

Supplementary Materials for a fully-open structure-guided RNA foundation model for robust structural and functional inference

Heqin Zhu^{1,2}, Ruifeng Li³, Feng Zhang^{1,2}, Fenghe Tang^{1,2},
Tong Ye^{1,2}, Xin Li^{1,2}, Yunjie Gu^{1,2}, Peng Xiong^{1,2*},
S. Kevin Zhou^{1,2,4,5*}

^{1*}School of Biomedical Engineering, Division of Life Sciences and
Medicine, University of Science and Technology of China (USTC), Hefei,
Anhui, 230026, China.

^{2*}Center for Medical Imaging, Robotics, Analytic Computing & Learning
(MIRACLE), Suzhou Institute for Advance Research, USTC, Suzhou,
Jiangsu, 215123, China.

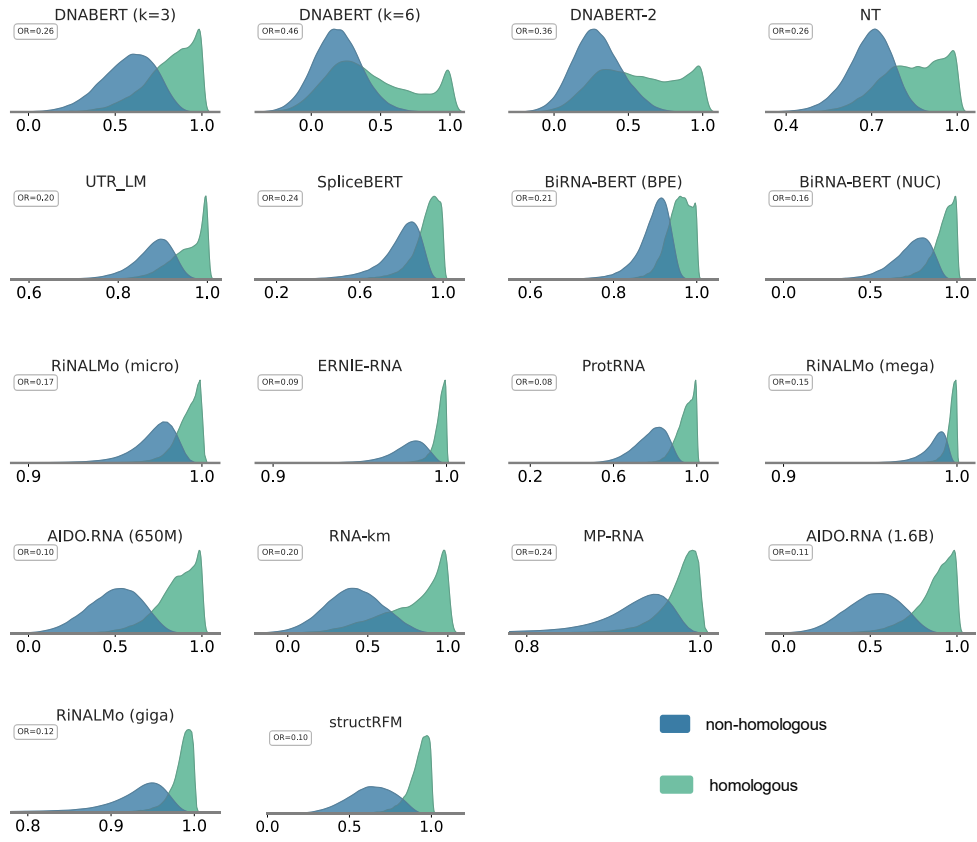
³College of Computer Science and Technology, Zhejiang University,
Hangzhou, Zhejiang, 310027, China.

⁴Jiangsu Provincial Key Laboratory of Multimodal Digital Twin
Technology, Suzhou, Jiangsu, 215123, China.

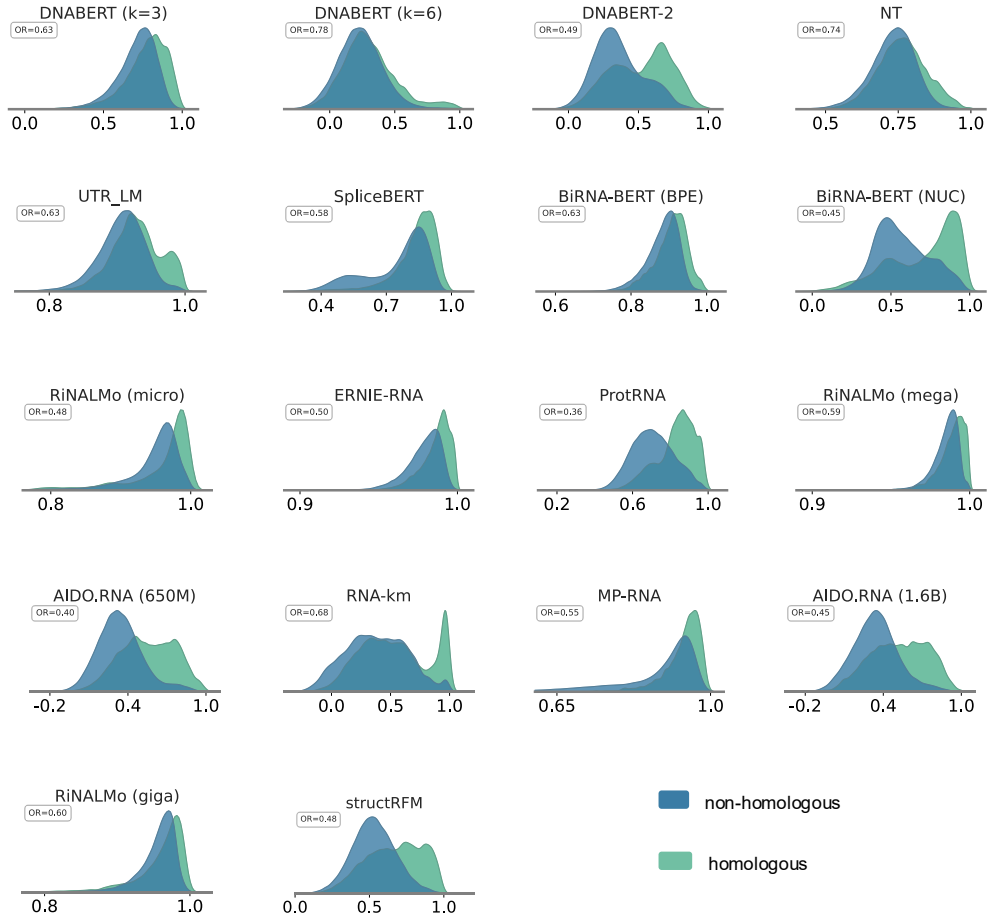
⁵State Key Laboratory of Precision and Intelligent Chemistry, USTC,
Hefei, Anhui, 230026, China.

*Corresponding author(s). E-mail(s): xiongap@ustc.edu.cn;
skevinzhou@ustc.edu.cn;

1 Supplementary figures



Supplementary Fig. 1: Cosine similarity distributions of homologous and non-homologous sequences for Rfam (24,523 RNAs) dataset.



Supplementary Fig. 2: Cosine similarity distributions of homologous and non-homologous sequences for ArchiveII600 (3,611 RNAs) dataset.

2 Supplementary tables

Supplementary Table 1: Summary of used datasets.

Dataset	Task	Phase	# Sequences	Length
RNAcentral [1]	Pretraining	-	21,477,078	≤ 512
Rfam [2, 3]	Zero-shot homology classification	test	24,607	19-4,002
ArchiveII [4]	Zero-shot homology classification	test	3,864	28-1,800
PDB [5]	Zero-shot secondary structure prediction	test	116	32-355
RNAstrAlign [6]	Secondary structure prediction	train	19,313	30-600
bpRNA-1m [7]	Secondary structure prediction	train	12,114	33-498
bpRNA-TS0 [7]	Secondary structure prediction	test	1,305	22-499
ArchiveII600 [4]	Secondary structure prediction	test	3,911	28-600
PDB [5]	Tertiary structure prediction	train	3,633	11-960
CASP15 [8]	Tertiary structure prediction	test	12	30-720
RNA Puzzles [9, 10]	Tertiary structure prediction	test	20	37-189
Splice [11]	Splice site prediction	train	162,706	100
Splice [11]	Splice site prediction	test	16,505	100
IRES [12]	IRES identification	train	12,998	174
IRES [12]	IRES identification	test	1,164	174
nRC [13]	ncRNA classification	train	6,320	38-1,136
nRC [13]	ncRNA classification	test	2,600	39-1,182

Supplementary Table 2: Evaluation of finetuned structRFM for secondary structure prediction on ArchiveII600 (3,911 RNAs) and bpRNA-TS0 (1,305 RNAs) datasets. The baseline performances are borrowed from RNAErnie [14]

Method	ArchiveII600			bpRNA-TS0		
	Precision	Recall	F1	Precision	Recall	F1
RNAstructure	0.563	0.615	0.585	0.494	0.622	0.533
RNAsoft	0.665	0.594	0.622	0.497	0.626	0.535
RNAfold	0.565	0.627	0.592	0.494	0.631	0.536
MXfold2	0.788	0.760	0.768	0.519	0.646	0.558
Mfold	0.428	0.383	0.401	0.501	0.627	0.538
LinearFold	0.641	0.617	0.621	0.561	0.581	0.550
Eternafold	0.667	0.622	0.636	0.516	0.666	0.563
E2Efold	0.738	0.665	0.690	0.140	0.129	0.130
CONTRAFold	0.607	0.679	0.638	0.528	0.655	0.567
Contextfold	0.873	0.821	0.842	0.529	0.607	0.546
RNAErnie	0.884	0.869	0.873	0.576	0.668	0.608
RNABERT	0.694	0.669	0.677	0.554	0.581	0.548
RNA-MSM	0.608	0.585	0.593	0.548	0.612	0.560
RNA-FM	0.857	0.844	0.847	0.596	0.716	0.639
structRFM _F ⁻	0.846	0.846	0.843	0.581	0.704	0.622
structRFM ⁻	0.850	0.857	0.850	0.587	0.691	0.621
structRFM _F	0.862	0.867	0.861	0.606	0.707	0.641
structRFM	0.875	0.876	0.873	0.602	0.712	0.641

Supplementary Table 3: The detailed performance of structRFM and structRFM⁻ for secondary structure prediction on ArchiveII600 (3,911 RNAs) across nine families.

Family	structRFM			structRFM ⁻		
	Precision	Recall	F1	Precision	Recall	F1
16s	0.684	0.688	0.683	0.725	0.718	0.717
23s	0.332	0.295	0.311	0.552	0.487	0.516
5s	0.975	0.985	0.980	0.962	0.971	0.966
RNaseP	0.848	0.825	0.834	0.818	0.793	0.802
grp1	0.670	0.693	0.677	0.642	0.689	0.659
srp	0.794	0.821	0.803	0.752	0.800	0.771
tRNA	0.960	0.993	0.975	0.962	0.980	0.969
telomerase	0.578	0.619	0.595	0.435	0.475	0.453
tmRNA	0.810	0.703	0.748	0.737	0.670	0.698

Supplementary Table 4: Evaluation of finetuned structRFM for tertiary structure prediction on CASP15 (12 RNAs, 8 natural and 4 synthetic) and RNA Puzzles (20 RNAs) datasets.

Method	CASP15		CASP15-natural		RNA Puzzles	
	RMSD (Å)	F1	RMSD (Å)	F1	RMSD (Å)	F1
trRosettaRNA	26.0	0.625	18.4	0.659	7.28	0.646
structRFM _F ⁻	25.2	0.669	16.3	0.702	6.64	0.631
structRFM ⁻	26.0	0.660	15.9	0.707	7.52	0.580
structRFM _F	23.6	0.718	15.2	0.719	7.31	0.663
structRFM	24.1	0.740	14.8	0.763	6.82	0.665

Supplementary Table 5: Evaluation of finetuned structRFM for IRES identification on IRES dataset (1,164 RNAs).

Method	Accuracy	Recall	Precision	F1	MCC	AUROC
IRESfinder	0.475	0.474	0.475	0.475	-0.050	0.449
DeepIRES	0.507	0.196	0.518	0.284	0.018	0.584
structRFM _F ⁻	0.651	0.658	0.649	0.653	0.302	0.687
structRFM ⁻	0.650	0.737	0.628	0.678	0.305	0.684
structRFM _F	0.650	0.747	0.626	0.681	0.307	0.679
structRFM	0.653	0.840	0.611	0.708	0.330	0.697

Supplementary Table 6: Evaluation of finetuned structRFM for ncRNA classification on nRC dataset (2,600 RNAs). The baseline performances are borrowed from RNAErnie [14]

Method	Accuracy	Recall	Precision	F1	MCC	AUROC
RNAcon	0.374	0.373	0.450	0.351	0.334	-
nRC	0.696	0.689	0.688	0.688	0.663	-
ncRFP	0.797	0.788	0.790	0.788	0.771	-
RNAGCN	0.857	0.861	0.988	0.856	0.846	-
ncRDeep	0.880	0.884	0.891	0.886	0.880	-
ncRDense	0.9510	0.951	0.953	0.951	0.947	-
RNAErnie	0.964	0.964	0.964	0.964	0.961	
RNABERT	0.747	0.747	0.757	0.750	0.727	0.951
RNA-MSM	0.933	0.933	0.935	0.934	0.928	0.995
RNA-FM	0.972	0.972	0.972	0.972	0.969	0.999
structRFM _F ⁻	0.862	0.862	0.876	0.863	0.851	0.989
structRFM ⁻	0.946	0.946	0.948	0.946	0.942	0.997
structRFM _F	0.876	0.876	0.876	0.875	0.866	0.991
structRFM	0.973	0.973	0.973	0.973	0.970	0.999

Supplementary References

- [1] Sweeney, B. A. *et al.* Rnacentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res.* **49**, D212–D220 (2021).
- [2] Kalvari, I. *et al.* Non-coding rna analysis using the rfam database. *Current protocols in bioinformatics* **62**, e51 (2018).
- [3] Kalvari, I. *et al.* Rfam 14: expanded coverage of metagenomic, viral and microrna families. *Nucleic acids research* **49**, D192–D200 (2021).
- [4] Sloma, M. F. & Mathews, D. H. Exact calculation of loop formation probability identifies folding motifs in rna secondary structures. *RNA* **22**, 1808–1818 (2016).
- [5] Berman, H. M. *et al.* The protein data bank. *Nucleic acids research* **28**, 235–242 (2000).
- [6] Tan, Z., Fu, Y., Sharma, G. & Mathews, D. H. Turbofold ii: Rna structural alignment and secondary structure prediction informed by multiple homologs. *Nucleic acids research* **45**, 11570–11581 (2017).
- [7] Danaee, P. *et al.* bprna: large-scale automated annotation and analysis of rna secondary structure. *Nucleic acids research* **46**, 5381–5394 (2018).
- [8] Das, R. *et al.* Assessment of three-dimensional rna structure prediction in casp15. *Proteins: Structure, Function, and Bioinformatics* **91**, 1747–1770 (2023).
- [9] Cruz, J. A. *et al.* Rna-puzzles: a casp-like evaluation of rna three-dimensional structure prediction. *Rna* **18**, 610–625 (2012).
- [10] Miao, Z. *et al.* Rna-puzzles round iv: 3d structure predictions of four ribozymes and two aptamers. *Rna* **26**, 982–995 (2020).
- [11] Jaganathan, K. *et al.* Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548 (2019).
- [12] Chen, C.-K. *et al.* Structured elements drive extensive circular rna translation. *Molecular Cell* **81**, 4300–4318.e13 (2021).
- [13] Fiannaca, A., La Rosa, M., La Paglia, L., Rizzo, R. & Urso, A. nrc: non-coding rna classifier based on structural features. *BioData mining* **10**, 1–18 (2017).
- [14] Wang, N. *et al.* Multi-purpose rna language modelling with motif-aware pre-training and type-guided fine-tuning. *Nature Machine Intelligence* **6**, 548–557 (2024).