Datasets for benchmarking RNA design algorithms

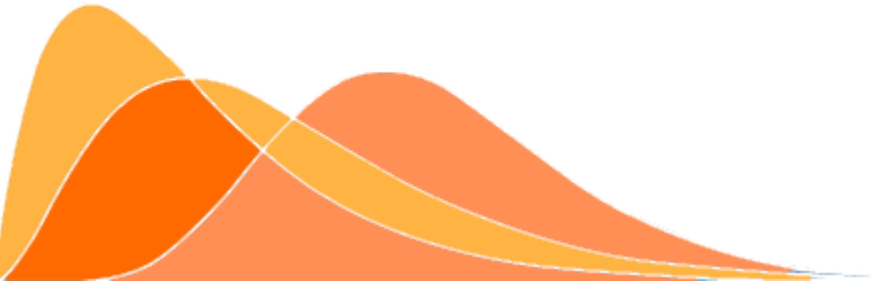
Agnieszka Rybarczyk[1,2]

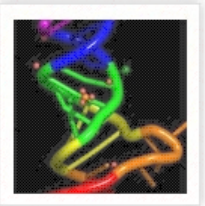
Jan Badura[1]

Tomasz Żok[1]

[1]*Institute of Computing Science, Poznan University of Technology*
[2]*Institute of Bioorganic Chemistry, Polish Academy of Sciences*

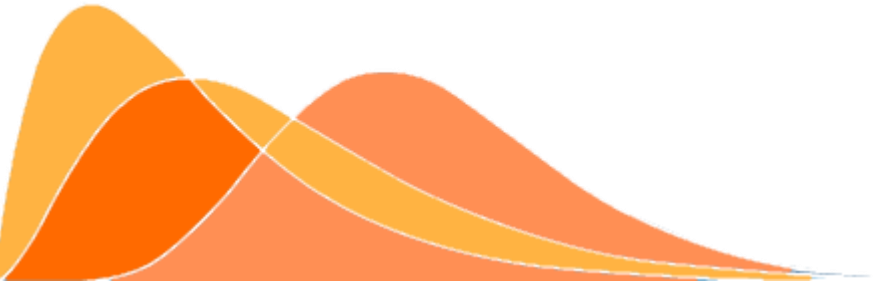
arybarczyk@cs.put.poznan.pl





Outline

- Motivation.
- Dataset sources and preparation pipeline.
- Evaluation and comparison of RNA design algorithms' performance.
- Conclusions.

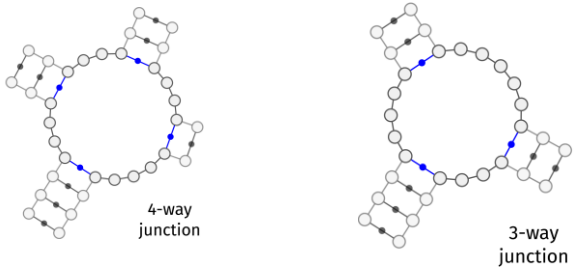


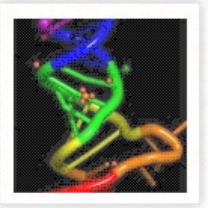
Motivation

- RNA design involves designing RNA sequences that fold into a desired structure to perform a specific function.
- The only data set available and recognized by the scientific community for this purpose is EteRNA100, a collection of structures assembled manually by experts:
  - 100 distinct secondary structure design challenges with lengths varying between 12 and 400 nucleotides and an average length of 127 nucleotides.
- Some algorithms managed to successfully solve most of the EteRNA100 design challenges.

Motivation

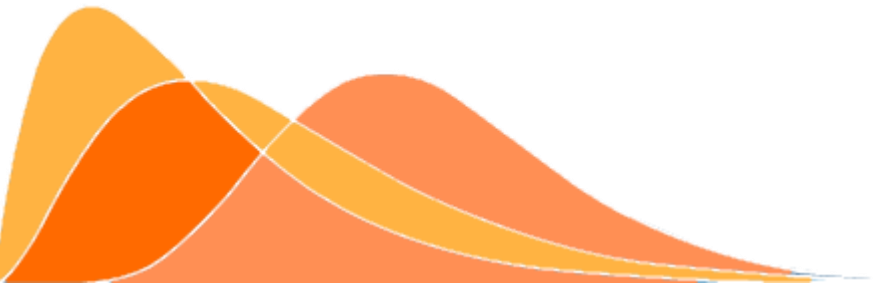
- Need for a new community-wide standard benchmark specifically designed for RNA design and RNA modeling algorithms.
- We created a very large, comprehensive and general-purpose dataset of over 15 million secondary structures with lengths ranging from 7 to 10,098.
- Our focus was mainly on multi-branched loops, which are often challenging to predict accurately.
- This dataset contains a diverse range of difficult-to-design motifs, from internal loops to n-way junctions (where n >= 3):
  - N-way junctions are substructures which have three or more helical “arms” (N) branching off.





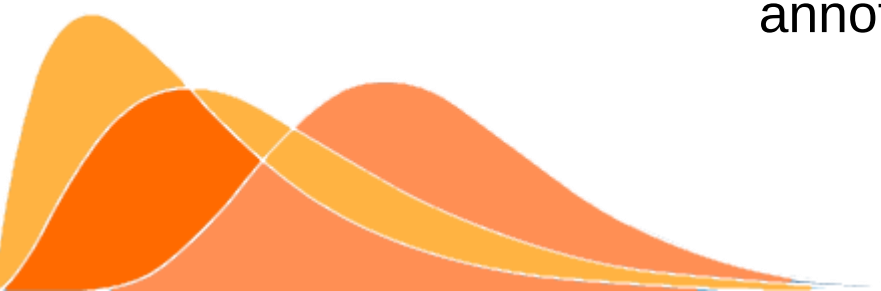
Data sources

- Separate structures from Rfam and RNAsolo provide complementary information that together allows for a more comprehensive and accurate understanding of RNA structure and function.
- Rfam 14 (https://rfam.org/)
  - Database being collection of RNA families.
  - Secondary structures help identify and characterize motifs such as loops, stem-loops, and other structural elements that are evolutionarily conserved and may have functional significance.

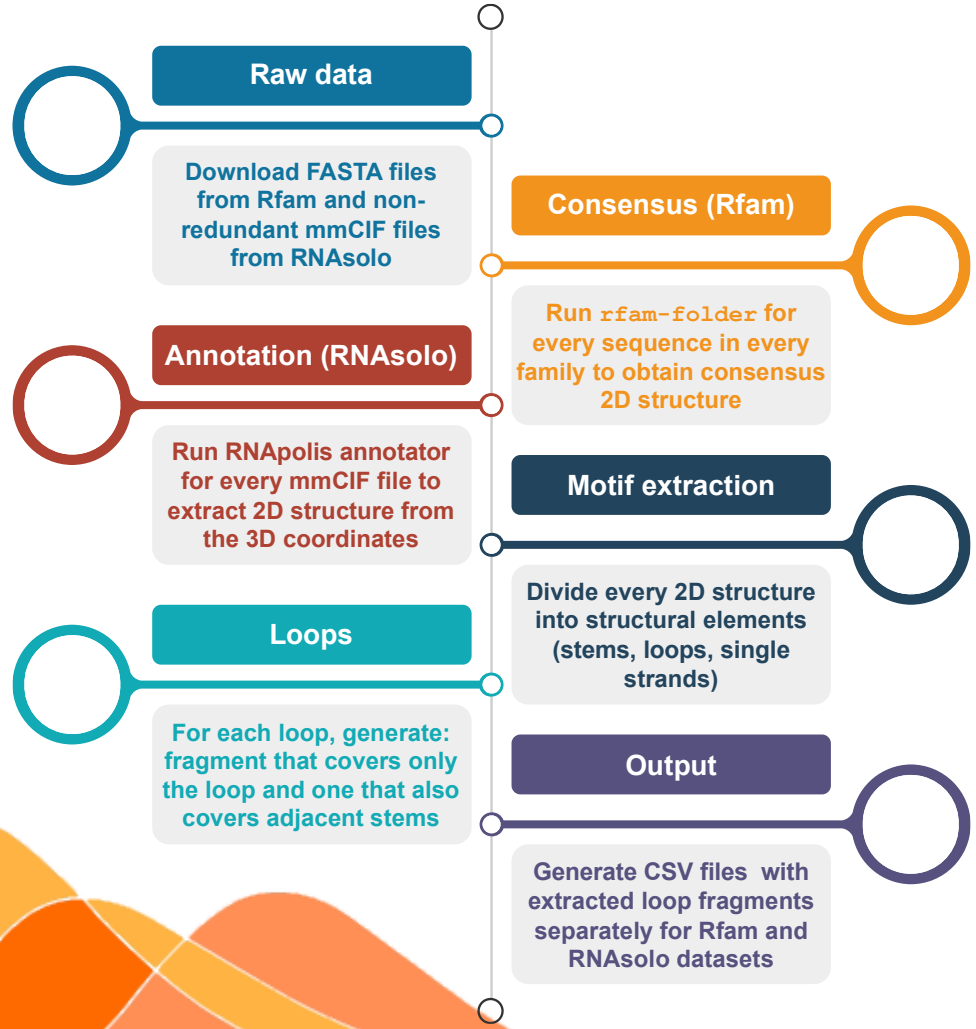


Data sources

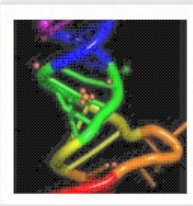
- RNAsolo (https://rnasolo.cs.put.poznan.pl/)
  - A self-updating database for experimentally determined RNA 3D structures, curated from the Protein Data Bank (PDB).
  - Cleans files from non-RNA data.
  - Offers downloads of various data subsets - whether clustered by resolution, source, or format
  - As of June 20, 2024 hosts 15,049 RNA structures, organized into 3,356 equivalence classes, each exemplified by a cluster representative.
  - We collected non-redundant 3D structures, which we then annotate for their canonical 2D representations.



Data preparation pipeline: from Rfam and RNAsolo to extracted loop motifs

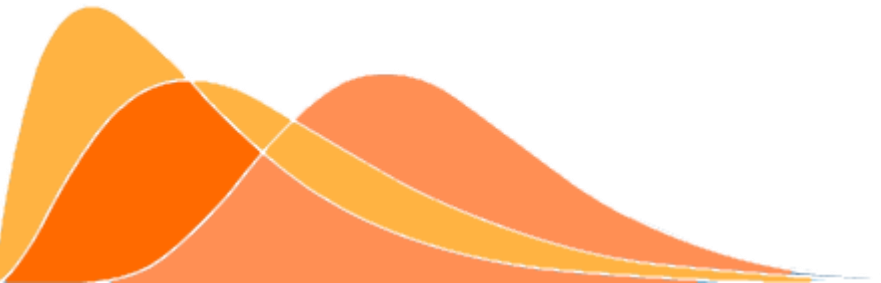


- Rfam 14
  - We collected covariance models and seed sequences for all RNA families from the Rfam 14 database.
  - We developed script rfam-folder (https://github.com/tzok/rnapolis-py) for generating consensus secondary structure for each RNA sequence in every Rfam family.
  - The textual results were transformed into standardized dot-bracket notation.
  - The resultant 2D structure is often underfolded, as it relies on strong signals from a large number of aligned sequences
  - To address this limitation, the rfam-folder runs RNAfold, treating the initial 2D structure as a hard constraint to fill unpaired regions with probable base pairs.



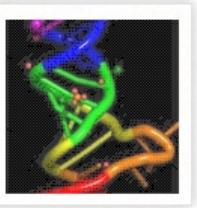
Data preparation pipeline: from Rfam and RNAsolo to extracted loop motifs

- The obtained 2D structures represent a more complete and realistic ones.
- For a more diversified dataset, we gathered results from both approaches: the straightforward unification of Infernal’s outputs and the refined structures generated by RNAfold with hard constraints.



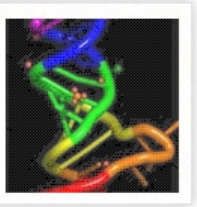
Data preparation pipeline: from Rfam and RNAsolo to extracted loop motifs

- RNAsolo
  - We used the annotator script from the RNApolis-py library for each PDBx/mmCIF file from the RNAsolo database to identify canonical base pairs and generate dot-bracket notation for entire structures.
  - We then integrated it with data from Rfam for comprehensive analysis in subsequent stages.
- Motif extraction
  - We dissected each 2D structure into following components: loops, stems, and single strands.
  - We used motif-extractor script from the RNApolis-py library.
  - It identifies and categorizes the structural fragments based on predefined rules e.g., recognizing adjacent base pairs as stems.



Data preparation pipeline: from Rfam and RNAsolo to extracted loop motifs

- To create effective RNA design targets, we focused on loops, which are often challenging to predict accurately.
- Loops removed from their structural context (e.g., the connecting stems) are energetically unstable and unlikely to be independently predicted by RNA design algorithms.
- Thus, for each identified loop motif, we generated two datasets:
  - The isolated loop fragment
  - The 2D structure of loop fragment extended with its connecting stems
- The final step in our data preparation pipeline consolidates the results into a CSV file.
- Each row corresponds to a loop, with columns identifying the motif's source and the sequence or dot-bracket encoded structure of the two mentioned instances.



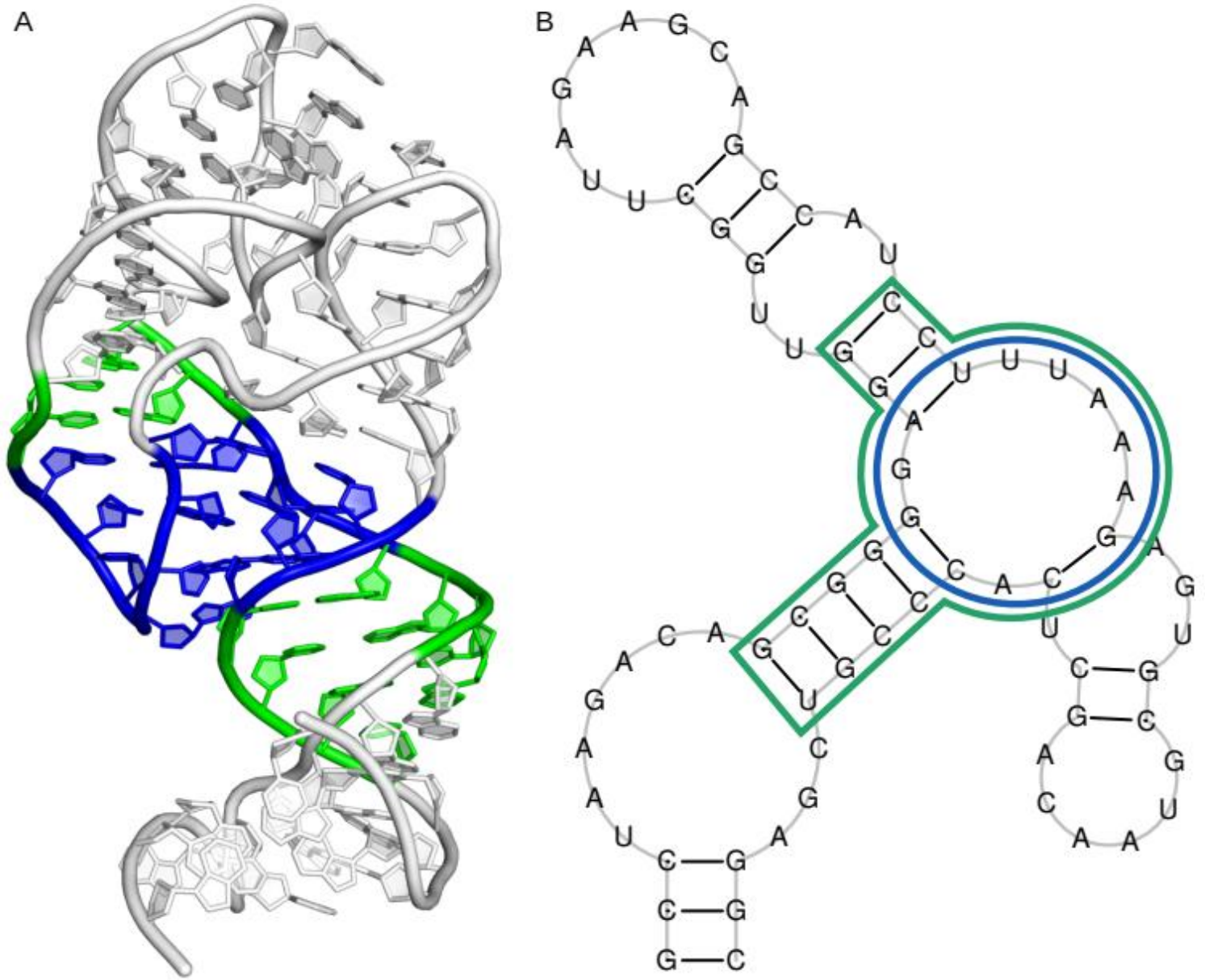
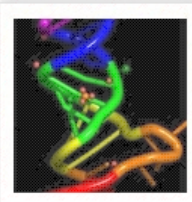
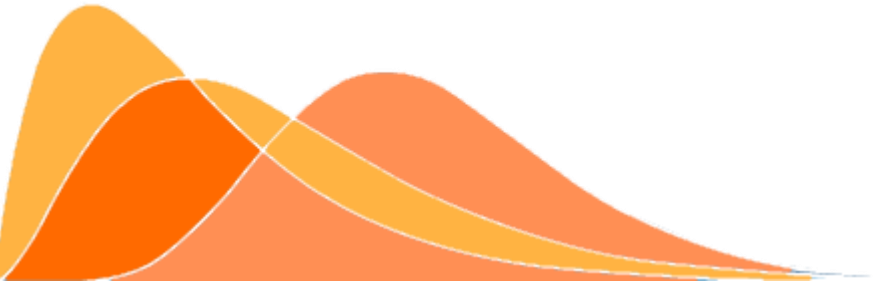


Fig. 2 Structure of the base of ribosomal P stalk (PDB id: 5D8H, chain A). A) 3D representation with the 3-way junction shown in blue and connecting stems shown in green. B) 2D representation colored the same way.



Data preparation pipeline: from Rfam and RNAsolo to extracted loop motifs

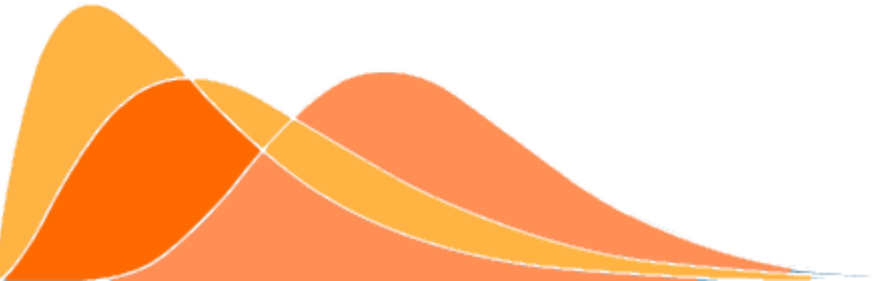
- The datasets are available at: https://zenodo.org/doi/10.5281/zenodo.12681122

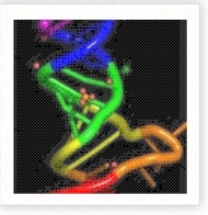


Statistics of loop motifs with connecting stems extracted from the RNAsolo database

- The dataset contains 8,746 loop motifs.
- Most of them (76%) are internal loops, about 78 nucleotides long on average, including the motif and connecting stems.
- 3-way and 4-way junctions each make up 9% of the dataset, with average lengths of 155 and 133 nucleotides.

Type	Count	Length			
		Min	Max	Mean	Std. Dev.
Internal loop	6678	7	3048	78.4	114.81
3-way junction	815	27	571	155.21	126.49
4-way junction	784	32	2089	133.22	215.9
5-way junction	265	12	1835	294.47	306.6
6-way junction	69	43	1510	250.26	258.95
7-way junction	47	49	2176	463.15	674.07
8-way junction	24	73	1982	602.62	677.44
9-way junction	19	46	3040	401.84	635.26
10-way junction	12	50	362	175.08	132.93
11-way junction	11	60	1390	939.0	534.18
12-way junction	4	98	1271	491.0	458.33
13-way junction	2	291	303	297.0	6.0
15-way junction	1	69	69	69.0	0.0
18-way junction	7	211	2824	621.71	899.32
19-way junction	1	275	275	275.0	0.0
21-way junction	4	2709	2927	2852.0	84.39
22-way junction	2	2880	3117	2998.5	118.5
25-way junction	1	3113	3113	3113.0	0.0
Total	**8746**				

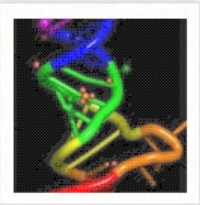




Statistics of loop motifs with connecting stems extracted from the Rfam database

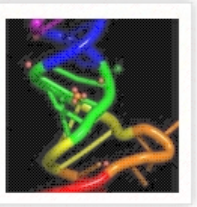
- The dataset contains 15 million loop motif instances.
- Similarly to the RNAsolo dataset, internal loop motifs dominate, (80%).
- 3-way and 4-way junctions make up 8% and 9% of instances respectively.
- The average lengths of these motifs: about 75 nts for internal loops, 121 nts for 3-way junctions, and 112 nts for 4-way junctions.

Type	Count	Length			
		Min	Max	Mean	Std. Dev.
Internal loop	12,101,168	9	7420	74.84	113.35
3-way junction	1,208,667	24	7470	121.04	110.85
4-way junction	1,301,478	33	7284	111.58	124.29
5-way junction	319,030	53	7415	240.87	248.51
6-way junction	63,607	69	10046	351.5	353.97
7-way junction	61,518	102	8331	434.92	379.33
8-way junction	38,785	129	9579	619.36	527.72
9-way junction	16,022	142	7092	603.64	536.48
10-way junction	9,292	180	7356	1210.02	770.44
11-way junction	6,365	202	9599	1777.76	1316.1
12-way junction	6,758	220	10098	2598.01	885.82
13-way junction	7,325	243	8178	2766.77	665.11
14-way junction	1,927	255	6895	2408.88	936.15
15-way junction	681	269	6752	2195.54	1010.57
16-way junction	689	284	7671	2295.81	1032.97
17-way junction	364	349	6406	2576.04	959.62
18-way junction	170	366	5754	2287.79	986.42
19-way junction	134	736	5113	2604.21	992.42
20-way junction	97	1018	5908	2767.38	1000.39
21-way junction	55	1088	8228	2672.84	1091.09
22-way junction	44	1104	4279	2292.52	808.39
23-way junction	28	1311	5290	2673.04	942.08
24-way junction	17	1143	3320	2439.71	708.18
25-way junction	9	1699	3343	2690.78	628.12
26-way junction	13	1493	5028	3213	785.58
27-way junction	6	1802	4903	3344.83	904.92
28-way junction	5	1801	3643	2464.6	820.94
29-way junction	8	2958	4425	3546	536.61
30-way junction	4	1763	3330	2825	625.34
31-way junction	3	2612	4780	3420	967.33
32-way junction	1	4064	4064	4064	0
33-way junction	3	3099	3459	3299	149.67
35-way junction	1	3108	3108	3108	0
36-way junction	1	3330	3330	3330	0
37-way junction	1	2614	2614	2614	0
Total	**15,144,276**				



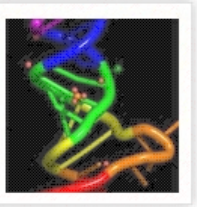
RNA design algorithms used for benchmarking and their evaluation

- We chose the following open-source RNA design algorithms:
  - RNAinverse, INFO-RNA, DSS-Opt, RNAfbinv, RNARedPrint, and DesiRNA.
- All tools were run using their default settings.
- https://github.com/jbadura/rna_design/
- For each sequence generated by the RNA design tool during testing, RNAfold was used to determine its secondary structure.
- To evaluate the results two metrics were used: RNAdistance and RNApdist.
- RNAdistance values were normalized - by dividing each RNAdistance value by the corresponding length of the RNA sequence, ensuring a more balanced comparison across different RNA sequences.
- The results are presented using violin plots.



RNA design algorithms used for benchmarking and their evaluation

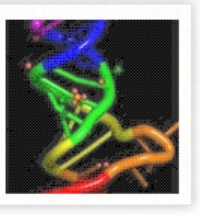
- The dataset was used to evaluate and compare the performance of selected RNA design tools: RNAinverse, INFO-RNA, DSS-Opt, RNAfbinv, RNARedPrint, and DesiRNA.
- The first test was performed using a dataset derived from the RNAsolo database.
- For the second one, due to the enormous size of the dataset derived from Rfam database, we decided to showcase its capabilities using a specific family, the glutamine riboswitch.
- This riboswitch, with its characteristic 3-way junction, presents significant modeling challenges.
- Due to the varying accuracy levels of different RNA design tools across cases of different lengths, an analysis was performed on the common instances addressed by all tools.



Benchmarking test case using a dataset of loop motifs derived from the RNAsolo database

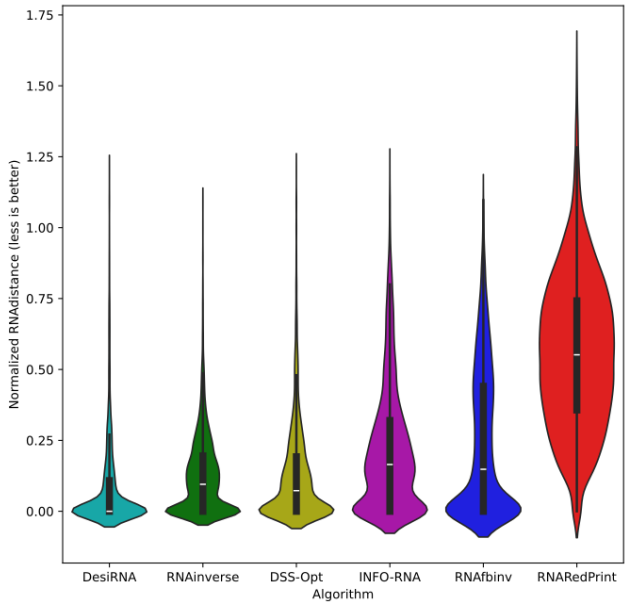
Table 4 RNA design benchmark results for the whole RNAsolo dataset.

RNA design algorithm	No of solved cases	Average computing time (s)	Normalized RNAdistance	RNApdist
Results for 8746 instances				
RNAinverse	7677	2.66	0.13	15.85
RNAfbinv	7206	8.24	0.26	15.93
INFO-RNA	7041	1.85	0.23	17.05
RNARedPrint	8746	0.11	0.61	46.12
DSS-Opt	8737	3.25	0.14	39.46
DesiRNA	8096	331.85	0.10	21.54
Results for 6037 instances successfully solved by each algorithm				
RNAinverse	6037	0.84	5.89	12.79
RNAfbinv	6037	8.18	13.64	15.28
INFO-RNA	6037	0.35	12.06	12.70
RNARedPrint	6037	0.10	28.97	18.32
DSS-Opt	6037	1.78	5.34	13.21
DesiRNA	6037	312.56	3.56	12.26

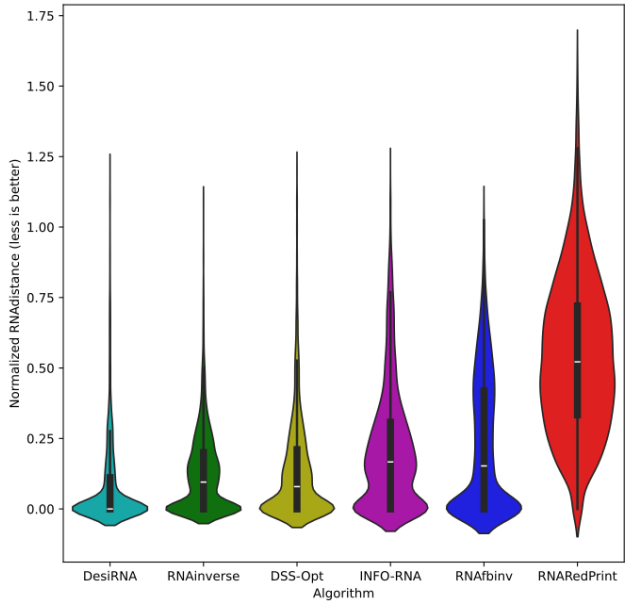


Benchmarking test case using a dataset of loop motifs derived from the RNAsolo database

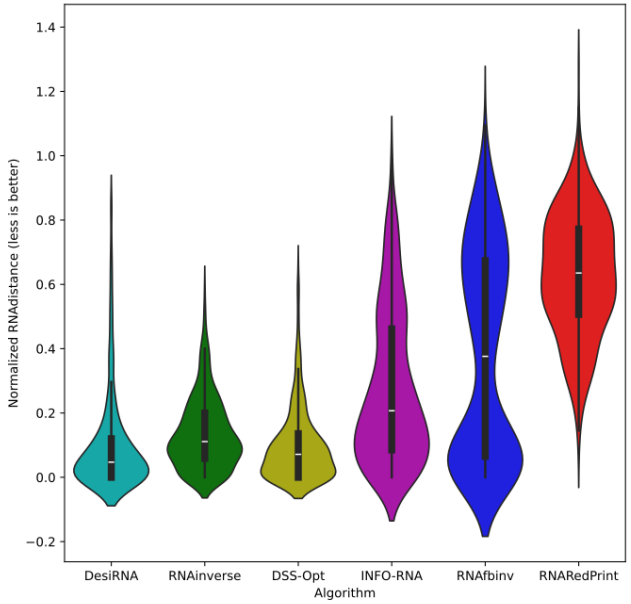
A



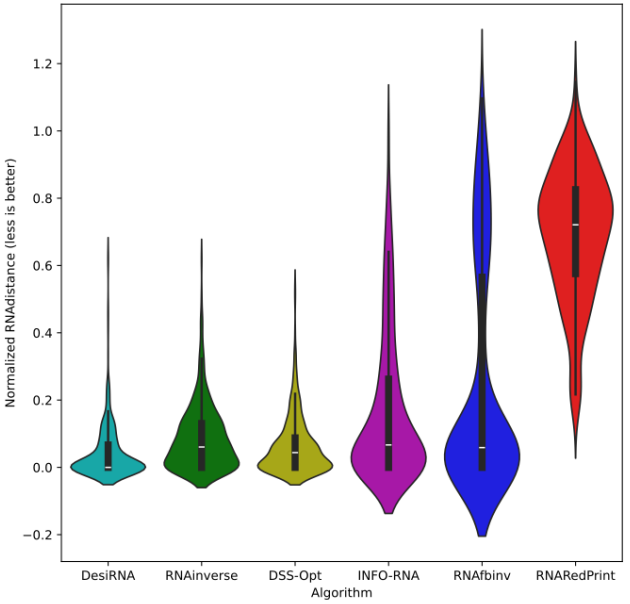
B

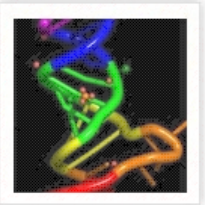


C



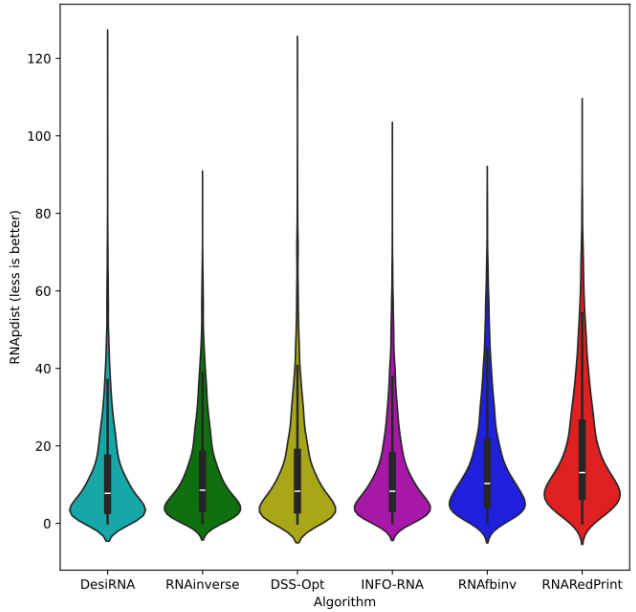
D



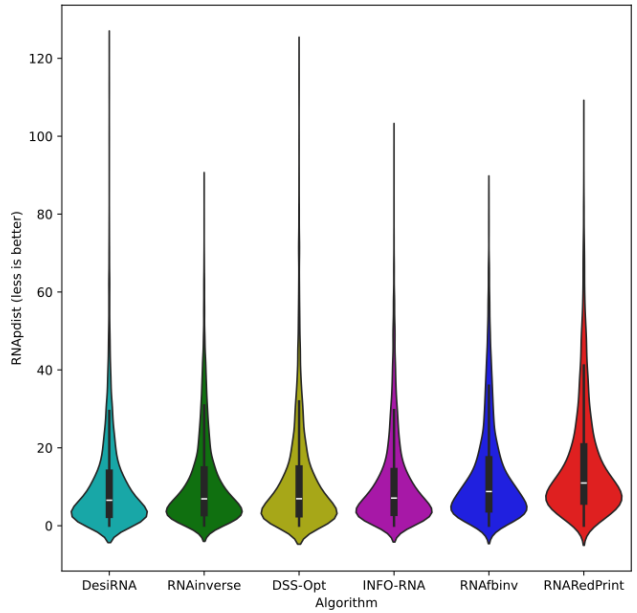


Benchmarking test case using a dataset of loop motifs derived from the RNAsolo database

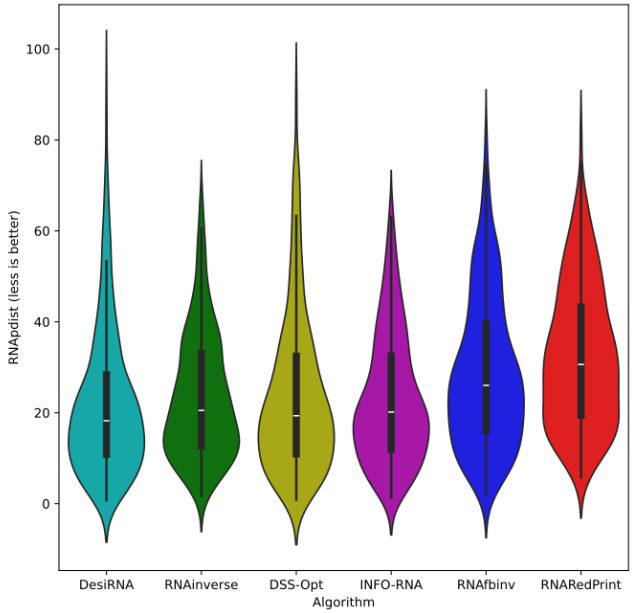
A



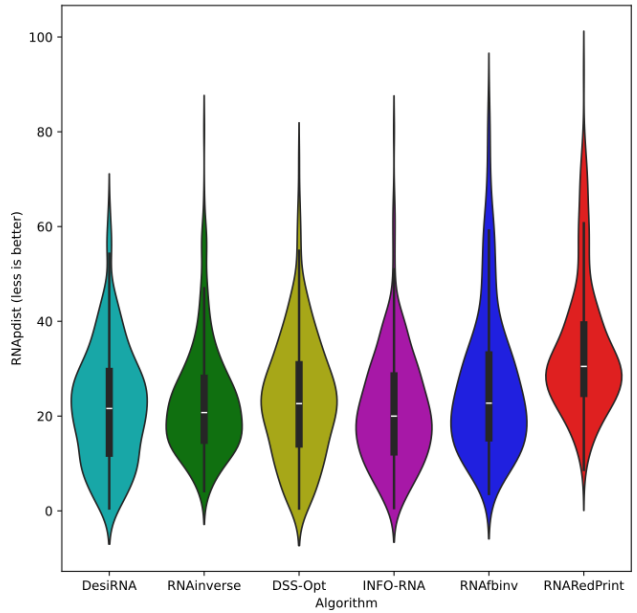
B

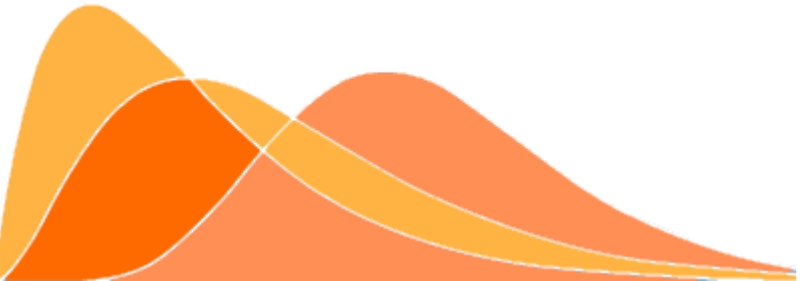


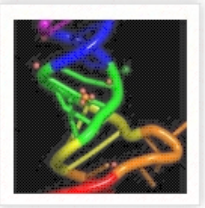
C



D

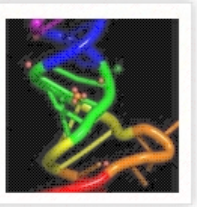






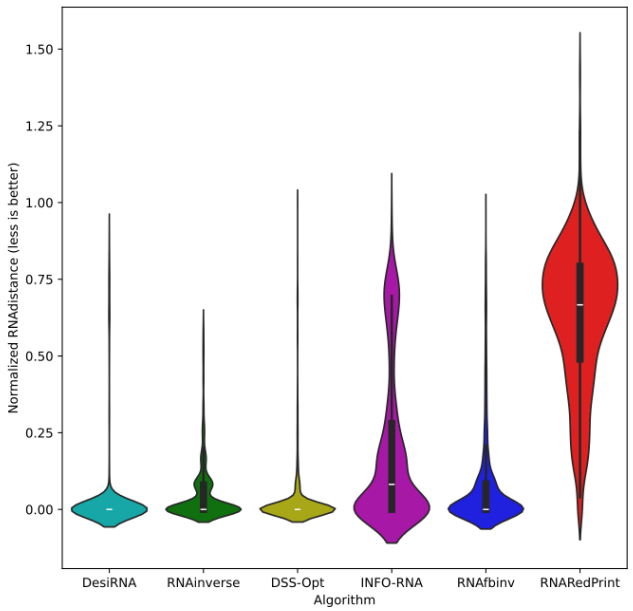
Benchmarking test case using a loop motifs dataset derived from the Rfam database, illustrated by the example of the glutamine riboswitch

- As an example of using a dataset derived from the Rfam database, we have chosen the RF01739 (glutamine riboswitch) Rfam family because it contains an important 3-way junction.
- The alignment consists of over 1700 sequences and includes more than 2200 loops.
- Similarly to the previous example, we used this set to evaluate and compare the performance of the following RNA design tools: RNAinverse, INFO-RNA, DSS-Opt, RNAfbinv, RNARedPrint, and DesiRNA.

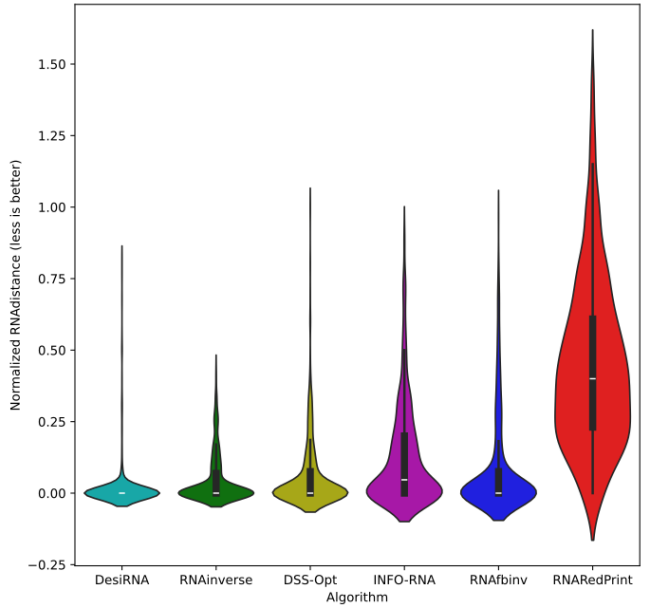


Benchmarking test case using a dataset of loop motifs derived from the Rfam database

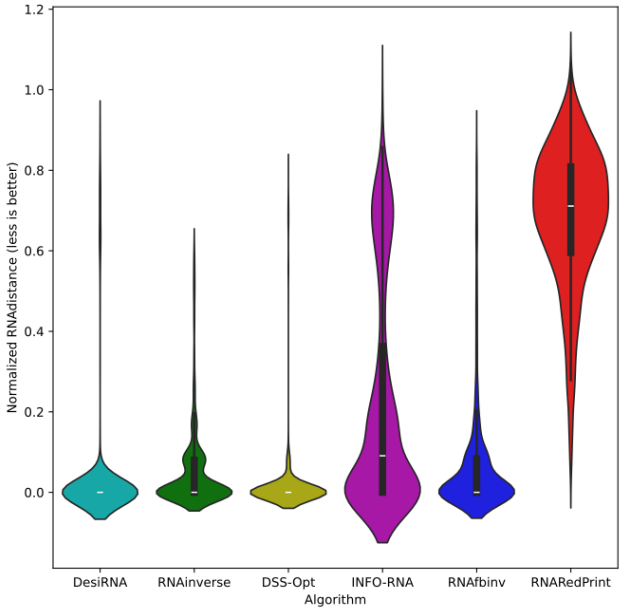
A



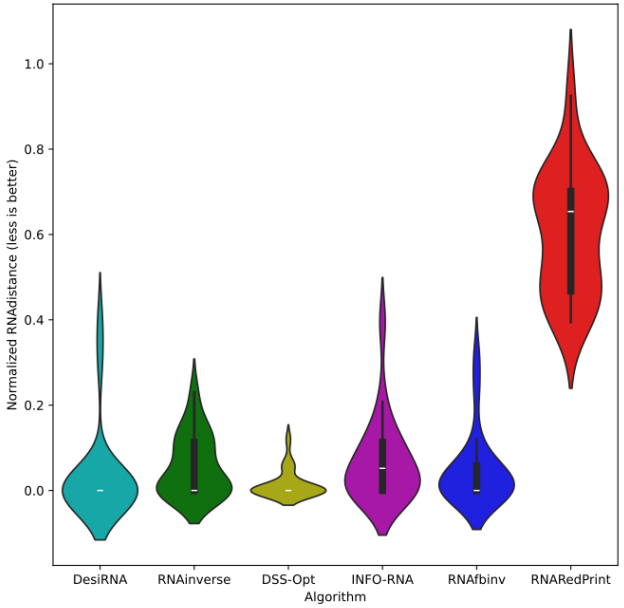
B

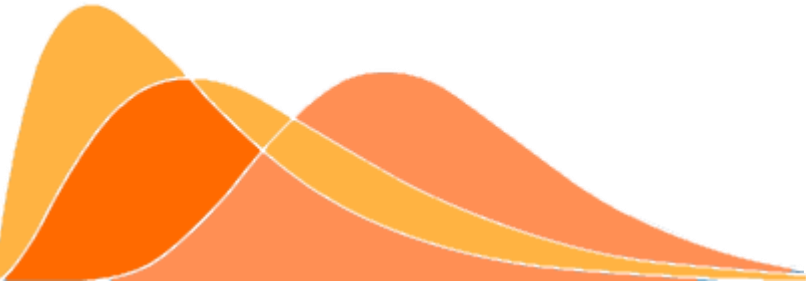


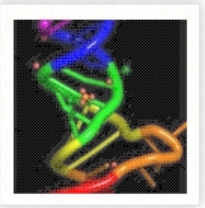
C



D

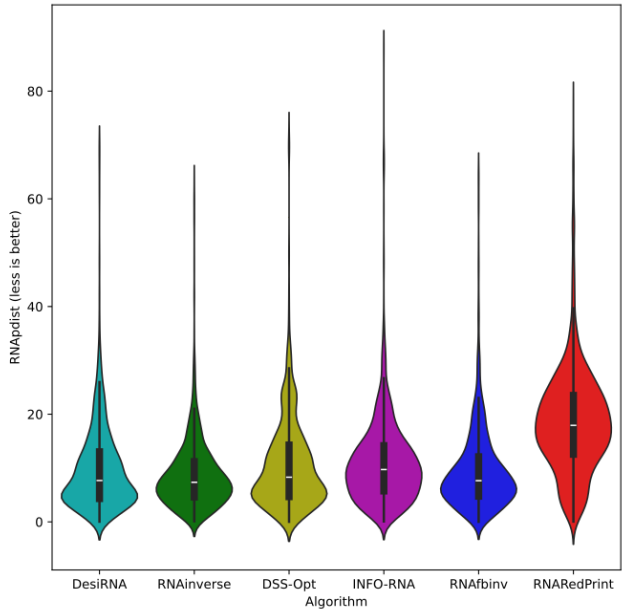




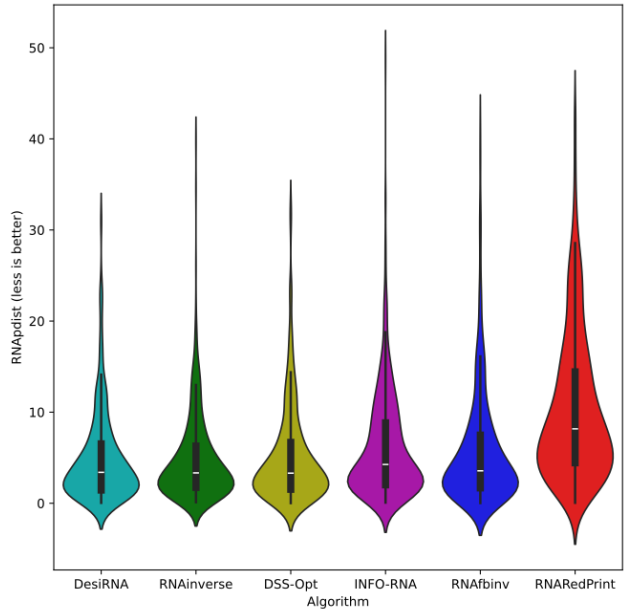


Benchmarking test case using a dataset of loop motifs derived from the RNAsolo database

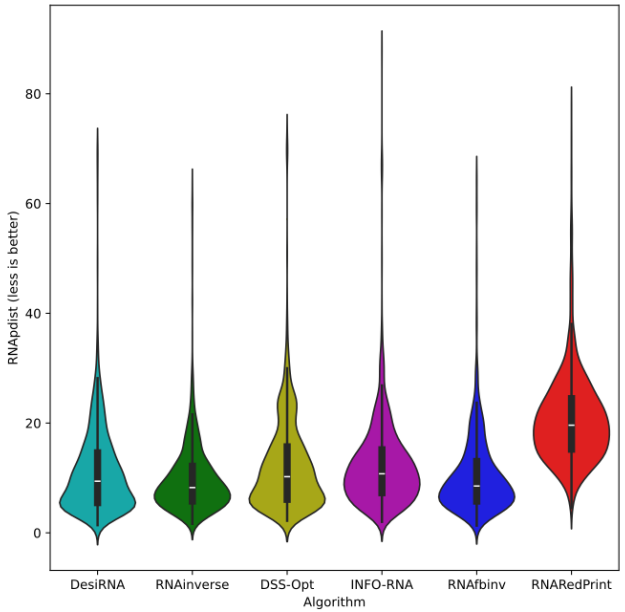
A



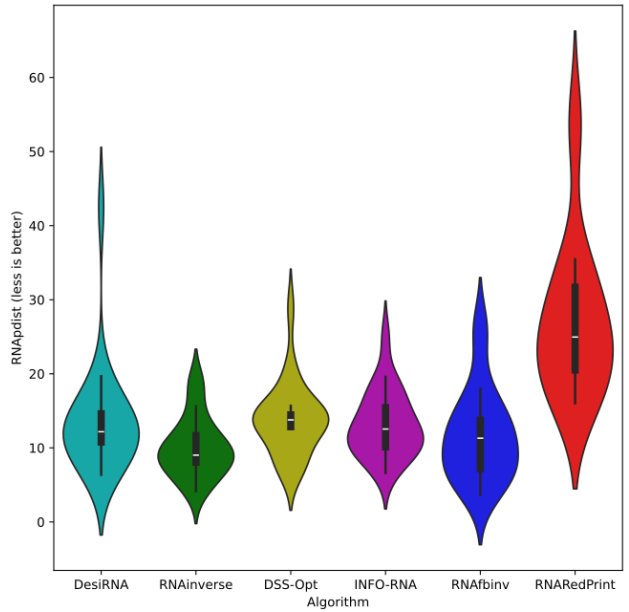
B

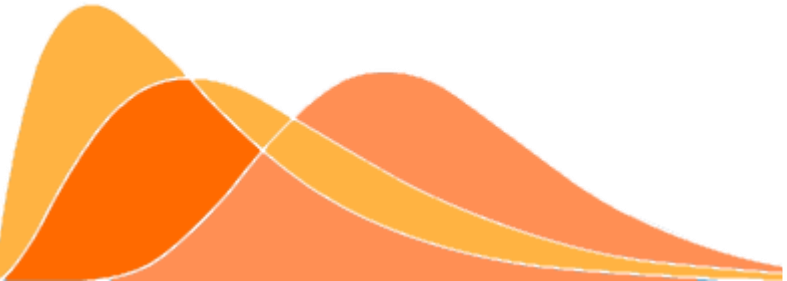


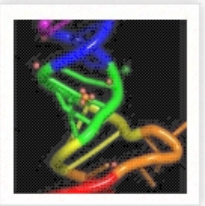
C



D

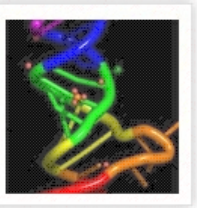






Benchmarking test case using a loop motifs dataset derived from the Rfam database, illustrated by the example of the glutamine riboswitch

- For predicting 3-way junction motifs, DesiRNA, RNAinverse and RNAfbinv showed very similar distributions, reflecting high accuracy and consistency, and achieving the best results.
- All algorithms, except for RNARedPrint, displayed relatively compact distributions with low median values.
- RNARedPrint, on the other hand, had a wide distribution and a noticeably higher median value, indicating more variability and less consistency in approximating the target structure.



Conclusions

- In the rapidly evolving field of RNA bioinformatics, the demand for high-quality data for use in benchmarking algorithms is increasing.
- To address the need, we have developed a comprehensive dataset of loop motifs in RNA structures.
- It combines information from experimentally solved 3D structures and the entire sequence repository of Rfam, a database of RNA families and their sequential alignments.
- It contains 15 million entries, ecompassing extracted internal loops, 3-way, 4-way, and higher cardinality junctions.
- These are not synthetic constructs, but rather motifs derived from experimentally verified data.

Conclusions

- This datasets can be used by researchers working on RNA design (also known as inverse folding) and also machine learning pipelines that incorporate both sequence and structural information.
- The versatility of our dataset is enhanced by its ability to describe each extracted motif either in isolation or within its structural context. This flexibility allows researchers to tailor their analyses to specific needs and objectives.
- To demonstrate the dataset's utility, we conducted extensive experiments evaluating the performance of various inverse folding algorithms using different metrics.