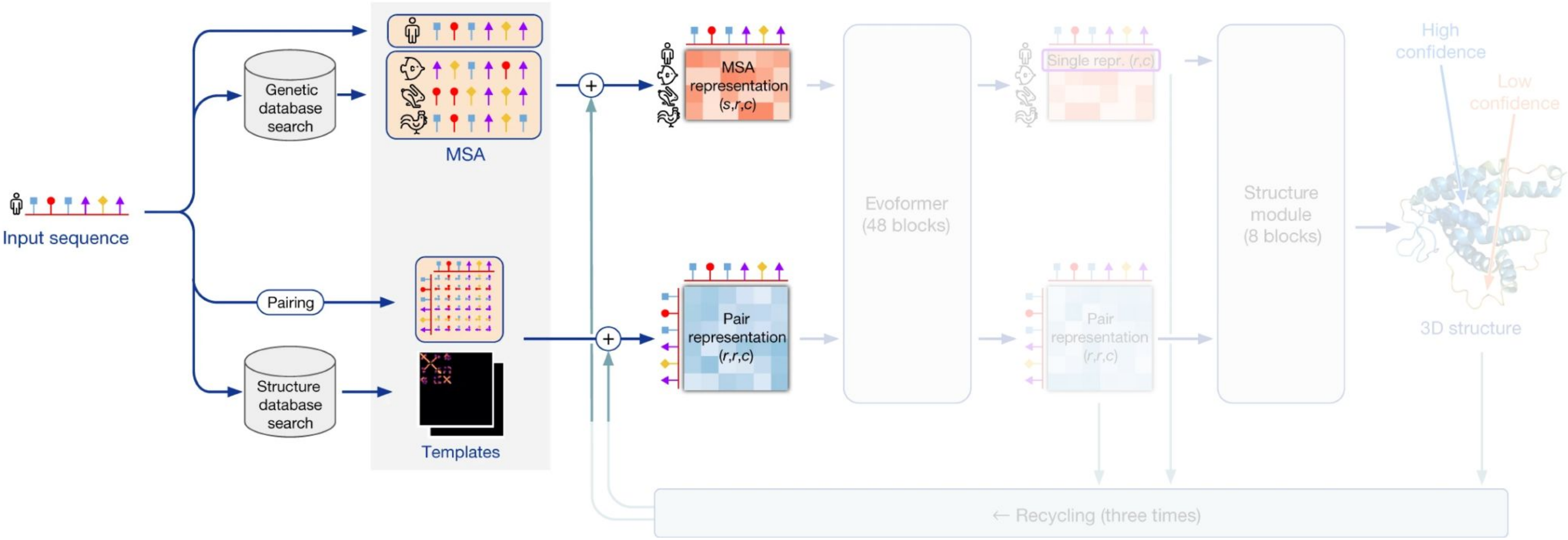

Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
+
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three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

*Figure adapted from [Jumper et al., 2021]

CREATING GROWTH, ENHANCING LIVES

Data preprocessing: multiple sequence alignment (MSA)

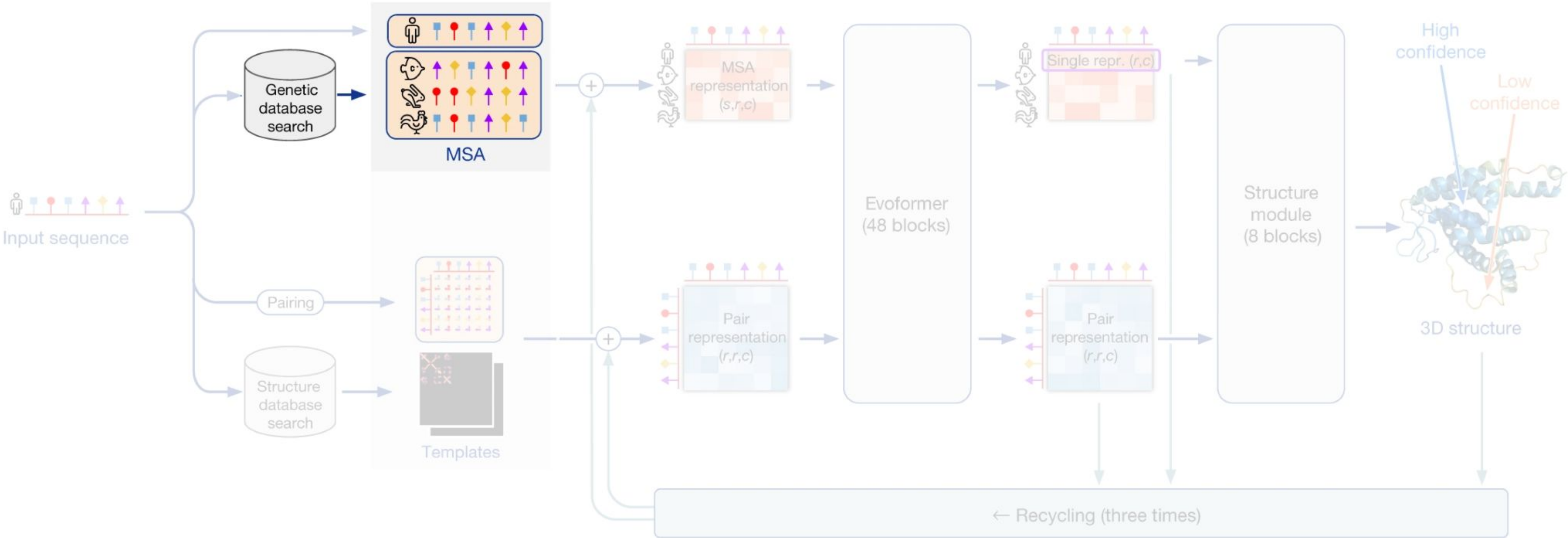
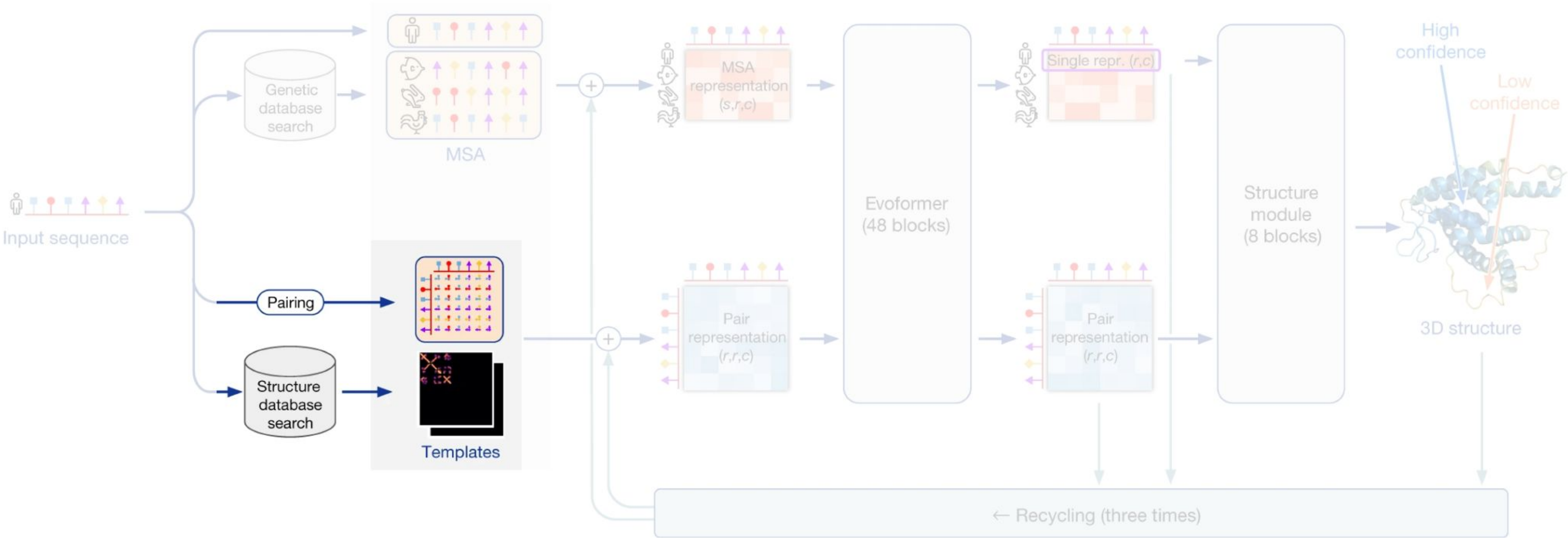

Genetic database search
MSA
Input sequence
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three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

*Figure adapted from [Jumper et al., 2021]

Data preprocessing: secondary structures (SS)



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

*Figure adapted from [Jumper et al., 2021]

End-to-end vs. predicting geometric restraints

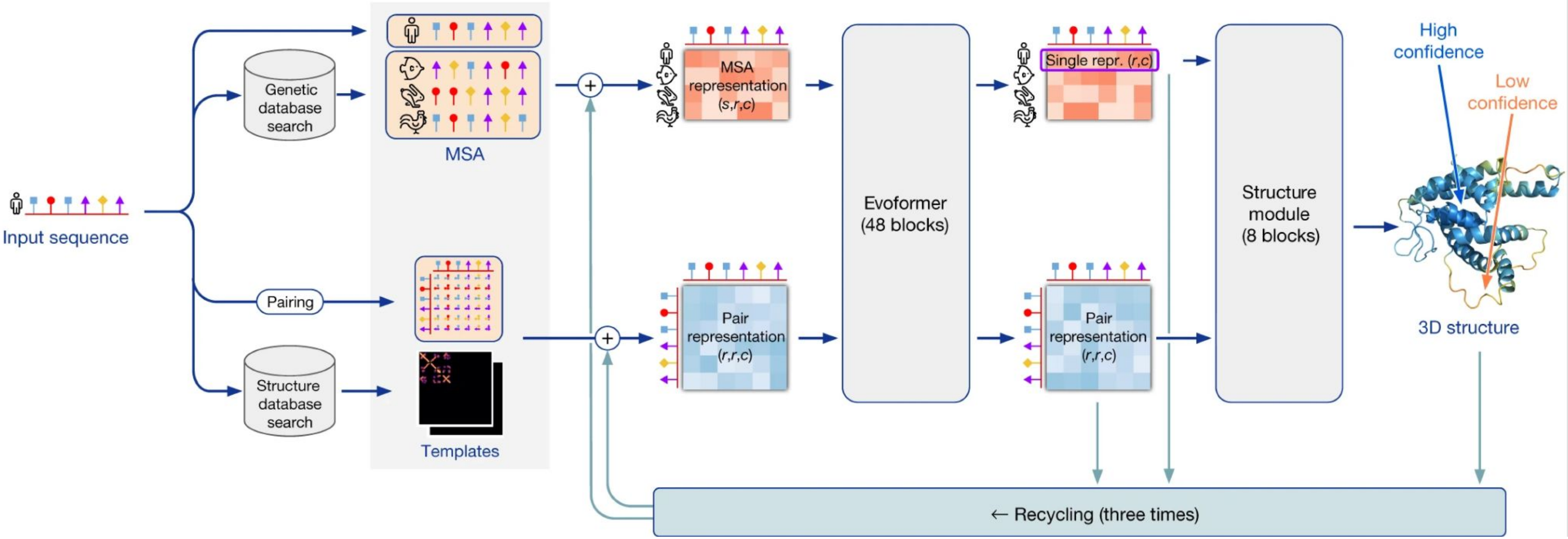

Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three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

*Figure adapted from [Jumper et al., 2021]

CREATING GROWTH, ENHANCING LIVES

End-to-end vs. predicting geometric restraints

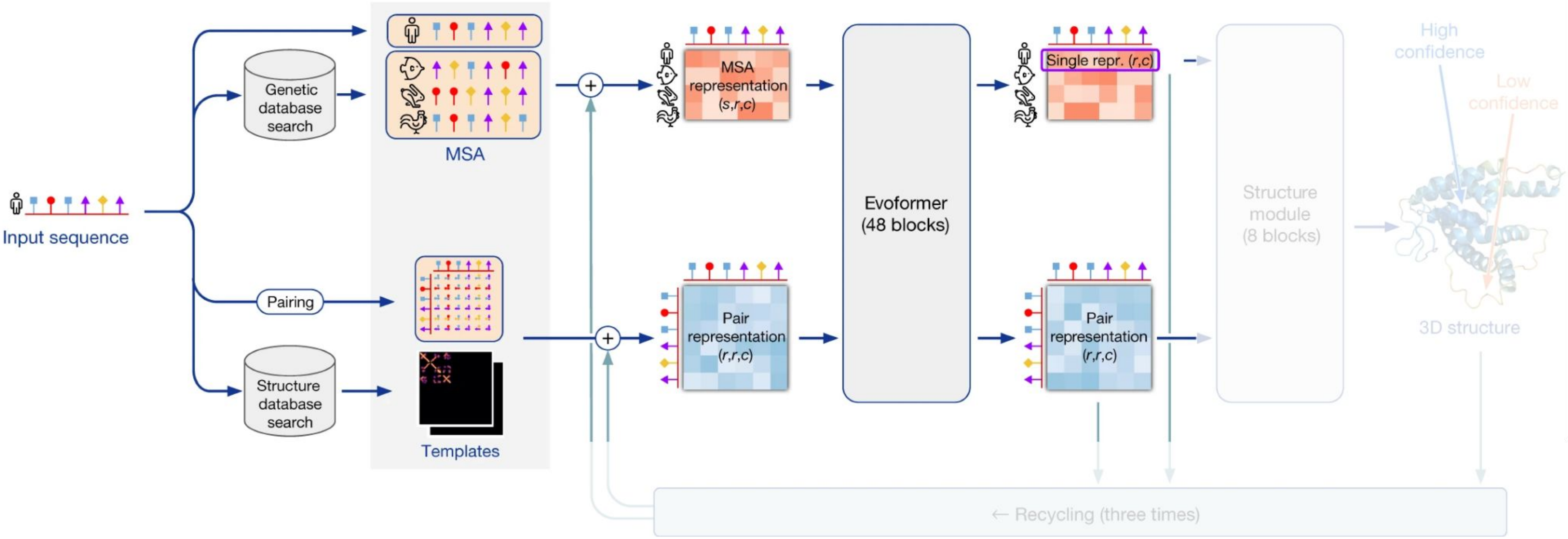

Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three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

*Figure adapted from [Jumper et al., 2021]

From geometric restraints to structural model

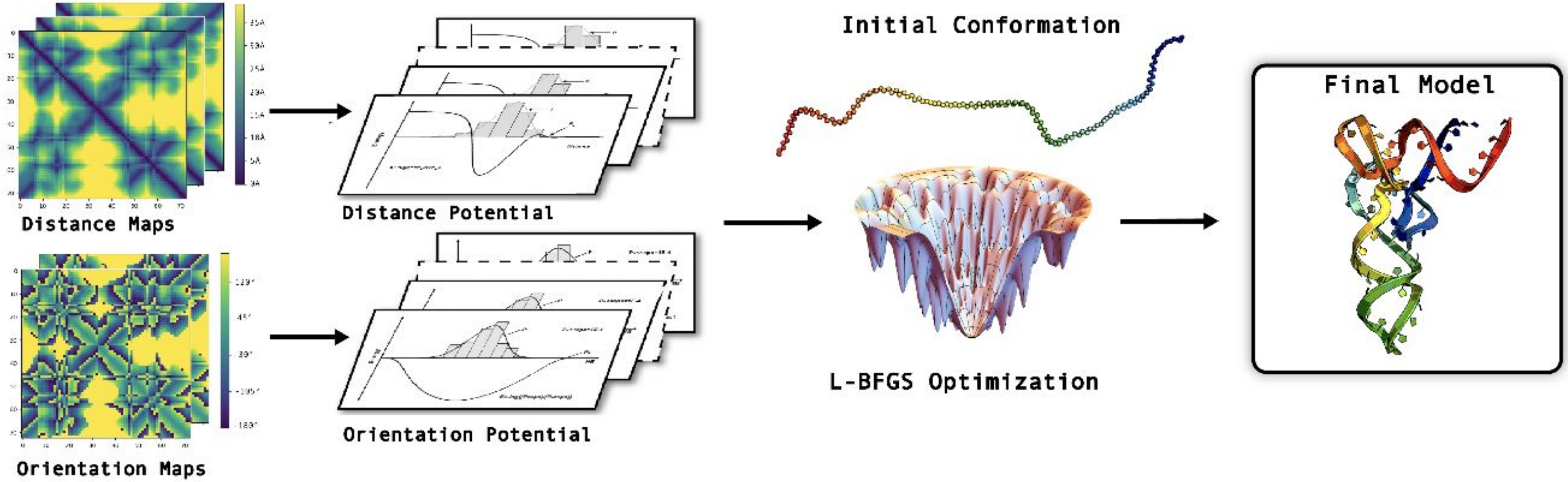


*Figure adapted from [Pearce et al., 2022]

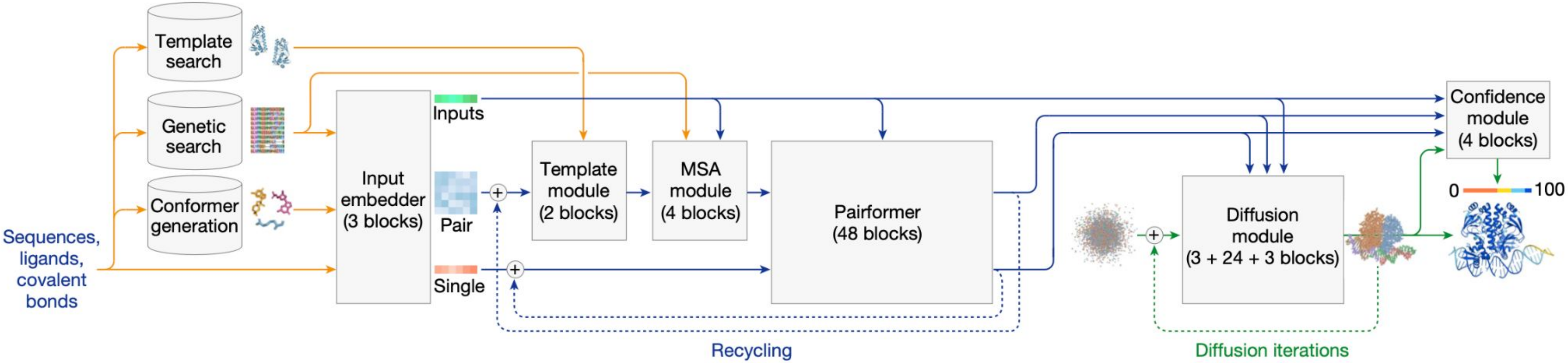
Similarities and differences between RNA models

Tool	end-to-end	uses MSA	uses SS	uses LM
DRfold	✅	❌	✅	❌
DeepFoldRNA	❌	✅	✅	❌
RhoFold	✅	✅	❌	✅
RoseTTAFoldNA	✅	✅	❌	❌
trRosettaRNA	❌	✅	✅	❌

CREATING GROWTH, ENHANCING LIVES

AlphaFold 3 - main differences

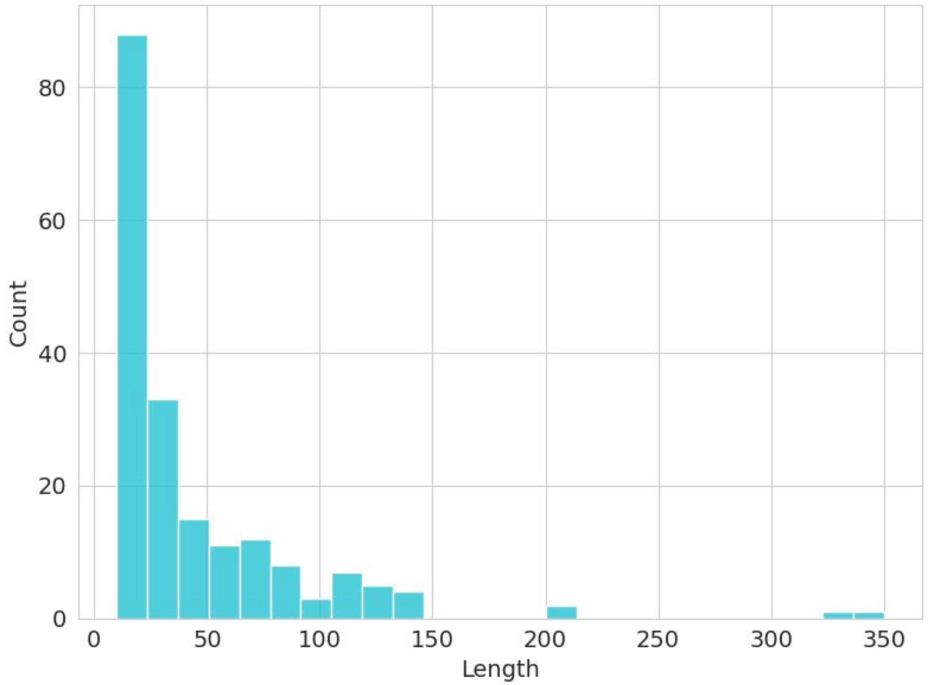
- not only for proteins, AF3 works with RNA chains, ligands and combinations of all three
- changed structure module - diffusion module
- working on atom level instead of residue level


Template search
Genetic search
Conformer generation
Sequences, ligands, covalent bonds
Input embedder (3 blocks)
Inputs
Pair
Single
Template module (2 blocks)
MSA module (4 blocks)
Pairformer (48 blocks)
Recycling
Diffusion module (3 + 24 + 3 blocks)
Diffusion iterations
Confidence module (4 blocks)
0
100

Datasets

- Dataset 1 = RNA Puzzles
  - 37 RNAs - puzzles 35 and 36 removed since they are in CASP15 dataset
- Dataset 2 = CASP15
  - 12 RNAs - 8 natural & 4 synthetic
- Dataset 3 = curated dataset from PDB
  - 190 RNAs
  - published after April 2022 in Protein Data Bank (PDB)
  - clustered with sequence identity 90% - only considering ones which are not similar to those prior to April 2022
  - filtering: length < 10nt, resolution > 9Å, %defined residues < 90%, sequences containing only 'N' or 'X'
  - 329 clusters => only 190 without errors for all tools

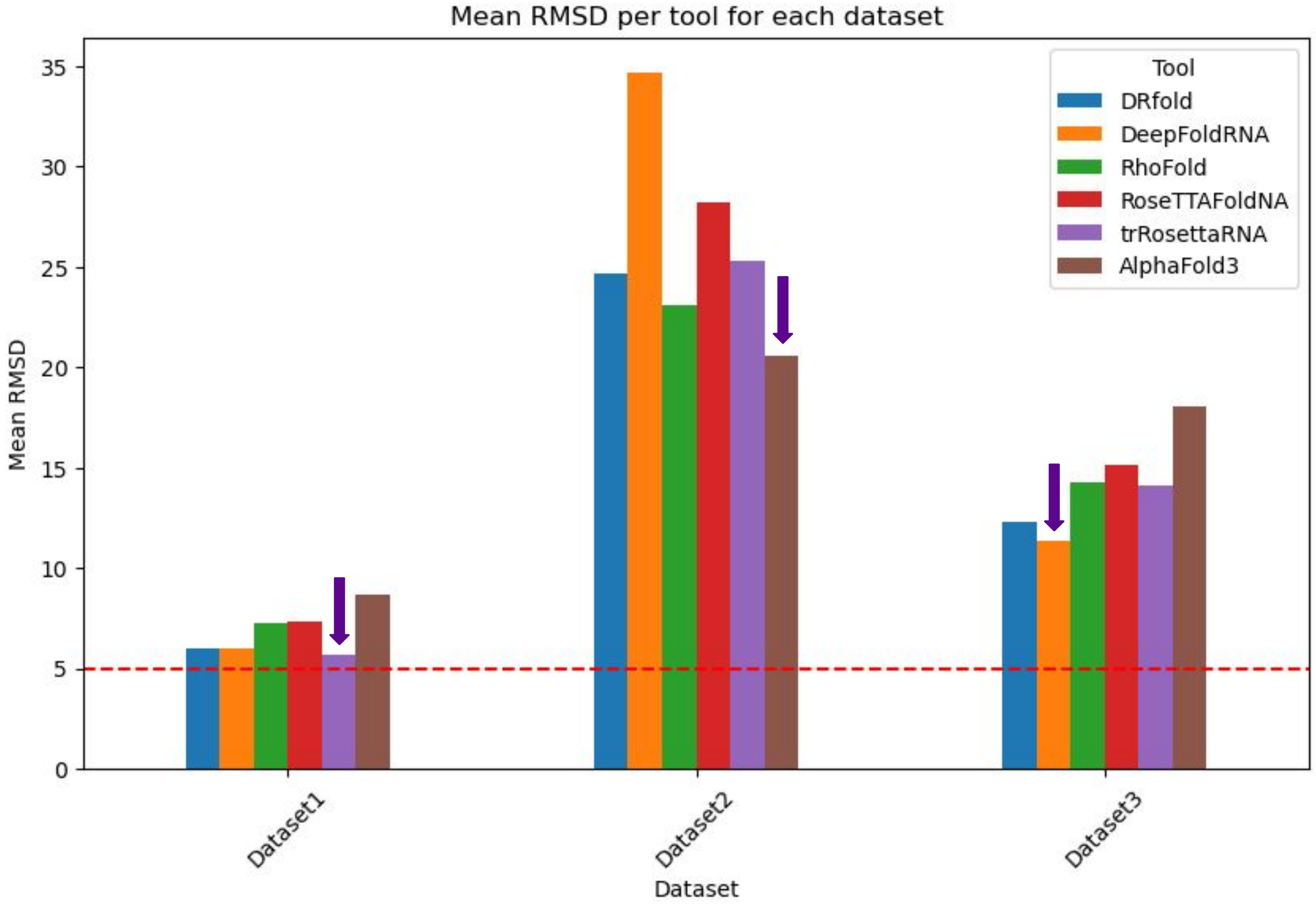
Length distribution for Dataset 3



CREATING GROWTH, ENHANCING LIVES

Results

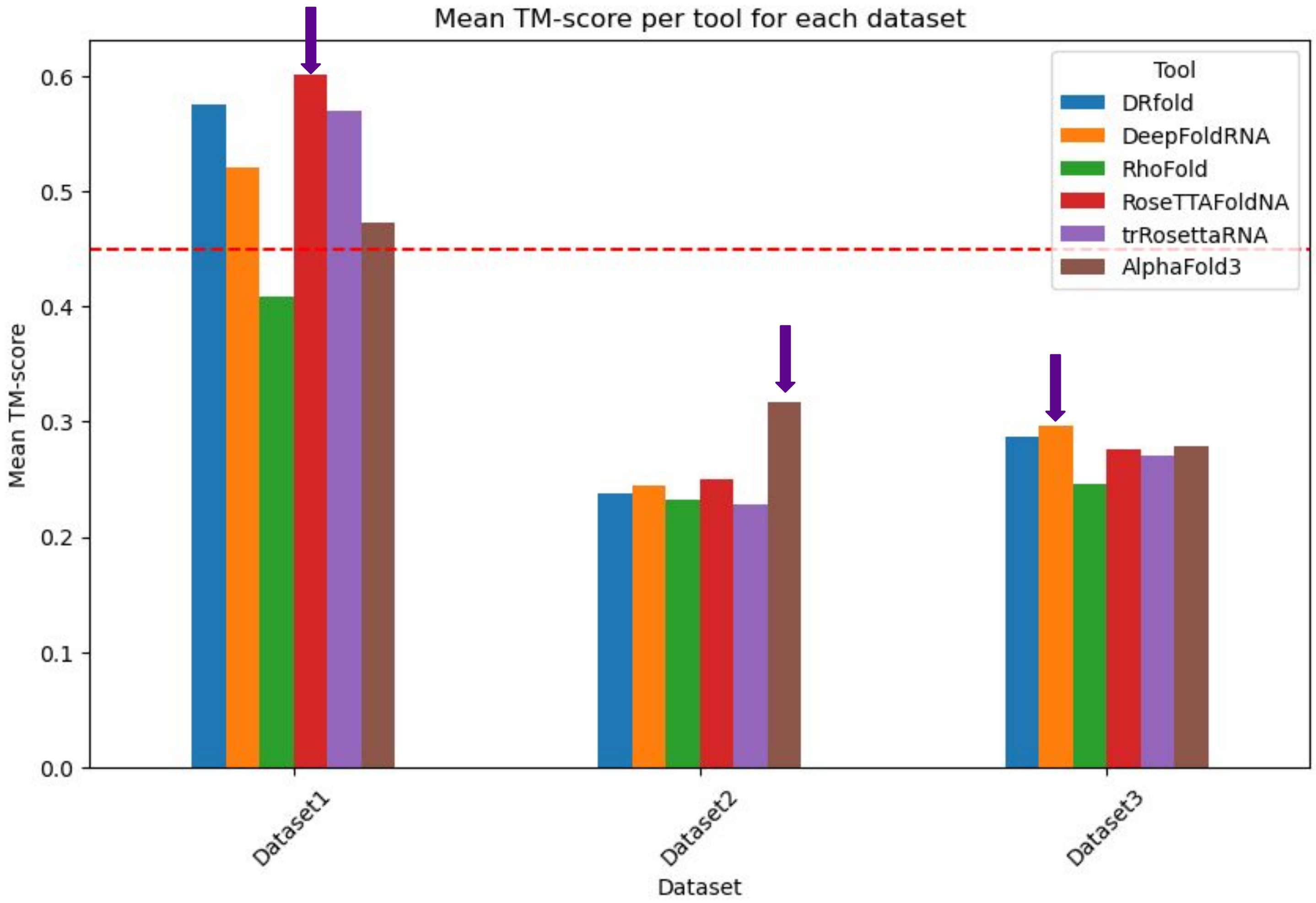
Results - RMSD



Note:
different datasets = different tool with the lowest RMSD

CREATING GROWTH, ENHANCING LIVES

Results - TM-score

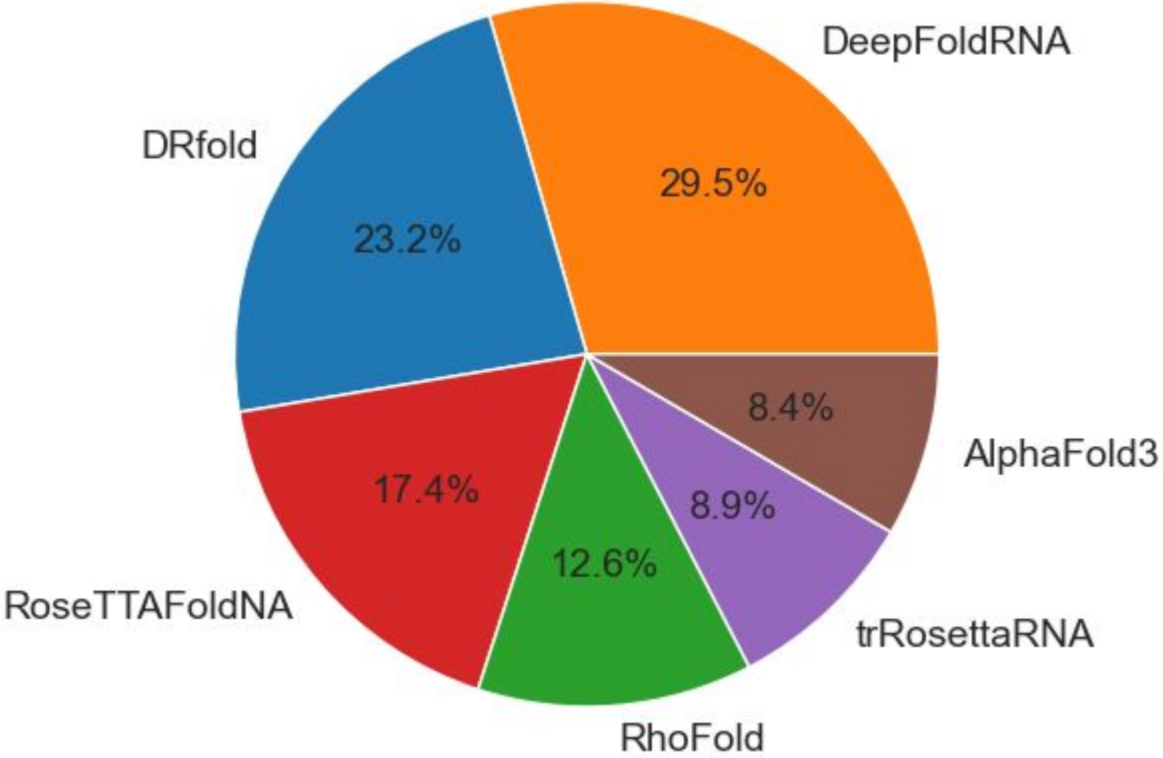


Note:
for Dataset 1 RoseTTAFoldNA is the best (has the highest TM-score), not trRosettaRNA (as in the case of RMSD)

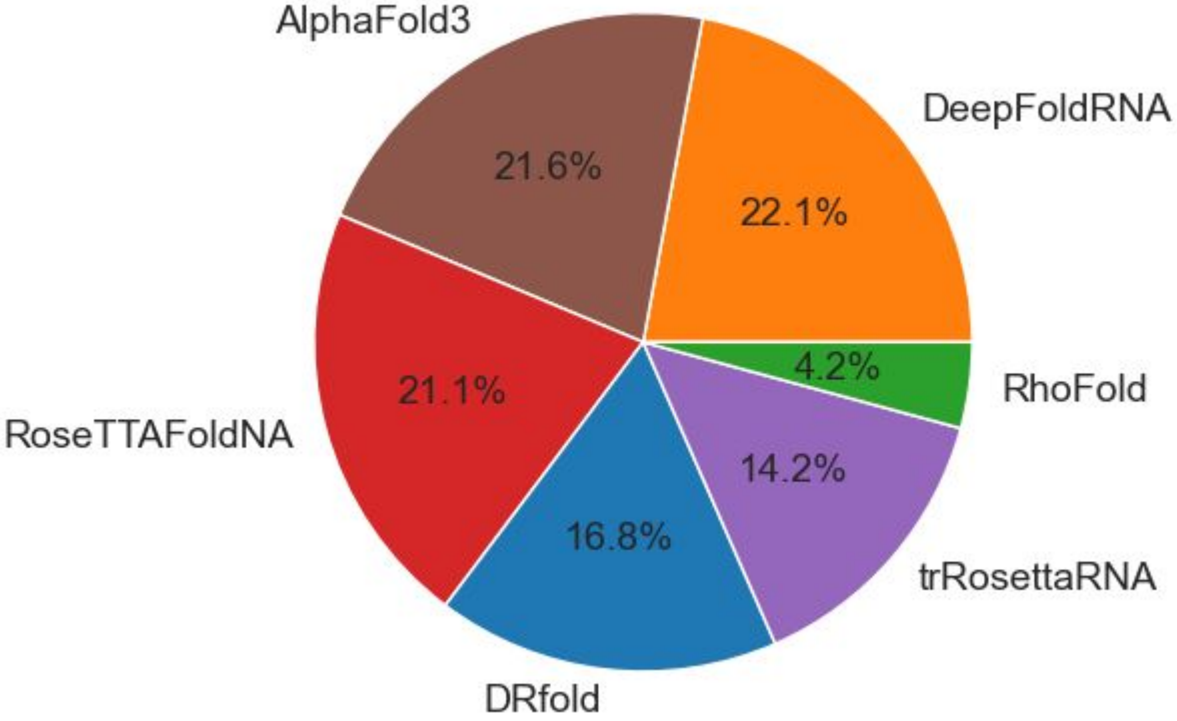
CREATING GROWTH, ENHANCING LIVES

Percentage of RNAs for which each tool was the best

according to RMSD:



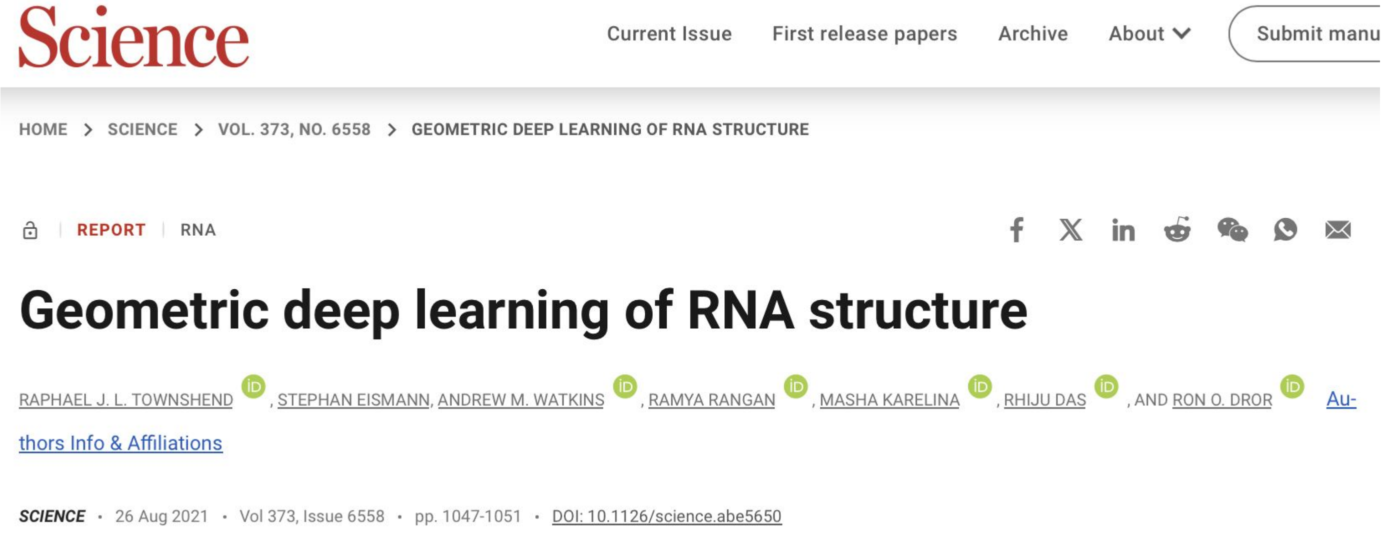
according to TM-score:



Scoring functions - ARES and Rosetta scores

- ARES is a deep learning method for scoring RNA structures
  (input = PDB file,
  output = score for given structure - lower values mean better structure)

- Rosetta score = rna_score
  from Rosetta toolkit



HOME > SCIENCE > VOL. 373, NO. 6558 > GEOMETRIC DEEP LEARNING OF RNA STRUCTURE

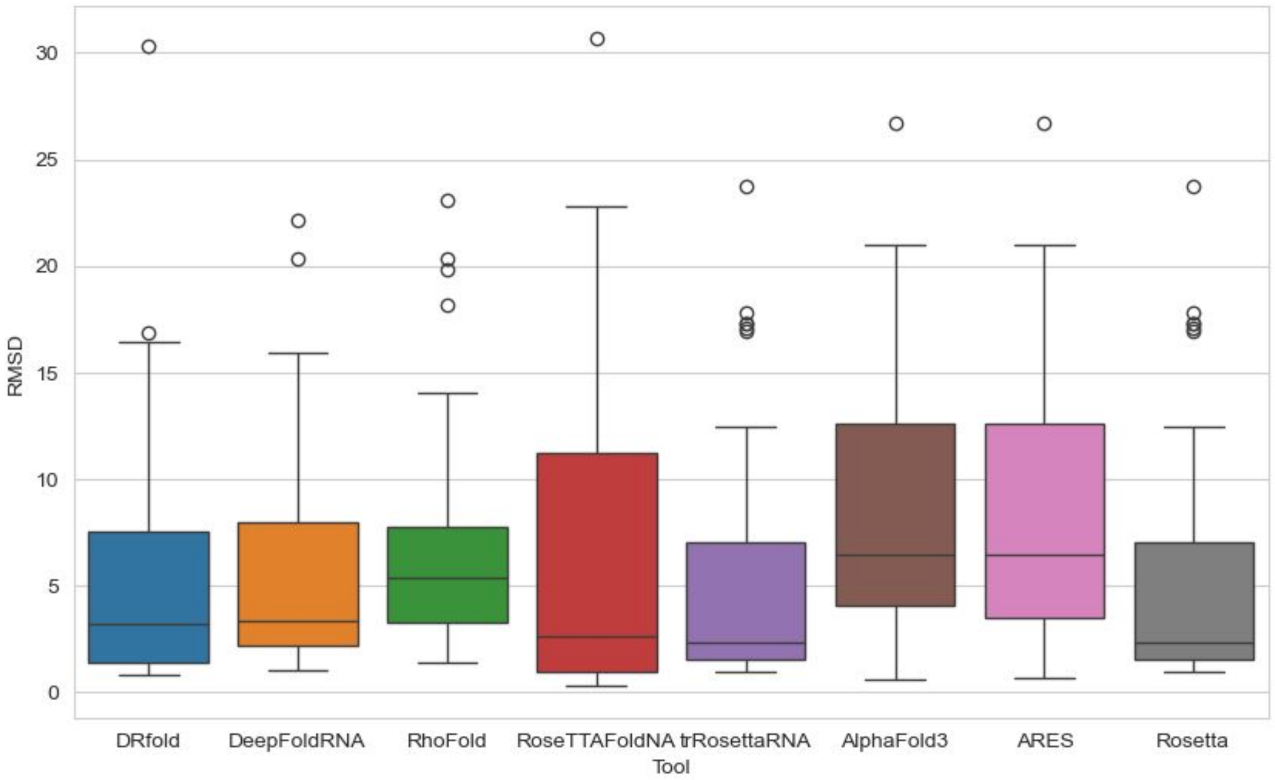
REPORT | RNA

Geometric deep learning of RNA structure

RAPHAEL J. L. TOWNSHEND, STEPHAN EISMANN, ANDREW M. WATKINS, RAMYA RANGAN, MASHA KARELINA, RHIJU DAS, AND RON O. DROR Authors Info & Affiliations

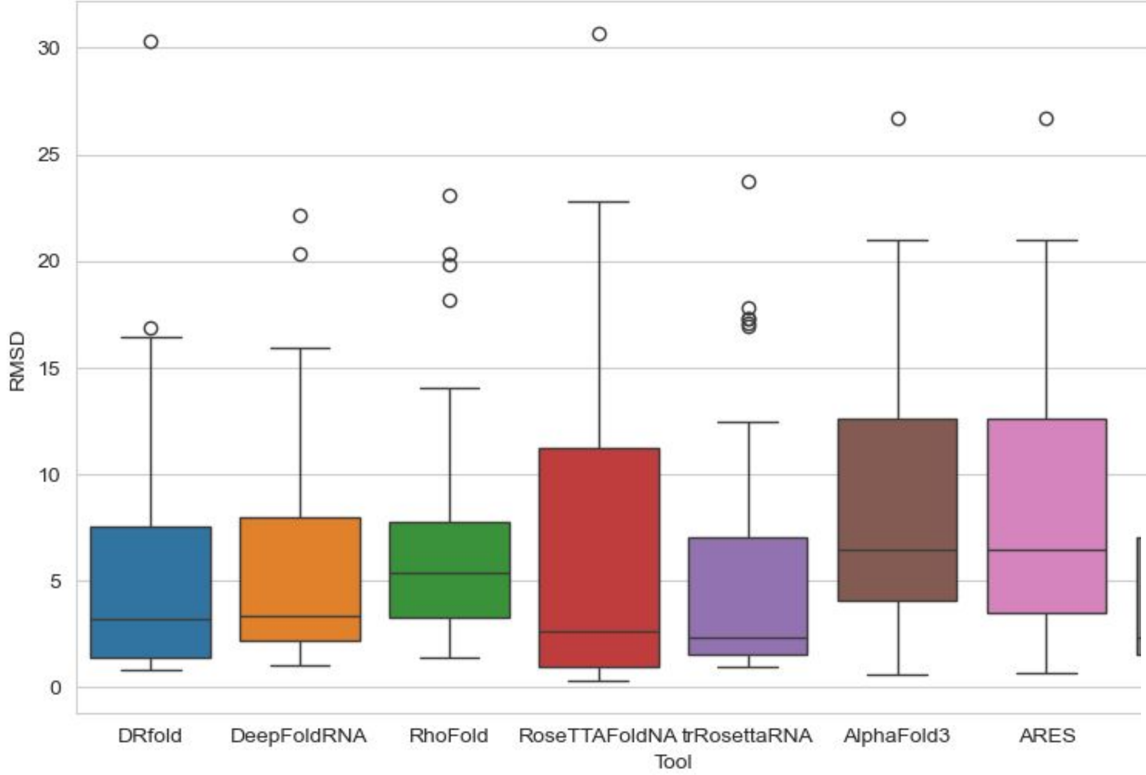
SCIENCE • 26 Aug 2021 • Vol 373, Issue 6558 • pp. 1047-1051 • DOI: 10.1126/science.abe5650

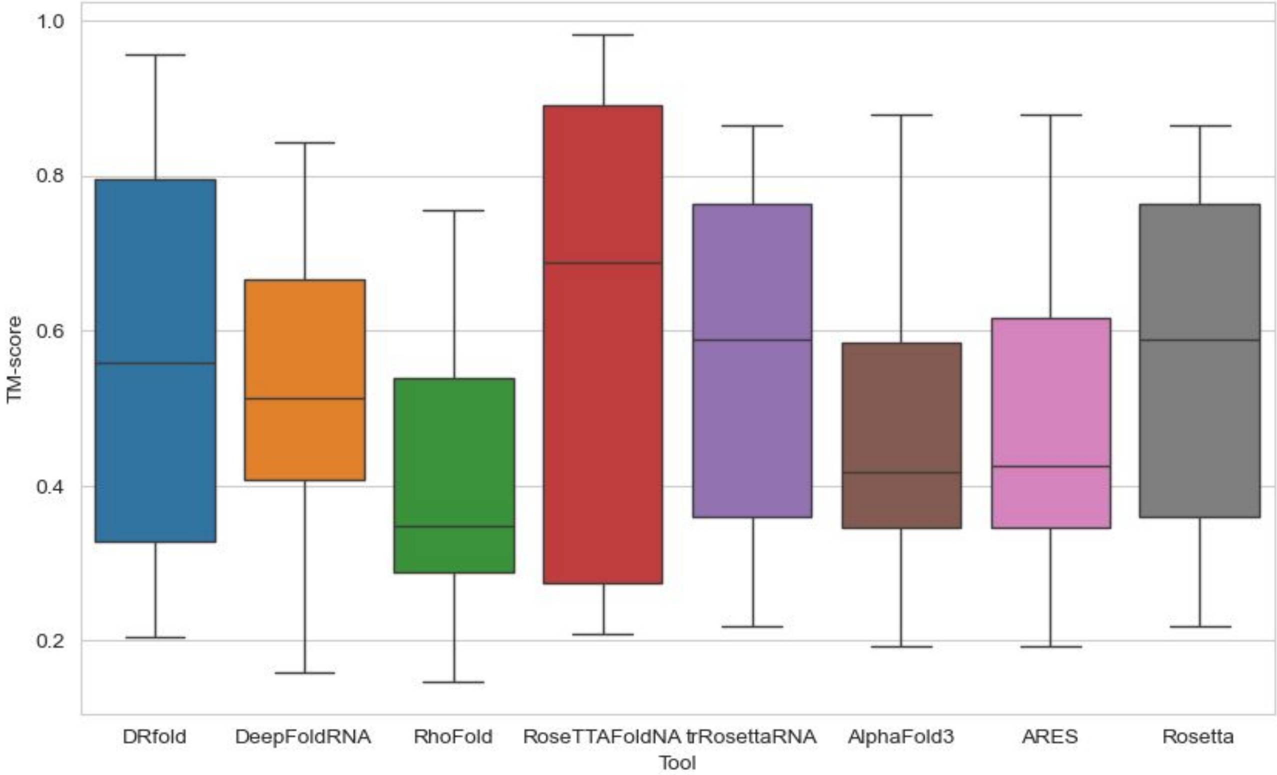
Dataset 1 - RNA Puzzles



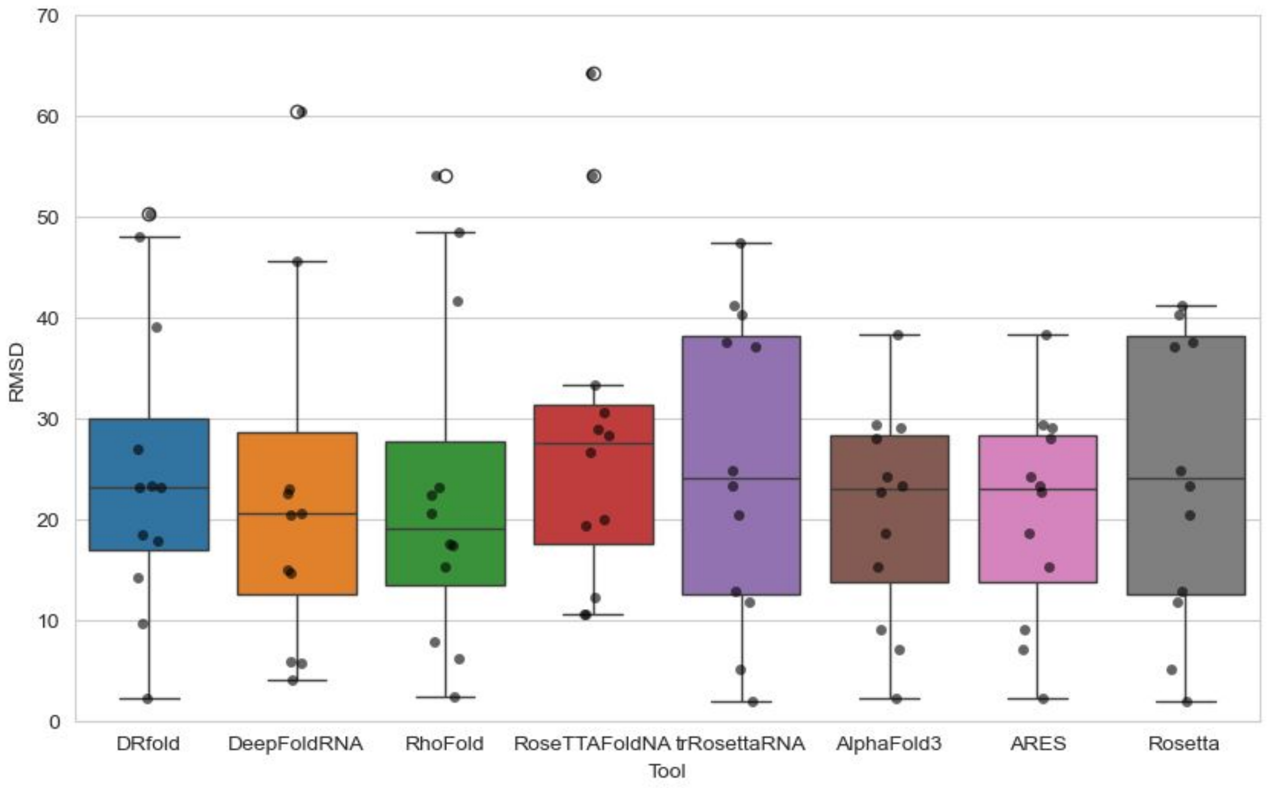
Note:
Rosetta score was the lowest for trRosettaRNA's model for all RNAs

Dataset 1 - RNA Puzzles





Note:
Rosetta score was the lowest for trRosettaRNA's model for all RNAs

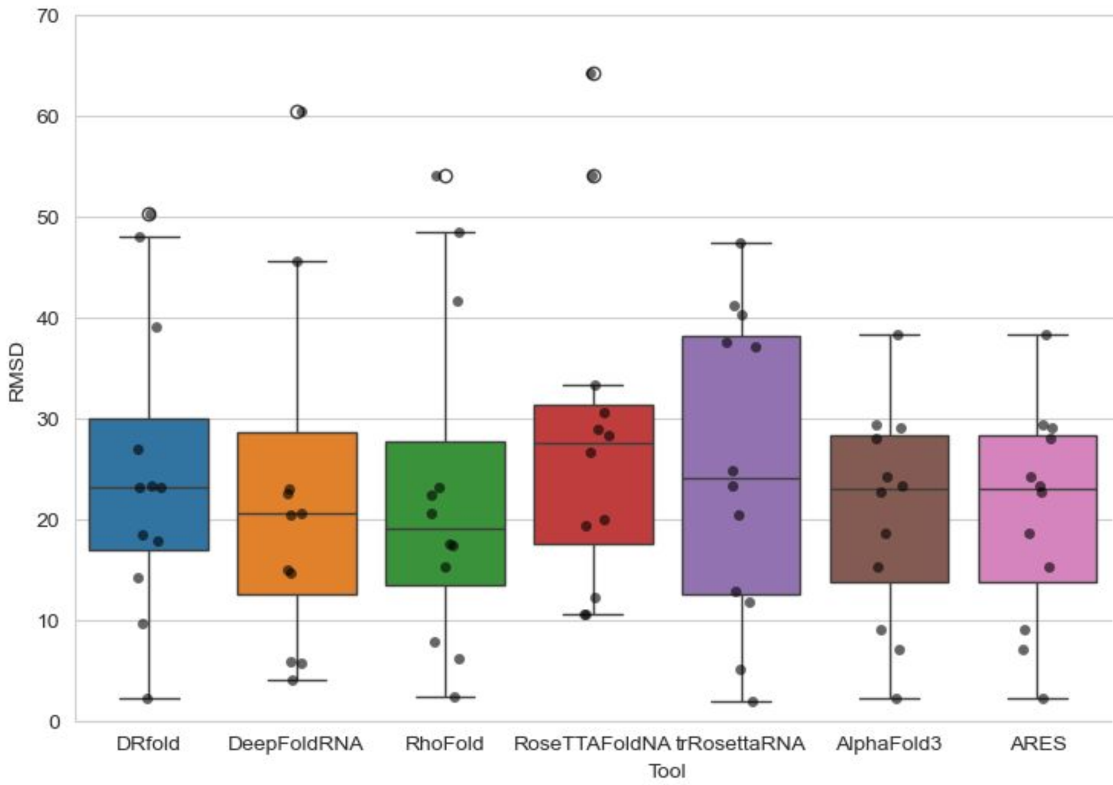


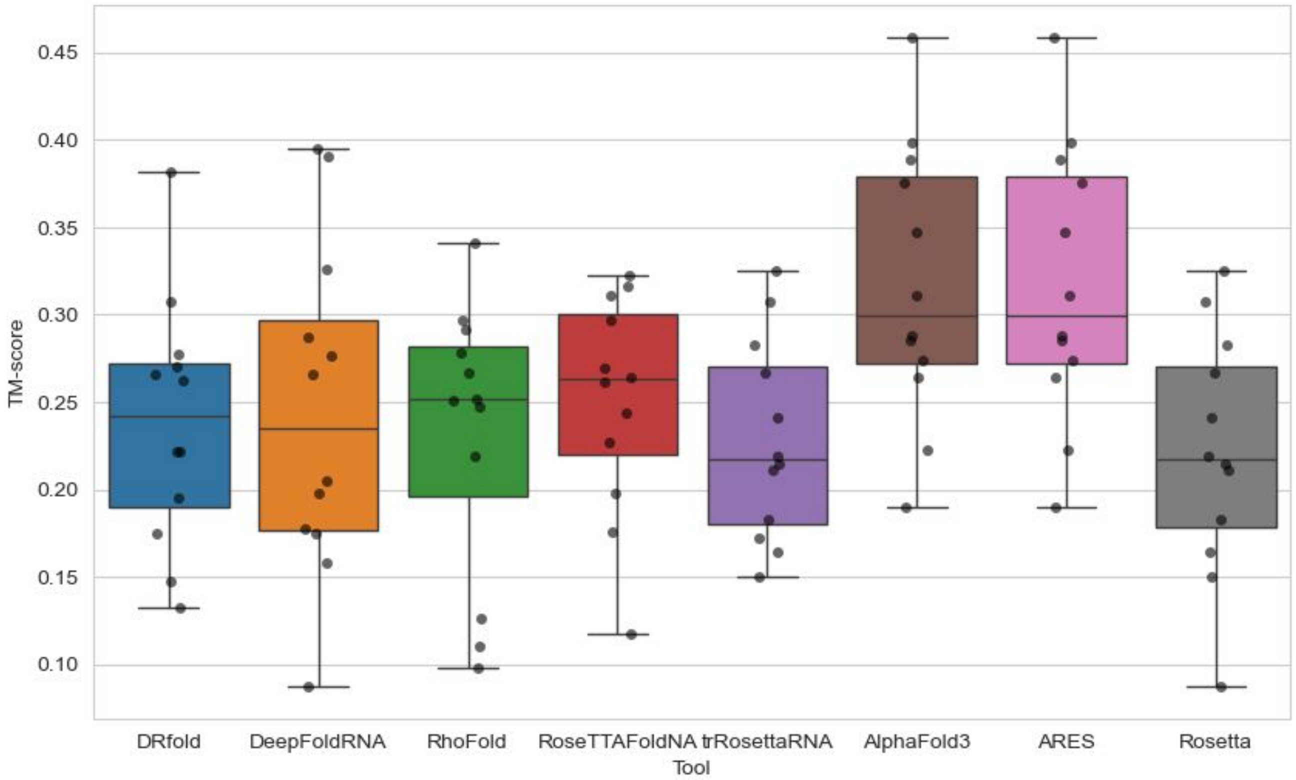
Dataset 2 - CASP15

Note:
ARES selects AF3 models as best for all RNA targets.

CREATING GROWTH, ENHANCING LIVES

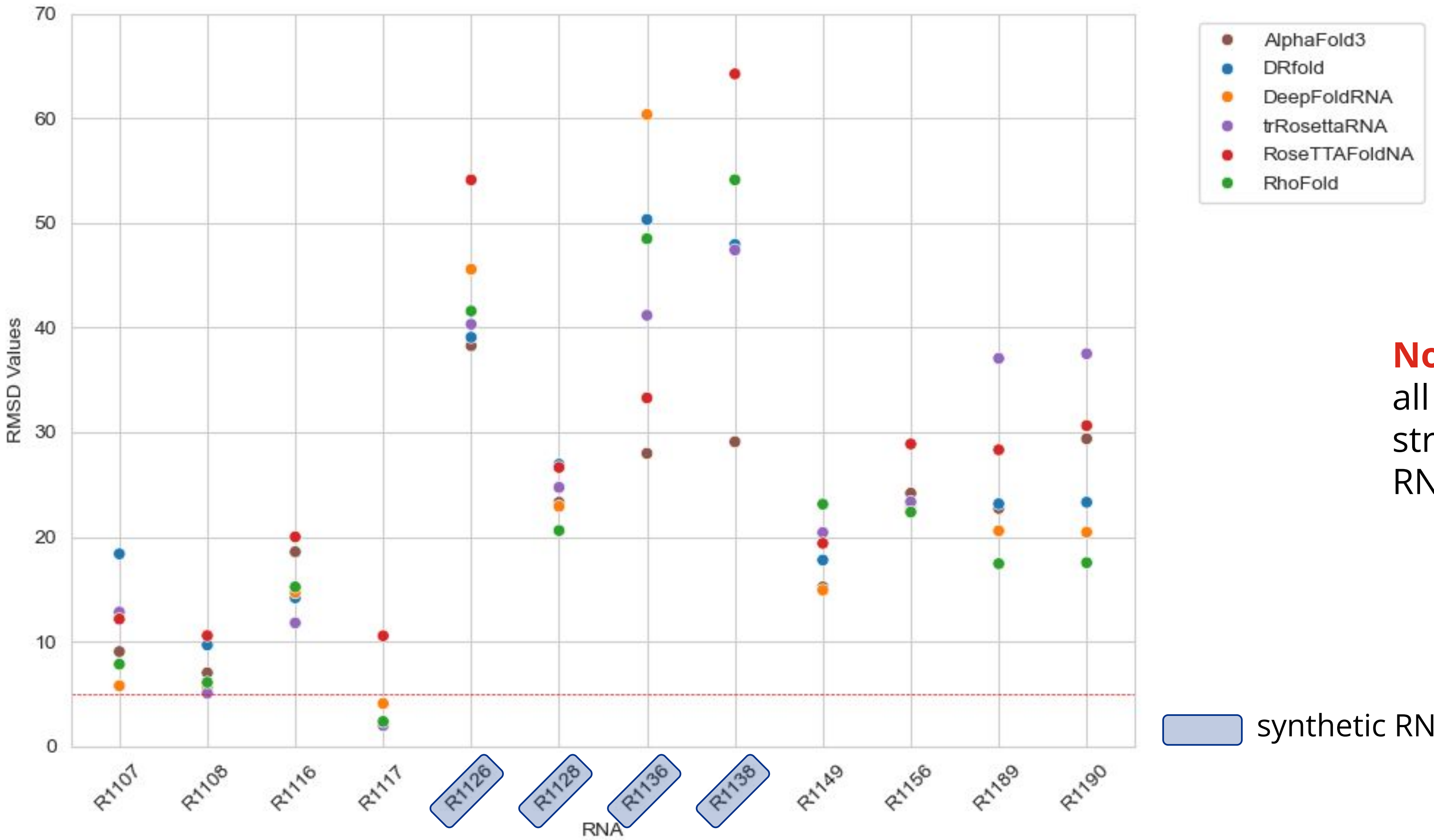
Dataset 2 - CASP15





Note:
ARES selects AF3 models as best for all RNA targets.

Dataset 2 - CASP15



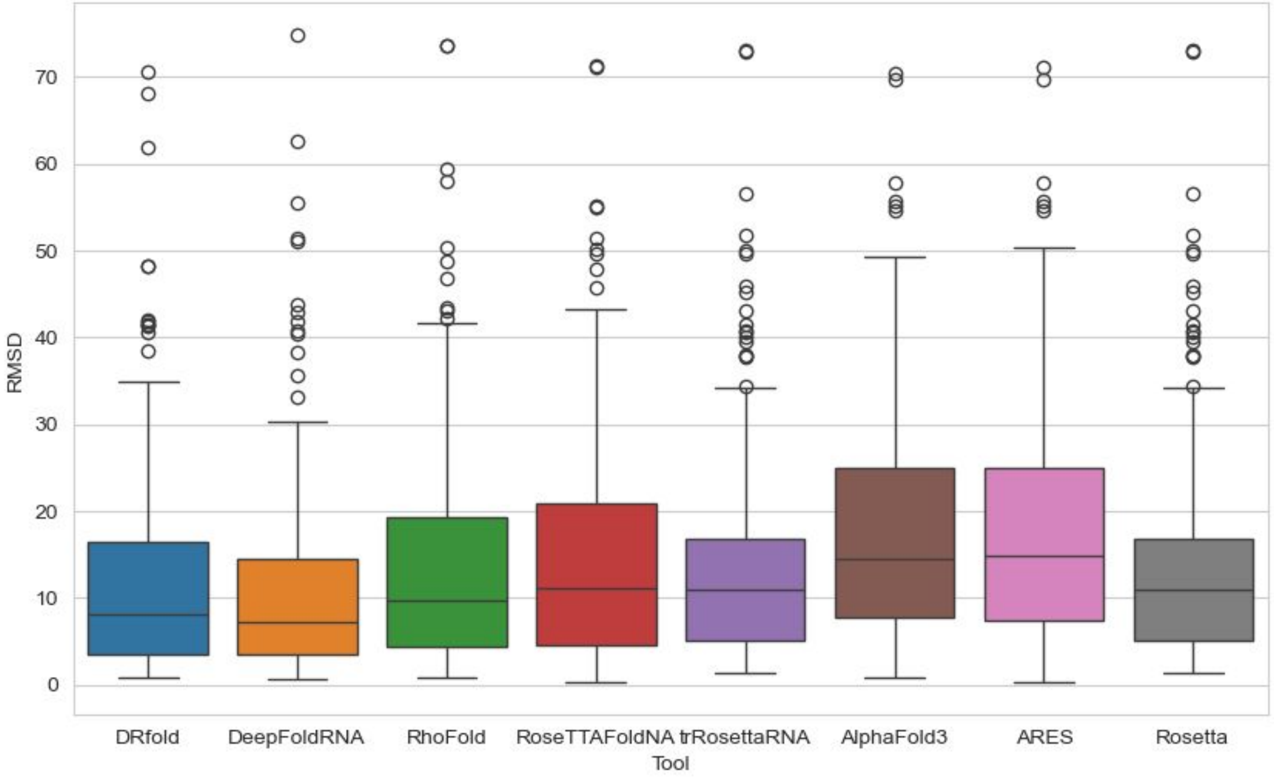
Note:
all DL-based methods struggle with synthetic RNAs

▭ synthetic RNA

CREATING GROWTH, ENHANCING LIVES

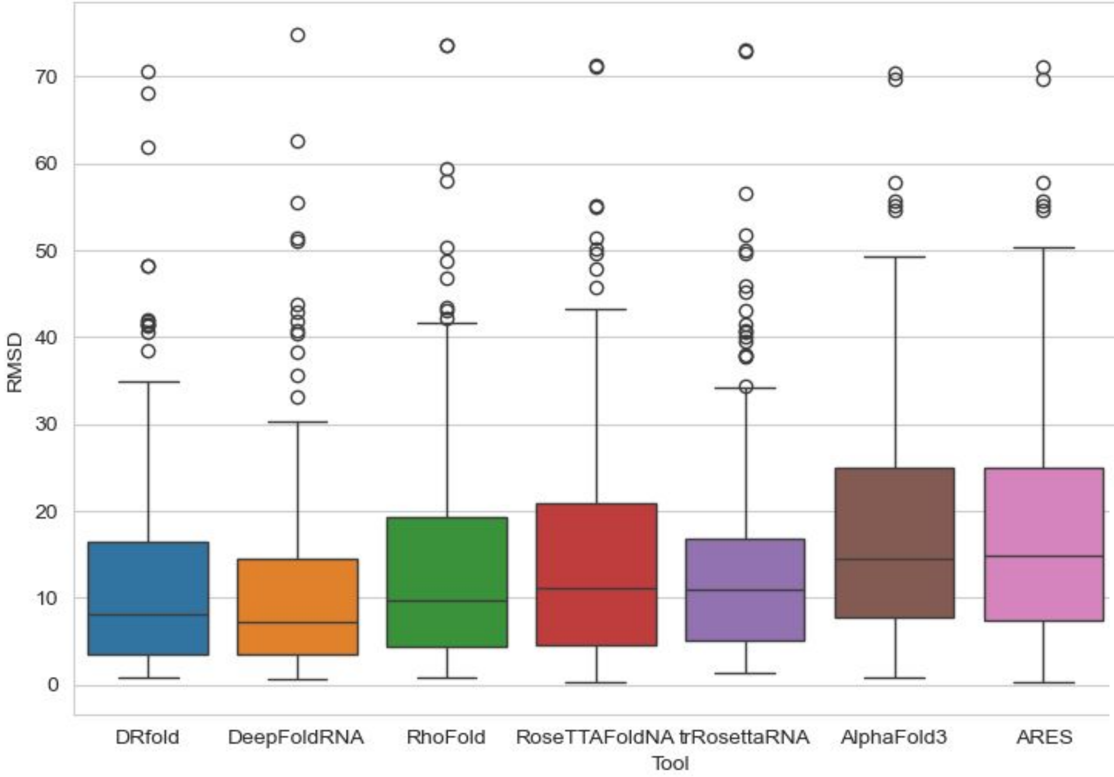
CREATING GROWTH, ENHANCING LIVES

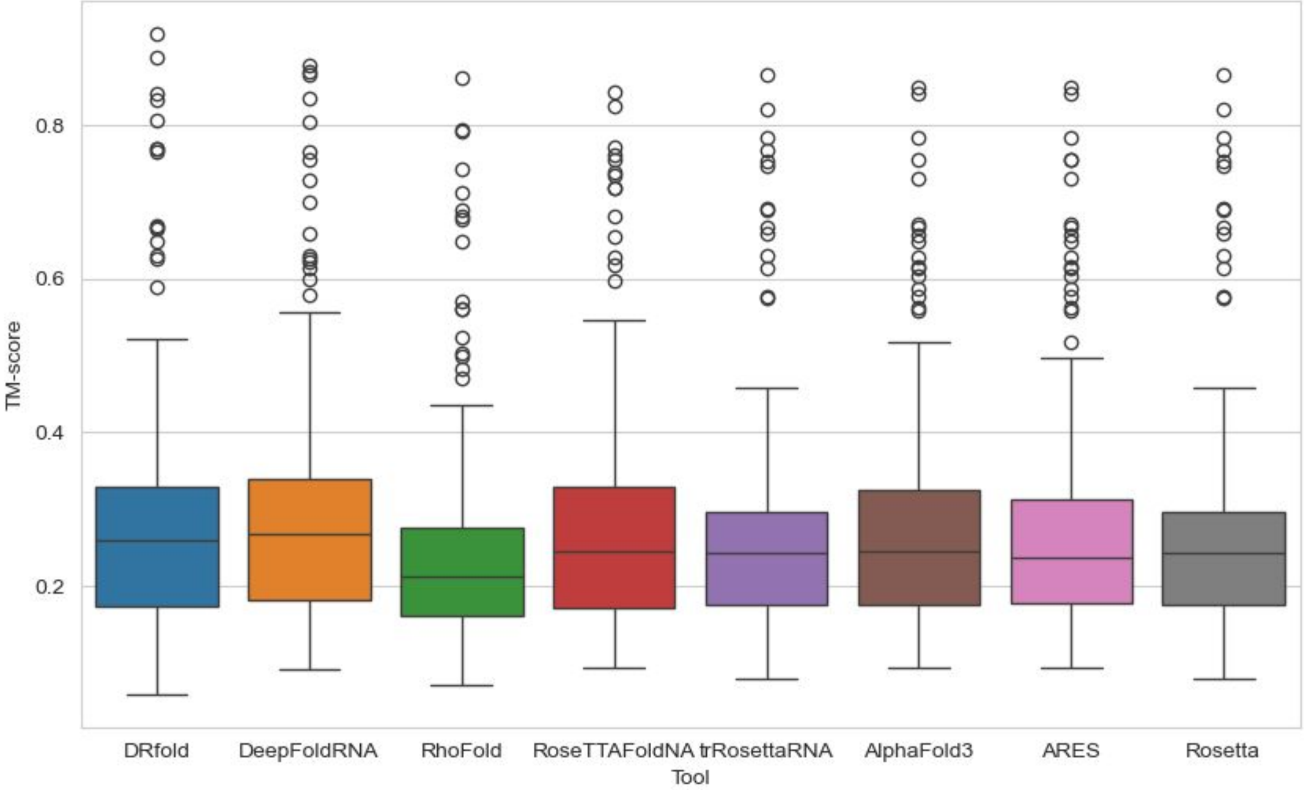
Dataset 3



Note:
again, Rosetta score was the lowest for trRosettaRNA's model for all RNAs

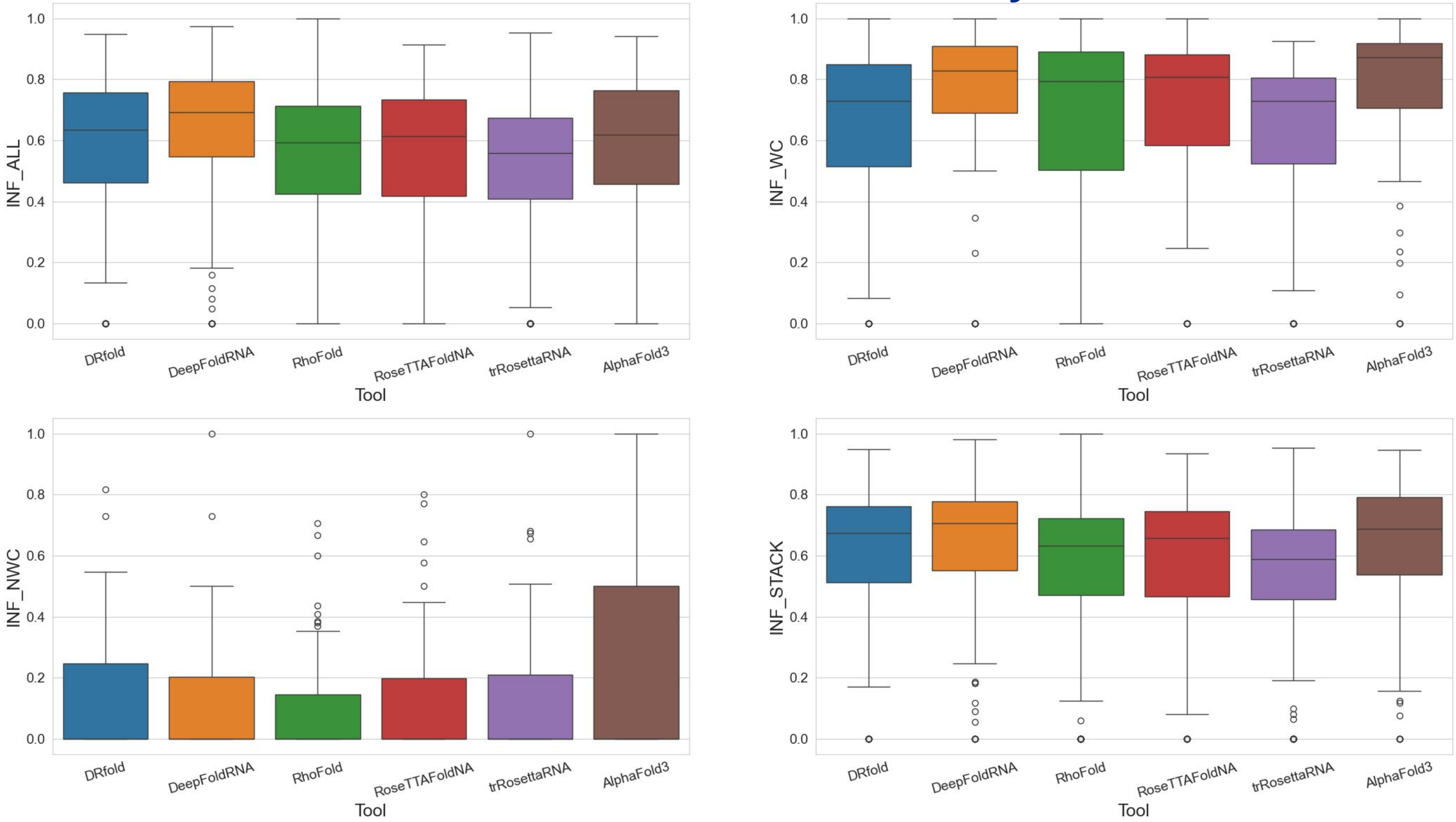
Dataset 3





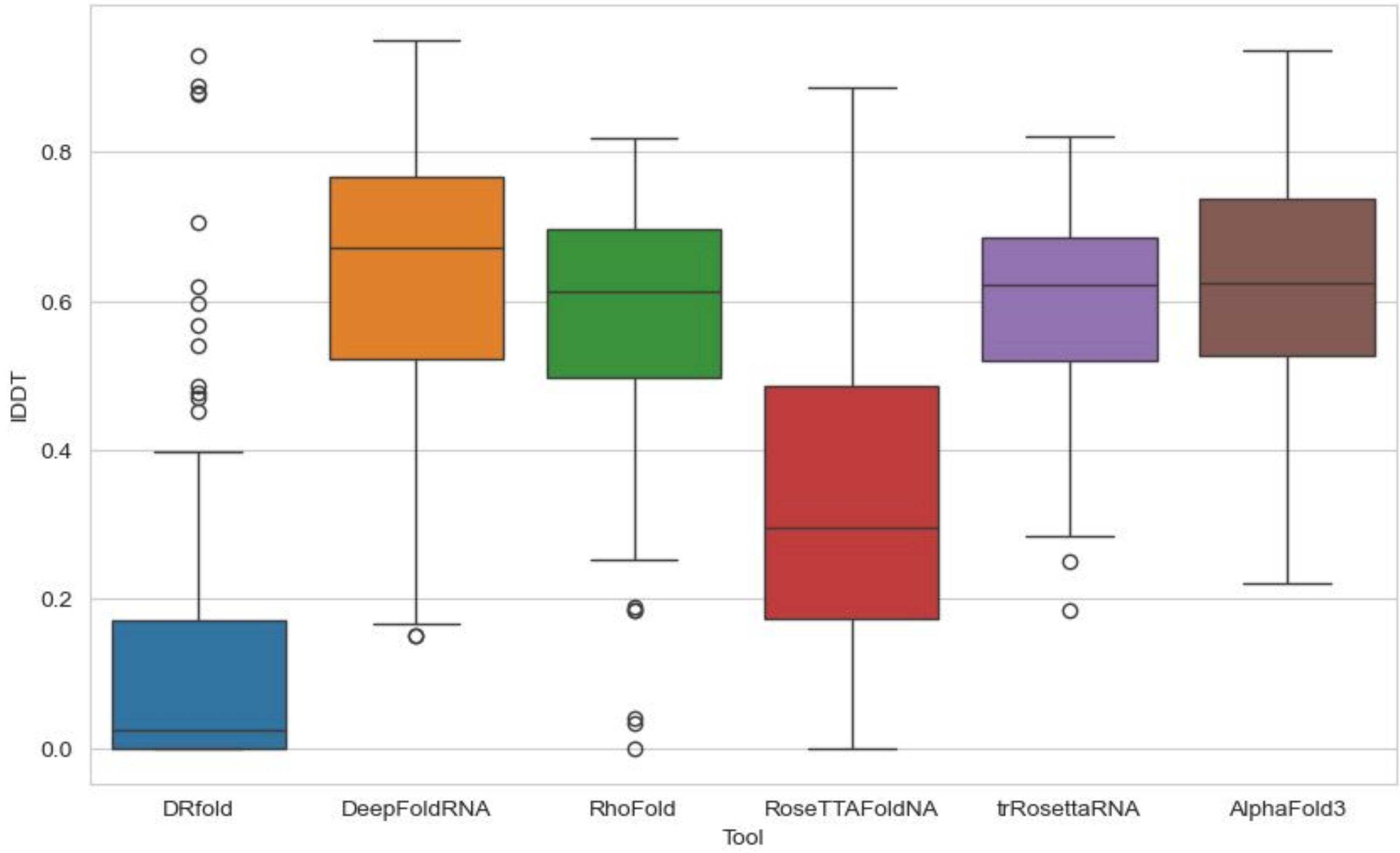
Note:
again, Rosetta score was the lowest for trRosettaRNA's model for all RNAs

Dataset 3: Interaction Network Fidelity (INF)

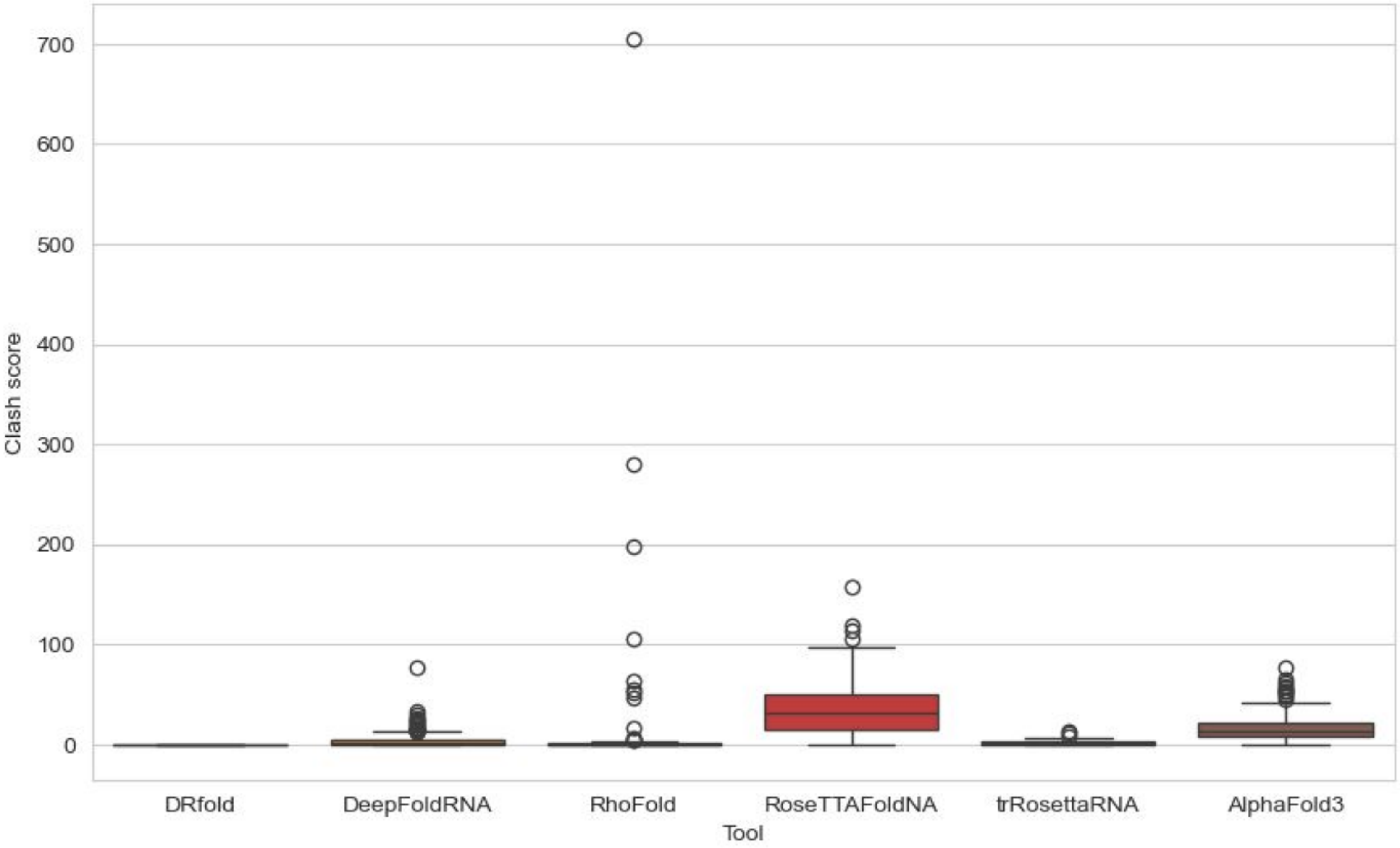


CREATING GROWTH, ENHANCING LIVES

Dataset 3: Local Distance Difference Test (lDDT)



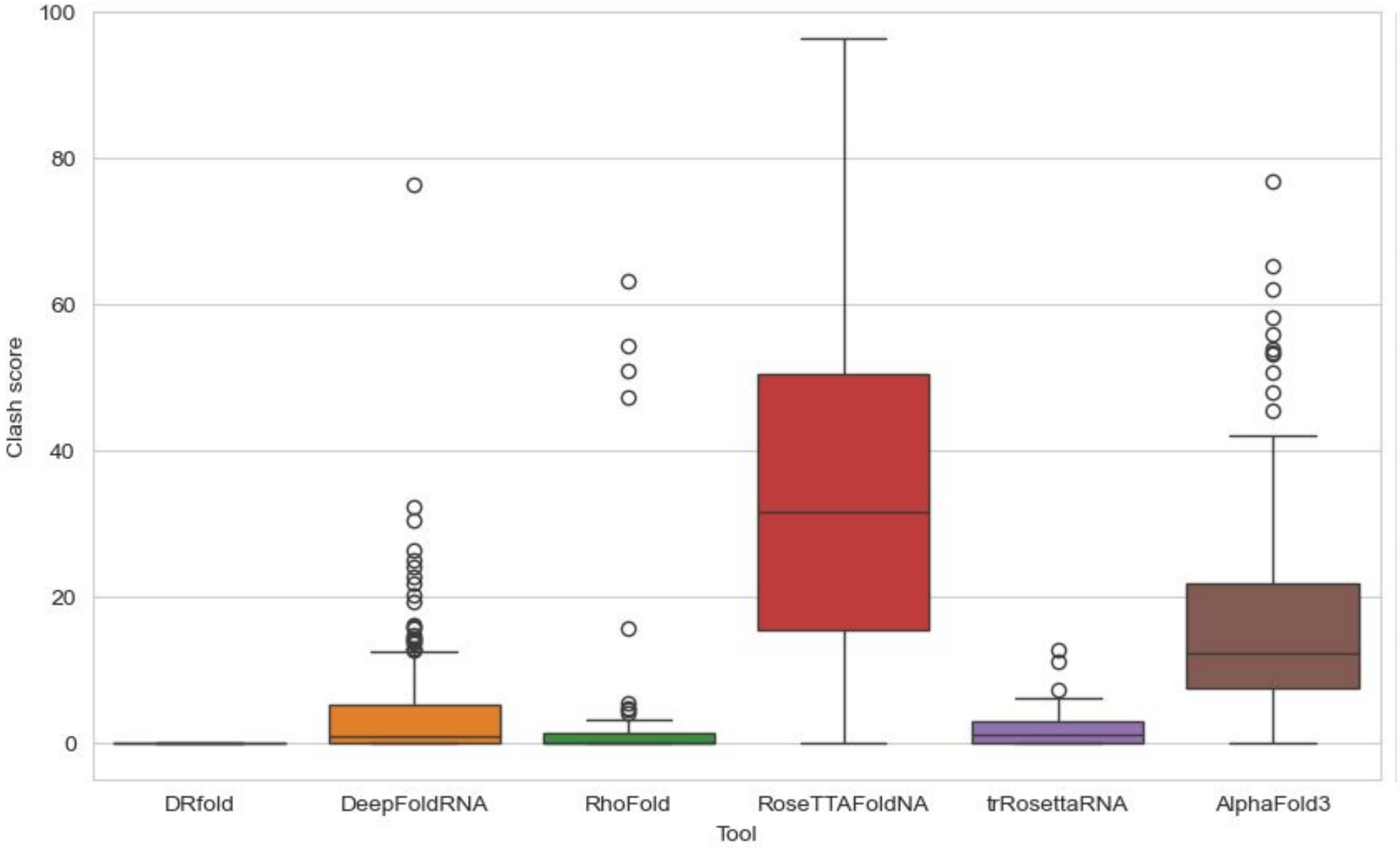
Dataset 3: Clash score



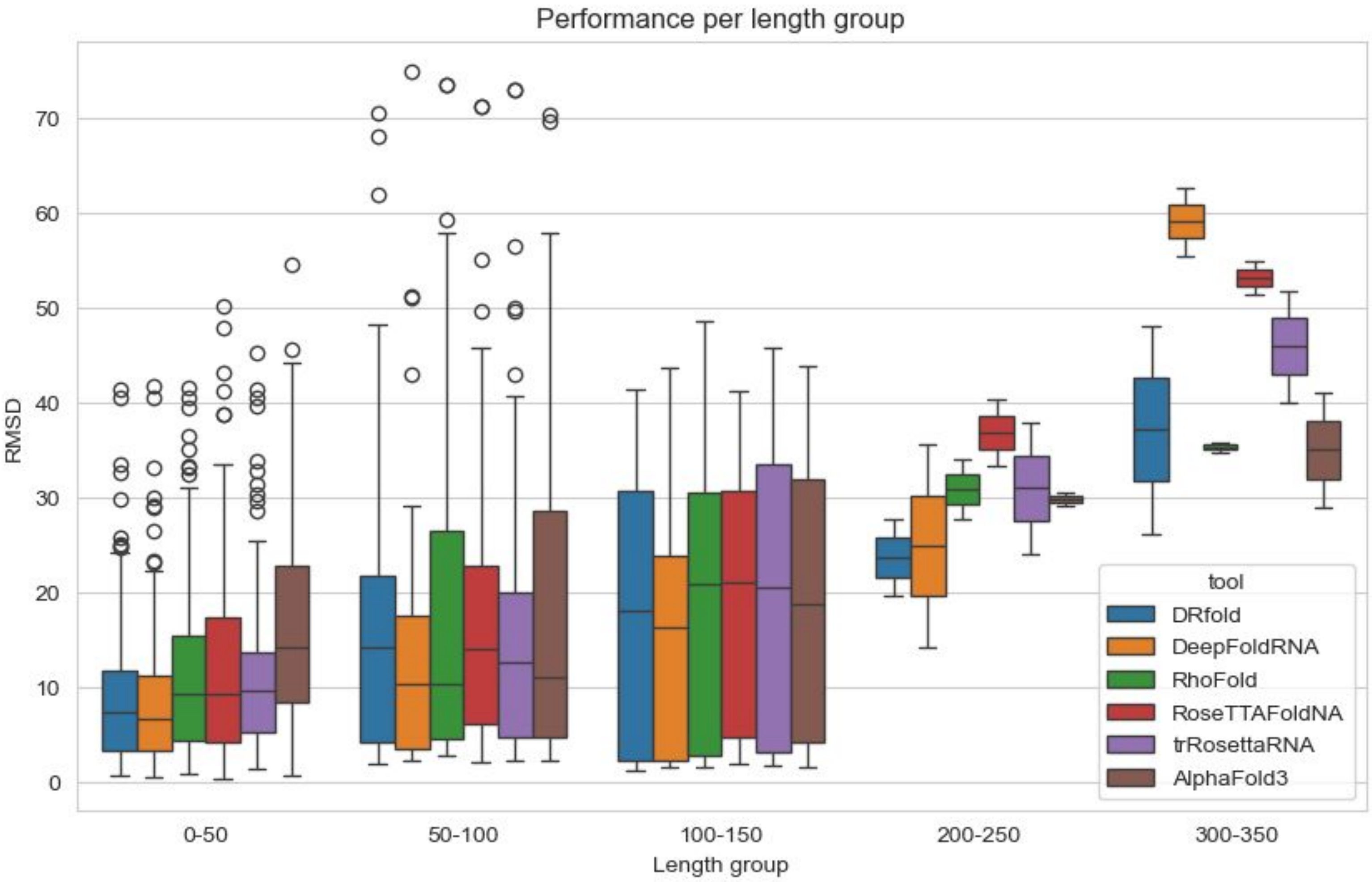
CREATING GROWTH, ENHANCING LIVES

CREATING GROWTH, ENHANCING LIVES

Dataset 3: Clash score



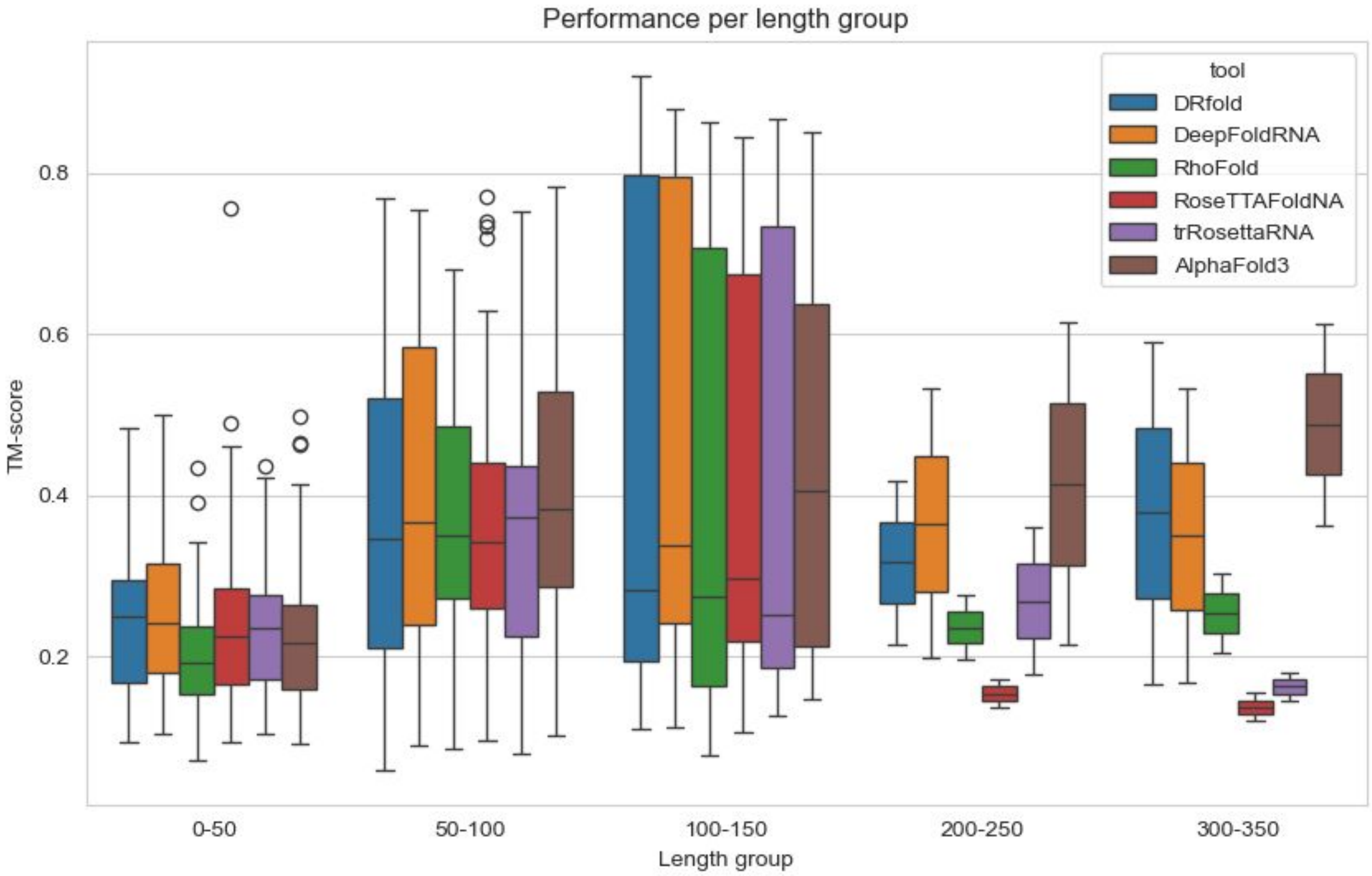
Performance (RMSD) per length group (Dataset 3)



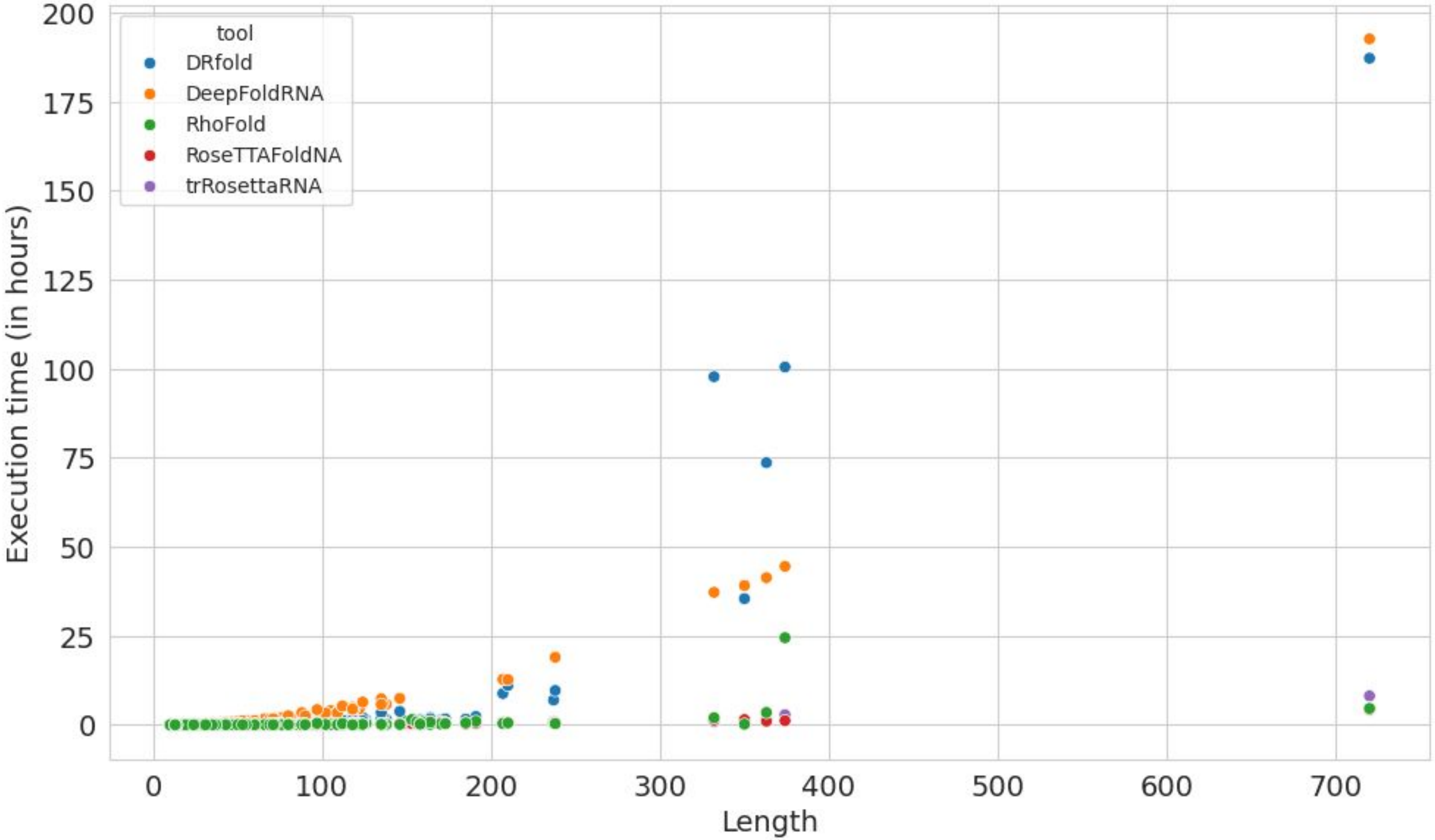
CREATING GROWTH, ENHANCING LIVES

Computational Approaches to RNA Structure and Function, Benasque 2024

Performance (TM-score) per length group (Dataset 3)

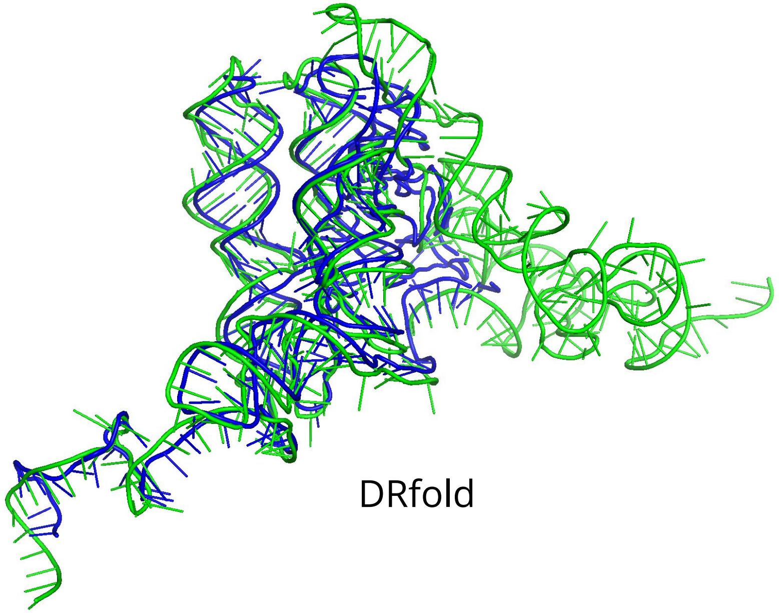


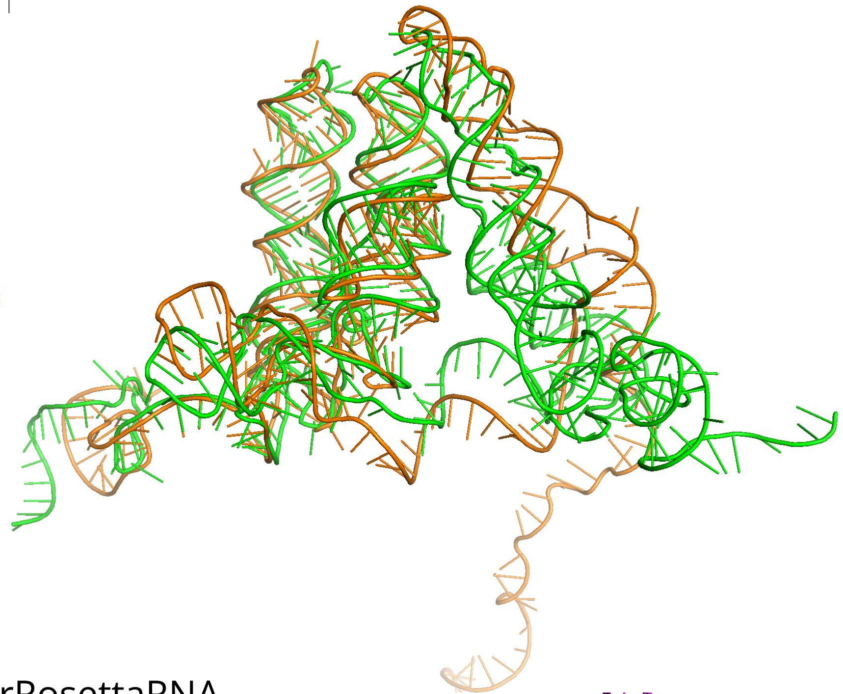
Execution time (for all three datasets)

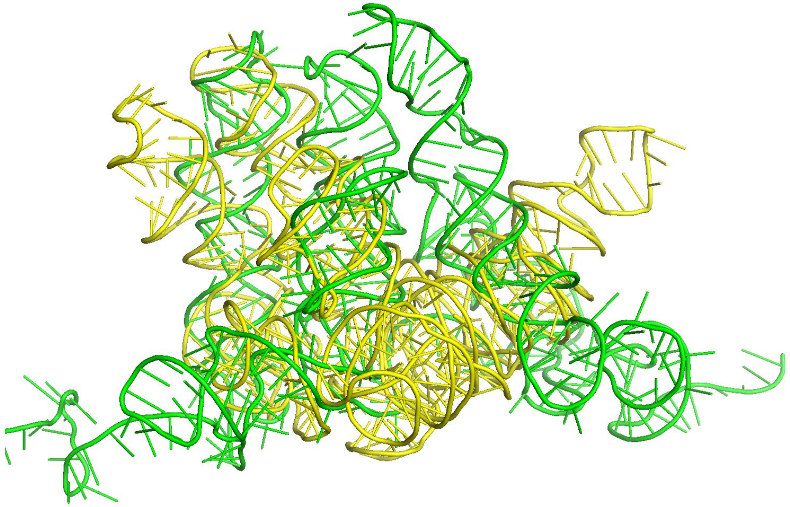


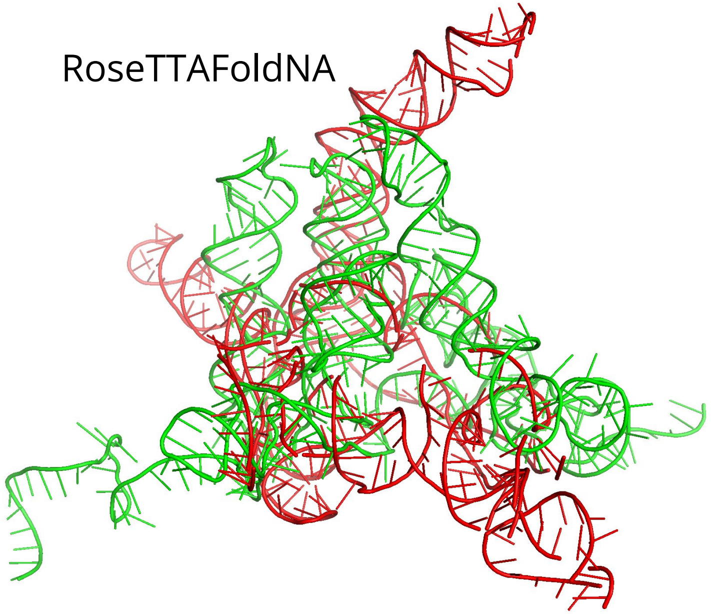
Examples

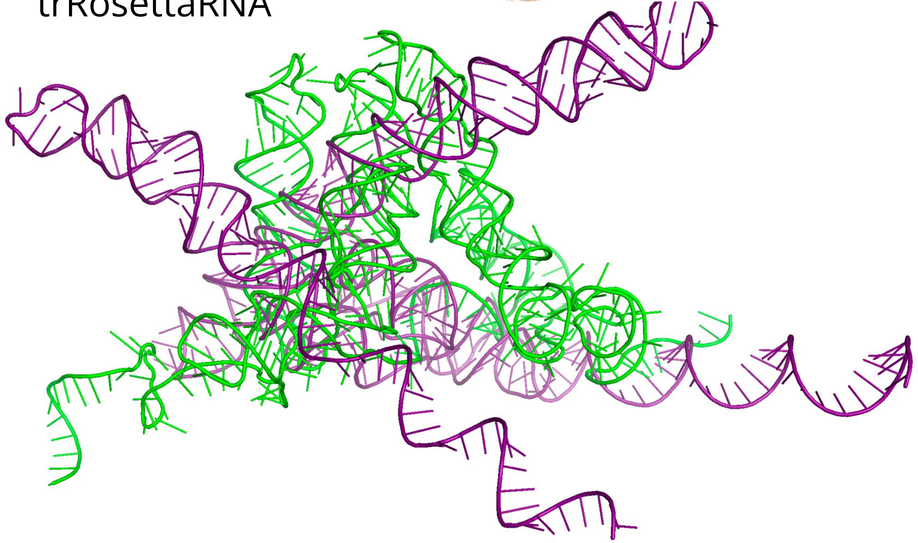
Example: 8A22_A8

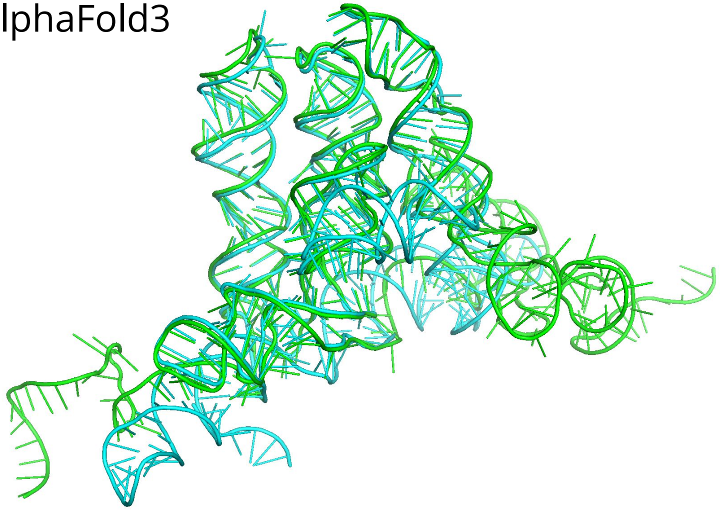

DRfold


DeepFoldRNA

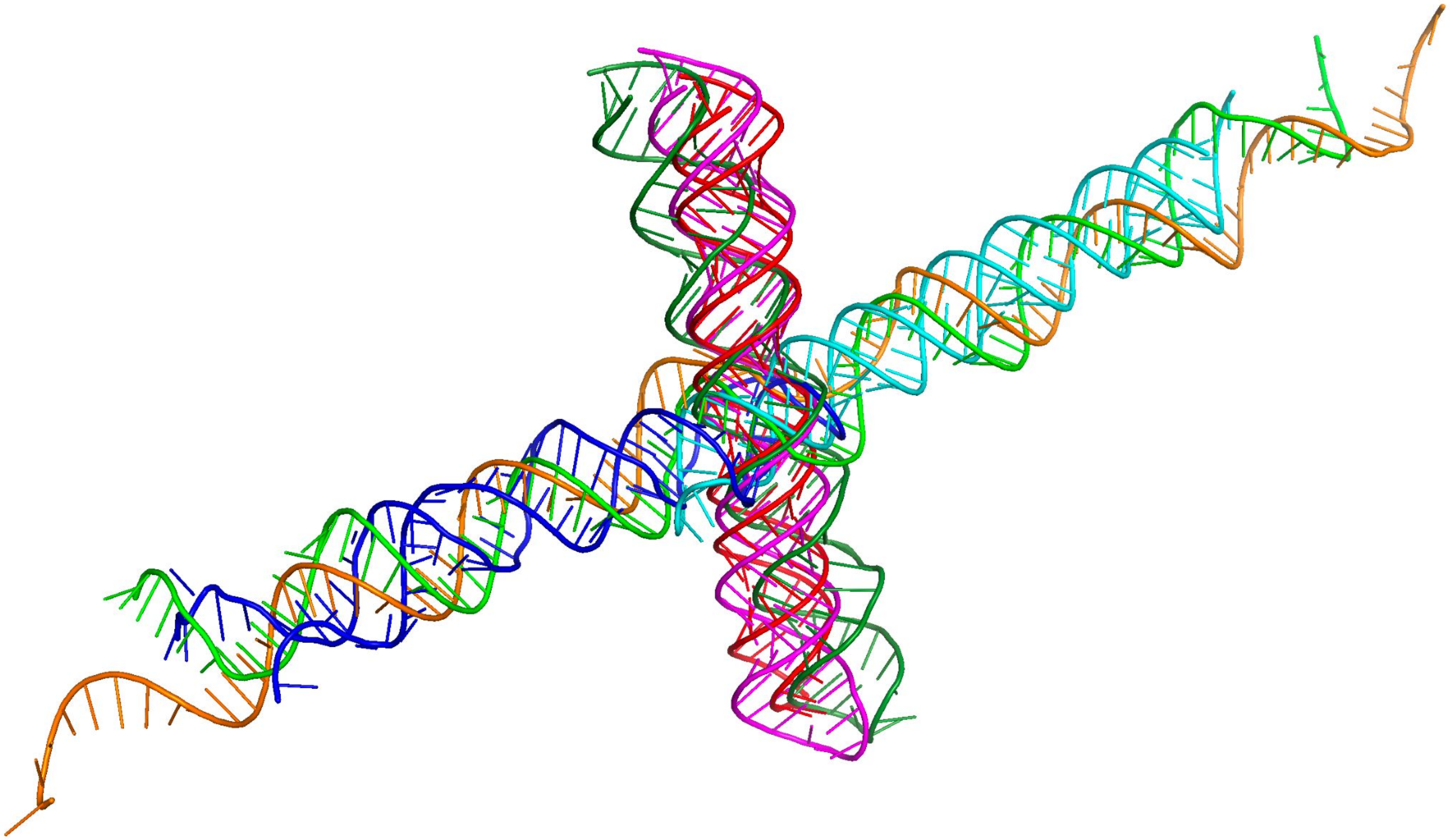

RhoFold

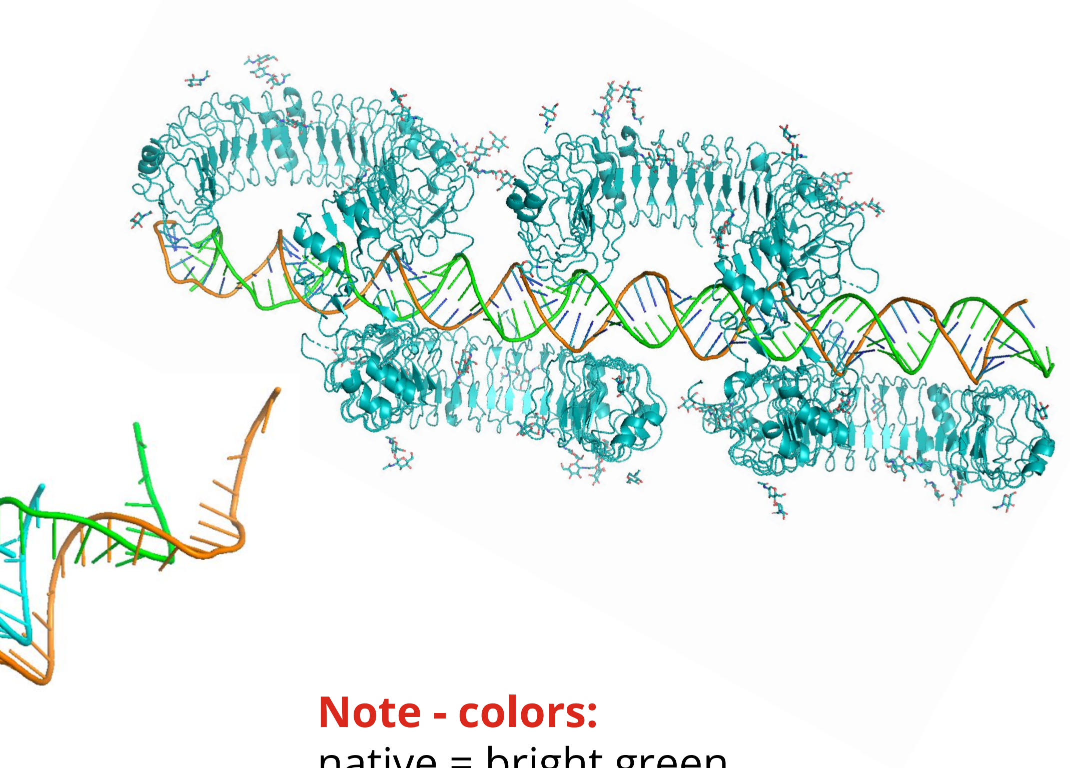

RoseTTAFoldNA


trRosettaRNA


AlphaFold3

Example: 7WM4_B



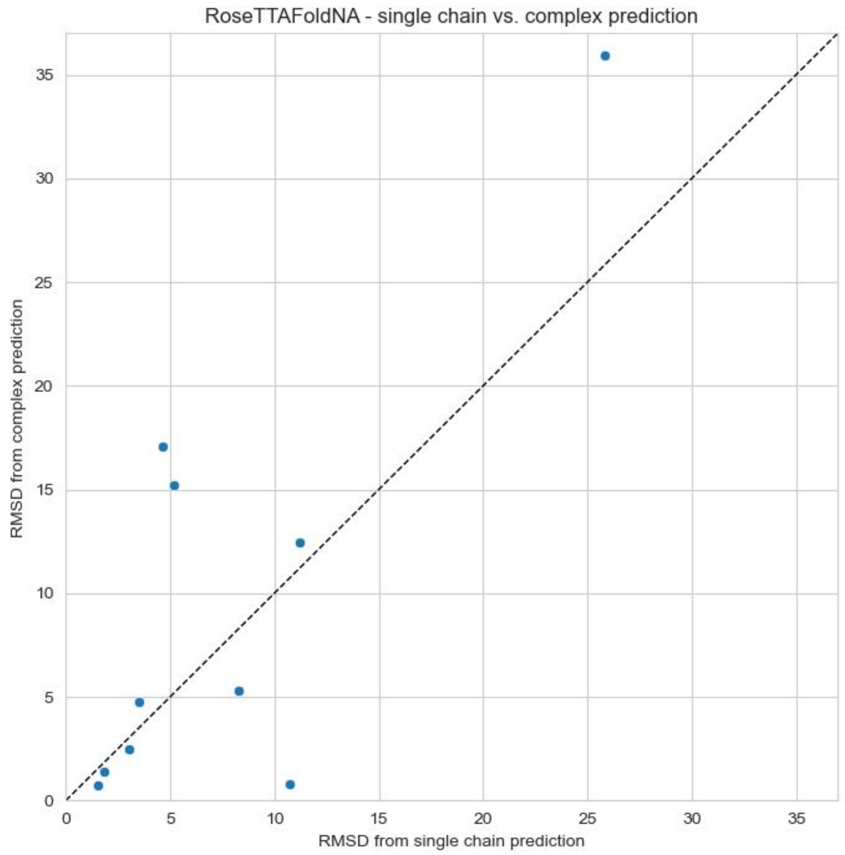


Note - colors:
native = bright green
DRfold = blue
DeepFoldRNA = orange
RhoFold = dark green
RF2NA = red
trRosettaRNA = purple
AF3 = cyan

RNA chain as a single chain vs. as part of the complex

- RoseTTAFoldNA has an option of providing protein sequences in which it still predicts only RNA chain(s), but in couple of examples where we tried this the resulting predictions were almost the same

- RoseTTAFoldNA tried out on 10 complexes - better only in 50% of cases, not by much



CREATING GROWTH, ENHANCING LIVES

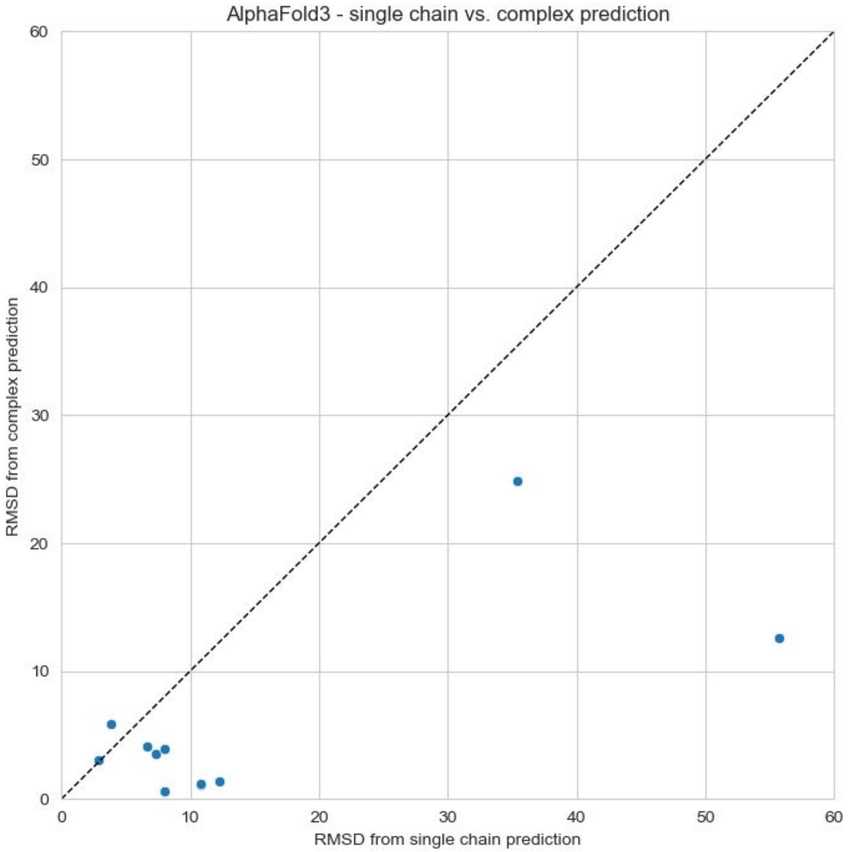
RNA chain as a single chain vs. as part of the complex

- RoseTTAFoldNA has an option of providing protein sequences in which it still predicts only RNA chain(s), but in couple of examples where we tried this the resulting predictions were almost the same

- RoseTTAFoldNA tried out on 10 complexes -
  better only in 50% of cases, not by much

- AlphaFold3 for these 10 complexes -
  better for 9/11 chains, for 7YCH_B
  91% lower RMSD in complex prediction

- Not sure how trustworthy AlphaFold3 results are
  (it could be trained on these examples)



Conclusion

- Q: Which of the tools performs best across different datasets and evaluation metrics?
  A: No unique tool, depends on the use case.

- Q: Do certain design choices and methodologies impact accuracy?
  A: Unable to answer, no direct connections.

- Q: How well these tools generalize to RNA sequences different from those used in their training?
  A: The best tool on generalization dataset (Dataset 3) is DeepFoldRNA across most metrics.

- Q: Can we choose the best predicted structure using ARES or Rosetta score?
  A: According to our tests, no.

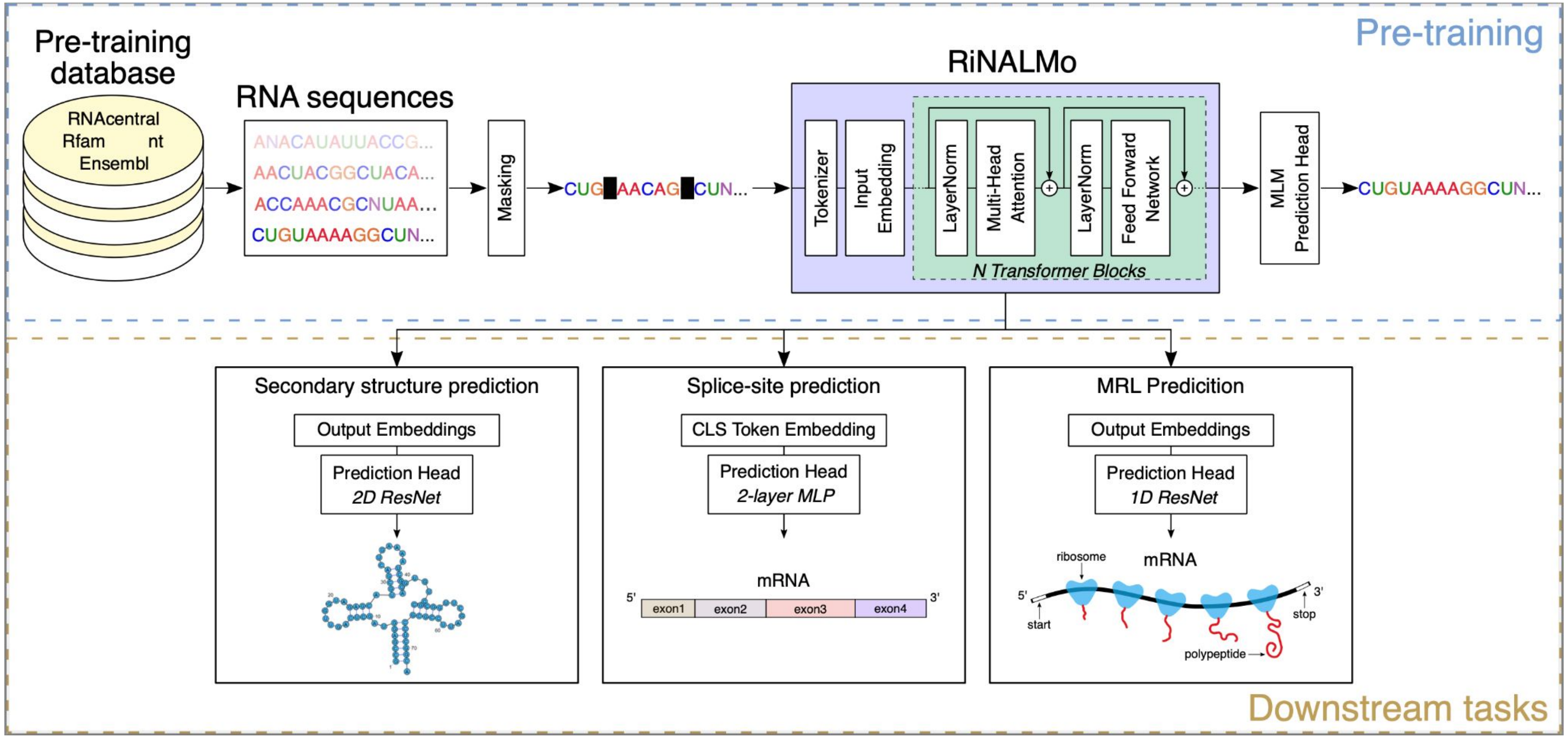
- Q: How much does having context help in structure prediction?
  A: In case of AlphaFold3, currently it seems a lot, but in case of RoseTTAFoldNA, not that much.

RiNALMo: RNA language model

Rafael Josip Penić

Tin Vlašić

- **Motivation:** success of protein language models
- BERT-style language model pretrained using masked language modeling
- 36M unannotated ncRNA sequences, 650M parameters

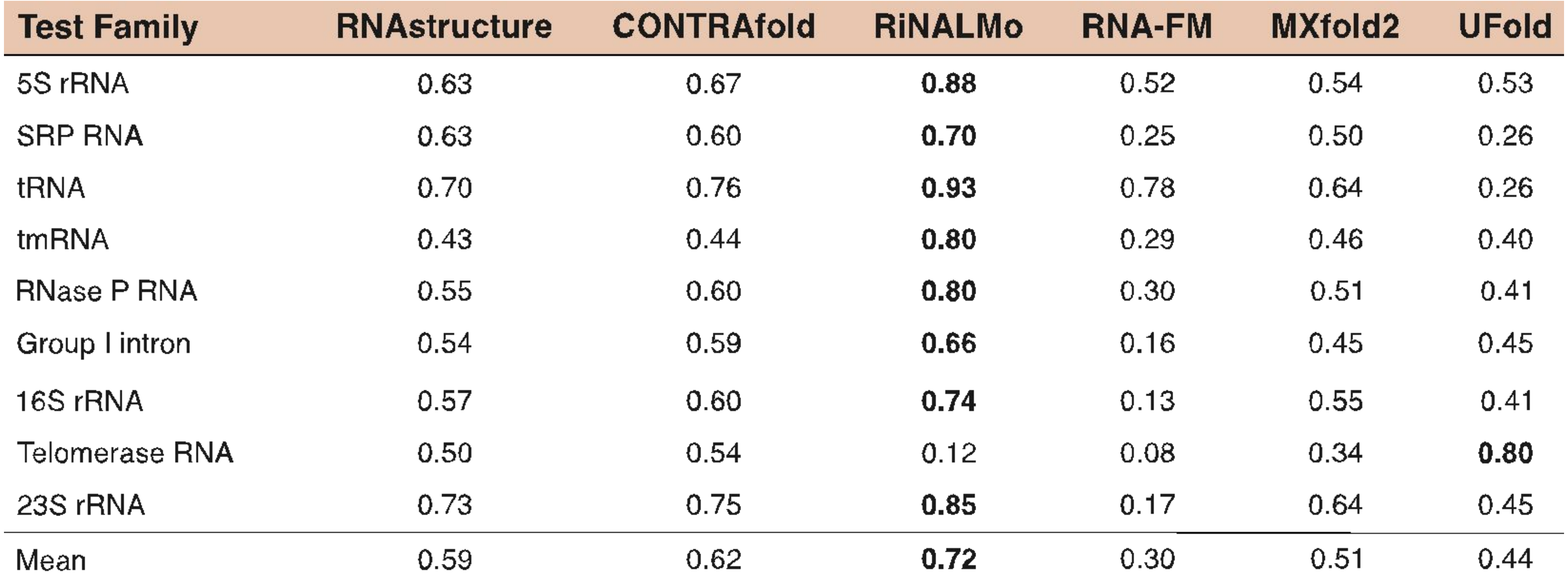


https://github.com/lbcb-sci/RiNALMo

Secondary structure prediction

- Finetuned RiNALMo embeddings + small CNN prediction head
- RiNALMo helps to generalize well on RNA families not seen in the training dataset unlike other deep learning methods

*Dataset obtained from [Szikszai et al., 2022]



Test Family	RNAstructure	CONTRAfold	RiNALMo	RNA-FM	MXfold2	UFold
5S rRNA	0.63	0.67	**0.88**	0.52	0.54	0.53
SRP RNA	0.63	0.60	**0.70**	0.25	0.50	0.26
tRNA	0.70	0.76	**0.93**	0.78	0.64	0.26
tmRNA	0.43	0.44	**0.80**	0.29	0.46	0.40
RNase P RNA	0.55	0.60	**0.80**	0.30	0.51	0.41
Group I intron	0.54	0.59	**0.66**	0.16	0.45	0.45
16S rRNA	0.57	0.60	**0.74**	0.13	0.55	0.41
Telomerase RNA	0.50	0.54	0.12	0.08	0.34	**0.80**
23S rRNA	0.73	0.75	**0.85**	0.17	0.64	0.45
Mean	0.59	0.62	**0.72**	0.30	0.51	0.44

CREATING GROWTH, ENHANCING LIVES

Future directions

- Currently pretraining a 1.6B parameter RiNALMo on ~100M RNA sequences
- Multimodal pretraining including chemical probing data
- 3D structure prediction model leveraging RiNALMo sequence embeddings

CREATING GROWTH, ENHANCING LIVES

Collaborators:

WAN Yue, GIS

Roland G. HUBER, BII

Agency for Science, Technology and Research

SINGAPORE

Mile SIKIC,
Group Leader

THANK YOU

www.a-star.edu.sg

Šikić Lab

AI IN GENOMICS

Contacts:

✉ martinovici@gis.a-star.edu.sg

✉ miles@gis.a-star.edu.sg

https://sikic-lab.github.io