RNA-Puzzles in the AI era

what we learned, what should change

Zhichao (Chichau) Miao
Guangzhou National Laboratory

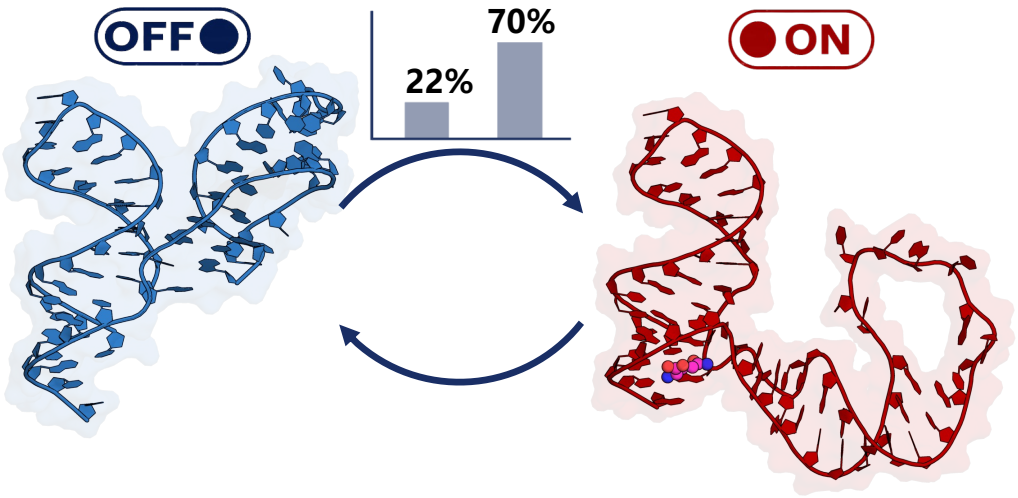
Eric Westhof
WIUCAS, Wenzhou

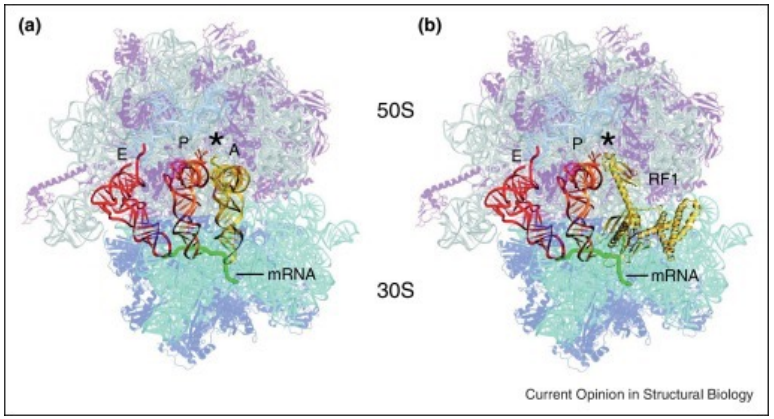
RNA-Puzzles should remain the community backbone, but the AI era requires clearer comparison classes, explicit leakage control, and more mechanism-aware reporting.

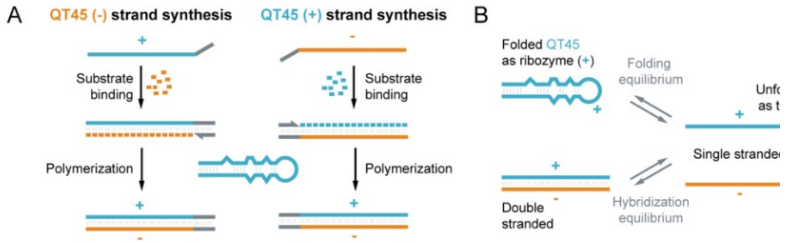
Fast background

Why RNA 3D prediction matters

Function still depends on 3D architecture when experiment is slow, sparse, or ambiguous.







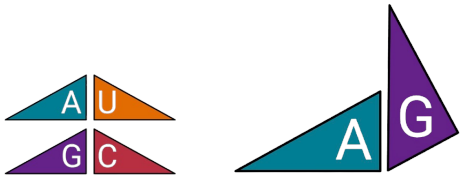
Science 2026;391(6789)1022-1028

Prediction is often the fastest route from sequence to a mechanistic question.

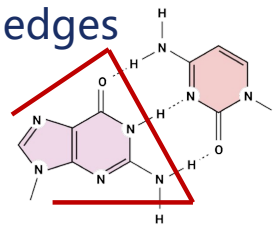
Why RNA Prediction is still hard

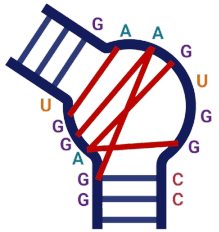
RNA is not a single-score coordinate-fitting problem.

Non-Watson-Crick contacts

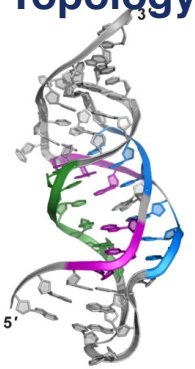


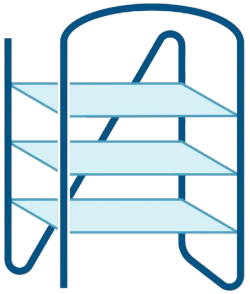
three edges

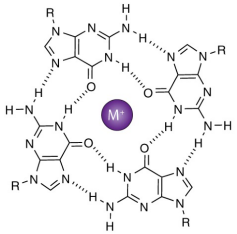




Topology and long-range packing







Ions, ligands, crowded interfaces

Biological RNAs are stabilized by context that simple scoring functions still miss.

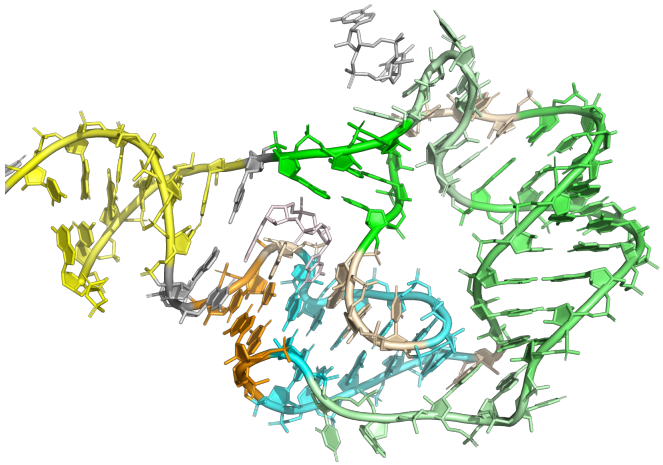
This is why a reasonable-looking RNA model can still be biologically misleading.

Fast background

2D is not always enough

3D structure

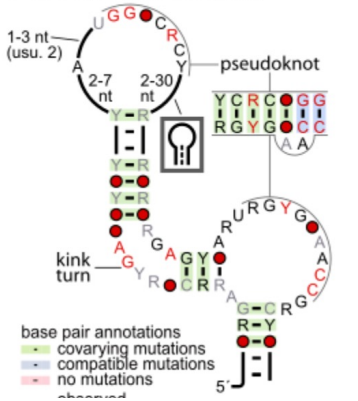
Real 3D structure (major helices colored) 3Q3Z A



Helix colors P1 P2 P3 P4

The native architecture is stabilized by interactions beyond the standard nested MFE model.

2D prediction



Real 2D

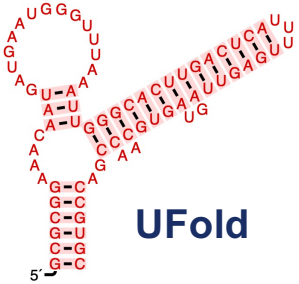
Best exact F1 = 0.56

BPfold recovers only 14 / 27 truth pairs

RNAfold / LinearFold / IPknot / UFold each recover only 11 / 27.

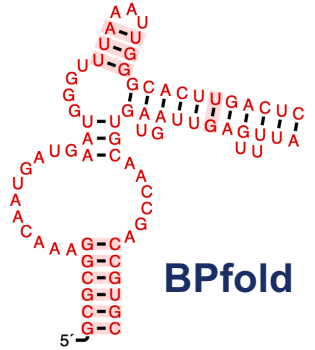
MetaFold-RNA recovers 10 / 27.

RNAfold

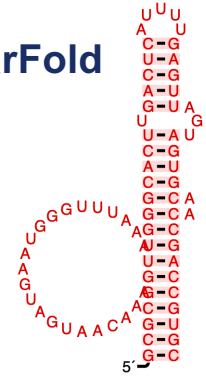


UFold

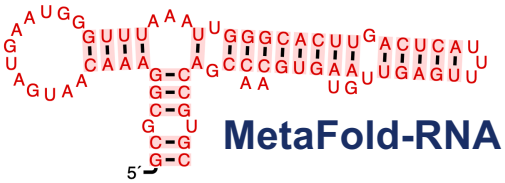
IPknot



BPfold



LinearFold



MetaFold-RNA

Bottom up prediction

Sequence based alignment

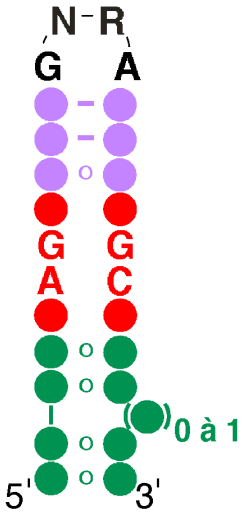
```
Ref.  GCUGGAGGGAAGCAAUUUAGCACGGC
Seq1  GCUUCAGUAGAGCGAUCUAGCGGAAGU
Seq2  AGGACAGUGGAGUAAUCCAGCAUCGCU
Seq3  CUUCGAGGUUCGAAAGAGUGCAGAGG
```

Structure based alignment

```
      ((((....(((....)))....)).))
Ref.  GCUGGAGGGAAGCAAUUUAGCACG-GC
Seq1  GCUUCAGUAGAGCGAUCUAGCGGAAGU
Seq2  AGGACAGUGGAGUAAUCCAGCAUCGCU
Seq3  CUUCGAGGUUCGAAAGAGUGCAGA-GG
```

Ref. struct

consensus



We need to understand the base pairs to achieve a good modelling.

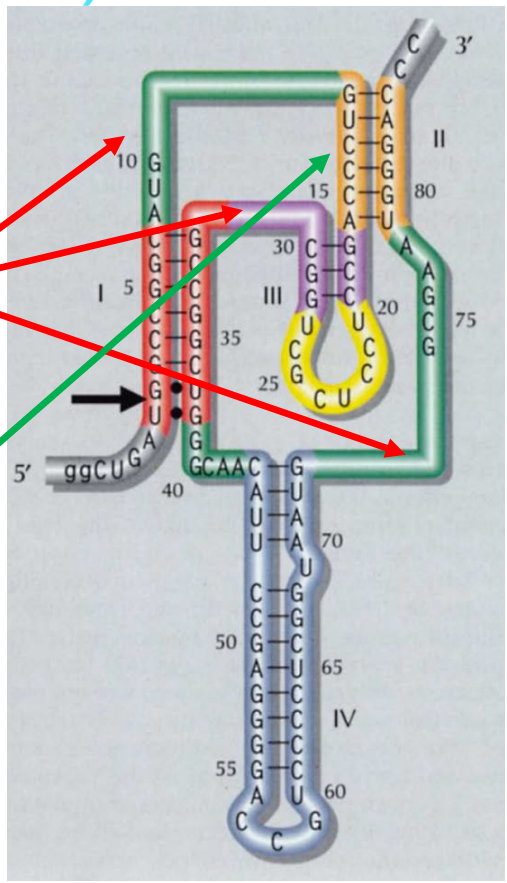
Success examples: HDV

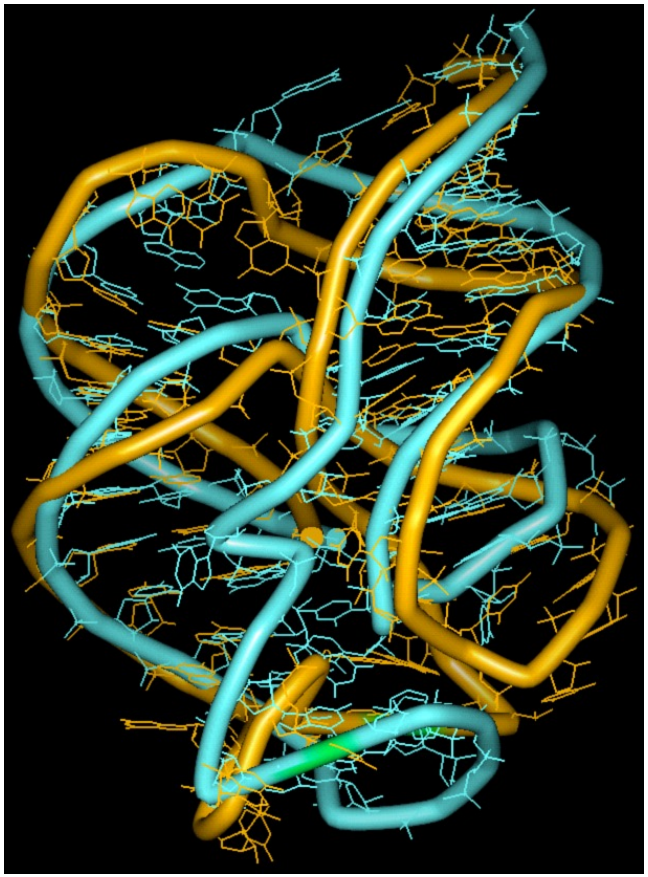
Ferré d'Amaré et al. Nature 395, 567 (1998)
Tanner et al. Current Biol. 4, 488 (1994)

X-ray vs. Model (r.m.s. 8.0 Å)

Strong linkage constraints :
short single-stranded segments

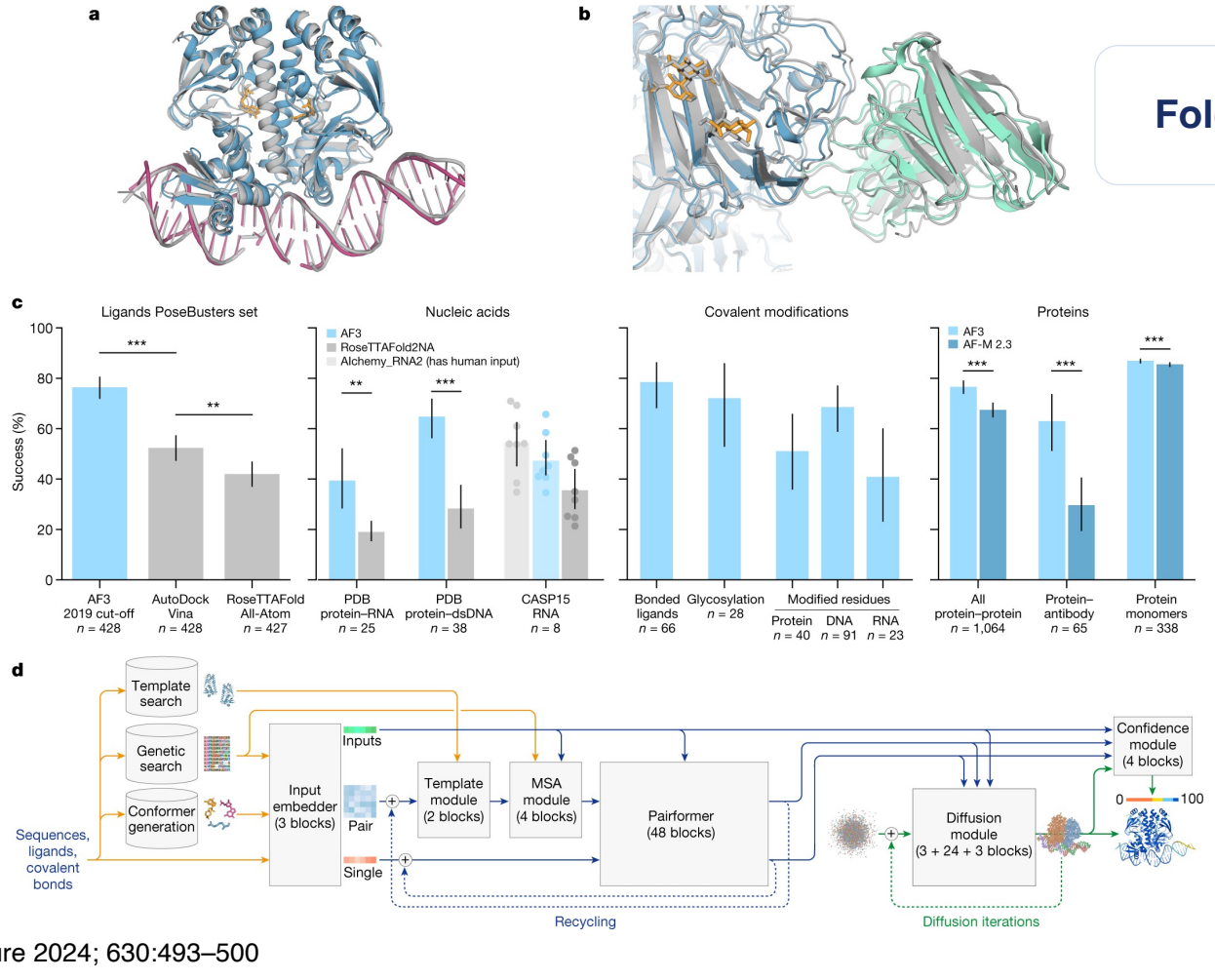
Strong linkage constraint :
Presence of a pseudoknot





Fast background

Data driven prediction: AF3

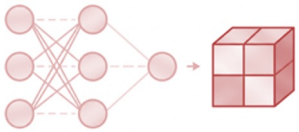


Folding rules can be learnt from data.

BUT available data is very limited compared to protein

Nature 2024; 630:493–500

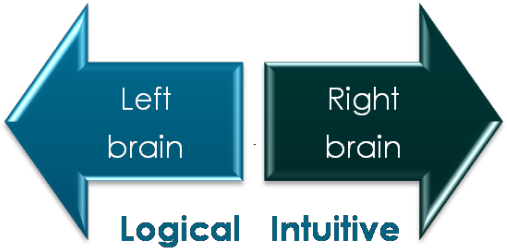
Reductionism vs. Holism



AI reasoning

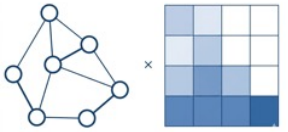
Reductionism

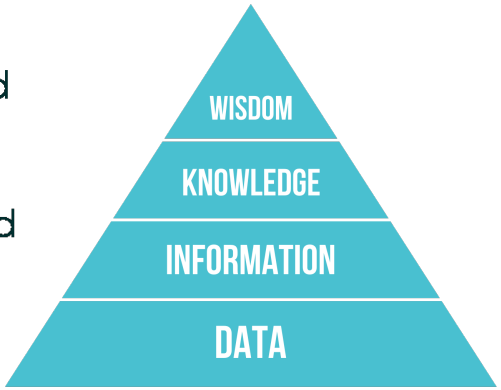
Parts
Structured
Rational
Prove it!
Hierarchy
Categories
Seperate
Future/past
Precise
Static
Male
Nosy
Seperate notes
Mechanic



Holism

Whole
Creative
Intuitive
Open mind
Synergy
Individuals
Connected
Now
Chaotic
Dynamic
Female
Selfcorrecting
Harmony
Organic



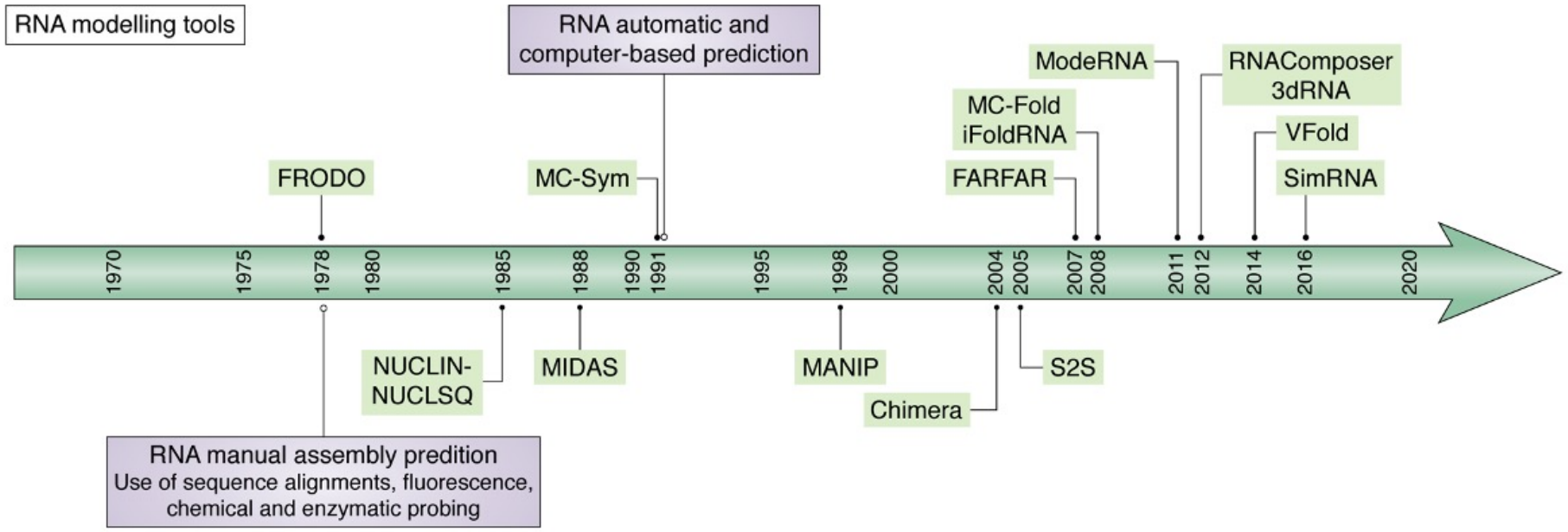


Today's prediction landscape

Template-guided

Assembly / physics

End-to-end AI

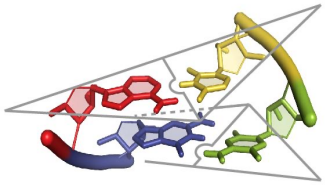


These are not interchangeable error models: they fail on different targets, for different reasons.

J. Biol. Chem. 2021;296:100555

Framework

What RNA-Puzzles is



Benchmarks shape incentives.

Prediction method → Blind target → Assessment → What the field learns

Blind target

Unpublished or protected structures are released as sequence-based prediction challenges.

Community submissions

Human groups and servers submit models under common timing and formatting rules.

Independent assessment

Released structures are used to compare methods with RNA-aware metrics.

RNA-Puzzles is infrastructure for the field: not only a contest, but a shared language for capability, failure, and progress.

That is exactly why its rules now need to become more explicit for the AI era.

Current workflow and timing

Structure recruitment → Preparation → Prediction → Submission → Assessment → Release

Automated servers

48 hours
Fully automatic; breadth and speed are the strength.

Human experts

3–4 weeks
Deeper intervention and broader reasoning are allowed.

Rule detail

Up to 5 models per method.

Interpretation

The same scoreboard mixes different effort regimes.

Consequence

Direct ranking becomes hard to explain cleanly.

What to fix

Keep one benchmark, but label comparison classes explicitly.

Framework

What RNA-Puzzles evaluates beyond RMSD

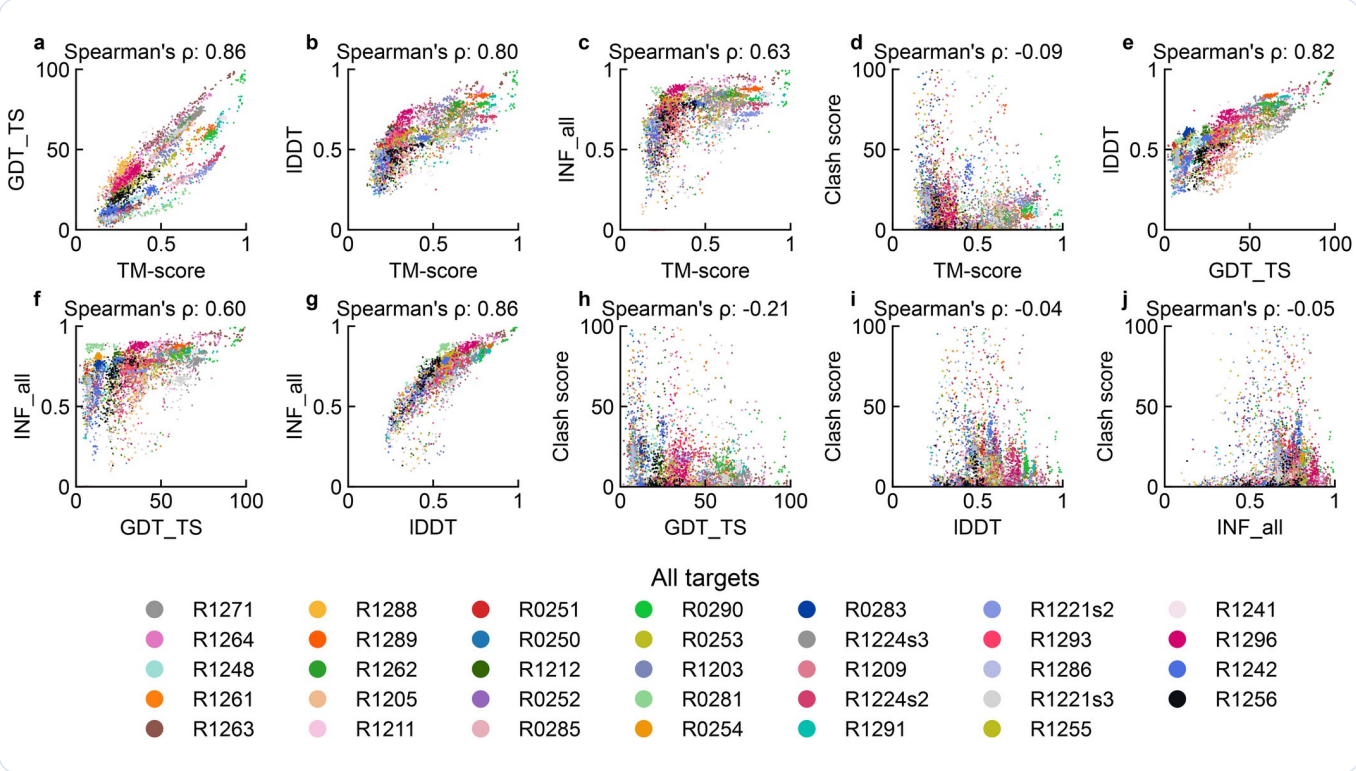
Overall shape is necessary, but not sufficient for RNA mechanism.

RMSD
Geometry
Global coordinate deviation.

INF
Network
Interaction fidelity and base-pair logic.

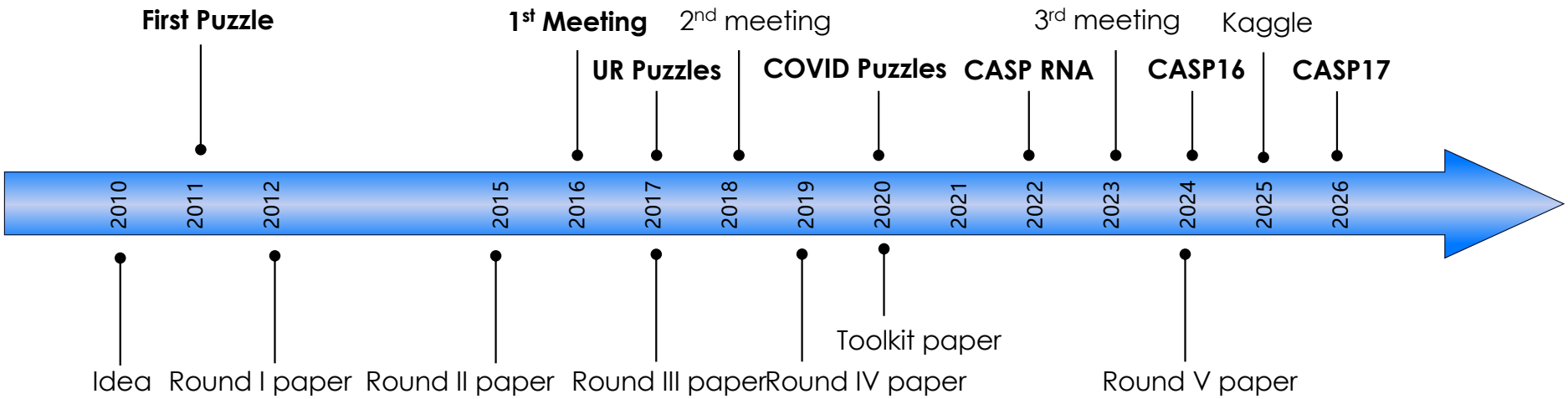
DI / local
Deformation
Local distortion that RMSD can hide.

P-value
Statistics
Whether a score is unusually good.



These metrics correlate only partially: similar fold-level scores can hide very different interaction logic.

Overview of RNA-Puzzles



Round I	**Cruz, et al.**	**RNA**	**2012**
Round II	**Miao, et al.**	**RNA**	**2015**
Round III	**Miao, et al.**	**RNA**	**2017**
Round IV	**Miao, et al.**	**RNA**	**2020**
CASP15 RNA	**Das, et al.**	**Proteins**	**2023**
Round V	**Bu, et al.**	**Nature Methods**	**2024**
CASP16 RNA	**Westhof, et al.**	**Proteins**	**2025**

Rules

What the current framework already gets right

Confidentiality

The benchmark protects structural biologists and makes participation possible.

Standardized submission

Common format and capped model count make assessment tractable.

Community continuity

Rounds accumulate into a long-term memory of what the field can and cannot do.

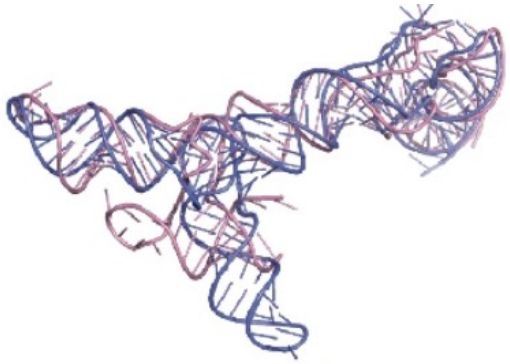
Rich assessment culture

RNA-specific metrics and manual inspection go beyond a single geometric score.

The point is to preserve these strengths while making AI-era interpretation cleaner.

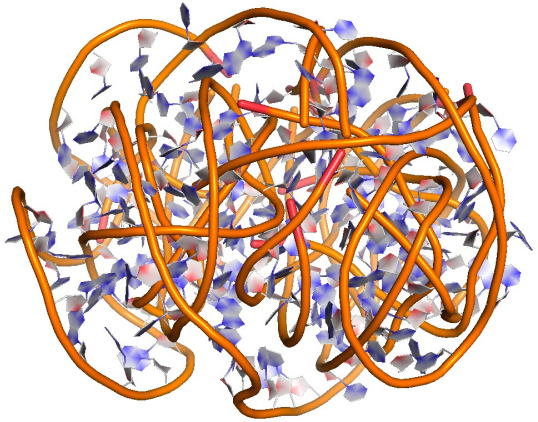
Rules

data leakage is now a real AI-era risk



RhoFold+
r.m.s.d.: 8.92 Å
TM score: 0.551
GDT-TS: 0.42
LDDT: 0.658

Learnt from Data



Never seen

Put into different categories

Template based

De novo prediction

Nature Methods 2024;21:2287–2298

AF3 is not much better

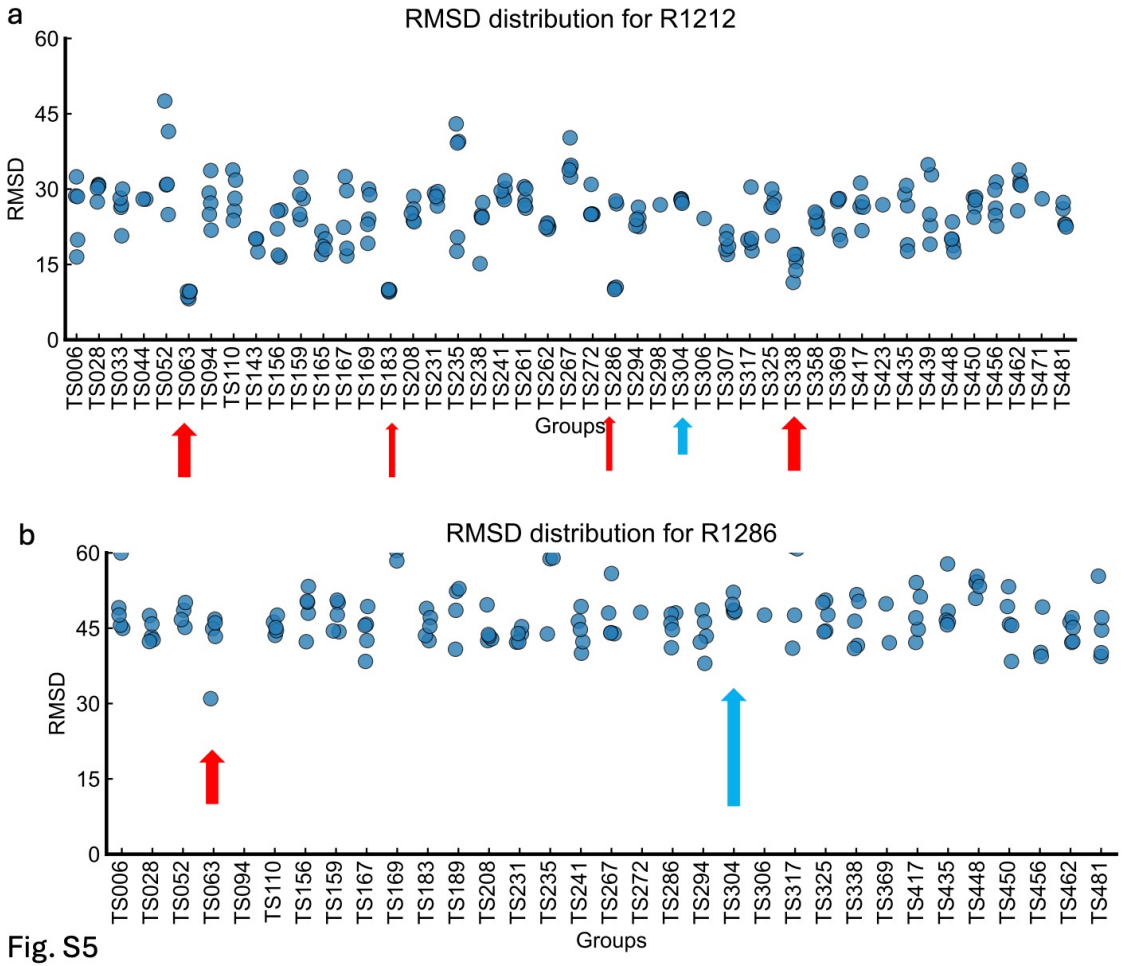
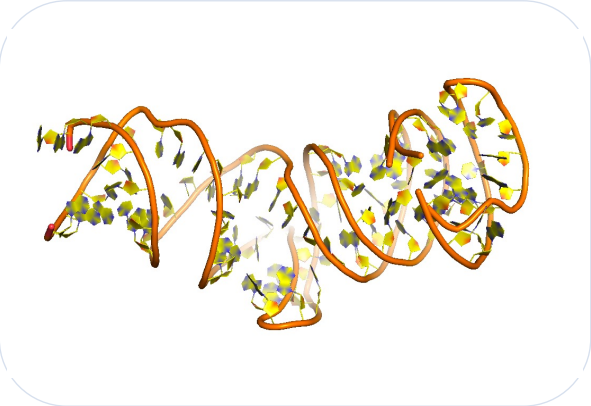


Fig. S5

Proteins: 2026;94(1):218–229

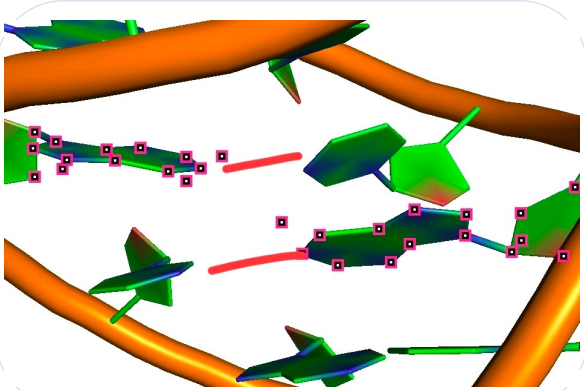
What recent programs still fail on

Topology, noncanonical interaction recovery, and physical realism remain persistent failure modes.



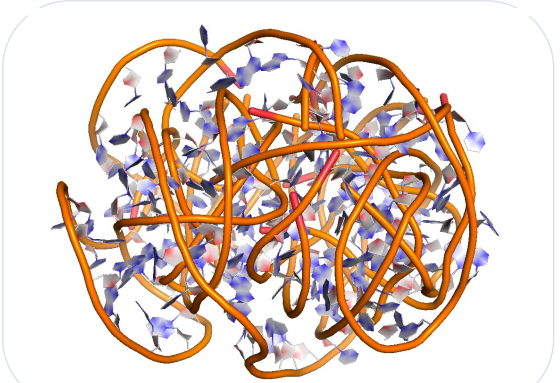
Global packing

A plausible backbone path can still conceal wrong long-range organization.



Noncanonical network

A reasonable fold does not guarantee the right contact logic.

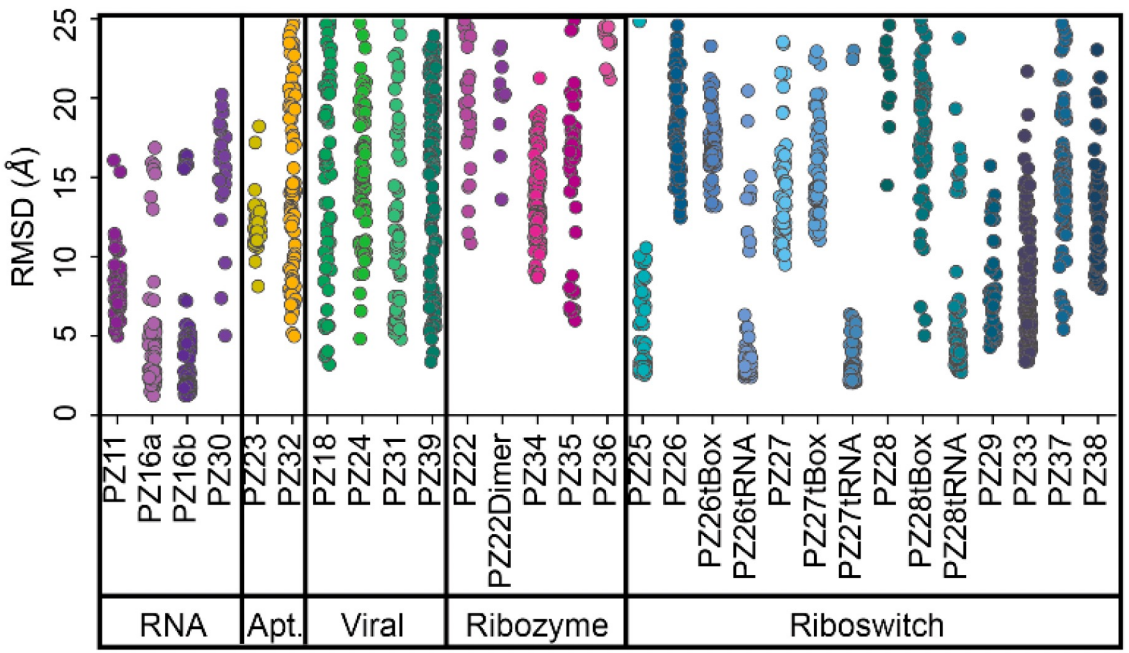


Physical realism

Knots, clashes, or chain artifacts still appear on difficult targets.

Recent puzzles snapshot

The value of recent puzzles is not who ranked first, but which target classes still expose blind spots.



Target mix still matters

Viral RNAs, ribozymes, riboswitches, and interfaces do not fail in the same way.

No universal winner

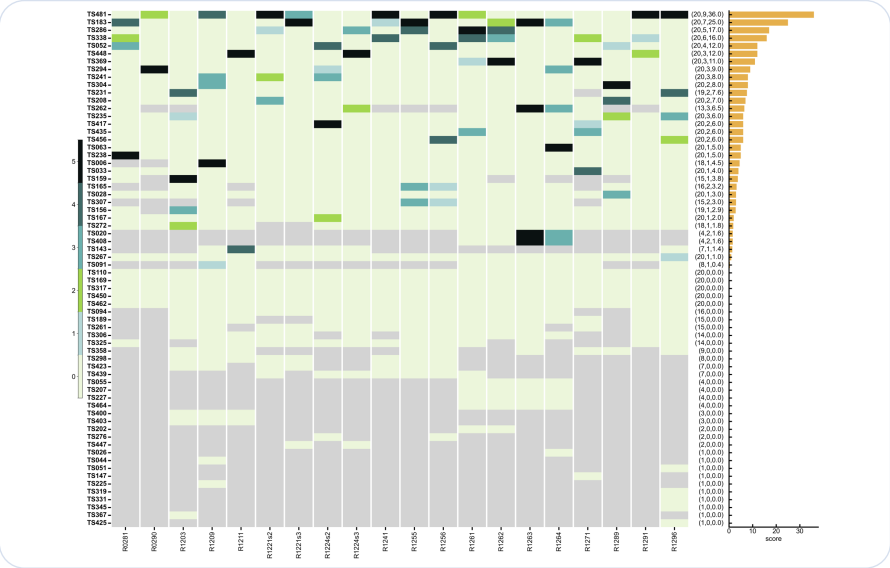
Success still depends strongly on puzzle type and available priors.

Blind spots are persistent

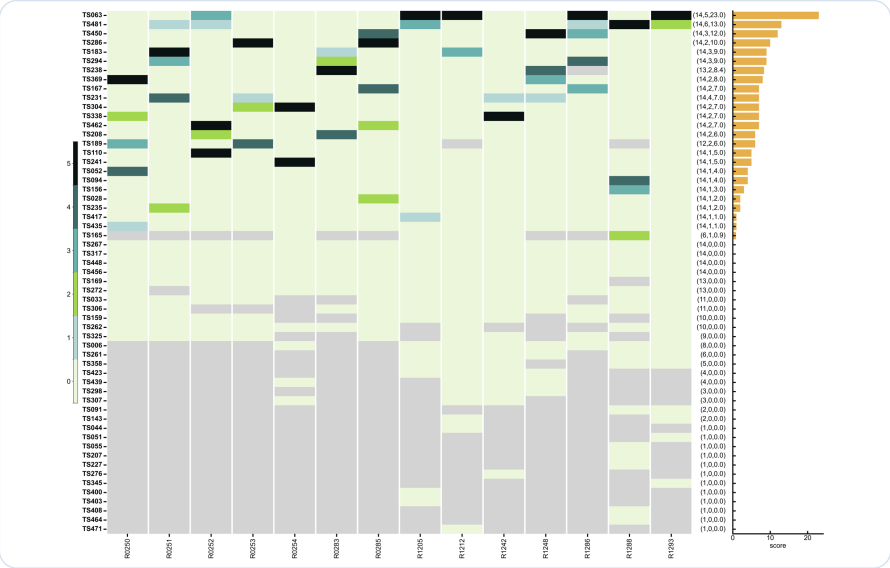
The hardest classes remain information-poor and mechanism-rich.

Template in recent puzzles

Outcome patterns already suggest that comparison classes should be separated.



With template



Without template

Servers improve breadth and speed; hard no-template cases still require explicit labeling of what kind of prediction was actually achieved.

Biological and modeling lessons distilled

1. Templates still matter

Much of the strongest practical performance still comes from reusing structural precedent.

2. Local motifs are easier than full mechanism

Contact fragments can improve while state and topology remain wrong.

3. Topology and state remain hard

The hardest problems are global arrangement, switching logic, and context dependence.

4. Evaluation must follow biology

RNA needs reporting that matches mechanism, not just coordinate proximity.

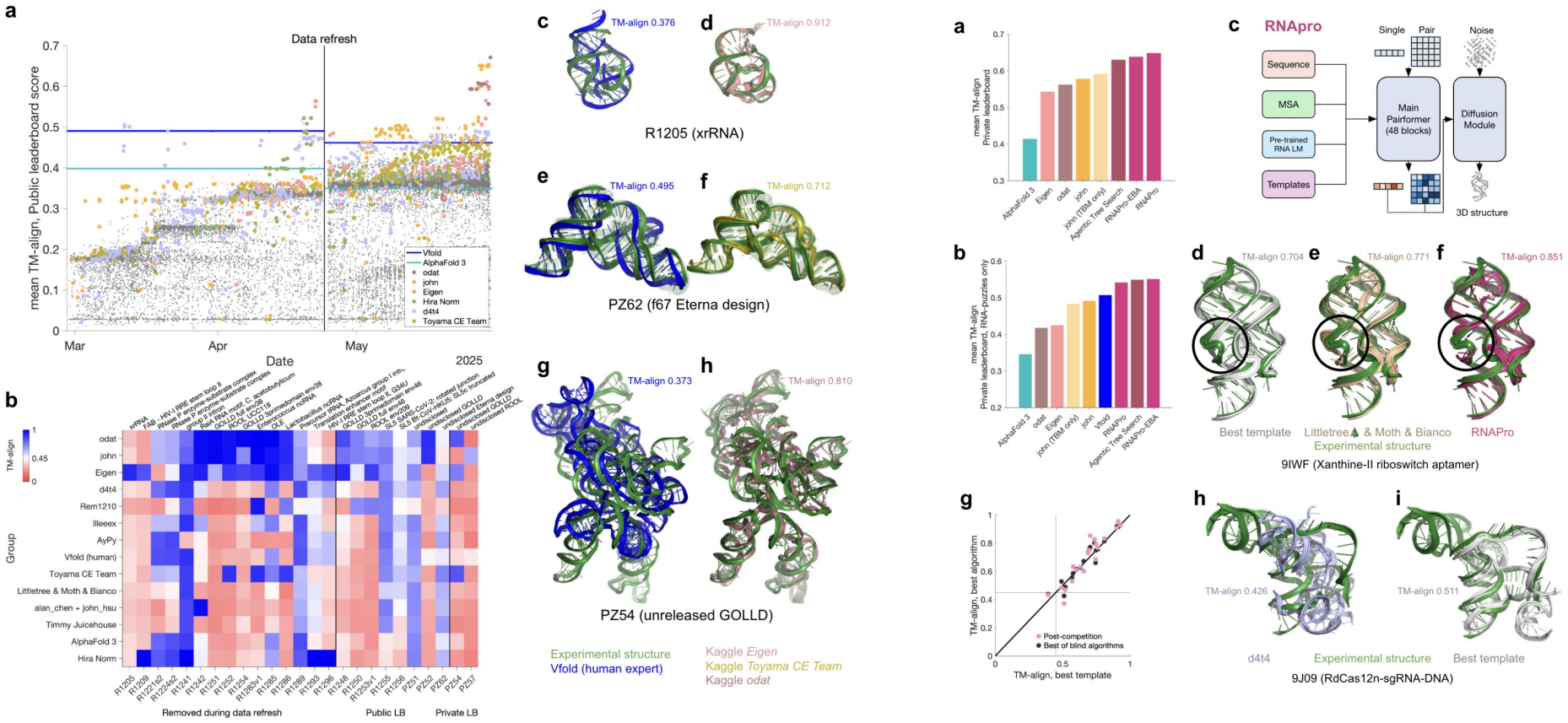
Recent puzzles still teach the same core lesson: useful RNA prediction means modeling interaction logic, topology, and state.

Recent prediction programs

Method	Venue	Year	Core idea
AlphaFold 3	Nature	2024	Diffusion model for joint protein-RNA-ligand-ion complexes.
trRosettaRNA2	Nat Mach Intell	2026	Pretrained 2D model + structure-aware attention; predicts conformers.
RNAbpFlow	Nat Methods	2026	Base-pair-conditioned SE(3) flow matching for RNA ensemble generation.
CaCoFold-R3D	Nat Methods	2025	Covariation-driven secondary + 3D motif prediction in one fold.
RhoFold+	Nat Methods	2024	RNA language model + MSA/data augmentation for end-to-end folding.
RoseTTAFoldNA	Nat Methods	2024	Single network for protein-DNA and protein-RNA complex prediction.
NuFold	Nat Commun	2025	End-to-end RNA folding with nucleobase-center all-atom representation.
DRfold	Nat Commun	2023	Local frames + geometric restraints + energy-guided assembly.
trRosettaRNA	Nat Commun	2023	Transformer-predicted geometries followed by restrained folding.
DeepFoldRNA	bioRxiv	2022	Deep geometric potentials + simulation-based atomic assembly.
OpenFold3	GitHub/Zenodo	2025	Open AF3-style multimolecule reproduction for NA complexes.
Boltz-2	bioRxiv	2025	General biomolecular model; improved structure plus affinity prediction.
Protenix-v1	bioRxiv/GitHub	2026	Open AF3-class predictor with template and RNA-MSA support.

Trend: geometric-restraint pipelines -> RNA language / flow models -> general multimolecule foundation models.

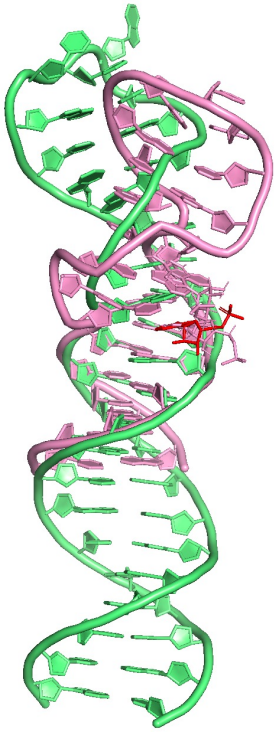
A lesson from recent Kaggle



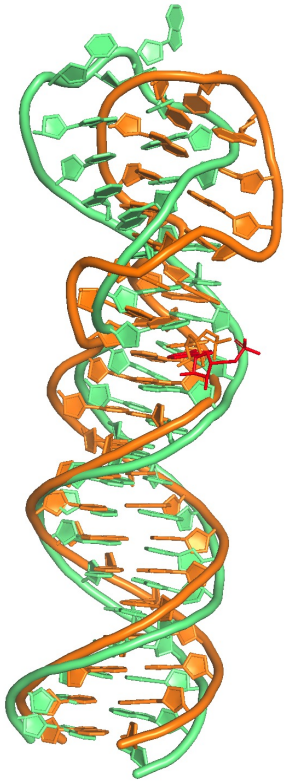
Template based method (RNApro) outperforms others

doi.org/10.64898/2025.12.30.696949

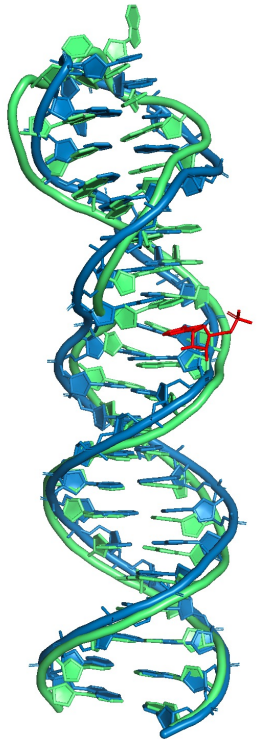
Sometimes it can be better than template



Solution
1AW3-NMR
RMSD = 4.192 Å (PYMOL)



Solution
NuFold
RMSD = 4.675 Å

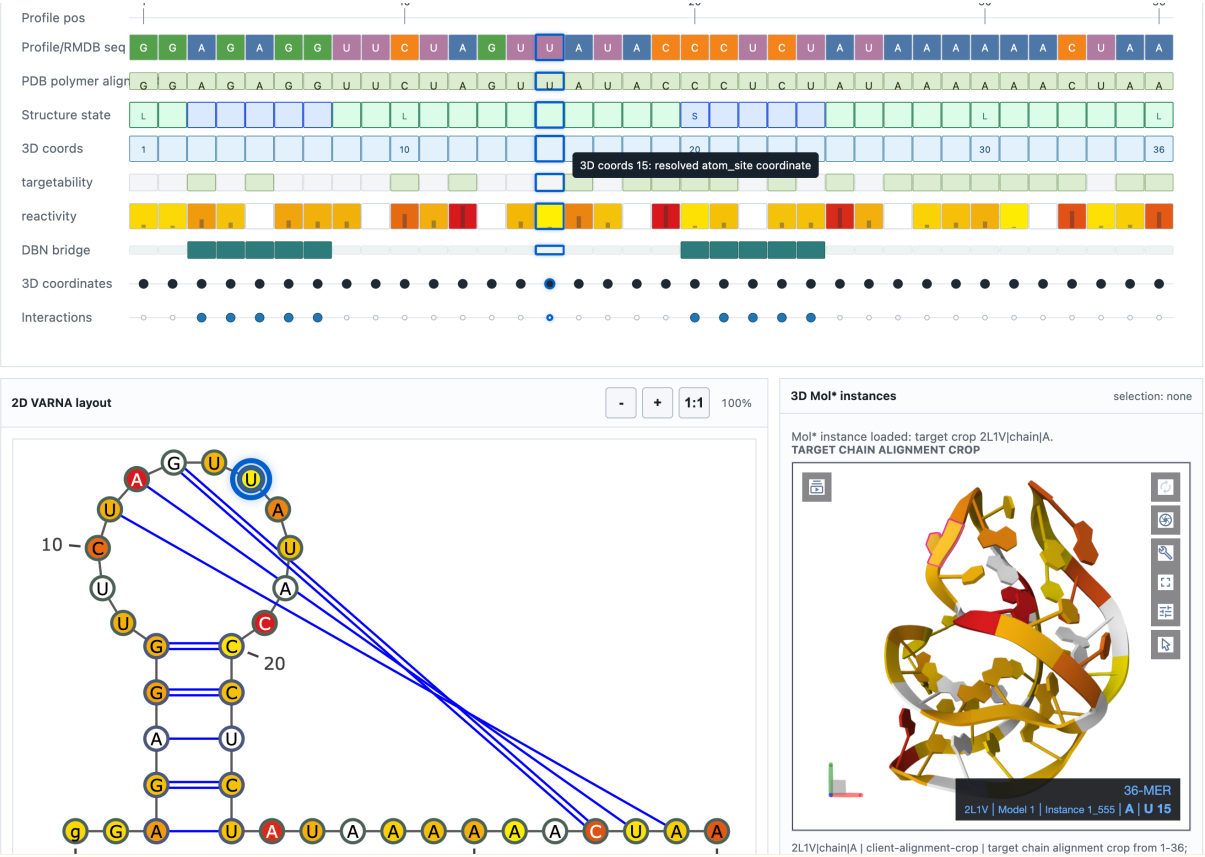


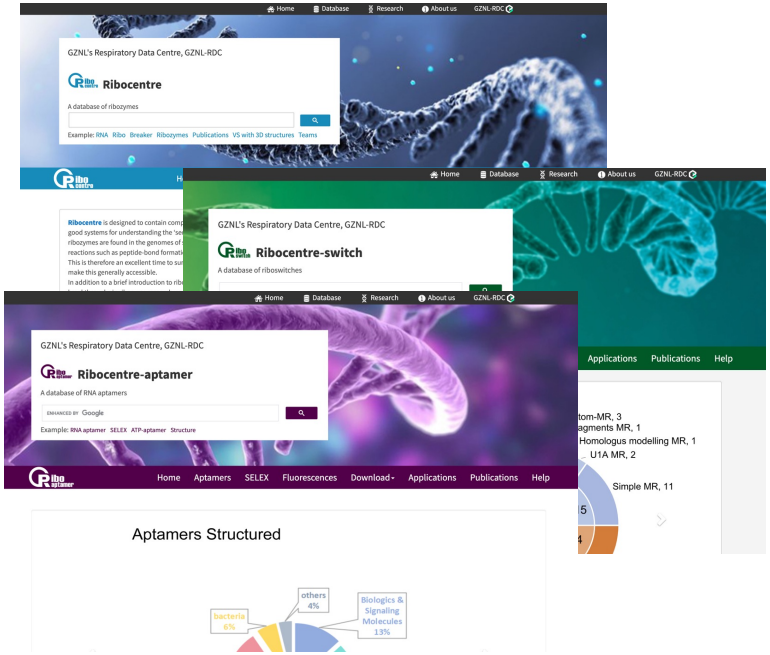
Solution
Boltz2
RMSD = 2.437 Å

direction

Better understand data

https://foldbridge.ribocentre.org/

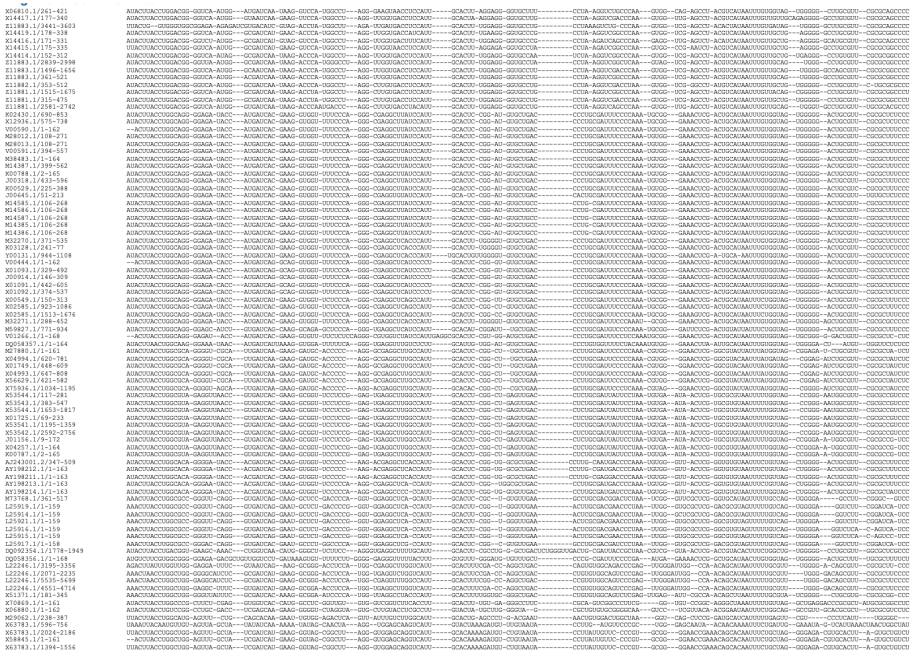




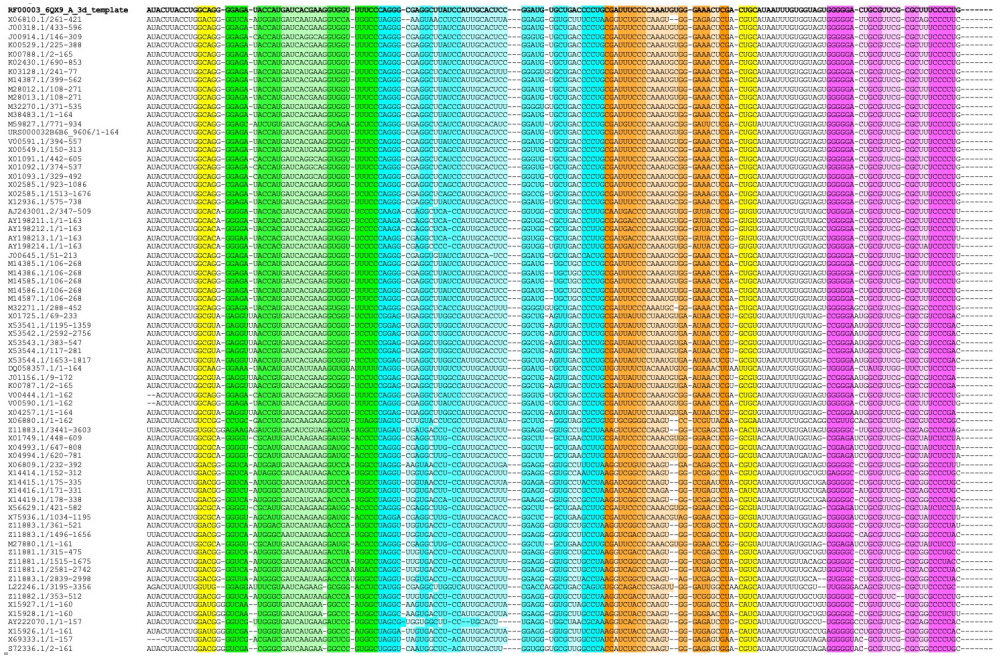
ribozyme：www.ribocentre.org
riboswitch：switch.ribocentre.org
Aptamer：aptamer.ribocentre.org
GlycoRNA：glycornadb.com

direction

Improve sequence alignment



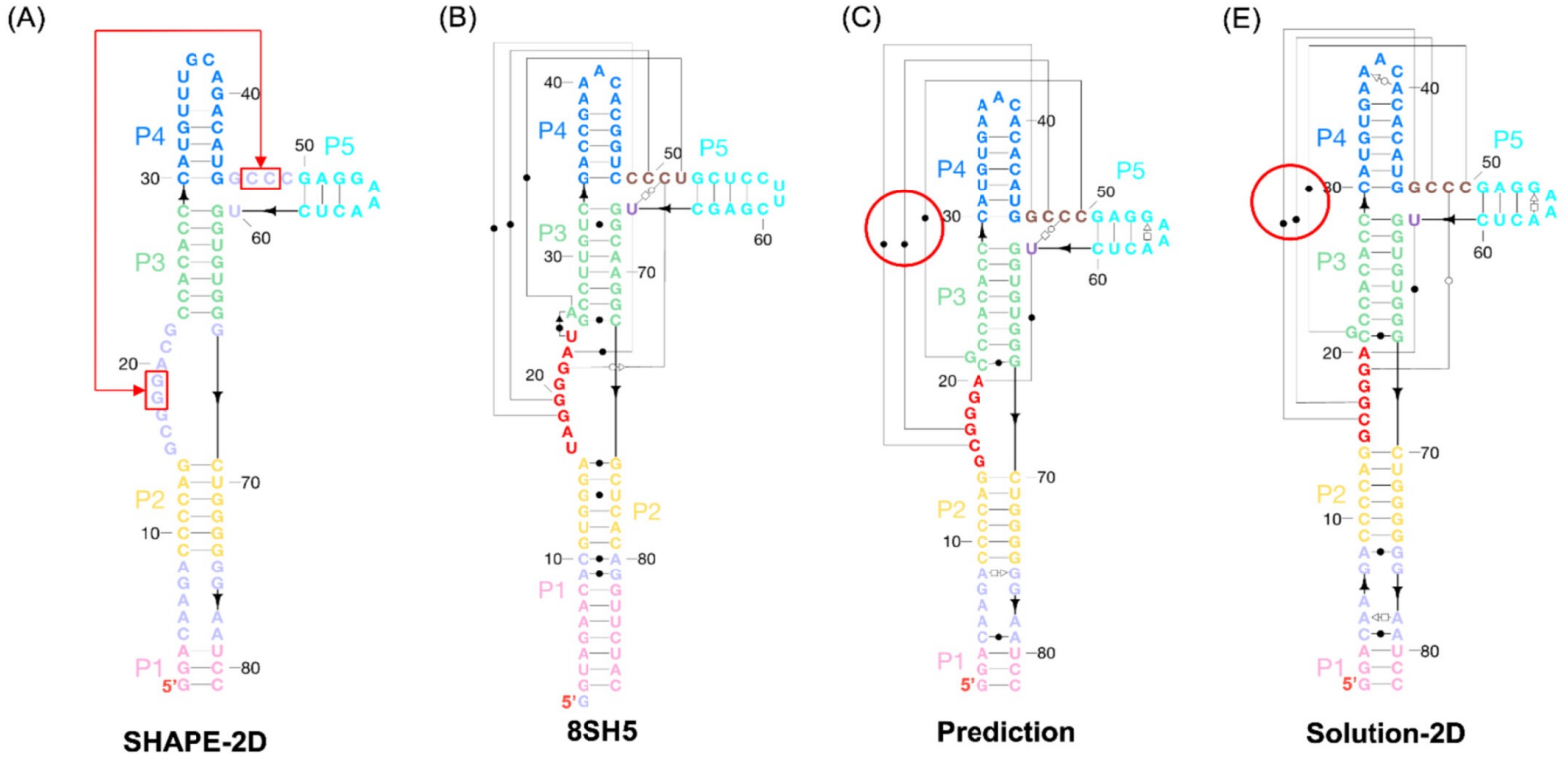
Rfam seed alignment



tried improvement

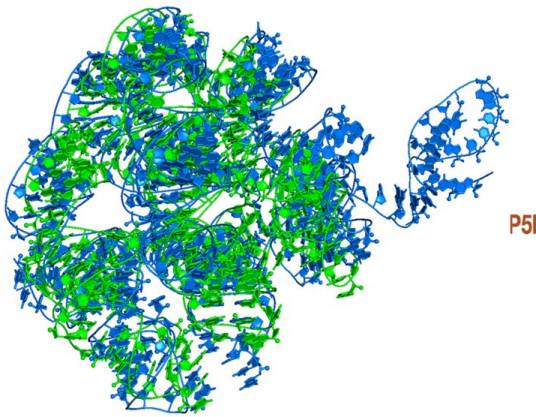
direction

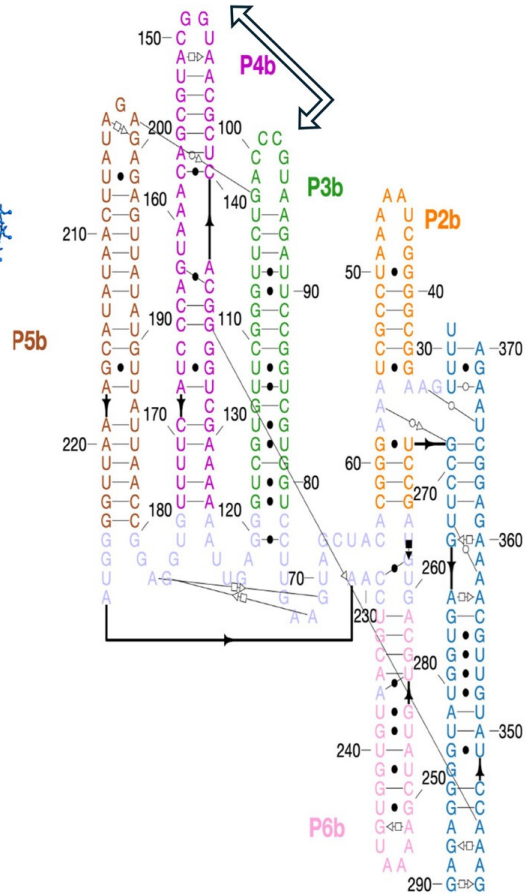
Use homology template information

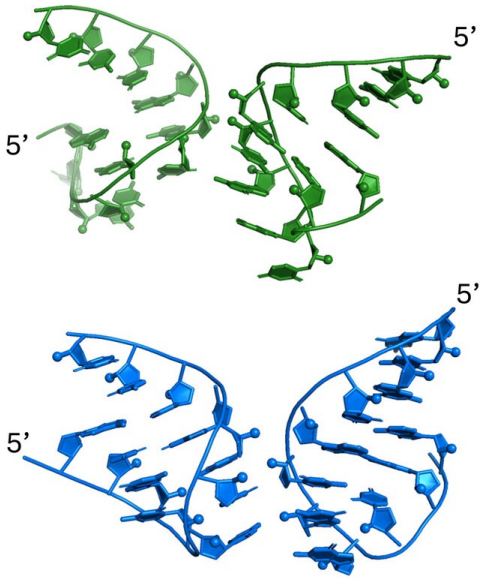


direction

Predict key interactions





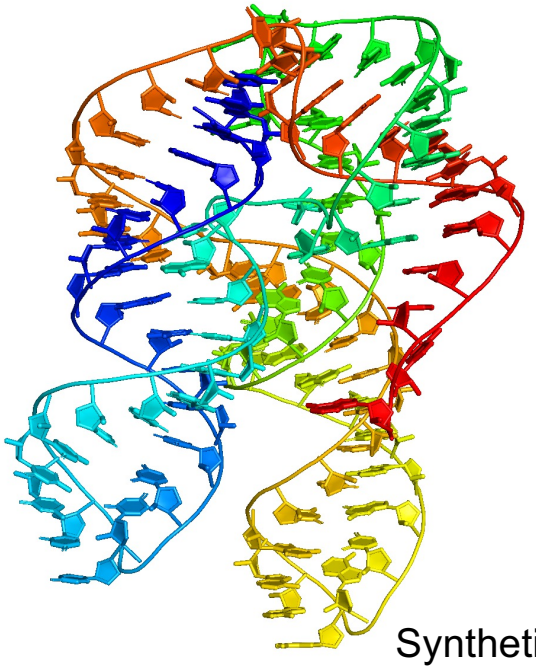


direction

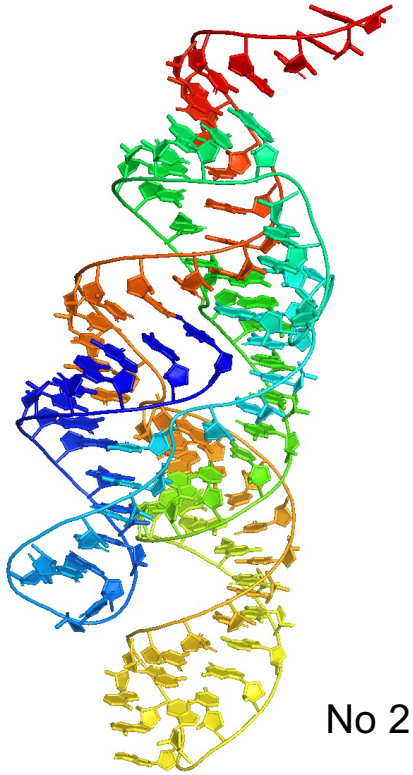
2.5D is not always helping

DNA WC F1	RNA WC F1	overall F1	DNA non-WC F1	RNA non-WC F1
0.9432	**0.9117**	**0.6512**	**0.1765**	**0.2413**

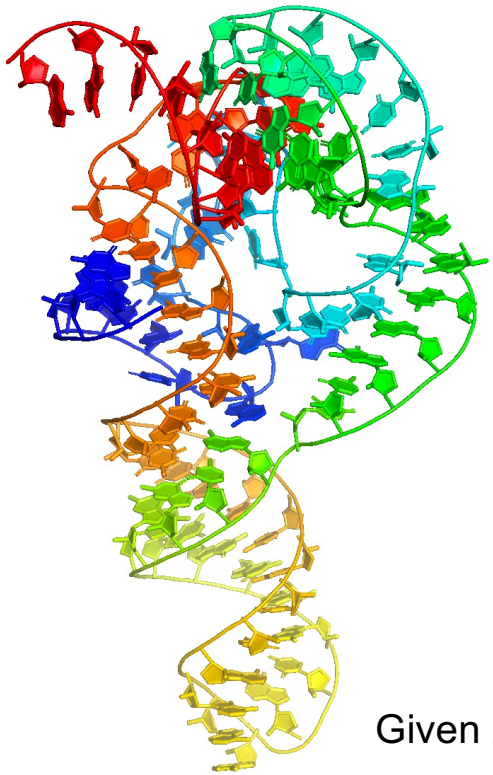
However the **overall conformations** are not as smooth as when no 2.5D info added



Synthetic target



No 2.5D



Given 2.5D

Take Home

keep strict, benchmark prediction in detail

1. Separate classes

Human, server, hybrid, and assisted predictions should not be silently merged into one interpretation.

2. Audit leakage

AI-era trust requires simple pre-release and post-hoc checks.

3. Report by mechanism

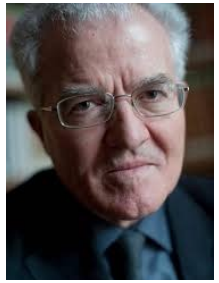
Global fold, interaction network, topology, and state should be visible separately.

Then RNA-Puzzles can remain the backbone benchmark for RNA structure prediction in the AI era.

Blind evaluation of RNA 3D structure prediction (RNA-Puzzles, CASP, Kaggle) are important to understanding the prediction.

acknowledgements

Thanks to …



Eric Westhof

Crystallographers:

- Adrian Ferre-D'Amare, Jinwei Zhang (NIH)
- Dinshaw Patel(MSKCC, New-York)
- Thomas Hermann (UCSD)
- David M J Lilley (Dundee)
- Jeffrey S. Kieft (UC denver)
- Barbara L. Golden (Purdue)
- Robert T. Batey (U colorado boulder)
- Joseph A Piccirilli (U chicago)
- Alexander Serganov (NYU, New-York)
- Anne-Catherine Dock-Bregeon (ENSP, Paris)
- Benoît Masquida (CNRS, Strasbourg)
- Amy Ren (Zhejiang University)
- Lin Huang (SYS University)
- Yijing Liu (Tianjin University)
- Kelly Nguyen (MRC-LMB, Cambridge)