Lab	Num	RMSD	DI all	INF all	INF wc	INF nwc	INF stacking	Clash Score	P-value	mcq	TM-score	best
SimRNA	1	3.396	4.658	0.729	0.559	0.000	0.857	144.300	6.66e-16	17.63	0	
Dfold2	1	4.204	6.946	0.605	0.648	0.218	0.640	104.770	2.33e-14			
Dokholyan	1	4.831	7.370	0.655	0.667	0.236	0.701	132.0				
Bujnicki	1	5.303	7.738	0.685	0.688	0.000	0.730					
Xiao	1	5.418	7.265	0.746	0.775	-						
Nithin	1	5.538	8.800	0.629	0							
Ares	1	6.058	7.914									
Farfar2	1	6.085										
Chen	1											
Perez												
Da												

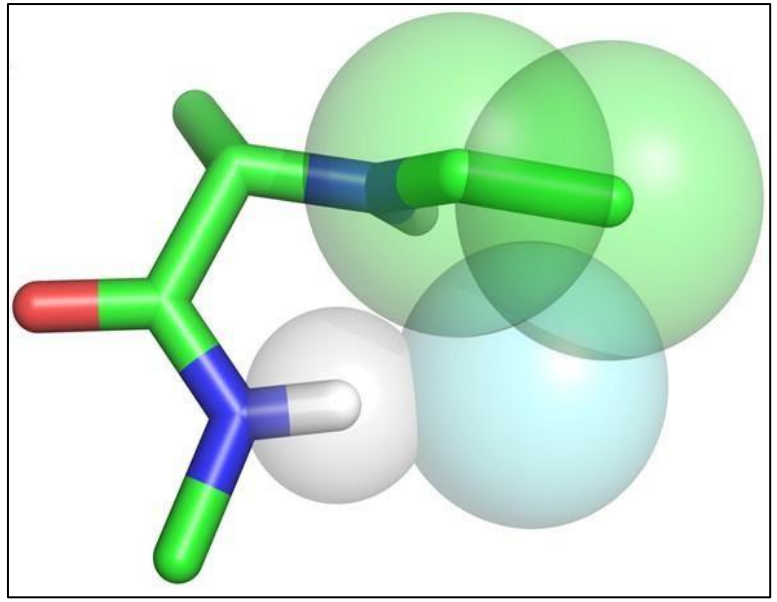
RNA Benasque, 7-20.08.2022

Evaluation of predictions.
The geometry of RNA 3D models.

Marta Szachniuk

mszachniuk@cs.put.poznan.pl

Clash score



- The measure of structure quality assessment determined by MolProbity[1]
- Controls the quality of the model geometry (geometric parameters)
- Checks all atomic overlays: <0.4 Å
- Global assessment of structure S:

$$ClashScore(S) = 1000 \cdot \frac{Number\ of\ bad\ overlaps}{Number\ of\ atoms\ in\ S}$$

[1]Williams et al. (2018) MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science* 27: 293-315.

What about other geometric parameters?

- Close contacts
- Bond lengths
- Bond angles
- Planarity
- Chirality
- Polymer linkage
- Handedness of helices
- Base-pairing geometry

Bond and dihedral angles

RCSB MAXIT

Bond angles

Atom 1	Atom 2	Type	Minimum	Maximum	Reference	Standard dev
P	OP1	s	1.383	1.587	1.485	0.017
P	OP2	s	1.383	1.587	1.485	0.017
P	OP3	s	1.535	1.679	1.607	0.012
P	O5'	s	1.533	1.653	1.593	0.010
O5'	C5'	(P)O5'-C5'	1.344	1.536	1.440	0.016
O5'	C5'	(H)C2'-endo	1.328	1.520	1.424	0.016
O5'	C5'	(H)C3'-endo	1.366	1.474	1.420	0.009

Dihedral angles for nucleic acids

Atom 1	Atom 2	Atom 3	Atom 4	Type	Minimum	Maximum	Reference	Standard dev
O3'	P	C5'			226.5	344.1	285.3	9.8
					8.4	153.6	81.0	12.1
P	O5'	C5'	C4'		105.5	261.5	183.5	13.0
O5'	C5'	C4'	C3'		18.3	86.7	52.5	5.7
					141.0	217.8	179.4	6.4
					219.1	6.7	292.9	12.3
C4'	C3'	O3'	P		162.4	265.6	214.0	8.6
C3'	O3'	P	O5'		260.4	318.0	289.2	4.8
					354.9	166.5	80.7	14.3
Sugar								
C5'	C4'	C3'	O3'	C2'-endo	117.9	176.7	147.3	4.9
O4'	C4'	C3'	O3'	C2'-endo	236.3	299.9	268.1	5.3

Close contacts, polimer linkage

RCSB MAXIT

Conditions for close contact identification (MAXIT)

	Two adjacent residues in the chain	Two nonadjacent residues in the chain				
		D – D	D – H	D/H – Any	{*} – Any	Any – Any
Distance [Å]	< 1.80	< 1.415	< 1.35	< 1.60	< 1.70	< 2.195

D – Deuterium
{*} – {Na, Cu, Ca, Mn, Zn, Mg, Ni, Co, Hg, Pt, Ba}
Any – any type of atom except for D, H, and atoms in {*}

Polymer linkage values, bond lengths in Angstroms [Å] and angles in degrees [°].

Atom 1	Atom 2	Atom 3	Type	Minimum	Maximum	Reference	Standard dev
Bond length							
O3'	P		s	1.535	1.679	1.607	0.012
Angles							
C3'	O3'	P	s	112.5	126.9	119.7	1.2
O3'	P	O5'	s	92.6	115.4	104.0	1.9
O3'	P	OP2	large	103.9	117.1	110.5	1.1
O3'	P	OP2	small	92.0	118.4	105.2	2.2
O3'	P	OP1	large	103.9	117.1	110.5	1.1
O3'	P	OP1	small	92.0	118.4	105.2	2.2

Geometry validation by MolProbity

Example MolProbity output for some 3D structure data

```
GGGUCGUGACUGGCGAACAGGUGGGAAACCACCGGGGAGCGACCCGCCGCCCGCCUGGGC
(((((((...[.[[[[....((((((....))))))[[....)))))))(((.]]]]]].])))) =
21 bps
```

All-Atom Contacts	Clashscore, all atoms:	34.39		11th percentile* (N=1784, all resolutions)
	Clashscore is the number of serious steric overlaps (> 0.4 ?) per 1000 atoms.			
Nucleic Acid Geometry	Probably wrong sugar puckers:	0	0.00%	Goal: 0
	Bad backbone conformations#:	19	31.67%	Goal: <= 5%
	Bad bonds:	17 / 1448	1.17%	Goal: 0%
	Bad angles:	201 / 2257	8.91%	Goal: <0.1%

Do RNA-Puzzles predictions pass the test?

Data from the github repo[2]

https://github.com/RNA-Puzzles

22 challenges

- **22** target RNA structures (X-ray)
- **1028** predicted RNA 3D models
  - 233 predictions in webserver category
  - 795 predictions in human category

[2]Magnus et al. (2020) *Nucleic Acids Res* 48(2):576–588

Do RNA-Puzzles predictions pass the test?

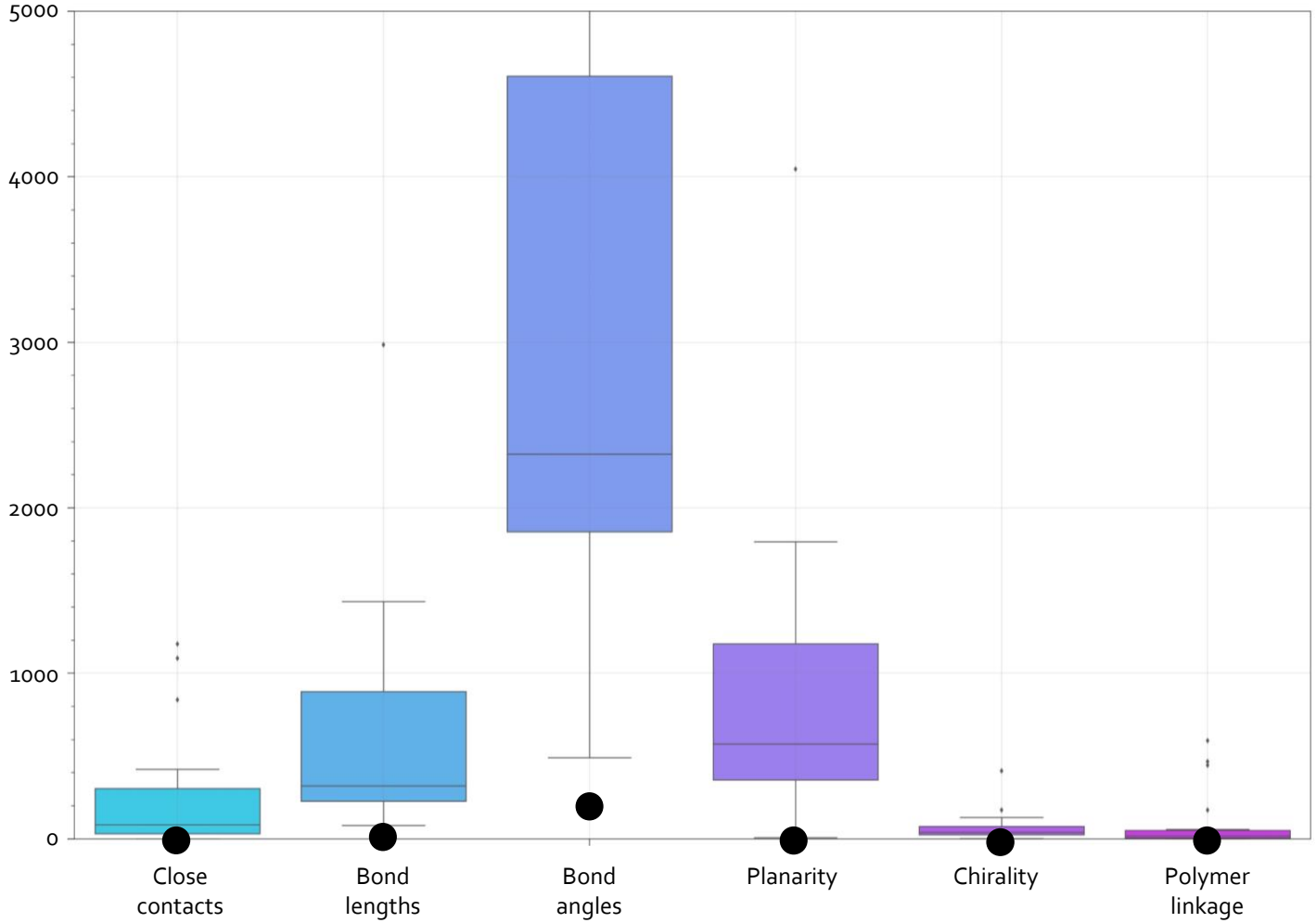
Data from the github repo[2]

https://github.com/RNA-Puzzles

22 challenges

- **22** target RNA structures (X-ray)
- **1028** predicted RNA 3D models
  - 233 predictions in webserver category
  - 795 predictions in human category

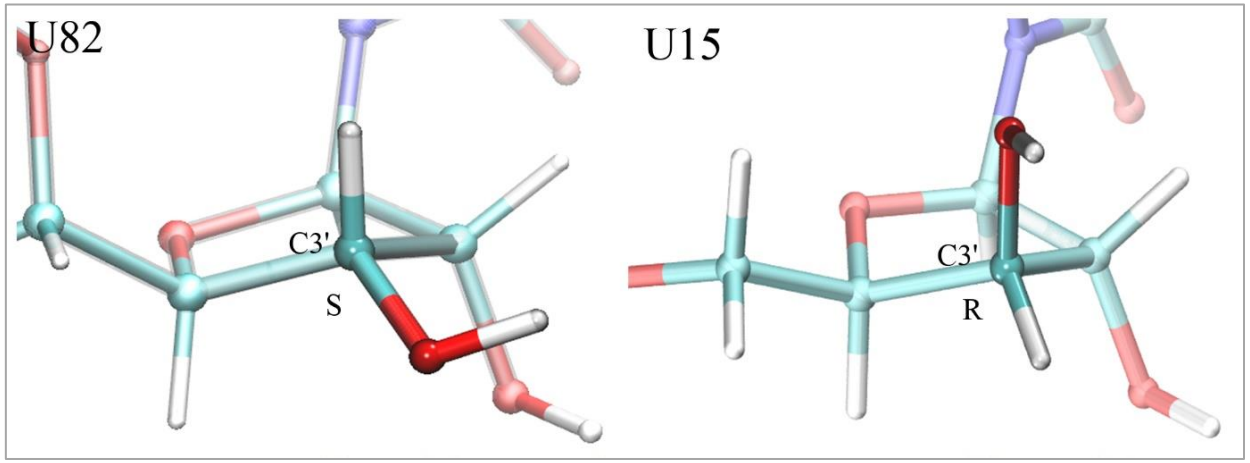
Total numer of errors in 1028 predicted models



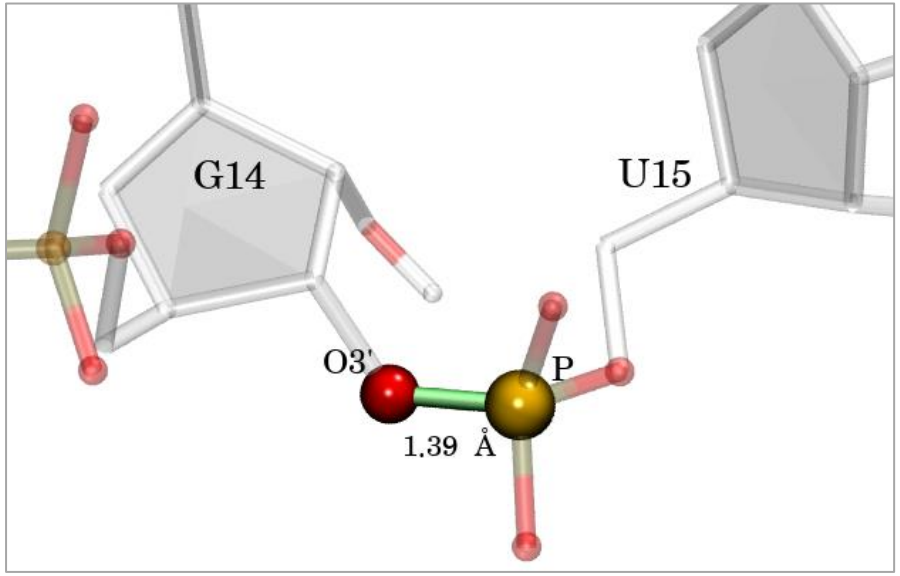
	Close contacts	Bond lengths	Bond angles	Planarity	Chirality	Polymer linkage
In targets	54	32	183	9	0	9

[2]Magnus et al. (2020) *Nucleic Acids Res* 48(2):576–588

Example bugs

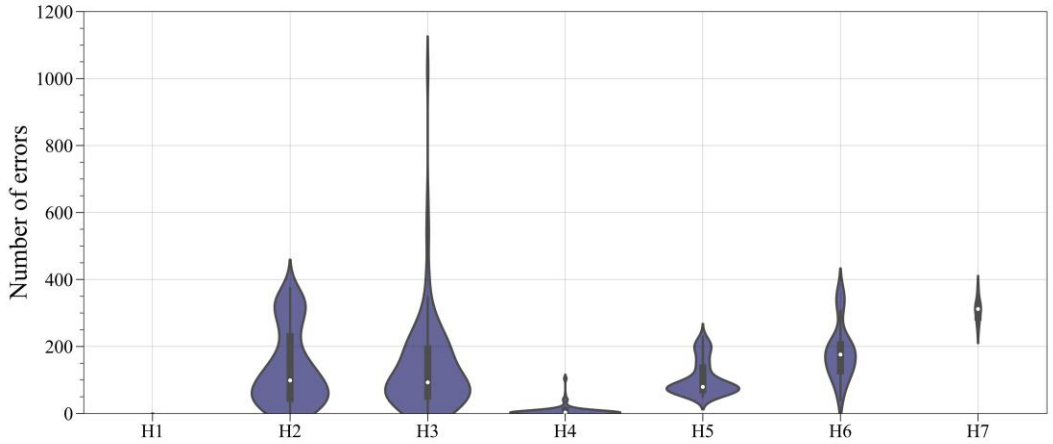


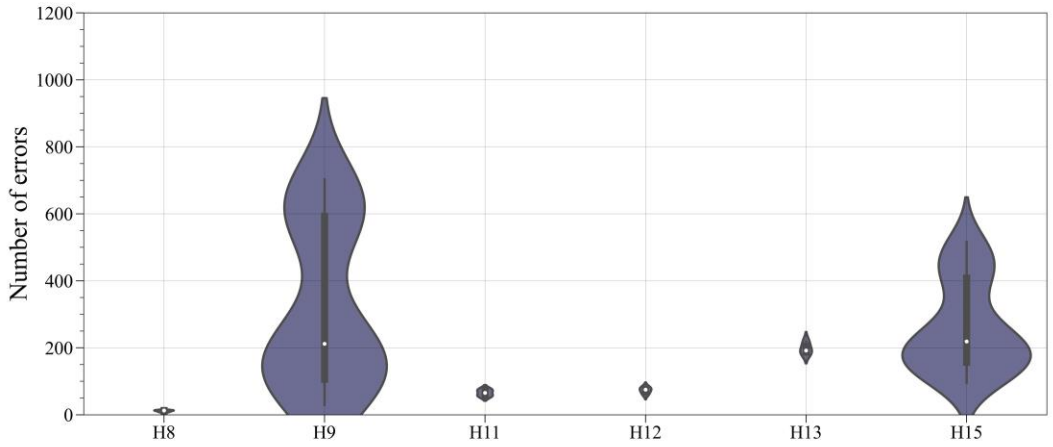
Chiral errors
Left: C3' atom in U82 with correct chiral centre;
Right: U15 with incorrect chiral inversion at carbon atom C3', changing a ribose to a xylose moiety.
H3 model from PZ07



Polymer linkage abnormality
found between G14 and U15. The distance should be ~1.6Å (with a sigma ~0.01).
H7 model from PZ24

Total numer of errors by participants

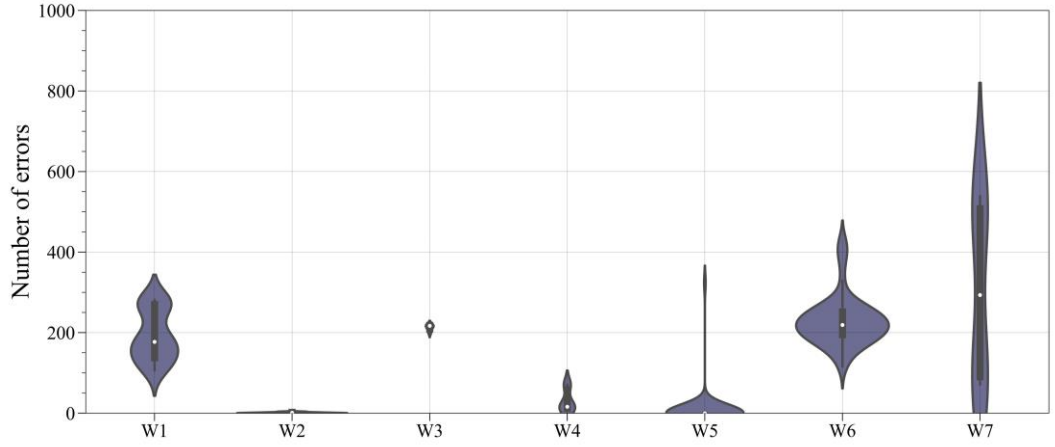




Human participants

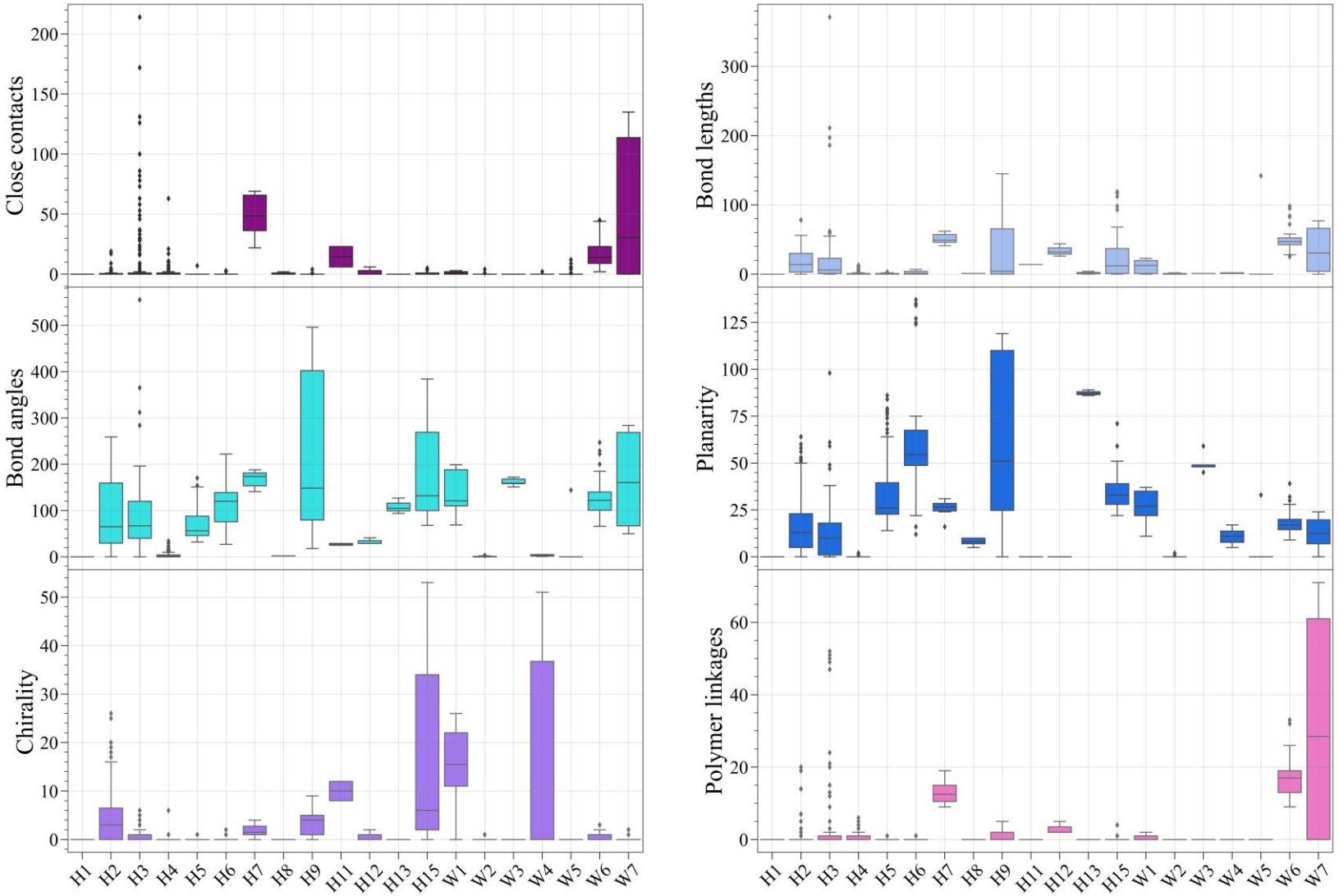
The number of stereochemical errors in all (1028) predicted models from the benchmark set by participants.

The white dot - the median
The black bar - interquartile range
The 1st and 3rd quartile - wicks up and down from the interquartile range
The violin shape shows error distribution.



Webserver participants

Error types by participants



Has the number of errors decreased over 10 yrs of RNA-Puzzles?

NO!

More details:

Carrascoza *et al*. (2022) Evaluation of the stereochemical quality of predicted RNA 3D models in the RNA-Puzzles submissions, *RNA* 28(2), 250-262.

- Stereochemical bugs occur in models predicted in human and webserver categories.
- Even the best models are not free of errors.
- We should check and ensure the correctness of models, preferably in an automated way.
- We have tools to identify these bugs in RNA models, and to correct some of them.

Acknowledgments

People behind this work:

- Maciej Antczak
- Francisco Carrascoza
- Chichau Miao
- Marta Szachniuk
- Eric Westhof

Funding

- National Science Center, Poland [2019/35/B/ST6/03074]
- Poznan University of Technology [0311/SBAD/0705]
- Institute of Bioorganic Chemistry, PAS

We thank Zukang Feng (RCSB PDB) for explaining the details of MAXIT performance.