

Supplementary Information

A Large-Scale Foundation Model for RNA Enables Diverse Function and Structure Prediction

Shuxian Zou^{1,2*}, Tianhua Tao^{1,4*}, Sazan Mahbub^{1,3*}, Caleb N. Ellington¹
Robin Algayres[†], Dian Li¹, Yonghao Zhuang^{1,3*}, Hongyi Wang¹
Le Song^{1,2‡}, Eric P. Xing^{1,2,3‡}

¹GenBio AI

²Mohamed bin Zayed University of Artificial Intelligence

³Carnegie Mellon University

⁴University of Washington

April 14, 2025

*Work done during internship at GenBio AI.

†Work done while at GenBio AI.

‡Corresponding authors: le.song@genbio.ai, eric.xing@genbio.ai.

FM	Architecture	Model size	Pre-training dataset	Tokenization	Pre-training objective
RNABERT (1)	encoder-only transformer	<10M	RNAcentral (76,237 human-derived small ncRNA sequences) & Rfam alignment	base	MLM(15%) + others
RNA-FM (2)	encoder-only transformer	100M	RNAcentral v19(23M ncRNA sequences)	base	MLM(15%)
RNA-MSM (3)	MSA transformer	~200M	Rfam v14.7	base	MLM(20%)
Uni-RNA (4)	encoder-only transformer	≤ 400 M	RNAcentral & nt & Genome Warehouse (1B potential RNA sequences)	base	MLM
SpliceBERT (5)	encoder-only transformer	20M	2M pre-mRNA sequences from UCSC genome browser	base	MLM (15%)
CodonBERT (6)	encoder-only transformer	~100M	NCBI (10M mRNA coding sequences)	codon	MLM(15%) + others
3UTRBERT (7)	encoder-only transformer	~100M	86k 3'UTRs curated from GRCh38.p13, Release 40	k-mer	MLM (15%)
UTR-LM (8)	encoder-only transformer	<10M	Ensembl & eGFP & mCherry & Cao (700K 5'UTRs)	base	MLM + others
ATOM-1 (9)	encoder-decoder transformer	unknown	chemical mapping data	base	unknown
RiNALMo (10)	encoder-only transformer	650M	RNAcentral & Rfam & nt & Ensembl (36M unique ncRNA sequences)	base	MLM
CaLM (11)	encoder-only transformer	86M	European Nucleotide Archive (9M cDNA sequences)	codon	MLM (25%)
RNAErnie (12)	encoder-only transformer	~100M	RNAcentral (23M ncRNA sequences)	base	MLM(15%) + others
AIDO.RNA (ours)	encoder-only transformer	1.6B	RNAcentral v24.0 (42M ncRNA sequences)	base	MLM (15%)

Supplementary Table 1 | Related work of RNA foundation models.

RNA type	# train	# valid	# test	train ratio	valid ratio	test ratio
rRNA	28,913,600	10,000	10,000	69.7%	12.5%	12.5%
tRNA	3,543,484	10,000	10,000	8.5%	12.5%	12.5%
lncRNA	3,531,218	10,000	10,000	8.5%	12.5%	12.5%
misc_RNA	2,463,079	10,000	10,000	5.9%	12.5%	12.5%
sRNA	468,084	5,000	5,000	1.1%	6.3%	6.3%
pre_miRNA	360,711	5,000	5,000	0.9%	6.3%	6.3%
ncRNA	348,184	5,000	5,000	0.8%	6.3%	6.3%
snRNA	302,658	5,000	5,000	0.7%	6.3%	6.3%
snoRNA	281,057	5,000	5,000	0.7%	6.3%	6.3%
piRNA	209,806	5,000	5,000	0.5%	6.3%	6.3%
SRP_RNA	200,854	5,000	5,000	0.5%	6.3%	6.3%
others	879,659	5,000	5,000	2.1%	6.3%	6.3%
total	41,502,394	80,000	80,000	100.0%	100.0%	100.0%

Supplementary Table 2 | RNA type distribution of training, validation, and test set.

RNA family	Mean length	Total count
5S rRNA	119	1283
SRP RNA	180	918
tRNA	77	557
tmRNA	366	462
RNase P RNA	332	454
Group I Intron	375	74
16S rRNAa	317	67
Telomerase RNA	438	35
23S rRNAa	326	15
Mean	281	429
Total	-	3865

Supplementary Table 3 | Family-wise statistics of RNA sequences in filtered ArchiveII.

PDB ID	Description	gRNAde	gRNAde + AIDO.RNA-cDiffusion
1CSL	RRE high affinity site	50.0	53.846
1ET4	Vitamin B12 binding RNA aptamer	42.857	45.714
1F27	Biotin-binding RNA pseudoknot	50.0	44.444
1L2X	Viral RNA pseudoknot	59.259	62.963
1LNT	RNA internal loop of SRP	65.0	60.0
1Q9A	Sarcin/ricin domain from E.coli 23S rRNA	88.889	92.593
4FE5	Guanine riboswitch aptamer	41.791	46.269
1X9C	All-RNA hairpin ribozyme	40.0	48.333
1XPE	HIV-1 B RNA dimerization initiation site	60.87	65.217
2GCS	Pre-cleavage state of glmS ribozyme	45.902	47.541
2GDI	Thiamine pyrophosphate-specific riboswitch	58.974	60.256
20EU	Junctionless hairpin ribozyme	38.095	33.333
2R8S	Tetrahymena ribozyme P4-P6 domain	70.886	68.987
354D	Loop E from E. coli 5S rRNA	60.0	65.0
Average		55.18	56.74

Supplementary Table 4 | Inverse folding performance comparison on 14 RNA structures of interest identified by Das *et al.* (13). All of these 14 RNAs belong to the test set. The results by gRNAde (14) were produced by the model checkpoint provided in the official github repository (source: <https://github.com/chaitjo/geometric-rna-design>, checkpoint file name gRNAde_ARv1_1state_all.h5. Last accessed: Sept 25, 2024).

Task	Dataset	Metric	RiNALMo	AIDO.RNA-650M (ours)
Secondary structure	bpRNA TS0	F1-score	0.772	0.778
	Archive-II	F1-score	0.720	0.743
Translation efficiency	Muscle	Spearman CC	0.711	0.697
	PC3	Spearman CC	0.698	0.699
	HEK	Spearman CC	0.661	0.664
Expression level	Muscle	Spearman CC	0.698	0.688
	PC3	Spearman CC	0.672	0.683
	HEK	Spearman CC	0.697	0.707

Supplementary Table 5 | Comparision of downstream task performance between AIDO.RNA-650M and RiNALMo.

References

1. M. Akiyama, Y. Sakakibara, *NAR genomics and bioinformatics* **4**, lqac012 (2022).
2. J. Chen, *et al.*, *bioRxiv* pp. 2022–08 (2022).
3. Y. Zhang, *et al.*, *bioRxiv* pp. 2023–03 (2023).
4. X. Wang, *et al.*, *bioRxiv* pp. 2023–07 (2023).
5. K. Chen, *et al.*, *bioRxiv* pp. 2023–01 (2023).
6. S. Li, *et al.*, *bioRxiv* pp. 2023–09 (2023).
7. Y. Yang, *et al.*, *bioRxiv* pp. 2023–09 (2023).
8. Y. Chu, *et al.*, *Nature Machine Intelligence* pp. 1–12 (2024).
9. N. Boyd, *et al.*, *bioRxiv* pp. 2023–12 (2023).
10. R. Josip Penić, T. Vlašić, R. G. Huber, Y. Wan, M. Šikić, *arXiv e-prints* pp. arXiv–2403 (2024).
11. C. Outeiral, C. M. Deane, *Nature Machine Intelligence* **6**, 170 (2024).
12. N. Wang, *et al.*, *Nature Machine Intelligence* pp. 1–10 (2024).
13. R. Das, J. Karanicolas, D. Baker, *Nature methods* **7**, 291 (2010).
14. C. K. Joshi, *et al.*, *bioRxiv* (2024).