RFamLlama: an efficient conditional language model for RNA sequence generation across diverse structural families

Jinyuan Sun [1] **Han Li** [2] **Yifan Deng** [3]

Abstract

The ability to efficiently generate specific RNA sequences on demand has significant implications for both scientific research and therapeutic applications. In this context, we introduce RFamLlama, a conditional language model that is specifically optimized for generating RNA sequences across diverse families. This model was trained on RNA sequences representing over 4,000 distinct families, each augmented with control tags to denote the specific family. We have shown that the inclusion of family-specific tags substantially enhances the capabilities of our model in zero-shot fitness prediction of RNA molecules. Additionally, this model supports a conditional generation approach, allowing for the generation of RNA sequences by using Rfam IDs as input prompts, thereby eliminating the need for further functional-specific fine-tuning. Consequently, RFamLlama is poised to be an effective and widely applicable tool for the zero-shot fitness prediction and generation of RNA sequences, potentially pushing the boundaries of what can be achieved beyond natural evolutionary processes.

1. Introduction

The engineering and design of RNA with various functions are opening new avenues in synthetic molecular systems (Dykstra et al., 2022), leading to advancements in next-generation RNA-based therapeutics (Zhu et al., 2022; Pardi et al., 2018). The increasing attention towards RNA-based therapeutics has already had a significant impact on global health, as evidenced by the 2023 Nobel Prize awarded to Katalin Karikó and Drew Weissman for their work on modified nucleosides (Karikó et al., 2005), crucial for developing effective mRNA vaccines against COVID-19 (Nobel Prize Outreach AB, 2024). Traditionally, RNA design methods have relied on directed evolution (Beaudry & Joyce, 1992), which might be costly and inefficient due to only a small fraction of fitness landscape can be explored (Romero & Arnold, 2009). Hence, new methods are needed to speed up this process, such as generating a large pool of synthetic RNA molecules and using high-throughput DNA synthesis and screening techniques (Sumi et al., 2024).

Deep generative models have shown promise in creating realistic data, including text and images (Bond-Taylor et al., 2021). These models learn from real data and generate new samples accordingly. Therefore, developing a generative model for RNA to produce sequences meeting specific structural or functional criteria is feasible with enough training data. Luckily, the vast RNA sequence space sampled over millions of years of evolution (Joyce, 1989) provides ample data for training. Recent research has demonstrated that deep learning models trained on natural sequences can generate artificial ribozyme sequences, many of which are proven active through massively parallel assay (Sumi et al., 2024).

In this study, we aimed to broaden the scope from one structural family to encompass all structural families by training a language model on all classified sequences. To facilitate the controlled generation of RNA sequences belonging to desired structural families, we used Rfam IDs as prefixes for RNA sequences. We call this series of models RFamLlama, indicating their training data source (Rfam database (Kalvari et al., 2021)) and architecture (Llama (Touvron et al., 2023)). Our results show that RFamLlama is an efficient model with only 35 million parameters, yet it outperforms larger models in predicting mutation fitness. Additionally, our models can generate RNA sequences across diverse RNA families, which fold as expected based on predicted structures from AlphaFold 3 (Abramson et al., 2024). We have made three RFamLlama models of different sizes publicly available at the provided URL to assist RNA biochemists in the design of next-generation RNA-based bioparts, vaccines, and gene editors (Liu et al., 2024).

[1]Shanghai Mayoo Technology Co., Ltd [2]Independent Researcher [3]Syneron Technology, Guangzhou 510000, China. Correspondence to: Jinyuan Sun <jinyuansun98@gmail.com>.

Proceedings of the ICML 2024 Workshop on Accessible and Efficient Foundation Models for Biological Discovery, Vienna, Austria.

2. Related works

2.1. Deep learning for RNA design

The design of RNA sequences typically aims to achieve functions such as protein synthesis or gene regulation. Research teams use the structure of the target RNA to perform inverse folding, obtaining the desired sequences (Obonyo et al., 2024; Rubio-Largo et al., 2023). Recently, RfamGen (Sumi et al., 2024) introduced a sequence generation model that combines a Variational Autoencoder (VAE) and a Covariance Model (CM). This model overcomes the limitations of flexibility and generality in RNA inverse folding and generates sequence representations that are semantically meaningful. However, this method still relies on Multiple Sequence Alignments (MSAs), which means that the model requires homologous sequence alignments collected from various sequence databases, constrained by the availability of homologous sequences in the databases and requires training from scratch for each family therefore hindered broader use in the biochemical community.

2.2. Conditional generative models for molecules

Molecular conditional generation models focus on generating molecules with desired properties under specific conditions. In the field of proteins, ProGen (Madani et al., 2023) is analogous to an English language model, using control tags to generate new sequences within a given protein family (e.g., immunoglobulins, lysozymes), enabling the *de novo* design of proteins and demonstrating significant potential applications. Furthermore, in the design of functional enzymes, ZymCTRL (Munsamy et al., 2022) can generate new enzymes based on user-defined catalytic reactions. It also allows fine-tuning with specific enzyme datasets, enhancing the model's confidence in particular enzyme classes. However, most conditional sequence generation models are currently applied to proteins, with few extending to RNA sequences.

3. Methods

3.1. Preparation of datasets

The Rfam 14.10 (Kalvari et al., 2021) database was downloaded from the official ftp site. Sequence clustering of each family was performed using `easy-cluster` command of MMseqs2 (Steinegger & Söding, 2017) with the `--min-seq-id` option set to 0.9, `-c` set to 0.8, and `--cov-mode` set to be 1.

3.2. Model training

The Llama architecture was adopted for RFamLlama without modification except for the number of layers and the dimension. We used the transformers library to train the model with flash attention enabled. Training stopped at the lowest validation loss.

3.3. Zero-shot fitness prediction

For RFamLlama models, we calculated the model predicted log likelihood of a mutant to assign the predicted fitness. For each trained model, we predicted both log likelihoods with or without the Rfam ID of the mutant. For the RNA-FM model, we also calculated the log likelihood of each mutant. For RfamGen and EVmutation, the results from the RfamGen paper (Sumi et al., 2024) were taken.

3.4. Sequence Generation

When generating sequences of a family, the sampling parameters were not modified and the Rfam ID was used as the only prompt. The structure predictions were carried out using the AlphaFold 3 server (Abramson et al., 2024).

4. Results

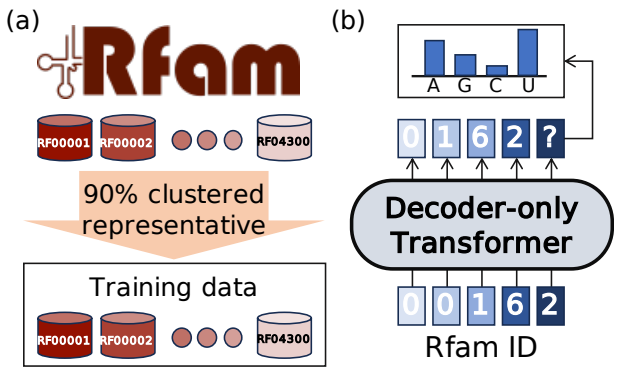

Figure 1. Training data and model architecture of RFamLlama. (a) Training data are representative sequences of each Rfam family, clustered by 90% identity. (b) The RFamLlama is a decoder-only Transformer trained to predict the next token based on previous context.

4.1. The Rfam database and training dataset

The Rfam database represents a sophisticated repository of structural RNA families, encompassing non-coding RNAs equipped with cis-regulatory elements. Families within the Rfam database are established through a rigorous curatorial process that includes seed alignments and secondary structure annotations aimed at identifying homologous sequences (Kalvari et al., 2021). Consequently, an Rfam ID encapsulates a high-level abstraction, reflecting the secondary structure and presumed identical functions among a group of RNAs. Nevertheless, the distribution of data within Rfam

is markedly uneven in terms of family sizes Figure 3. For instance, the largest family, RF00005, which corresponds to the tRNA family, comprises over 100,000 members. In contrast, several smaller families contain fewer than ten members each. To cultivate a balanced training dataset, we implemented clustering for each family at 90% sequence identity, effectively minimizing the overrepresentation of larger families (Figure 1). Following this clustering process, approximately 676,000 sequences remained. We ensured that at least one sequence from each family were designated for validation.

4.2. The design of RFamLlama

RFamLlama is designed to synthesize RNA sequences on demand using a language model built on the Llama architecture, which was specifically trained on the dataset built from the Rfam database to predict the next token in a sequence (Figure 1). RNA sequences were tokenized at the single nucleotide level, with permissible tokens being `A`, `T`, `C`, `G`, and `N`. In the sequences, all occurrences of `U` were replaced with `T` to maintain consistency. Additionally, we integrated further contextual information into the training process. Firstly, specific tokens were used to denote the 5' (`<|5|>`) and 3' (`<|3|>`) ends of RNA sequences, enhancing the model's understanding of sequence orientation. Secondly, the Rfam ID for each sequence was embedded at the beginning of sequences using the tokens `<|tag_start|>` and `<|tag_end|>`, signaling the presence of a tagged segment. Rfam IDs were also tokenized, digit by digit, from 0 to 9. Given that each Rfam ID consists of five digits, this method allows for a more complex representation space, facilitating deeper learning compared to a simplistic one-hot encoding approach. The development of RFamLlama involved the training of three models, varying in size from 13 million to 88 million parameters.

4.3. Zero-shot fitness prediction

The abliity of RFamLlama to predict the effects of mutations was evaluated by using model-predicted likelihood scores as a proxy for fitness. This approach does not rely on any direct knowledge of experimentally measured effects, thereby qualifying these predictions as zero-shot. We benchmarked RFamLlama's zero-shot fitness prediction abilities against other models, including family-specific models such as RfamGen and EVmutation, as well as universally pretrained models like RNA-FM (Chen et al., 2022) on five deep mutational scanning (DMS) datasets. Our model demonstrated superior performance, achieving the highest Spearman's correlation coefficients in two out of five datasets, a result comparable to that of EVmutation (Table 1). It is noteworthy that both EVmutation and RfamGen, which are family-specific models, require training on MSA for each respective family. Compared to another general-purpose pretrained RNA language model, RFamLlama consistently exhibited higher Spearman's coefficients across all datasets.

Furthermore, our findings highlight the critical role of the family tag in enhancing zero-shot prediction accuracy. For instance, the average Spearman's correlation coefficient for the RFamLlama-base model was 0.372 when predictions included the Rfam ID, but this figure significantly decreased to 0.190 without the family tag. This observation underscores the importance of the family tag in guiding the model to recall and apply the specific fitness landscape associated with each family, thereby improving prediction accuracy.

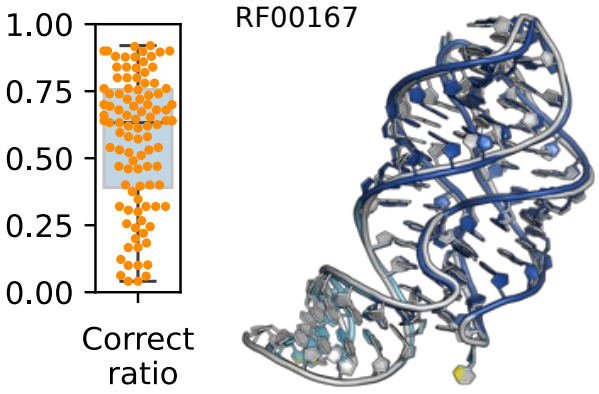

Figure 2. The left is a swarmplot with boxplot indicating the distribution of correct ratio of each tested Rfam IDs with each orange dot representing a Rfam ID. The right is a comparison of predicted structure (colored with pLDDT, blue indicating pLDDT higher than 90, cyan indicating pLDDT lower than 90 but higher than 70) and X-ray structure of a template of this family (PDB ID: 4TZX, Chain X) colored in grey. Structural figure prepared using Pymol (DeLano et al., 2002).

4.4. *De novo* generation of RNA sequences

We selected 100 Rfam IDs and utilized RFamLlama-base to generate 50 sequences for each, with the Rfam IDs serving as the sole prompt for generation. The Rfam IDs of the generated RNA sequences were identified using cmscan against the Rfam database. The average Rfam ID recovery ratio was 0.57, with the highest recovery ratio observed for the Purine riboswitch (RF00167), reaching 0.92 (refer to Figure Figure 2). Furthermore, all tested Rfam IDs had at least two samples correctly assigned to the corresponding prompt Rfam ID. The sequence with the highest predicted fitness for the generated Purine riboswitch was subjected to structural prediction using the AlphaFold 3 server. The overall predicted Local Distance Difference Test (pLDDT) score exceeded 70, indicating a confident prediction suitable for further analysis. Additionally, a Purine riboswitch

Table 1. Comparison of different models for various DMS datasets

Method	tRNA (Li et al., 2016)	glmS ribozyme 1 (Andreasson et al., 2020)	glmS ribozyme 2 (Sumi et al., 2024)	drz-agam-2-1 ribozyme (Kobori & Yokobayashi, 2018)	Twister ribozyme P1 (Kobori & Yokobayashi, 2016)
RfamGen (Sumi et al., 2024)	**0.556**	0.546	0.371	0.035	0.425
EVmutation (Hopf et al., 2017)	0.493	**0.657**	0.321	-0.121	**0.548**
RFamLlama-small	0.503	0.475	0.397	0.049	0.391
RFamLlama-base	0.460	0.518	**0.443**	0.016	0.421
RFamLlama-large	0.427	0.584	0.407	**0.077**	0.269
RNA-FM (Chen et al., 2022)	0.445	0.207	0.227	-0.033	0.055
RNAErnie (Wang et al., 2024)	0.135	0.048	0.322	-0.047	-0.054
RFamLlama-small-notag	0.227	-0.021	0.306	0.013	0.038
RFamLlama-base-notag	0.288	0.104	0.370	0.016	0.173
RFamLlama-large-notag	0.293	-0.141	0.265	**0.077**	0.075

crystal structure from the PDB database (PDB ID: 4TZX, Chain X) served as a reference template. The root-mean-square deviation (RMSD) between the predicted and template structures was calculated to be 1.284 Å, suggesting a potentially successful generation. These findings demonstrate that RFamLlama can generate sequences meeting CM constraints and produce structurally accurate and potentially functional RNA sequences.

5. Disscussion

In this study, we introduced RFamLlama, an efficient conditional language model designed for RNA sequence generation across a wide range of families. The model exhibited strong zero-shot fitness prediction capabilities, underscoring the vital role of function tags in boosting prediction accuracy. The absence of these tags significantly impaired performance. RFamLlama has also demonstrated significant promise in generating artificial RNA sequences. These capabilities were further validated through structural predictions conducted using the AlphaFold 3 server, providing solid evidence of the model's effectiveness. With a compact architecture of only 32 million parameters, RFamLlama-base stands as a foundational model for RNA sequence generation. It offers a resource-efficient tool that could significantly aid research conducted by wet-lab biochemists.

Following the methodology suggested by (Zhang et al., 2024), we propose that RNA language models might store the co-evolutionary statistics for all families in the Rfam database. The Rfam database contains approximately 4,000 families. Assuming an average length of 150 nucleotides per sequence, where each position potentially forms three contacts—two with sequence neighbors and one Watson-Crick pairing if mismatches are allowed—the total parameter count is estimated using the formula:

$$\text{Total Parameters} = 4000 \times 150 \times 3 \times 4 \times 4 = 28,800,000$$

This estimate closely approximates the parameter count of our best-performing model, RFamLlama-base, which contains 32 million parameters. This similarity might elucidate RFamLlama's superior efficiency compared to other pre-trained models that lack family-specific tags. RFamLlama-base possesses sufficient parameters to store co-evolutionary information, enhancing its predictive accuracy. Moreover, the inclusion of pre-classified Rfam ID tags provides contextual information that helps reduce sequence hallucinations. However, this hypothesis necessitates further investigation within the domain of protein language models, a field that offers more comprehensive benchmarks than those available for RNA, to draw definitive conclusions.

Although not tested in this work, it is quite obvious that additional supervised finetuning on in-house dataset will further enhance the performance of RFamLlama. In the future, we expect that RFamLlama could be combined with high throughput experiment to help design enhanced RNA molecules more effeicently than evolution in the wild or the lab, to help develop better RNA-based bioparts, vaccines, and gene editors.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pp. 1–3, 2024.

Andreasson, J. O., Savinov, A., Block, S. M., and Greenleaf, W. J. Comprehensive sequence-to-function mapping of cofactor-dependent rna catalysis in the glms ribozyme. *Nature communications*, 11(1):1663, 2020.

Beaudry, A. A. and Joyce, G. F. Directed evolution of an rna enzyme. *Science*, 257(5070):635–641, 1992.

Bond-Taylor, S., Leach, A., Long, Y., and Willcocks, C. G. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7327–7347, 2021.

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

DeLano, W. L. et al. Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr*, 40(1): 82–92, 2002.

Dykstra, P. B., Kaplan, M., and Smolke, C. D. Engineering synthetic rna devices for cell control. *Nature Reviews Genetics*, 23(4):215–228, 2022.

Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P., Springer, M., Sander, C., and Marks, D. S. Mutation effects predicted from sequence co-variation. *Nature biotechnology*, 35(2):128–135, 2017.

Joyce, G. F. Rna evolution and the origins of life. *Nature*, 338(6212):217–224, 1989.

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

Karikó, K., Buckstein, M., Ni, H., and Weissman, D. Suppression of rna recognition by toll-like receptors: the impact of nucleoside modification and the evolutionary origin of rna. *Immunity*, 23(2):165–175, 2005.

Kobori, S. and Yokobayashi, Y. High-throughput mutational analysis of a twister ribozyme. *Angewandte Chemie International Edition*, 55(35):10354–10357, 2016.

Kobori, S. and Yokobayashi, Y. Analyzing and tuning ribozyme activity by deep sequencing to modulate gene expression level in mammalian cells. *ACS Synthetic Biology*, 7(2):371–376, 2018.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, C., Qian, W., Maclean, C. J., and Zhang, J. The fitness landscape of a trna gene. *Science*, 352(6287):837–840, 2016.

Liu, Z.-X., Zhang, S., Zhu, H.-Z., Chen, Z.-H., Yang, Y., Li, L.-Q., Lei, Y., Liu, Y., Li, D.-Y., Sun, A., et al. Hydrolytic endonucleolytic ribozyme (hyer) is programmable for sequence-specific dna cleavage. *Science*, 383(6682): eadh4859, 2024.

Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos, J. L., Xiong, C., Sun, Z. Z., Socher, R., et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, 2023.

Munsamy, G., Lindner, S., Lorenz, P., and Ferruz, N. Zymctrl: a conditional language model for the controllable generation of artificial enzymes. In *NeurIPS Machine Learning in Structural Biology Workshop*, 2022.

Nobel Prize Outreach AB. Press release. NobelPrize.org, May 2024. URL https://www.nobelprize.org/prizes/medicine/2023/press-release/. Accessed: 2024-05-19.

Obonyo, S., Jouandeau, N., and Owuor, D. Self-playing rna inverse folding. *SN Computer Science*, 5(4):403, 2024.

Pardi, N., Hogan, M. J., Porter, F. W., and Weissman, D. mrna vaccines—a new era in vaccinology. *Nature reviews Drug discovery*, 17(4):261–279, 2018.

Romero, P. A. and Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nature reviews Molecular cell biology*, 10(12):866–876, 2009.

Rubio-Largo, Á., Lozano-García, N., Granado-Criado, J. M., and Vega-Rodríguez, M. A. Solving the rna inverse folding problem through target structure decomposition and multiobjective evolutionary computation. *Applied Soft Computing*, pp. 110779, 2023.

Steinegger, M. and Söding, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

Sumi, S., Hamada, M., and Saito, H. Deep generative design of rna family sequences. *Nature Methods*, pp. 1–9, 2024.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Wang, N., Bian, J., Li, Y., Li, X., Mumtaz, S., Kong, L., and Xiong, H. Multi-purpose rna language modelling with motif-aware pretraining and type-guided fine-tuning. *Nature Machine Intelligence*, pp. 1–10, 2024.

Zhang, Z., Wayment-Steele, H. K., Brixi, G., Wang, H., Dal Peraro, M., Kern, D., and Ovchinnikov, S. Protein language models learn evolutionary statistics of interacting sequence motifs. *bioRxiv*, pp. 2024–01, 2024.

Zhu, Y., Zhu, L., Wang, X., and Jin, H. Rna-based therapeutics: an overview and prospectus. *Cell death & disease*, 13(7):644, 2022.

A. Appendix

A.1. Training data format

All sequences were prefixed with its Rfam IDs in this way:

```
<|bos|><|tag_start|>00050<|tag_end|><|5|>ATCG<|3|><|eos|>
```

in the line, the real Rfam ID should be `RF00050`. The tokenized text will be:

```
<|bos|> <|tag_start|> 0 0 0 5 0 <|tag_end|> <|5|> A T C G <|3|> <|eos|>
```

A.2. Model size and hyper-parameters for training

	RFamLlama-small	**RFamLlama-base**	**RFamLlama-large**
Number of layers	6	8	10
Number of heads	32	32	32
Embedding dim	384	512	768
Learning rate	3×10^{-4}	3×10^{-4}	3×10^{-4}
Weight decay	0.1	0.1	0.1
Total number of parameters (M)	13.5	33.5	88.5

Table 2. Comparison of RFamLlama models

A.3. Data in Rfam database

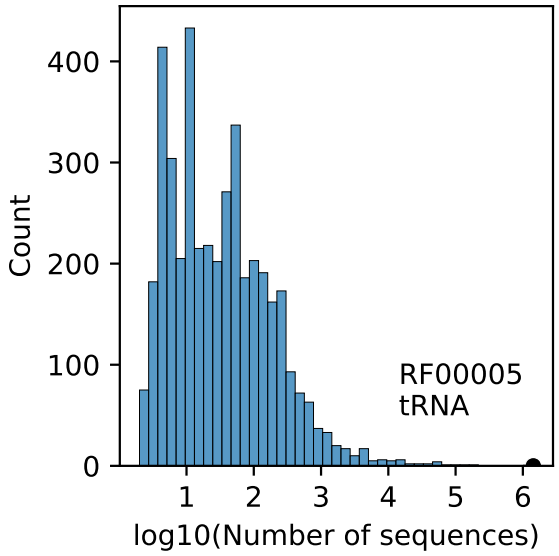

Figure 3. Sequence number length distribution of each family in the Rfam 14.10 database. The tRNA (RF00005) family has the most members.

A.4. Code availability

The code for the work is available at: https://github.com/JinyuanSun/RFamLlama.