From Base Pairs to Functions: Rich RNA Representations via Multimodal Language Modeling

Joel Bonnie[1,2*†] **Tin Vlašić**[1*‡] **Mile Šikić**[1,3]
[1]Genome Institute of Singapore [2]University of British Columbia [3]University of Zagreb
tin_vlasic@a-star.edu.sg

Abstract

RNA foundation models have recently emerged as powerful tools for learning from large sequence databases, yet their embeddings often fall short in simple probing setups, necessitating additional finetuning. Most existing models are pretrained solely on sequences, assuming that structural information will emerge implicitly. We introduce RABONA, a multimodal RNA language model jointly pretrained on sequence-structure pairs using modality-specific masking, designed for both generative and understanding tasks. It produces embeddings that form clearer family-specific clusters and shows stronger attention alignment with RNA base pairs compared to other RNA language models. In this paper, we focus on RABONA's predictive capabilities and show that it consistently outperforms larger baselines across diverse downstream tasks in both finetuning and linear probing setups, demonstrating that incorporating structure during pretraining yields richer RNA embeddings and enables more efficient foundation models.

1 Introduction

RNA is a versatile biomolecule whose structure encodes the key to its diverse regulatory functions in biological systems and cellular processes [Doudna and Cech, 2002, Morris and Mattick, 2014]. Due to the scarcity of RNA structural data [Schneider et al., 2023] and an enormous amount of unlabeled RNA sequence databases, RNA foundation models emerged in an attempt to crack the code of RNA. However, to take full advantage of them, RNA foundation models often need to be finetuned on small curated datasets specialized in scope.

The existing general-purpose RNA foundation models, such as RNA-FM [Chen et al., 2022], Uni-RNA [Wang et al., 2023], RiNALMo [Penić et al., 2025], and AIDO.RNA [Zou et al., 2024], are all encoder-only Transformers [Vaswani et al., 2017] pretrained on non-coding RNA (ncRNA) sequences only, using vanilla BERT-style masked language modeling [Devlin et al., 2019]. These models differ slightly in architectural choices and mostly in size, ranging from 100M to 1.6B parameters. Most often, they were pretrained using different versions of the RNAcentral database [RNAcentral Consortium, 2021] and sometimes augmented with additional smaller databases. When finetuned, RNA foundation models often achieve state-of-the-art results on various downstream tasks, from secondary structure to splice-site prediction. However, when employed in a linear probing setup, which denotes training only a lightweight prediction head for the downstream task while keeping the encoder parameters frozen, they often underperform. This questions the quality of existing RNA foundation models' embeddings and whether there is a way to obtain more versatile and richer sequence representations.

*Equal contribution.
†Work done during an internship at the Genome Institute of Singapore.
‡Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences.

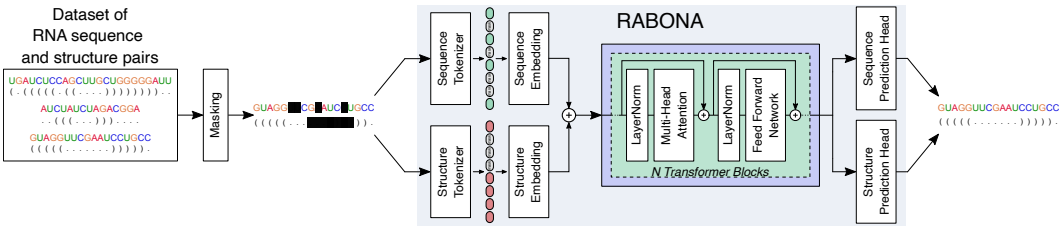

Figure 1: We pretrained RABONA on a collected dataset of RNA sequence and secondary structure pairs. After masking, the input sequence and structure pass through separate tokenizer and embedding layers before being element-wise added. The combined embeddings are processed by the Transformer encoder and passed through separate sequence and structure prediction heads.

We propose RABONA, a multimodal RNA language model, pretrained on more than 1M ncRNA sequence and secondary structure pairs obtained from expert databases and using R2DT [McCann et al., 2025]. RABONA's architecture and pretraining were designed to support both generative and understanding tasks, and in this paper, we focus on leveraging its representations to predict RNA function and behavior. RABONA, pretrained on a fraction of the data compared to the aforementioned RNA foundation models, yields richer representations, outperforming even $5\times$ larger models on most of the downstream tasks, both finetuned or employed in a linear probing setup.

2 Multimodal RNA Language Model

RABONA is a multimodal RNA language model pretrained on ncRNA sequences and their corresponding secondary structures. The sequence-structure pairs were partially collected from the bpRNA-1m [Danaee et al., 2018], ArchiveII [Mathews, 2019], and RNAStrAlign [Tan et al., 2017] databases, and partially by predicting secondary structures for the sequences from the Rfam 14.7 database [Kalvari et al., 2021] using R2DT [McCann et al., 2025]. That gave us around 1M sequence-structure pairs that were preprocessed by removing sequences outside the $[8, 1022]$ nucleotide range, removing sequence-structure duplicates, and clustering. Data preprocessing is explained in detail in Appendix A.

Inspired by the ESM3 multimodal pretraining approach [Hayes et al., 2025], we separately masked each modality. Sequence tokens are masked independently and identically, with the masking probability sampled from a $\beta(3, 7)$ distribution. For structure tokens, masking is performed in two modes: in 20% of cases the entire structure is masked, while in 80% of cases a single contiguous span is masked such that the masked fraction of tokens follows $PDF_{ss}(x) = 2x$ for $x \in [0, 1]$. RABONA pretraining is illustrated in Figure 1. RABONA is a 33.5M parameter language model with 12 Transformer blocks. Its 2-layer multi-layer perception (MLP) prediction heads are used to separately reconstruct the masked sequence and structure tokens from 480-dimensional output embeddings. More details on pretraining and its parameters are given in Appendix A.

We compared classification token embeddings from RiNALMo-33M, RNA-FM, and our RABONA on 21 ncRNA families from the Rfam database, visualized with t-SNE in Figure 2. RABONA's embeddings form clearer clusters than RiNALMo-33M and RNA-FM, reflecting its ability to capture family-specific structural and functional properties beyond sequence, yielding informative representations. In the same figure, we show an example where RABONA's attention weights align most strongly with RNA base-pairs. By contrast, RiNALMo-33M does not capture rRNA structure, while the $3\times$ larger RNA-FM shows partial base-pair awareness. Additional analyses in Appendix B demonstrate that the $5\times$ larger RiNALMo-150M begins to recover rRNA structure signals. These results support our approach of explicitly providing secondary structures during pretraining, which leads to more effective training and parameter utilization and yields more informative embeddings.

3 Results

RABONA's embeddings amplify predictive performance while fine-tuning or linear probing for several essential downstream tasks. We evaluated two modes of operation: when we provide the model with both sequence and secondary structure, and when we provide the model with only the

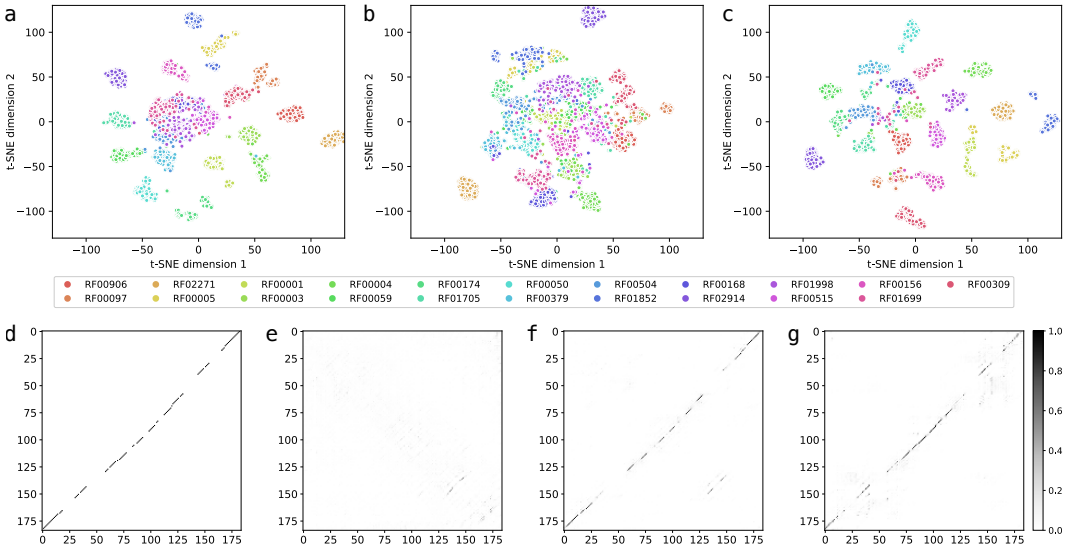

Figure 2: t-SNEs of Rfam sequence embeddings from RiNALMo (a), RNA-FM (b), and RABONA (c). Common base-pair matrix (d) of 50 unique rRNAs and their average attentions from RiNALMo (e), RNA-FM (f), and RABONA (g), computed over the heads most aligned with base-pairs.

sequence. We compare its performance with state-of-the-art language models such as RiNALMo-33M and RiNALMo-150M [Penić et al., 2025], as well as RNA-FM [Chen et al., 2022]. More detailed explanations for data preparation, model training, and evaluation can be found in Appendices C-F.

Fine-tuning RABONA for classification tasks. RABONA can be fine-tuned to predict RNA function through Rfam family classification. This can be modeled as a multi-class classification task. We experimented with 0% boundary noise, that is, the original sequence, as well as 200% boundary noise, which added the original sequence length of random nucleotides to either side. Secondary structures for RABONA were predicted from the input sequence using RNAFold [Lorenz et al., 2011]. We report models' accuracy for both 0% (ACC_0) and 200% (ACC_200) boundary noise.

When fine-tuned, RABONA performs comparably to other state-of-the-art foundational models (see Table 1). However, RABONA representations excel when only a simple two-layer MLP is trained, seen in Table 2, resulting in a significant increase in accuracy. This stems from RABONA's enhanced understanding of RNA functions and structures, resulting in richer embeddings. What we find interesting is the increase in accuracy when replacing predicted RNAFold structures with `[MASK]` tokens (RABONA_MASK). RABONA performs remarkably even without input secondary structures. The decrease in performance when using predicted secondary structures highlights the necessity of accurate secondary structures, as incorrect structures pose the risk of misleading RABONA. Accuracy for RABONA with 200% noise is similar to RiNALMo-33M and ahead of RNA-FM when finetuned. Linear probing with noise leads to decreased accuracy, which may stem from the random nucleotides prompting RNAFold to predict ill-formed secondary structures.

Table 1: ncRNA Classification - Finetuning

MODEL	ACC_0 ↑	ACC_200 ↑
RABONA	0.976	0.979
RABONA_MASK	0.975	0.974
RiNALMo-33M	0.980	0.977
RiNALMo-150M	0.982	0.985
RNA-FM	0.919	0.951

Table 2: ncRNA Classification - Linear Probing

MODEL	ACC_0 ↑	ACC_200 ↑
RABONA	0.909	0.488
RABONA_MASK	0.921	0.537
RiNALMo-33M	0.861	0.541
RiNALMo-150M	0.896	0.541
RNA-FM	0.791	0.462

Fine-tuning RABONA for predicting nucleotide reactivity. Chemical reactivity of RNA nucleotides is closely associated with the secondary and tertiary structures into which the molecule

folds. To achieve optimal predictive performance, model representations must capture a robust understanding of RNA structure. RABONA leverages a subset of the Ribonanza dataset [He et al., 2024] derived from Rfam to predict per-nucleotide reactivity values. Reactivity values per nucleotide were predicted for two chemical probing reagents: 2A3 and DMS. A simple MLP with one hidden layer was used as the prediction head to shift importance to the large language model representations. Reactivity prediction results are given in Table 3.

RABONA outperforms other foundational models in both the respective frozen and unfrozen cases. Highlighting RABONA embeddings further, we notice RABONA, even with masked out secondary structures, still outperforms RiNALMo-33M and RNA-FM. A comparison of foundational models for linear probing can be found in Appendix D.

Table 3: Reactivity - Finetuning

MODEL	RMSE ↓	MAE ↓
RABONA	0.413	0.255
RABONA_MASK	0.427	0.262
RiNALMo-33M	0.438	0.271
RiNALMo-150M	0.427	0.264
RNA-FM	0.471	0.287

Table 4: OpenVaccine - Finetuning

MODEL	RMSE ↓	MAE ↓
RABONA	0.306	0.169
RABONA_MASK	0.376	0.215
RABONA_FROZEN	0.365	0.198
RiNALMo-33M	0.395	0.230
RiNALMo-150M	0.375	0.215
RNA-FM	0.408	0.231

Fine-tuning RABONA for degradation prediction. RABONA can be fine-tuned to predict nucleotide degradation and reactivity values using the OpenVaccine dataset [Wayment-Steele et al., 2022]. The three prediction targets comprised reactivity values for structure inference, as well as for predicting the likelihood of degradation after Magnesium incubation under either high temperature (50 degrees Celsius) or alkaline conditions (pH 10). Again, we opted to use a simple one-hidden-layer MLP as the prediction head to emphasize LLM embeddings.

RABONA's representations excel, significantly outperforming other foundational models as shown in Table 4. Compared to RiNALMo-150M, a model 5 times larger, RABONA achieves an 18% decrease in RMSE and a 21% decrease in MAE. RABONA during linear probing surpasses not only the other foundational models under the same setting, but even their fine-tuned counterparts. These results again illustrate the immense potential of RABONA's representations, enriched through multimodal language modeling.

Fine-tuning RABONA for MRL prediction. Mean Ribosome Load (MRL) is a regression task associated with translational efficiency. Similar to the ncRNA classification task, we used RNAFold to predict secondary structures from the primary nucleotide sequence. For the prediction head, we used a 1D ResNet, as in RiNALMo.

Table 5: Ribosome Loading - Finetuning

MODEL	R^2 ↑	MAE ↓
RABONA	0.769	0.409
RABONA_MASK	0.772	0.406
RiNALMo-33M	0.811	0.377
RiNALMo-150M	0.844	0.342
RNA-FM	0.719	0.455

As shown in Table 5, RABONA underperforms RiNALMo but outperforms RNA-FM. Its pretraining was limited to 1M ncRNA sequences without mRNA untranslated regions (UTRs) and their structures. Additionally, UTR folds were approximated using RNAFold, as the ground-truth structures were not provided in the dataset, whose reliability is limited. This task highlights that accurate secondary structures, whether during pretraining, fine-tuning, or inference, are crucial for RABONA's optimal performance.

4 Conclusion and Future Work

We introduced RABONA, a structure-aware multimodal RNA language model whose rich representations prove useful across multiple downstream tasks, outperforming larger baselines. RABONA shows particular strength in linear probing tasks where other models struggle to perform well.

While this is still a work in progress, our results establish a strong foundation for future work. We plan to expand training with additional sequence-structure datasets and scale up the model to better exploit this data. This will lead to better generalization capabilities in downstream tasks such as MRL prediction. Beyond discriminative tasks, we will explore RABONA's generative capabilities, such as secondary structure prediction and inverse folding, which are naturally supported by its design.

Acknowledgments and Disclosure of Funding

J.B. acknowledges support from the Agency for Science, Technology and Research (A*STAR) through the Singapore International Pre-Graduate Award (SIPGA). T.V. acknowledges support from the Genome Institute of Singapore (GIS) through the GIS Early Career Researcher Grant Award. T.V. and M.Š. acknowledge support from the Croatian Science Foundation under Project *Deep Learning-Based RNA Tertiary Structure Prediction and Generation* (IP-2024-05-1554).

The authors would like to thank Rafael Josip Penić, Sara Bakić, and Ivona Martinović for fruitful discussions and useful comments during the development of this work.

References

J. Chen, Z. Hu, S. Sun, Q. Tan, Y. Wang, Q. Yu, L. Zong, L. Hong, J. Xiao, T. Shen, et al. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *arXiv preprint arXiv:2204.00300*, 2022.

P. Danaee, M. Rouches, M. Wiley, D. Deng, L. Huang, and D. Hendrix. bpRNA: large-scale automated annotation and analysis of RNA secondary structure. *Nucleic Acids Research*, 46(11): 5381–5394, 2018.

T. Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

R. Das, H. Wayment-Steele, D. S. Kim, C. Choe, B. Tunguz, W. Reade, and M. Demkin. OpenVaccine: COVID-19 mRNA Vaccine Degradation Prediction. `https://kaggle.com/competitions/stanford-covid-vaccine`, 2020. Kaggle.

R. Das, S. He, R. Huang, J. Townley, R. Kretsch, T. Karagianes, J. Nicol, G. Nye, C. Choe, J. Romano, M. Demkin, W. Reade, and E. players. Stanford Ribonanza RNA Folding. `https://kaggle.com/competitions/stanford-ribonanza-rna-folding`, 2023. Kaggle.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional Transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

J. A. Doudna and T. R. Cech. The chemical repertoire of natural ribozymes. *Nature*, 418(6894): 222–228, 2002.

T. Hayes, R. Rao, H. Akin, N. J. Sofroniew, D. Oktay, Z. Lin, R. Verkuil, V. Q. Tran, J. Deaton, M. Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387 (6736):850–858, 2025.

S. He, R. Huang, J. Townley, R. C. Kretsch, T. G. Karagianes, D. B. Cox, H. Blair, D. Penzar, V. Vyaltsev, E. Aristova, et al. Ribonanza: deep learning of RNA structure through dual crowdsourcing. *bioRxiv*, 2024.

I. Kalvari, E. P. Nawrocki, N. Ontiveros-Palacios, J. Argasinska, K. Lamkiewicz, M. Marz, S. Griffiths-Jones, C. Toffano-Nioche, D. Gautheret, Z. Weinberg, et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research*, 49(D1):D192–D200, 2021.

R. Lorenz, S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. ViennaRNA Package 2.0. *Algorithms for molecular biology*, 6(1):26, 2011.

D. H. Mathews. How to benchmark RNA secondary structure prediction accuracy. *Methods*, 162: 60–67, 2019.

H. McCann, C. D. Meade, L. D. Williams, A. S. Petrov, P. Z. Johnson, A. E. Simon, D. Hoksza, E. P. Nawrocki, P. P. Chan, T. M. Lowe, et al. R2DT: a comprehensive platform for visualizing RNA secondary structure. *Nucleic Acids Research*, 53(4):gkaf032, 2025.

K. V. Morris and J. S. Mattick. The rise of regulatory RNA. *Nature Reviews Genetics*, 15(6):423–437, 2014.

T. M. R. Noviello, F. Ceccarelli, M. Ceccarelli, and L. Cerulo. Deep learning predicts short non-coding rna functions from only raw sequence data. *PLoS computational biology*, 16(11):e1008415, 2020.

R. J. Penić, T. Vlašić, R. G. Huber, Y. Wan, and M. Šikić. RiNALMo: General-purpose RNA language models can generalize well on structure prediction tasks. *Nature Communications*, 16(1): 5671, 2025.

RNAcentral Consortium. RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Research*, 49(D1):D212–D220, 2021.

P. J. Sample, B. Wang, D. W. Reid, V. Presnyak, I. J. McFadyen, D. R. Morris, and G. Seelig. Human 5' utr design and variant effect prediction from a massively parallel translation assay. *Nature biotechnology*, 37(7):803–809, 2019.

B. Schneider, B. A. Sweeney, A. Bateman, J. Cerny, T. Zok, and M. Szachniuk. When will RNA get its AlphaFold moment? *Nucleic Acids Research*, 51(18):9522–9532, 2023.

N. Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

M. Steinegger and J. Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, 2024. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.127063. URL `https://www.sciencedirect.com/science/article/pii/S0925231223011864`.

Z. Tan, Y. Fu, G. Sharma, and D. H. Mathews. TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs. *Nucleic Acids Research*, 45(20):11570–11581, 2017.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

X. Wang, R. Gu, Z. Chen, Y. Li, X. Ji, G. Ke, and H. Wen. UNI-RNA: universal pre-trained models revolutionize RNA research. *bioRxiv*, pages 2023–07, 2023.

H. K. Wayment-Steele, W. Kladwang, A. M. Watkins, D. S. Kim, B. Tunguz, W. Reade, M. Demkin, J. Romano, R. Wellington-Oguri, J. J. Nicol, et al. Deep learning models for predicting RNA degradation via dual crowdsourcing. *Nature Machine Intelligence*, 4(12):1174–1184, 2022.

S. Zou, T. Tao, S. Mahbub, C. N. Ellington, R. Algayres, D. Li, Y. Zhuang, H. Wang, L. Song, and E. P. Xing. A large-scale foundation model for RNA function and structure prediction. *bioRxiv*, pages 2024–11, 2024.

A Multimodal RNA Language Modeling

Data Preprocessing We collected around 145K ncRNA sequence-structure pairs from publicly available datasets ArchiveII [Mathews, 2019], bpRNA-1m [Danaee et al., 2018], and RNAStrAlign [Tan et al., 2017]. We augmented the data with around 1.4M ncRNA sequence-structure pairs obtained from the Rfam 14.7 dataset and using the R2DT secondary structure prediction tool [McCann et al., 2025].

We implemented a multistep preparation pipeline. First, we removed sequences outside the $[8, 1022]$ nucleotide range. Furthermore, we removed examples that have identical both sequence and structure using `seqkit rmdup`. This resulted in $1,006,320$ unique sequence-structure pairs. We then clustered the sequences from the sequence-structure pairs with MMSeqs2 [Steinegger and Söding, 2017] using `mmseqs easy-cluster` and the following options `--min-seq-id 0.95` and `-c 0.8`. This resulted in $623,778$ clusters. We divided the clusters into training and validation clusters with a $99 - 1$ train-validation split. Our sampling strategy was controlled such that at least one high-quality sequence-structure pair from either ArchiveII, bpRNA-1m, or RNAStrAlign had to exist in each cluster ending in the validation set. This way, we ensured that the validation set is quality-controlled and sequentially different from the training set, allowing us to evaluate the model properly during pretraining. Furthermore, we merged the training set and again performed MMSeqs2 clustering of only the training data, this time with a minimum sequence identity of 0.99 and coverage 0.9. In the end, we were left with $996,857$ training samples clustered into $943,522$ clusters and $9,463$ validation samples clustered into $6,237$ clusters. During training epochs, we randomly sampled a single sequence-structure pair from each training cluster to ensure sequence diversity in each batch and balance the data. Our validation set was sampled such that it consisted of only high-quality sequence-structure pairs from either ArchiveII, bpRNA-1m, or RNAStrAlign.

Tokenization We used separate tokenizers for RNA sequence and structure. During sequence tokenization, each nucleotide was treated as a single token. We replaced all "U"s in the sequences with "T"s, and our sequence vocabulary consisted of the following standard nucleotide codes: "A", "T", "G", "C", and "N", where "N" stands for "any nucleotide" token. The sequence vocabulary additionally comprised the following special tokens: `[CLS]`, `[EOS]`, `[PAD]`, and `[MASK]`.

All RNA secondary structures were denoted in a dot-bracket format, perfectly suited for structure tokenization. Each character in the structure string corresponds to a nucleotide in the sequence and was treated as a single token. Dots "." indicate unpaired bases, while an opening parenthesis "(" indicates a paired base, and its corresponding closing parenthesis ")" shows the base it is paired with. The structure vocabulary consisted of the following codes: ".", "(", ")", "[", "]", "{", and "}", where the last two matched bracket types were extensions of the original notation to allow representation of pseudoknots. Similar to the sequence vocabulary, structure vocabulary additionally comprised the following special tokens: `[CLS]`, `[EOS]`, `[PAD]`, and `[MASK]`.

During the masking procedure, we changed standard nucleotides from both vocabularies with the modality-specific `[MASK]` tokens. Tokens `[CLS]` and `[EOS]`, from the sequence and structure vocabulary, were added at the beginning and end of the sequence and structure, respectively. The `[PAD]` tokens were appended at the end of shorter sequences and the corresponding structures to have all the sequence-structure pairs in a batch of the same length.

Language Model Architecture We adapted the encoder-only Transformer architecture from Penić et al. [2025] for their RiNALMo-33M model. First, sequence and structure are tokenized using separate sequence and structure tokenizers and turned into 480-dimensional vector using separate input embedding layers. Sequence and structure input embeddings are element-wise summed before being passed to the Transformer encoder. The Transformer comprises 12 Transformer blocks, each consisting of a multi-head attention with 20 heads and a feed-forward network (FFN). Similar to RiNALMo-33M, we employed RoPE [Su et al., 2024], SwiGLU activation function [Shazeer, 2020], and FlashAttention-2 [Dao, 2023]. The hidden size of the FFN layers was set to $1,280$. The residual connections and layer normalizations are integrated as illustrated in Figure 1. RABONA employs two prediction heads, one for sequence and one for structure, allowing their independent prediction from the output embeddings.

Pretraining Our pretraining strategy was inspired by the ESM3 pretraining approach [Hayes et al., 2025]. We separately masked sequence and structure modalities and pretrained RABONA using masked language modeling. As explained in Section 2, we employed different masking strategies for sequence and structure. For each sequence and structure in a batch, we independently chose their masking probabilities.

The maximum context of RABONA was set to $1,024$ tokens, reserved for the classification `[CLS]` token, $1,022$ sequence and structure tokens, and the end-of-sequence `[EOS]` token. During pretraining, we randomly sampled each sequence-structure pair in a batch from a different training cluster. Thus, in each epoch, RABONA saw $943,522$ sequence-structure pairs.

We pretrained RABONA using a single H100 GPU of 80 GB for 55 hours. The batch size was set to 512, and the total number of steps to $100,000$. In contrast, RiNALMo was pretrained for $77,000$ steps with a batch size of $1,344$. We adopted the cosine annealing learning rate schedule with a linear warm-up. During the warm-up period, the learning rate increases from 0 to 10^{-4} for $2,000$ steps. For the cosine annealing schedule, the minimum learning rate was set to 10^{-5}.

B Additional Attention Weights Analyses

In Section 2, we showed that RABONA's attention heads are the best aligned with base-pairs when compared to the same size RiNALMo-33M and the $3\times$ larger RNA-FM. We noticed that the $5\times$ larger RiNALMo-150M exhibits base-pair awareness, as shown in Figure 3. As we can see, RiNALMo-150M is able to understand the secondary structure implicitly; however, it requires $5\times$ more parameters to achieve this.

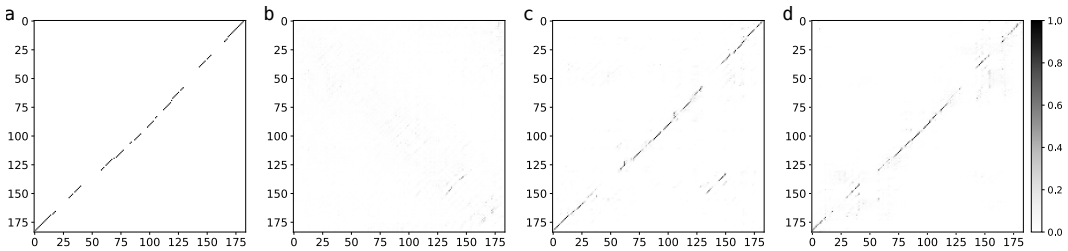

Figure 3: Common base-pair matrix (a) of 50 unique rRNAs and their average attentions from RiNALMo-33M (b), RiNALMo-150M (c), and RABONA (d), computed over the heads most aligned with base-pairs.

By explicitly providing secondary structure and leveraging it during self-supervised pretraining, RABONA is forced to learn the correct base pairing. This way, we get a structure-aware RNA language model for one-fifth of the parameters of the large RiNALMo-150M model.

We provide additional attention-weight visualizations for 50 unique RNAs with the same secondary structure from the tRNA family in Figure 4.

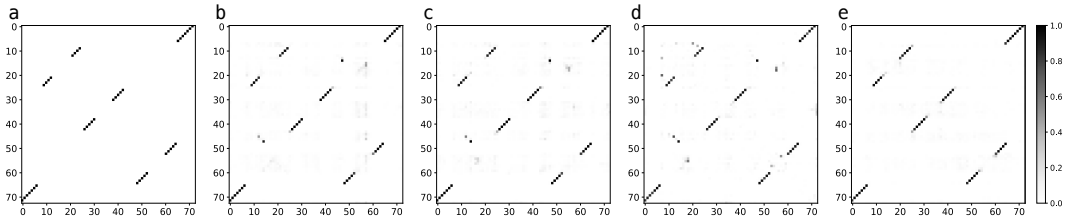

Figure 4: Common base-pair matrix (a) of 50 unique tRNAs and their average attentions from RiNALMo-33M (b), RiNALMo-150M (c), RNA-FM (d), and RABONA (e) computed over the heads most aligned with base-pairs.

We see that for tRNAs, which are shorter and usually most prevalent in ncRNA databases, even RiNALMo-33M captures base-pairing signals. We conclude that smaller RNA language models are

good enough to capture local information between a few tens of neighboring nucleotides; however, they struggle with longer dependencies. We support this with additional visualizations of attention weights for tmRNA family examples in Figure 5.

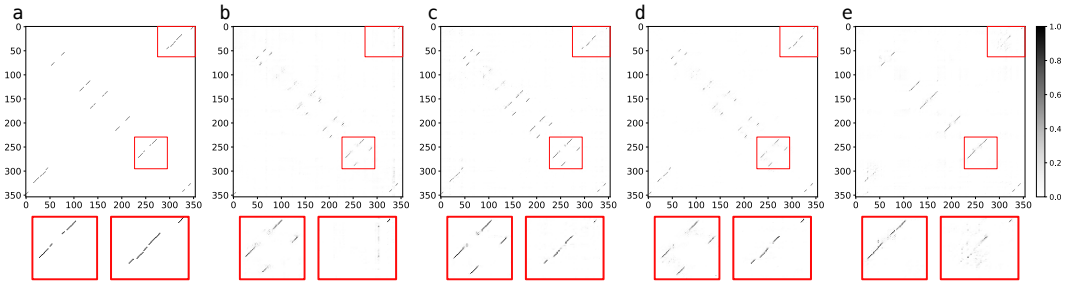

Figure 5: Common base-pair matrix (a) of 25 unique tmRNAs and their average attentions from RiNALMo-33M (b), RiNALMo-150M (c), RNA-FM (d), and RABONA (e) computed over the heads most aligned with base-pairs. The red rectangles below the subfigures are zoomed-in regions of interest in the attention maps.

From the top-right zoomed-in region, we observe that smaller models, particularly RiNALMo-33M, struggle to capture long-range dependencies, such as those in tmRNA, where the 5' end pairs with the 3' end nearly 300 nucleotides apart. In contrast, RiNALMo-150M and RNA-FM successfully model these distant interactions. The bottom-right zoomed-in region further shows that sequence-only RNA language models tend to focus on shorter, fragmented nucleotide relations that misalign with the true base-pairing pattern. RABONA, by explicitly incorporating secondary structure, aligns its attention with base pairs, producing richer, structure-aware RNA representations.

C ncRNA Classification

Dataset preparation for non-coding RNA classification followed the setup used in RiNALMo, specified by Noviello et al. [2020]. The dataset, obtained from Rfam, contained a total of $306,016$ sequences across 88 target classes after processing. An $84-8-8$ train-validation-test split was applied for each family, and the train and validation sets were resampled to address class imbalance. The final dataset contained $105,864$ training, $17,324$ validation and $25,342$ test sequences. We fine-tuned the model on two versions of the dataset: one with 0% noise, that is, the original sequences, and one with 200% noise. For the 200% noise dataset, random nucleotides equal in length to the original sequence were appended to both sides of each sequence.

The prediction head was a multi-layer perceptron (MLP) with a hidden layer dimension of 128 and a GELU activation. The model was fine-tuned for 25 epochs with a batch size of 64, using a constant learning rate of 10^{-5}. For linear probing, the learning rate was linearly decreased from 5×10^{-4} to 10^{-5} over 7500 steps and stayed constant for the rest of the training. Optimization employed AdamW, applying a weight decay of 0.01. Cross-entropy loss was used as the prediction loss. This configuration was kept consistent across the other foundational models. Model weights were selected based on the best validation performance.

D Reactivity

For reactivity prediction, nucleotide reactivity values were predicted under two probing reagents: 2A3 and DMS. An Rfam subset of the Ribonanza dataset [Das et al., 2023] was utilized for reactivity prediction. Both sequence and structure were provided in the dataset. Only entries having a signal-to-noise ratio greater than or equal to 1, and with reads greater than or equal to 100 for both probing reagents, were kept. After filtering, we had a total of $18,876$ RNA sequences of length 177. The reactivity values less than 0 in the ground truth or outside the region of interest were masked out and considered invalid. Predicted values were not clipped to be between 0 and 1. We performed a random $80-10-10$ train-validation-test split for fine-tuning.

The prediction head consists of a two-layer MLP with a hidden layer dimension of 128, and the ReLU activation function. The model was fine-tuned for $15,000$ training steps with only the prediction head being trained for the first 5 epochs. An AdamW optimizer was utilized with a weight decay of 0.01. The learning rate was linearly decayed from 10^{-3} to 2×10^{-5} over 1200 training steps and then kept constant. Huber Loss was used as the prediction loss with a delta value of 0.1. Batch sizes were set to 64. Other foundational language models were fine-tuned similarly, keeping identical prediction heads.

Among the results for linear probing, we observe a similar trend as when fine-tuned. RABONA outperforms RNA-FM and RiNALMo-150M, foundational models that are three times and five times larger, respectively. We also note the necessity of foundational models for this task, with one-hot encoded baselines having significantly larger errors.

Table 6: Reactivity - Finetuning

MODEL	RMSE ↓	MAE ↓
RABONA	0.413	0.255
RABONA_MASK	0.427	0.262
RiNALMo-33M	0.438	0.271
RiNALMo-150M	0.427	0.264
RNA-FM	0.471	0.287
One-Hot Sequence	0.551	0.336
One-Hot Sequence + SS	0.538	0.324

Table 7: Reactivity - Linear Probing

MODEL	RMSE ↓	MAE ↓
RABONA	0.478	0.290
RABONA_MASK	0.500	0.306
RiNALMo-33M	0.502	0.306
RiNALMo-150M	0.495	0.301
RNA-FM	0.511	0.312

E OpenVaccine

The three targets were reactivity values for structure inference and for predicting the degradation likelihood after Magnesium incubation at either high temperature (50 degrees Celsius) or high pH (pH 10).

Datasets were prepared by combining the public training and private test suites from Kaggle [Das et al., 2020]. Sequences were filtered using the SN_Filter variable, retaining only those with a mean signal-to-noise ratio greater than one and having a minimum target value greater than -0.5. This led to a total number of $1,589$ and $2,493$ sequence-structure pairs for the public training and private test sets, respectively. The sequences in the public training dataset were shorter (107 nucleotides) than those in the private test dataset (130 nucleotides). Sequences from the private test set had been clustered at a sequence similarity of less than 50%. From the private test set, RNA sequences from 400 singleton clusters were sampled to construct the test set. This ensured that the test set contains diverse sequences to evaluate generalization accurately. For validation, 100 clusters with two members were sampled from the private test, yielding 200 sequences in total. For training, both the public Kaggle training set and the private test suite were utilized, excluding sequences assigned for testing and validation. Each dataset was then clustered separately using MMSeqs2 with a minimum sequence identity and minimum coverage of 0.8. From each cluster, three sequences were sampled with replacement, ensuring that a representative cluster member was always selected, to form the final training set of $3,825$ sequence-structure pairs. This preprocessing approach resulted in a well-balanced dataset with sequentially distinct training, validation, and test sets to evaluate the generalization capabilities of the models.

Similar to the Reactivity prediction head, the OpenVaccine consists of a one-hidden-layer MLP with a hidden layer dimension of 128 and ReLU activation. The model was fine-tuned for $10,000$ training steps. For the first 5 epochs, RABONA was frozen, and only the prediction head was trained. We used an AdamW optimizer applying a weight decay of 0.01. The learning rate was linearly decayed from 10^{-3} to 2×10^{-5} over 300 training steps. The Huber Loss was used with a delta value of 0.1. The batch size was fixed at 64. Once again, other foundational models such as RiNALMo and RNA-FM were fine-tuned identically.

RABONA representations excel in the linear probing tests, achieving lower errors than all other frozen and fine-tuned foundational models. We once again observe one-hot encoded baselines having much higher errors, with adding secondary structures reducing RMSE by 9% and MAE by 13%.

Table 8: OpenVaccine - Finetuning

MODEL	RMSE ↓	MAE ↓
RABONA	0.306	0.169
RABONA_MASK	0.376	0.215
RiNALMo-33M	0.395	0.230
RiNALMo-150M	0.375	0.215
RNA-FM	0.408	0.231
One-Hot Sequence	0.518	0.304
One-Hot Sequence + SS	0.468	0.263

Table 9: OpenVaccine - Linear Probing

MODEL	RMSE ↓	MAE ↓
RABONA	0.365	0.198
RABONA_MASK	0.428	0.246
RiNALMo-33M	0.423	0.245
RiNALMo-150M	0.420	0.239
RNA-FM	0.434	0.251

Table 10: Mean Ribosome Load - Linear Probing

MODEL	R^2 ↑	MAE ↓
RABONA	0.641	0.512
RABONA_MASK	0.624	0.516
RiNALMo-33M	0.661	0.498
RiNALMo-150M	0.672	0.489
RNA-FM	0.654	0.498

F Mean Ribosome Load

Fine-tuning for 5' UTR Mean Ribosome Load (MRL) prediction followed procedures used in RiNALMo. Data preparation was carried out using the methodology explained by Sample et al. [2019]. The original $83,919$ UTR sequences were filtered for sufficient read coverage. We produced two evaluation datasets: Human7600 and Random7600, each containing $7,600$ sequences for human and random 5' UTRs, respectively. We used Human7600 as a test set, while Random7600 was used as a validation dataset. The training set consisted of the remaining UTR sequences. Further, MRL targets were standardized relative to the mean and standard deviation of those present in the training dataset.

The prediction head consisted of 6 ResNet blocks, with the head embedding dimension being set to 32. Each block included two 1D convolution layers, along with instance normalization and the ELU activation function. The model was fine-tuned for 50 epochs with a batch size of 64. The learning rate was linearly decayed from 10^{-4} to 10^{-5} across the first $5,000$ training steps. During the first 5 epochs, the language model was kept frozen and only the prediction head was trained. Mean Squared Error was used as the prediction loss. The same procedure was repeated for the other foundational models.