MP-RNA: Unleashing Multi-species RNA Foundation Model via Calibrated Secondary Structure Prediction

Heng Yang[1], **Ke Li**[1]
[1]Department of Computer Science, University of Exeter, EX4 4QF, Exeter, UK
{hy345, k.li}@exeter.ac.uk

Abstract

RNA foundation models (FMs) have been extensively used to interpret genomic sequences and address a wide range of in-silico genomic tasks. However, current RNA FMs often overlook the incorporation of secondary structures in the pretraining of FMs, which impedes the effectiveness in various genomic tasks. To address this problem, we leverage filtered high-fidelity structure annotations for structure pretraining to enhance the modeling ability of FMs in single nucleotide resolution tasks. Experimental evaluations across four comprehensive genomic benchmarks demonstrate that our FM (**MP-RNA**) consistently outperforms existing RNA FMs, achieving a 40% improvement in RNA secondary structure prediction and obtaining top-tier results on DNA genomic benchmarks even though it has not been pretrained on any DNA genome. We release the code and tutorials[1] and models to encourage further research to bridge the gap between in-silico predictions and biological reality.

1 Introduction

RNA serves as a critical molecule in a variety of important cellular processes and controls the flow of genetic information from DNA to protein (Wang et al., 2024a). With the development of high-throughput RNA sequencing (Siegel et al., 2011), understanding the vast of RNA composed of nucleotide sequences reaches the efficiency and performance bottleneck of bioinformatics techniques. Recent studies leveraged genomic foundation models (FMs) to understand DNA sequences (Nguyen et al., 2023; Mendoza-Revilla et al., 2023; Dalla-Torre et al., 2023; Nguyen et al., 2024; Yin et al., 2024) and RNA (Chen et al., 2022; Yang et al., 2023; Wang et al., 2023; Zhang et al., 2024; Chu et al., 2024a) sequences and address a broad spectrum of in-silico genomic tasks such as mRNA vaccine design (Corbett et al., 2020; Runge et al., 2023), translation efficiency prediction (Avsec et al., 2021; Chu et al., 2024a), and gene expression prediction (Avsec et al., 2021; Mendoza-Revilla et al., 2023). This is because the generalization of language modeling from natural sentences to RNA sequences in the RNA 'language' (Nguyen et al., 2023) is intuitive and the FMs can efficiently model genomic sequences. Therefore, in the era of high-throughput genome sequencing, FMs are the vital bridge to learning intriguing genomic information from the tremendous RNA and DNA sequence databases.

Despite preliminary results of the previous DNA and RNA FMs to handle diverse genomic tasks (Mendoza-Revilla et al., 2023; Dalla-Torre et al., 2023; Nguyen et al., 2024), existing FMs usually struggle in challenging genomic tasks, e.g., single nucleotide resolution tasks, hindering wide acknowledgment from bio-scientists and wide applications of FMs. Since it has been acknowledged that *"the functionality and stability of RNA significantly depend on complex structures, e.g., secondary structures, in molecular biology"* (Ganser et al., 2019), this difficulty is probably because current FMs overlooked RNA secondary structures and straightforwardly took the techniques from language modeling to RNA language learning, leading to unimpressive performance on various genomes types (e.g., RNA and DNA, etc.). Moreover, this incapability in existing FMs limits the transferability to diversified unknown species because the genomes of millions of species are technically represented in numerous genomic 'language' variants, which is very different from the natural language modeling area. As the existing FMs struggle with challenging genomic tasks, we aim to propose a new multi-species plant RNA (**MP-RNA**) foundational model and prove it is robust and effective in addressing universal RNA genomic tasks. We elaborate on the main challenges that we met and

[1]https://github.com/yangheng95/OmniGenomeBench

Findings of the Association for Computational Linguistics: EACL 2024, pages 5278–5296

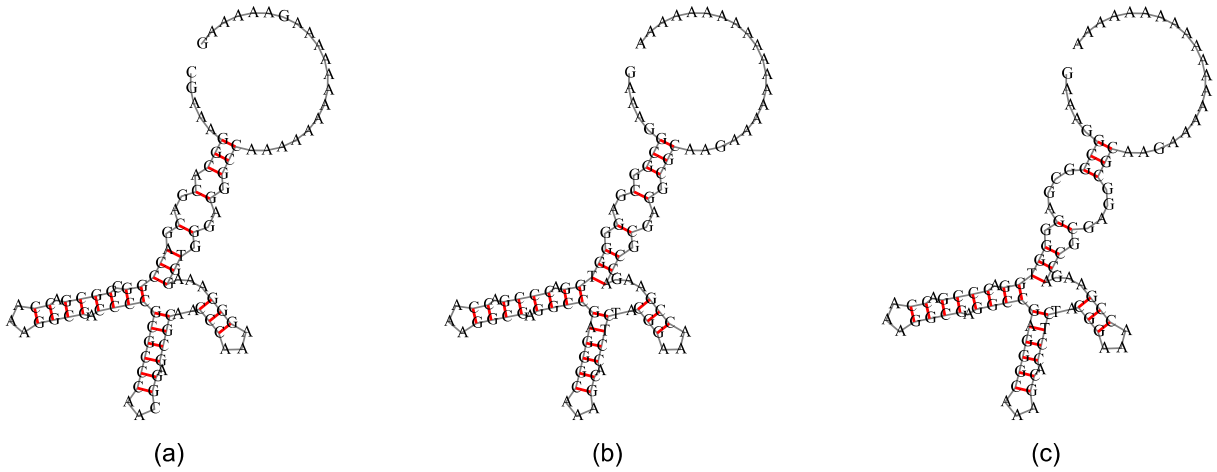

Figure 1: A virtual example for in-silico RNA secondary structure prediction. The sub-figures (a) and (c) indicate the same sequence with different structures. The sub-figures (b) and (c) denote an identical structure that can be folded from different sequences.

our solutions one by one.

The first challenge is to leverage secondary structures in RNA modeling. Learning RNA secondary structures is extremely challenging (Bonnet et al., 2020; Saman Booy et al., 2022) because the sequence information is generally sparse (Shastry, 2002) and structure candidates of specific sequences grow exponentially with increasing sequence lengths. Besides, an identical sequence may fold into different secondary structures because the folding principles of RNA sequences depend on a variety of biological factors (Tinoco Jr and Bustamante, 1999), such as dynamics mechanism (Mustoe et al., 2014). We show an example in Fig. 1 that two secondary structure variants belong to an identical RNA sequence. One simple but effective method to mitigate this problem is to exploit large-scale structure annotations (Tan et al., 2017; Danaee et al., 2018; Mathews, 2019; Kalvari et al., 2021) to supervise the FMs in structure prediction. Unfortunately, we cannot obtain enough RNA secondary structure annotations for pretraining because the structure annotations are complicated and expensive (Gong et al., 2024). To compile a large-scale database containing both sequence-structure pairs for FM pretraining, we utilize a well-known secondary structure prediction (SSP) tool, ViennaRNA (Lorenz et al., 2011), to obtain millions of plausible secondary structures with corresponding sequences. We show that pretraining on millions of secondary structures yields an effective and robust SSP, and it can improve the performance on various RNA modeling tasks even without structures involved as demonstrated in our experiments.

The second challenge of structure pretraining comes from the low fidelity of the secondary structure annotations. Although the secondary structure annotations calculated by ViennaRNA improve our FM's performance on extensive in-silico experiments, our study reveals a substantial problem, i.e., the secondary structures based on calculation, such as minimum free energy (Juan and Wilson, 1999; Mathews and Turner, 2006), fail to reflect the genuine in vivo structures. For example, the structure prediction F1 score of ViennaRNA on the bpRNA (Danaee et al., 2018) dataset is only $\approx 28\%$. The misalignment between vast sequences and plausible secondary structures casts a shadow on the structure pretraining, as well as various downstream genomic tasks. Although it is impossible to collect large-scale biologically verified structures at this stage, we realize that the FM, fine-tuned on the formal secondary structure datasets (Kalvari et al., 2021), yields the structures almost there, i.e., close to the genuine structures. Additionally, we have to make sure the structure predictions can be trusted before we employ them, so we calibrate the predictions from the perspective of uncertainty and perform incremental pretraining to enhance the FM. We adopt the popular temperature scaling (Guo et al., 2017), as it does not influence the model's performance, to calibrate the predicted structures as high-fidelity structure annotations to incrementally pretrain the FM and refine structure modeling ability. The incremental pretraining uses the top 10% of the RNA samples compared to the first pretraining to enhance the FM. The experimental results show that incremen-

tal pretraining comprehensively improves structure prediction tasks as well as four comprehensive genomic benchmarks.

The third challenge is to refine the RNA sequence modeling granularity, as our study indicates that the existing FMs are usually designed for sequence-level tasks, e.g. sequence classifications. However, most of the in vivo genomic analysis focus on single nucleotide resolution, such as single nucleotide variant (Miladi et al., 2020) (SNV) and single nucleotide polymorphism (Rafalski, 2002) (SNP), mutation detection and repair (Dalla-Torre et al., 2023; Nguyen et al., 2024). Besides, single nucleotide (SN) base changes can influence RNA folding results (Miladi et al., 2020), structure stability and molecular functionality, etc. To address this obstacle, we include single nucleotide mutation repair in the pretraining objectives, i.e., we pretrain the FM to find the mutation sites and predict the original bases in the RNA sequences. This SN mutation modeling is a pioneering effort to understand the dynamic connections between RNA sequences, structures, and SN mutations (Denny and Greenleaf, 2019).

To benefit future research of RNA FMs, we will release the pretrained models, benchmarks, and in-silico RNA modeling cases in this work.

2 Evaluations and Findings

Research on RNA FMs remains in a state of beginning, as the related theoretical and empirical studies are far from comprehensively completing the jigsaw of genomic sequence modeling. We have implemented comprehensive benchmarks to evaluate **MP-RNA**. As insights for future works, we summarize our findings as follows:

- **MP-RNA** outperforms state-of-the-art FMs on all 6 challenging tasks in the RNA genomic benchmark, obtaining up to 40% improvement for secondary structure prediction and top-1 performance of mRNA degrade rate regression. The experimental results show that vast plausible structure annotations can be utilized to improve RNA modeling tasks, providing a hint of **making use of other available genomic information, e.g., structure stability**.

- Apart from RGB, **MP-RNA** reveals the generalizability of the RNA pretraining paradigm on three comprehensive DNA genomic benchmarks, i.e., PGB, GUE, and GB. Overall, **MP-RNA** achieves **top-tier** performance on most of the DNA tasks, even though **these tasks have no secondary structures involved and DNA genome data were never seen in the MP-RNA's pretraining**. This observation indicates a promising FM research prospective of multi-modal (i.e., multi-genome and multi-species) modeling.

- Even though k-mers (Dalla-Torre et al., 2023; Mendoza-Revilla et al., 2023) and BPE (Zhou et al., 2023) have been widely utilized in the existing FMs because of the efficiency for long sequence modeling compared to single-nucleotide tokenization (SNT), these **coarse-grained tokenization methods ignore the SN-level interactions** in genomic sequences. The oversight of SN-level interactions results in unsatisfactory performance on the diverse set of SN-level tasks, e.g., structure prediction and mutation prediction. Our empirical experience indicates that adopting **mixed tokenizations** based on the existing tokenizations may balance the performance and efficiency of multi-granular tasks.

3 Methodology

In this section, we delve into the methodology of our pretraining paradigm step by step.

3.1 RNA Sequence Tokenization

The performance of genomic FMs highly depends on the implementation of RNA sequence tokenization (Zhou et al., 2023; Nguyen et al., 2023; Chen et al., 2023). The most popular tokenization method in previous works is k-mers (Yang et al., 2023; Dalla-Torre et al., 2023), including overlapping and non-overlapping variants. Byte Pair Encoding (Devlin et al., 2019) (BPE) was used in DNABERT2 (Zhou et al., 2023) to address the token misalignment (refer to Fig. 1 in Zhou et al. (2023)) in k-mers between similar RNA sequences. However, both the k-mers and BPE represent the tokens in multiple nucleotides, leading to coarse-grained RNA sequence modeling which can fail the SN-level genomic modeling tasks, such as RNA secondary structure prediction and RNA sequence design. SN-level tasks significantly require the SN-level modeling resolution and SN-level alignment between the inputs and outputs. To address these two problems, we adopt SNT and represent the input tokens as individual bases (Nguyen et al., 2023; Chen et al., 2023). We have prepared an illustrative example to depict the tokenization results and

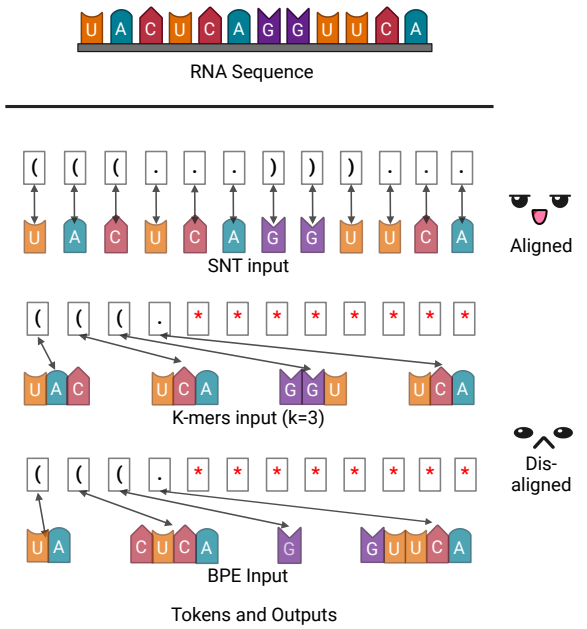

Figure 2: A virtual tokenization example of k-mers, BPE, and SNT. In the SN-level tasks, the token inputs and outputs of the k-mers and BPE are not aligned due to entangled tokens containing multiple bases. '*' indicates the outputs are predicted based on padding tokens.

SN-level alignment in Fig. 2. We also append the secondary structures to the sequence in a small proportion (10%) of examples for masked language modeling, which aims to enable our FM to predict the masked bases given the context of structures. Accordingly, we adopted a vocabulary {'A', 'T', 'C', 'G', 'U', 'N', '(', ')', '.'} to unify the tokenization of both nucleotide bases and RNA secondary structure information.

3.2 Pretraining Paradigm

We further introduce the self-supervised pretraining paradigm on large-scale RNA sequence databases, particularly for three objectives with implementations. Our pretraining objectives are induced based on genomic domain knowledge, secondary structure prediction (SSP), single nucleotide mutation repair (SNMR), and masked RNA language modeling (MRLM), tailored to sequence modeling given the transformer architecture. These objectives were combined during the pretraining of our framework.

RNA Secondary Structure Pretraining

The existing works regard SSP as a fine-tuning-based downstream task, while the scales of the downstream datasets are trivial to train the FM to predict structures in high fidelity. Therefore, we aim to pre-train an FM available for SSP based on large-scale RNA sequences to improve both sequence and structure modeling capacity. Due to the lack of verified structures, we collect large-scale sequences and calculate plausible secondary structure annotations using ViennaRNA. This secondary structure prediction is a token-level classification task so we utilize a cross-entropy as follows:

$$\mathcal{L}_{\text{SSP}} := -\sum_{i=1}^{C} \left[p_i \log\left(\tilde{p}_i\right) + q_i \log\left(\tilde{q}_i\right) \right], \quad (1)$$

where C is the number of strcuture labels, i.e., {'(', ')', '.' }. p and $\tilde{p}$ respectively indicate the true and predicted probability distributions of the structure label, while q and $\hat{q}$ represent any incorrect predictions and their likelihood, respectively.

Single-Nucleotide Mutation Repair

The SNV and SNP are important subjects to study in genomics, while existing FMs have not explored explicit modeling of SNV and SNP as it is very challenging. More specifically, it is resource-intense to directly model the SNV and SNP sequences because these sequences could be sparse with one-base differences in multiple sequence variants, e.g., it is estimated that SNPs occur every 1 in 1000 base pairs in the human genome (Shastry, 2002). Consequently, to train the FM to be sensible to SNV and SNP, we reformulate the SNV and SNP as the SN mutation detection and repair task. To overcome the SN mutation database scarcity, we synthesize single nucleotide mutation sites in natural RNA sequences and utilize the FM to reconstruct the original bases that have been mutated. According to empirical observation, we synthesized 5% mutation sites in the natural RNA sequences for the SNMR objective. The SNMR is implemented as a token-level classification task, employing cross-entropy as the loss function. The $\mathcal{L}_{\text{SNMR}}$ is a simple loss function so we omit its formula here.

Masked RNA Language Modeling

The MRLM is a simple but effective generalization from masked language modeling (Devlin et al., 2019) to masked genomic language modeling, aiming to understand the RNA sequences based on unsupervised pretraining. Following the previous works, we randomly mask 15% of the bases as well as structure tokens (e.g., '(', ')', '.'). MRLM enables the learning of implicit base-wise dependen-

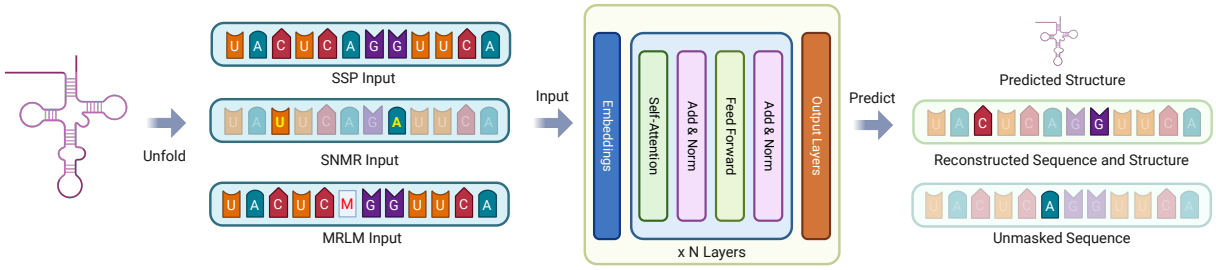

Figure 3: The pretraing paradigm of **MP-RNA**. The collected RNA sequences are prepared for three pretraining objectives, i.e., SSP, SNMR and MRLM, respectively. For the SSP objective, we feed the RNA sequences into the FM and predict the structures. The mutation bases are highlighted in boldface for the SNMR objective, and only the mutation bases are calculated in the loss function. For the MRLM objective, the masked bases are indicated as M, and only the masked bases are calculated in the loss function. 'N' is the number of transformer layers in the FM and can be 16 and 32 for **MP-RNA**-52M and **MP-RNA**-186M, respectively.

cies based on deep-contextualized sequence modeling. The MRLM objective employs the cross-entropy as the loss function. The $\mathcal{L}_{\text{MRLM}}$ is a regular and well-known loss function so we omit its formula here.

3.3 Structure Calibration for Incremental Pretraining

Accurate structures can significantly enhance the performance of FMs on genomic tasks. Although tools like ViennaRNA (Lorenz et al., 2011) can calculate secondary structures, the fidelity of computed structures can not be guaranteed due to the absence of consideration of biological molecule interactions across different species. As an alternative way, we propose leveraging FMs to improve the fidelity of secondary structures calibration based on temperature scaling (Guo et al., 2017), a method proven effective in aligning model predictions with actual probability distributions. This calibration process involves the following steps:

- Utilizing the training splits of previously established secondary structure prediction datasets, i.e., Rfam (Kalvari et al., 2021) to form a dataset for learning the temperature parameter.

- We adjust the softmax function of the secondary structure classifier in the FM by introducing a temperature parameter T to calibrate the probability:

$$Softmax_T(z_i) := \frac{e^{z_i/T}}{\sum_j e^{z_j/T}} \quad (2)$$

  where z_i are the logits output by the model, and T is the temperature parameter used to scale the logits before applying the softmax function.

- We optimize the temperature parameter T by minimizing the cross-entropy loss:

$$T^* = \arg\min_T CE(Y, Softmax_T(Z)) \quad (3)$$

  where Y denotes the true labels of structures, Z represents the logits provided by the model, and CE indicates the cross-entropy operator.

- The performance of structure prediction on the validation set is evaluated using the identified T^*. This calibration ensures that the model's predictions are more aligned with the true likelihoods, thus enhancing both the utility and reliability of the model in predicting RNA structures.

To reduce the computational budget of incremental pretraining, we calculate the expected calibration error (Platt et al., 1999) for all RNA structure predictions to filter the top 10% of the RNA samples with the minimum expected calibration error for incremental pretraining.

3.4 Pretraining RNA Database

Recent studies (Chen et al., 2023; Zhou et al., 2023) have shown that species diversity can enhance FM's performance with moderate model capacity. For the **MP-RNA** pretraining, we collected transcriptome data from the OneKP initiative (Carpenter et al., 2019), which compiles large-scale RNA sequence from 1,124 plant species. Because the information in raw sequences is sparse and noisy, they are far from ready for effective FM training.

To address this problem, we developed a four-step data curation protocol to improve data quality.

- Raw RNA sequences are often excessively long with thousands of bases. We first sliced them into segments with a window size of $1,024$ bases to provide a sufficient context window for RNA sequence understanding.
- To enhance training efficiency and reduce bias, we removed all duplicate sequences.
- To tackle incomplete transcriptome data and other noises, we discard sequences shorter than 50 bases.

3.5 Pretraining Implementation Details

We developed two Transformer-based versions of our framework for comparison and analysis, denoted as **MP-RNA**-52M and **MP-RNA**-186M, respectively, where specialized architectures and implementation details can be found in Table 1. The pretraining was performed on four Nvidia RTX 4090 GPUs over one month. We optimized the model architecture and hyperparameter configurations based on grid search and empirical experience, as the pretraining process is very time- and resource-intensive.

Pretraining Setup	MP-RNA-52M	MP-RNA-186M
# of Layers	16	32
Embed Dimension	480	720
Hidden Dimension	480	720
# of Heads	24	30
# of Parameters	52M	186M
Position Embedding	Rotary	
Dropout	0.0	
Learning Rate	$5e^{-5}$	
Weight Decay	0.01	
Optimizer	AdamW	
Optimizer Momentum	$\beta_1, \beta_2 = 0.9, 0.999$	
LR Scheduler	Linear Decay	
Batch Size	3072	
# of Training Epochs	1	
Sequence Length	1024	

Table 1: The model architecture and training configurations for **MP-RNA**. Please find our source code for more technical details.

4 Experiments

This section includes comprehensive in-silico experiments on four genomic benchmarks, i.e., RNA genomic benchmark (RGB), Plant genomic benchmark (Mendoza-Revilla et al., 2023) (PGB), Genomic Benchmark (Grešová et al., 2023) (GB) and Genomic Understanding Evaluation (Zhou et al., 2023) (GUE), for both RNA and DNA FMs, where the RGB aims to evaluate the performance of RNA sequences understanding and the rest of the benchmarks are used for generalizability evaluation. To avoid data leaks, all of the data are not included in **MP-RNA**'s pretraining database.

4.1 Evaluation Baselines

To evaluate the performance of **MP-RNA**, we made comparisons with the following baseline genomic FMs.

For further details such as tokenization methods, training databases, and types of genomic sequences, please refer to Table 2. While some FMs have been developed specifically for RNA modeling as well, such as RNA-FM (Chen et al., 2022), RNA-MSM (Zhang et al., 2024), Uni-RNA (Wang et al., 2023), 5UTR-LM (Chu et al., 2024a) and Evo (Nguyen et al., 2024), we cannot include the results as the original benchmark scripts or models checkpoints are not publicly accessible.

Please find brief introductions of the baseline FMs in Appendix B.

4.2 Fine-tuning Setup

In this section, we introduce the hyperparameters used for downstream task fine-tuning. To fairly compare **MP-RNA** with the baseline models, we carefully set proper and same hyperparameters for each experiment and recorded the average performance in three runs for all baseline FMs. Specifically, we used a learning rate of $2e^{-5}$, a batch size of 16, a L_2 regularization of $1e^{-5}$ and AdamW optimizer for all FMs. We train the FMs for 10 epochs on all datasets. These settings were applied consistently across all tasks to maintain experimental integrity.

4.3 RNA Genomic Benchmark

We first evaluate the performance of **MP-RNA** and baseline FMs on the RGB[2], a challenging benchmark dedicated to SN-resolution RNA genomic modeling capability evaluation. Due to resource limitations, we only perform incremental pretraining for **MP-RNA**-186M, i.e., we will not release the incrementally pretrained **MP-RNA**-52M.

These results in Table 3 indicate the results of **MP-RNA**. Overall, **MP-RNA**-186M outperforms all baseline FMs across nearly all SN-level tasks. The ablations of **MP-RNA**, i.e., **MP-RNA***-52M and

[2]Please find the details of RGB in the appendix, such as dataset statistics and input-output examples.

Model	Tokenization	# of Params	# of Tokens/Sequences	Species	Sequence Type
DNABERT-2	BPE	117M	32.49B Tokens	Human + 135 Species	DNA
NT-V2 (100M)	k-mers	96M	300B Tokens	Human + 850 Species	DNA
HyenaDNA (Large)	SNT	47M	3.2B Tokens	Human	DNA
Caduceus	SNT	1.9M	35B Tokens	Human	DNA
Agro-NT (1B)	k-mers	985M	472.5B Tokens	48 Edible Plants	DNA
SpliceBERT	SNT	19M	2M Sequences	Multi-Vertebrates	precursor-mRNA
CDSBERT	SNT	420M	4M Sequences	4069 RNA Families	CDS
3UTRBERT	k-mers	86M	20,362 Sequences	Multi-Species	mRNA 3'UTR
RNA-BERT	SNT	0.5M	4,069 RNA Families	Multi-Species	ncRNA
RNA-MSM	SNT	96M	4,069 RNA Families	Multi-Species	ncRNA
RNA-FM	SNT	96M	23M Sequences	Multi-Species	ncRNA
MP-RNA-52M (ours) **MP-RNA**-186M (ours)	SNT	52M 186M	54.2B Tokens	1124 Plant Species	mRNA, CDS, UTRs

Table 2: The brief statistics of baseline RNA and DNA FMs collected from original publications. The numbers of parameters of different FMs are approximately calculated.

Model	mRNA	SNMD	SNMR	Archive2	bpRNA	RNAStralign
	RMSE	AUC	F1	F1	F1	F1
ViennaRNA	N.A.	N.A.	N.A.	75.89	27.82	74.80
DNABERT-2	0.8158	49.94	15.86	59.82	43.40	65.49
HyenaDNA	0.8056	53.32	39.80	84.23	56.62	95.42
Caduceus	0.8026	57.01	39.59	91.37	68.76	97.28
NT-V2	0.7826	50.49	26.01	79.90	56.60	90.84
Agro-NT	0.7830	49.99	26.38	70.13	48.71	75.21
SpliceBERT	0.7340	58.11	46.44	89.05	69.10	96.97
3UTRBERT	0.7772	50.02	24.01	78.98	56.93	92.03
CDSBERT	0.7468	55.03	36.16	89.34	70.01	97.15
RNABERT	0.8087	51.32	29.14	24.66	83.68	47.96
RNA-MSM	0.7321	57.86	45.22	68.72	91.15	64.44
RNA-FM	0.7297	59.02	42.21	82.55	95.07	78.16
MP-RNA*-52M	0.7219	61.26	47.97	94.98	80.02	99.01
MP-RNA*-186M	0.7189	63.33	49.09	95.20	84.02	99.12
MP-RNA-186M	**0.7155**	**64.66**	**52.21**	**95.92**	**84.61**	**99.21**

Table 3: The performance of **MP-RNA** and baseline FMs on the RGB, with results averaged based on five random seeds. "N.A." indicates that ViennaRNA is not designed for other predictive genomic tasks, and missing entries where no data is available.

MP-RNA*-186M, also make an obvious performance difference to the baseline FMs. Even compared to large-scale FMs, such as Agro-NT and CDSBERT dedicated to genomic modeling, **MP-RNA** presents a consistent and significant (up to 25% on SSP task) improvement. This observation suggests the effectiveness of incremental pretraining on calibrated secondary structures. Another reason for the superiority of **MP-RNA** is mainly because the existing FMs usually adopt k-mers tokenization that cannot handle SN resolution tasks, including mutation site detection and repair.

4.4 Plant Genomic Benchmark

The second benchmark is PGB, a large-scale plant-oriented DNA genomic benchmark used for evaluating the transferability and generalizability of **MP-RNA**. We aim to leverage PGB to evaluate **MP-RNA** on multi-species tasks.

The results of **MP-RNA** in Table 4 reveal considerable performance improvement compared to even DNA-expertised FMs on various tasks, such as Polyadenylation, Splice Site, and Enhancer Region classification. Although we did not include any DNA genomes in the pretraining and the sequence lengths in PGB are usually greater than the max modeling length of **MP-RNA**, **MP-RNA** achieved the best F1 scores among the DNA tasks in PGB, which means the limitation of the modeling length in genome understanding may not be a critical problem. In comparisons, existing FMs, e.g., CDSBERT and Agro-NT, show lower performance with

Model	PolyA	LncRNA	Chrom Acc	Prom Str	Term Str	Splice	Gene Exp	Enhancer
	F1	F1	F1	RMSE	RMSE	F1	RMSE	F1
DNABERT-2	41.35	72.55	61.49	0.99	0.24	45.34	14.78	36.40
HyenaDNA	83.11	58.21	52.20	0.88	0.26	90.28	14.76	66.17
Caduceus	70.89	68.40	64.53	0.91	0.26	78.51	14.72	60.83
NT-V2	71.26	73.08	65.71	0.81	0.27	95.05	14.69	73.89
Agro-NT	78.89	67.24	63.27	0.94	0.78	88.45	15.56	62.83
SpliceBERT	65.23	71.88	63.62	0.75	0.22	96.45	14.70	69.71
3UTRBERT	76.48	70.75	63.71	1.04	0.36	94.44	14.87	71.67
CDSBERT	39.72	33.06	48.95	2.19	0.59	52.20	14.77	33.93
RNA-BERT	78.54	61.99	48.94	1.81	0.38	94.45	14.89	57.61
RNA-MSM	84.25	67.49	53.52	1.28	0.28	95.49	14.87	61.45
RNA-FM	84.94	68.75	54.92	0.95	0.27	95.95	14.83	57.14
MP-RNA-186M	**87.48**	**77.68**	**67.31**	**0.59**	**0.18**	**98.20**	**14.73**	**79.51**

Table 4: Performance of **MP-RNA** and baseline FMs on PGB. "PolyA" stands for Polyadenylation, "Chrom Acc" for Chromatin Accessibility, "Prom Str" for Promoter Strength, "Term Str" for Terminator Strength, "Splice" for Splice Site, "Gene Exp" for Gene Expression, and "Enh Reg" for Enhancer Region.

more parameters than **MP-RNA**. On the other hand, the results of PGB suggest that our pretraining paradigm, i.e., structure pretraining and SN mutation modeling, is adept at handling DNA classification and regression tasks. In short, **MP-RNA** has impressive generalizability from RNA to DNA genome modeling.

4.5 Benchmark Summary

The evaluation results of GB (Table 12) and GUE (Table 10) can be found in the appendix. Overall, the results of GB and GUE indicate that **MP-RNA** has good generalizability on various genomes and species. Intriguingly, our pretraining paradigm is dedicated to the SN-resolution RNA genomic tasks while obtaining top-tier performance on three DNA benchmarks, i.e., PGB, GB and GUE. This implicitly highlights the necessity of utilization of biological domain knowledge in future works.

The DNA FMs, such as Agro-NT, appear to be ineffective in transferring to RNA genomic tasks. We suspect that the modeling resolution matters in this phenomenon because the small SNT-based DNA FM, HyenaDNA, obtains considerable results compared to agro-NT which contains 1B parameters. Considering that **MP-RNA** is pre-trained on smaller sequence data scales than DNA FMs, such as NT and Agro-NT, which use over $300B$ training data, we hypothesize that the genetic information density of RNA may be greater than that of DNA, leading to the generalization ability from RNA to DNA genomic modeling. However, this hypothesis needs to be settled in future works.

5 Related Works

The field of biological sequence modeling, encompassing DNA, RNA, and proteins, has garnered increasing interest over recent years. While protein modeling has been extensively researched, as evidenced by projects such as AlphaFold (Jumper et al., 2021; Evans et al., 2021; Abramson et al., 2024) and ESM (Lin et al., 2022), DNA and RNA modeling have seen relatively less exploration. Among the earlier efforts in genomic sequence modeling, DNABERT (Ji et al., 2021) utilized BERT's architecture (Devlin et al., 2019) to tackle genomic sequence analysis, demonstrating initial success for in-silico genomic tasks. Its successor, DNABERT2 (Zhou et al., 2023), enhanced this approach by switching from k-mers to BPE tokenization, aiming to boost model performance across multi-species genomic data.

Exploring large-scale foundation models (FMs), such as nucleotide transformers (V1 and V2)(Dalla-Torre et al., 2023), AgroNT (Mendoza-Revilla et al., 2023), and SegmentNT (de Almeida et al., 2024), has proven fruitful. These models, with parameters in the billions, have shown considerable promise in DNA genomic modeling, handling model scales up to 2.5 billion and 1 billion parameters, respectively. Although Agro-NT (Mendoza-Revilla et al., 2023) was initially pre-trained on multi-species edible plant DNA sequences, it did not effectively adapt to RNA sequence modeling in subsequent tests. The challenge posed by the extensive lengths of genomes has increasingly shifted focus towards long-range sequence modeling and

the deployment of autoregressive FMs like HyenaDNA (Nguyen et al., 2023) and Evo (Nguyen et al., 2024).

In RNA genomic modeling, early endeavors such as scBERT (Yang et al., 2022), RNABERT (Akiyama and Sakakibara, 2022), RNA-FM (Chen et al., 2022), RNA-MSM (Zhang et al., 2023), and RNAErnie (Wang et al., 2024b) have emerged. These models, however, have only been trained on limited-scale databases due to the high cost of acquiring RNA sequences. Some focus narrowly on specific RNA sequence types like coding sequences (CDS) (Hallee et al., 2023), 5' untranslated regions (5'UTR) (Chu et al., 2024b), 3' untranslated regions (3'UTR) (Yang et al., 2023), or precursor mRNA sequences (Chen et al., 2023), restricting their ability to capture the full diversity of RNA sequences. Despite reports of effective performance, Uni-RNA (Wang et al., 2023) remains closed-source and thus unverifiable in comparative experiments.

Overall, existing FMs often overlook the critical issue of calibrated structure pretraining in RNA genomic modeling. While 5UTR-LM (Chu et al., 2024a) and ERNIE-RNA (Yin et al., 2024) integrate secondary structures into genomic modeling, they do not investigate the effects of the structure annotation quality.

6 Open-source Package

Genomic modeling is at a nascent stage. There is a domain knowledge gap between artificial intelligence and bio-science which results in a significant scarcity of research resources. Although some works for genomic sequence modeling have been proposed in recent years. However, most of these studies only release the model weights without accompanying training, fine-tuning, and benchmark evaluation scripts. To solve this problem, we have created a comprehensive open-source genomic modeling toolkit based on **`MP-RNA`**. This toolkit is designed to offer thorough fine-tuning tutorials and a standardized automated benchmark evaluation system. Please find the brief introduction of the package in Appendix A.

7 Conclusion

Our work focused on the refinement of RNA FMs through the rigorous integration of domain-specific knowledge, particularly secondary structures and single nucleotide mutations, into the pretraining phase. The introduction of a calibration mechanism using temperature scaling further enhanced the accuracy and reliability of our model predictions, aligning them more closely with biological realities.

Acknowledgment

This work was supported in part by the UKRI Future Leaders Fellowship under Grant MR/S017062/1 and MR/X011135/1; in part by NSFC under Grant 62376056 and 62076056; in part by the Royal Society under Grant IES/R2/212077; in part by the EPSRC under Grant 2404317; in part by the Kan Tong Po Fellowship (KTP\R1\231017); and in part by the Amazon Research Award and Alan Turing Fellowship.

Limitations

The limitations of our work are rooted in resource constraints and experimental scope. First, according to the data scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022; Muennighoff et al., 2023), we realize that the proposed RNA foundation model's scale remains small, and the parameterization may not be adequate to exploit the full potential of the OneKP database. Due to resource restrictions, we could not pretrain larger models. Moving forward, we aim to train larger-scale foundation models, anticipating that our contributions will somewhat accelerate the advancement of DNA and RNA foundation models.

Our study primarily relies on in-silico experiments and computational predictions. The absence of in-vivo experimental validation means that the biological relevance and efficacy of the model's predictions in real-world biological systems remain untested. Future research will need to integrate in-vivo experiments to confirm and refine the predictive capabilities of our foundation models in actual biological environments.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. 2024. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3.

Manato Akiyama and Yasubumi Sakakibara. 2022. Informative rna base embedding for rna structural align-

ment and clustering by deep representation learning. *NAR genomics and bioinformatics*, 4(1):lqac012.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. 2021. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203.

Édouard Bonnet, Paweł Rzążewski, and Florian Sikora. 2020. Designing rna secondary structures is hard. *Journal of Computational Biology*, 27(3):302–316.

Eric J. Carpenter, James H. Leebens-Mack, and Michael S. Barker et al. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, 574(7780):679–685.

Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, et al. 2022. Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. *bioRxiv*, pages 2022–08.

Ken Chen, Yue Zhou, Maolin Ding, Yu Wang, Zhixiang Ren, and Yuedong Yang. 2023. Self-supervised learning on millions of pre-mrna sequences improves sequence-based rna splicing prediction. *bioRxiv*, pages 2023–01.

Yanyi Chu, Dan Yu, Yupeng Li, Kaixuan Huang, Yue Shen, Le Cong, Jason Zhang, and Mengdi Wang. 2024a. A 5' utr language model for decoding untranslated regions of mrna and function predictions. *Nature Machine Intelligence*, pages 1–12.

Yanyi Chu, Dan Yu, Yupeng Li, Kaixuan Huang, Yue Shen, Le Cong, Jason Zhang, and Mengdi Wang. 2024b. A 5' utr language model for decoding untranslated regions of mrna and function predictions. *Nature Machine Intelligence*, pages 1–12.

Kizzmekia S Corbett, Darin K Edwards, Sarah R Leist, Olubukola M Abiona, Seyhan Boyoglu-Barnum, Rebecca A Gillespie, Sunny Himansu, Alexandra Schäfer, Cynthia T Ziwawo, Anthony T DiPiazza, et al. 2020. Sars-cov-2 mrna vaccine design enabled by prototype pathogen preparedness. *Nature*, 586(7830):567–571.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P de Almeida, Hassan Sirelkhatim, et al. 2023. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, pages 2023–01.

Padideh Danaee, Mason Rouches, Michelle Wiley, Dezhong Deng, Liang Huang, and David Hendrix. 2018. bprna: large-scale automated annotation and analysis of rna secondary structure. *Nucleic acids research*, 46(11):5381–5394.

Bernardo P de Almeida, Hugo Dalla-Torre, Guillaume Richard, Christopher Blum, Lorenz Hexemer, Maxence Gélard, Javier Mendoza-Revilla, Priyanka Pandey, Stefan Laurent, Marie Lopez, et al. 2024. Segmentnt: annotating the genome at single-nucleotide resolution with dna foundation models. *bioRxiv*, pages 2024–03.

Sarah K Denny and William J Greenleaf. 2019. Linking rna sequence, structure, and function on massively parallel high-throughput sequencers. *Cold Spring Harbor Perspectives in Biology*, 11(10):a032300.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, Olaf Ronneberger, Sebastian Bodenstein, Michal Zielinski, Alex Bridgland, Anna Potapenko, Andrew Cowie, Kathryn Tunyasuvunakool, Rishub Jain, Ellen Clancy, Pushmeet Kohli, John Jumper, and Demis Hassabis. 2021. Protein complex prediction with alphafold-multimer. *bioRxiv*.

Laura R Ganser, Megan L Kelly, Daniel Herschlag, and Hashim M Al-Hashimi. 2019. The roles of structural dynamics in the cellular functions of rnas. *Nature reviews Molecular cell biology*, 20(8):474–489.

Tiansu Gong, Fusong Ju, and Dongbo Bu. 2024. Accurate prediction of rna secondary structure including pseudoknots through solving minimum-cost flow with learned potentials. *Communications Biology*, 7(1):297.

Katarína Grešová, Vlastimil Martinek, David Čechák, Petr Šimeček, and Panagiotis Alexiou. 2023. Genomic benchmarks: a collection of datasets for genomic sequence classification. *BMC Genomic Data*, 24(1):25.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Logan Hallee, Nikolaos Rafailidis, and Jason P Gleghorn. 2023. cdsbert-extending protein language models with codon awareness. *bioRxiv*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals,

and Laurent Sifre. 2022. Training compute-optimal large language models. *CoRR*, abs/2203.15556.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V. Davuluri. 2021. DNABERT: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinform.*, 37(15):2112–2120.

Veronica Juan and Charles Wilson. 1999. Rna secondary structure prediction based on free energy and phylogenetic analysis. *Journal of molecular biology*, 289(4):935–947.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589.

Ioanna Kalvari, Eric P Nawrocki, Nancy Ontiveros-Palacios, Joanna Argasinska, Kevin Lamkiewicz, Manja Marz, Sam Griffiths-Jones, Claire Toffano-Nioche, Daniel Gautheret, Zasha Weinberg, et al. 2021. Rfam 14: expanded coverage of metagenomic, viral and microrna families. *Nucleic Acids Research*, 49(D1):D192–D200.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. 2022. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902.

Ronny Lorenz, Stephan H Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. 2011. Viennarna package 2.0. *Algorithms for molecular biology*, 6:1–14.

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

David H Mathews and Douglas H Turner. 2006. Prediction of rna secondary structure by free energy minimization. *Current opinion in structural biology*, 16(3):270–278.

Javier Mendoza-Revilla, Evan Trop, Liam Gonzalez, Masa Roller, Hugo Dalla-Torre, Bernardo P de Almeida, Guillaume Richard, Jonathan Caton, Nicolas Lopez Carranza, Marcin Skwark, et al. 2023. A foundational large language model for edible plant genomes. *bioRxiv*, pages 2023–10.

Milad Miladi, Martin Raden, Sven Diederichs, and Rolf Backofen. 2020. Mutarna: analysis and visualization of mutation-induced changes in rna structure. *Nucleic acids research*, 48(W1):W287–W291.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling data-constrained language models. *CoRR*, abs/2305.16264.

Anthony M Mustoe, Charles L Brooks, and Hashim M Al-Hashimi. 2014. Hierarchy of rna functional dynamics. *Annual review of biochemistry*, 83:441–466.

Eric Nguyen, Michael Poli, Matthew G Durrant, Armin W Thomas, Brian Kang, Jeremy Sullivan, Madelena Y Ng, Ashley Lewis, Aman Patel, Aaron Lou, et al. 2024. Sequence modeling and design from molecular to genome scale with evo. *bioRxiv*, pages 2024–02.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin W. Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton M. Rabideau, Stefano Massaroli, Yoshua Bengio, Stefano Ermon, Stephen A. Baccus, and Christopher Ré. 2023. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *CoRR*, abs/2306.15794.

John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Antoni Rafalski. 2002. Applications of single nucleotide polymorphisms in crop genetics. *Current opinion in plant biology*, 5(2):94–100.

Frederic Runge, Jörg K.H. Franke, Daniel Fertmann, Rolf Backofen, and Frank Hutter. 2023. Partial rna design. *bioRxiv*.

Mehdi Saman Booy, Alexander Ilin, and Pekka Orponen. 2022. Rna secondary structure prediction with convolutional neural networks. *BMC bioinformatics*, 23(1):58.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. 2024. Caduceus: Bi-directional equivariant long-range dna sequence modeling. *arXiv preprint arXiv:2403.03234*.

Barkur S Shastry. 2002. Snp alleles in human disease and evolution. *Journal of human genetics*, 47(11):561–566.

T Nicolai Siegel, Kapila Gunasekera, George AM Cross, and Torsten Ochsenreiter. 2011. Gene expression in trypanosoma brucei: lessons from high-throughput rna sequencing. *Trends in parasitology*, 27(10):434–441.

Zhen Tan, Yinghan Fu, Gaurav Sharma, and David H Mathews. 2017. Turbofold ii: Rna structural alignment and secondary structure prediction informed by multiple homologs. *Nucleic acids research*, 45(20):11570–11581.

Ignacio Tinoco Jr and Carlos Bustamante. 1999. How rna folds. *Journal of molecular biology*, 293(2):271–281.

Ning Wang, Jiang Bian, Yuchen Li, Xuhong Li, Shahid Mumtaz, Linghe Kong, and Haoyi Xiong. 2024a. Multi-purpose rna language modelling with motif-aware pretraining and type-guided fine-tuning. *Nature Machine Intelligence*, pages 1–10.

Ning Wang, Jiang Bian, Yuchen Li, Xuhong Li, Shahid Mumtaz, Linghe Kong, and Haoyi Xiong. 2024b. Multi-purpose rna language modelling with motif-aware pretraining and type-guided fine-tuning. *Nature Machine Intelligence*, pages 1–10.

Xi Wang, Ruichu Gu, Zhiyuan Chen, Yongge Li, Xiaohong Ji, Guolin Ke, and Han Wen. 2023. Unirna: universal pre-trained models revolutionize rna research. *bioRxiv*, pages 2023–07.

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. 2022. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nat. Mac. Intell.*, 4(10):852–866.

Yuning Yang, Gen Li, Kuan Pang, Wuxinhao Cao, Xiangtao Li, and Zhaolei Zhang. 2023. Deciphering 3'utr mediated gene regulation using interpretable deep representation learning. *bioRxiv*, pages 2023–09.

Weijie Yin, Zhaoyu Zhang, Liang He, Rui Jiang, Shuo Zhang, Gan Liu, Xuegong Zhang, Tao Qin, and Zhen Xie. 2024. Ernie-rna: An rna language model with structure-enhanced representations. *bioRxiv*, pages 2024–03.

Yikun Zhang, Mei Lang, Jiuhong Jiang, Zhiqiang Gao, Fan Xu, Thomas Litfin, Ke Chen, Jaswinder Singh, Xiansong Huang, Guoli Song, et al. 2024. Multiple sequence alignment-based rna language model and its application to structural inference. *Nucleic Acids Research*, 52(1):e3–e3.

Ying Zhang, Fang Ge, Fuyi Li, Xibei Yang, Jiangning Song, and Dong-Jun Yu. 2023. Prediction of multiple types of rna modifications via biological language model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana V. Davuluri, and Han Liu. 2023. DNABERT-2: efficient foundation model and benchmark for multi-species genome. *CoRR*, abs/2306.15006.

A Open-source Package

Genomic modeling is at a nascent stage. There is a domain knowledge gap between artificial intelligence and bio-science which results in a significant scarcity of research resources. Although some works for genomic sequence modeling have been proposed in recent years. However, most of these studies only release the model weights without accompanying training, fine-tuning, and benchmark evaluation scripts. To solve this problem, we have created a comprehensive open-source genomic modeling toolkit based on **MP-RNA**. This toolkit is designed to offer thorough fine-tuning tutorials and a standardized automated benchmark evaluation system. Please find the brief introduction of the package in Appendix A.

Here are the key features of the **MP-RNA** Package:

- **Fine-Tuning Tutorials**: Our tutorials guide users through the entire fine-tuning process for various downstream genomic modeling tasks, from dataset processing and model setup to actual training. These include a detailed example of secondary structure fine-tuning that demonstrates both training and application. Access these tutorials here: `https://github.com/yangheng95/OmniGenomeBench/tree/master/examples`

- **Automated Benchmark Evaluation**: Our toolkit includes an automated benchmark evaluation tool that features predefined configurations for benchmarking subtasks, including necessary hyperparameters. This tool facilitates the seamless and fair evaluation of future FMs and the integration of new benchmarks. Learn more about this process through our tutorial, available at: `https://github.com/yangheng95/OmniGenomeBench/tree/master/examples/benchmarks/run_rgb_auto_bench.py`

- **Genomic Repository Hub**: To address the issue of limited resources, we have developed a repository hub that hosts open-source licensed datasets, model checkpoints, and benchmark evaluations, alongside flexible interfaces for community sharing of datasets and models.

This hub enhances resource availability and collaboration. The hub will be accessible shortly.

We are finalizing the documentation and will officially launch this tool soon.

B Evaluation Baselines

To help understand the baseline FMs, we briefly summarize the FM in the following sections. Please find the method and experiment details of these FMs in the original publications.

- **DNABERT-2** (Zhou et al., 2023). This is a recent FM tailored for DNA sequence learning, utilizing BPE for RNA tokenization to improve DNABERT.
- **HyenaDNA** (Nguyen et al., 2023). Designed as an autoregressive FM for genome long-range genome modeling, HyenaDNA excels in processing extensive genomic data, especially for DNA, supporting sequence lengths up to $1M$ nucleotides.
- **Caduceus** (Schiff et al., 2024). Caduceus[3] is an advanced DNA language model built on the MambaDNA architecture, designed to address challenges in genomic sequence modelling, such as long-range token interactions and reverse complementarity (RC).
- **Nucleotide Transformers (NT)** (Dalla-Torre et al., 2023). The NT-series FMs are trained on comprehensive DNA genomic datasets, including the human reference genome and multi-species DNA sequences. These FMs are intended to discern patterns within nucleotide sequences across a spectrum of genomic tasks.
- **Agro-NT** (Mendoza-Revilla et al., 2023). Similar in scope to the Nucleotide Transformers, Argo-NT is a high-capacity DNA FM with $1B$ parameters, specifically concentrating on plant DNA.
- **SpliceBERT** (Chen et al., 2023). Specialized for RNA splicing, SpliceBERT is trained on $2M$ precursor messenger RNA (pre-mRNA) sequences, focusing on the intricate dynamics of pre-mRNA processing.
- **CDSBERT** (Hallee et al., 2023). Adapted from ProteinBERT, CDSBERT is trained extensively on CDS regions within genomes. It utilizes these datasets to predict protein structures and functions, benefiting from the alignment between RNA and protein sequences.
- **3UTRBERT** (Yang et al., 2023). This FM is trained on $20k$ sequences of 3'UTRs, tailored for tasks related to 3'UTR-mediated gene regulation. It distinguishes itself from **MP-RNA** by employing k-mers tokenization rather than SNT.
- **RNA-BERT** (Akiyama and Sakakibara, 2022). RNA-BERT is a BERT-style model pre-trained on a large corpus of non-coding RNA sequences. It uses masked language modelling (MLM) as its primary training objective. The model is designed to predict RNA structural alignments and can be fine-tuned for various RNA sequence classification and regression tasks
- **RNA-MSM** (Zhang et al., 2024) RNA-MSM is an unsupervised RNA language model based on multiple sequence alignment (MSA). It is the first model of its kind to produce embeddings and attention maps that directly correlate with RNA secondary structure and solvent accessibility. RNA-MSM is particularly effective for tasks involving evolutionary relationships in RNA sequences.
- **RNA-FM** (Chen et al., 2022) RNA-FM is a BERT-based RNA foundation model trained on a vast dataset of non-coding RNA sequences. The model excels in predicting RNA structure and function by leveraging masked language modelling (MLM) during pre-training. RNA-FM's training data is sourced from the RNAcentral database, providing it with extensive knowledge across diverse RNA species.
- **MP-RNA** and **MP-RNA***(ours): These are our FMs for RNA genomic modeling. **MP-RNA** is the formal variant of our FM that is obtained from incremental pretraining and **MP-RNA***is the variant without incremental pretraining.

C Zero-shot Sequence to Structure Prediction Evaluation

This subsection assesses **MP-RNA** in zero-shot secondary structure prediction. The experimental results are available in Table 5.

In Table 5, ViennaRNA performs well generally but is consistently outperformed by **MP-RNA**-186M, suggesting that **MP-RNA**-186M can provide superior performance on specific types of RNA structure predictions. The results demonstrate the effectiveness of **MP-RNA** in handling complex and diverse RNA structures across different SSP datasets, with

[3] https://huggingface.co/kuleshov-group/caduceus-ps_seqlen-131k_d_model-256_n_layer-16

Model	Archive2	bpRNA	RNAStralign	Rfam
ViennaRNA	**75.89**	**27.82**	62.80	75.32
MP-RNA*-52M	63.82	24.41	64.68	76.08
MP-RNA*-186M	65.57	24.71	73.07	78.60
MP-RNA-186M	68.01	25.41	**74.89**	**82.15**

Table 5: Performance in zero-shot RNA secondary structure prediction across various benchmarks, showcasing the capabilities of **MP-RNA** in zero-shot learning scenarios without any fine-tuning or domain adaptation.

consistent improvements over other FMs. In other words, our pretraining paradigm works well in unleashing the SN resolution genomic modeling tasks, and it can be adapted to future works on genomic FMs.

D Benchmark Results

D.1 RNA Genomic Benchmark

The detailed task descriptions for each nucleic acid and species, including the number of examples, classes, evaluation metric, and sequence length, are outlined in Table 6. Each task is carefully curated to reflect the complexity and variety inherent in genomic data, providing a robust framework for assessing the nuanced capabilities of state-of-the-art RNA FMs. RGB contains 6 SN-level tasks that are curated or collected from published articles. The purpose of RGB is to benchmark genomic FMs in challenging SN-level modeling tasks such as detection and repair of SN mutations, mRNA sequence degradation rates, and RNA secondary structure prediction. Due to the lack of a plant RNA benchmark dataset, RGB includes the modeling of RNA sequences from a variety of species, e.g., plant and human. The sequence length in RGB ranges from 107 to 512, which is sufficient for most RNA understanding tasks. In summary, these multi-species and SN-level tasks in RGB serve as the first comprehensive benchmark utilized to assess the RNA sequence modeling capabilities of **MP-RNA** and its baseline models. The brief introduction of the datasets in RGB is as follows:

- **Single-Nucleotide Mutation Detection (SNMD)**: We developed a plant RNA dataset synthesizing the single-nucleotide mutations. Focused on identifying potential single nucleotide changes, this task is essential for detecting mutations linked to genetic disorders. The SNMD dataset introduces up to 10 random mutations in the original sequences, regardless of variation ratios. Cross-entropy is utilized as the loss function for this binary token classification task.

- **Single-Nucleotide Mutation Repair (SNMR)**: This task challenges the model to suggest corrective actions at the single nucleotide level, aiding in gene therapy approaches. The SNMR dataset mirrors the SNMD dataset, with cross-entropy as the loss function, indicating a token 4-way (i.e., `A`, `U`, `C`, `G`) classification task.

- **mRNA Degrade Rate Prediction (mRNA)**: Estimating the decay rate of nucleotides in mRNA sequences, this task is vital for deciphering gene expression and regulation. The dataset originates from the Kaggle COVID-19 vaccine design competition[4], focusing solely on sequence-based degradation rate prediction and excluding RNA structures. It's a token regression task using MSE as the loss function, with the dataset resplit into training, validation, and testing sets for evaluation.

- **RNA Secondary Structure Prediction (bpRNA & Archive2 & RNAStralign & Rfam)**: Aiming to predict RNA folding into secondary structures, this task is fundamental to RNA functionality and interactions. We evaluated **MP-RNA** on four datasets, bpRNA (Danaee et al., 2018) (TR0, VL0, TS0 sets), ArchiveII (Mathews, 2019), RNAStralign (Tan et al., 2017) and Rfam (Kalvari et al., 2021). Following existing works, we have excluded sequences over 512 bases and complex structures, simplifying to three symbols: '`(`', '`.`', '`)`'Ṙesults may not directly compare with other studies due to these modifications. Cross-entropy serves as the loss function.

Please find the appendix for the input and output examples of each subtask in RGB. The detailed task descriptions for each nucleic acid and species, including the number of examples, classes, evaluation metric, and sequence length, are outlined in Table 6. Each task is carefully curated to reflect the complexity and variety inherent in genomic data, providing a robust framework for assessing the nuanced capabilities of state-of-the-art RNA FMs.

[4] https://www.kaggle.com/competitions/stanford-covid-vaccine

Task	Task Type	# of examples	# of classes	Metric	Sequence length	Source
SNMD	Token classification	$8,000/1,000/1,000$	2	AUC	200	This work
SNMR	Token classification	$8,000/1,000/1,000$	4	macro F1	200	This work
mRNA	Token regression	$1,735/193/192$	—	RMSE	107	Kaggle
bpRNA	Token classification	$10,814/1,300/1,305$	3	macro F1	≤ 512	(Danaee et al., 2018)
AchiveII	Token classification	$2278/285/285$	3	macro F1	≤ 500	(Mathews, 2019)
RNAStrAlign	Token classification	$17483/2186/2185$	3	macro F1	≤ 500	(Tan et al., 2017)
Rfam*	Token classification	$501376/62672/62672$	3	macro F1	≤ 512	(Kalvari et al., 2021)

Table 6: The brief statistics of subtasks in the RGB. These benchmark datasets are held out or not included in the pretraining database. The numbers of examples in training, validation and testing sets are separated by "/". * indicates the datasets are used for zero-shot performance evaluation only.

Table 7 show the virtual examples of different datasets in RGB. Please refer to our supplementary materials to find the datasets for more details.

D.2 Plant Genomic Benchmark

The Plant Genomic Benchmark (Mendoza-Revilla et al., 2023) (PGB) provides a comprehensive suite of datasets designed to evaluate and improve the predictive capabilities of genomic models in plant biology. This benchmark, as shown in Table 8, encompasses a range of critical genomic tasks[5], including binary classification, single and multi-variable regression, and multi-label classification, addressing various aspects of plant genomics such as RNA processing, gene expression, and chromatin accessibility. By integrating diverse genomic tasks, the PGB aims to facilitate advanced research and development in plant genomics, offering a robust platform for the assessment and enhancement of model performance across different plant species. To obtain a detailed description of PGB, please refer to Agro-NT (Mendoza-Revilla et al., 2023).

D.3 Genomic Understanding Evaluation

The Genome Understanding Evaluation (Zhou et al., 2023) serves as a DNA genomic benchmark, encompassing 36 datasets across nine crucial genome analysis tasks applicable to a variety of species. Similar to PGB and GB, it is used for evaluating the generalizability of **MP-RNA** on DNA genome benchmarking. To thoroughly assess the capabilities of genome foundation models across sequences of varying lengths, tasks have been chosen with input lengths spanning from 70 to $10,000$. The brief statistics for each dataset included in the GUE benchmark are displayed in Table 9, and the task descriptions are available in Zhang et al. (2023). Due to resource limitations, we do not include large-scale FMs in this benchmark, e.g., agro-NT and CDSBERT. Besides, we run the evaluation on a subset of GUE, where for each task we randomly select at most 10k samples from the original splits, e.g., training, testing and validation (if any) sets.

The benchmark results on GUE are available in Table 10. Although the performance of **MP-RNA**-186M is not the best on all datasets, we can still observe a clear conclusion that **MP-RNA**-186M achieves top-tier performance even without being pretrained on any DNA genome database. The performance on GUE suggests that while some FMs are tailored for specific genomic tasks (e.g., SpliceBERT for splice sites), **MP-RNA**-186M, an FM designed for RNA genome, provides robust across-the-board efficacy. The variation in performance across different tasks and species highlights that there could be strong generalizability among genomic tasks, only if we take the biological domain knowledge into the training of FMs.

D.4 Genomic Benchmarks

The genomic benchmark (GB) is also a DNA-oriented FM benchmark suite, which can be used for generalizability evaluation of **MP-RNA**-186M. It contains a well-curated collection of datasets designed for the classification of genomic sequences, focusing on regulatory elements across multiple model organisms. This collection facilitates robust comparative analysis and development of genomic FMs. The task names in the original repository are complex, we abbreviate the names as follows:

- DEM corresponds to "Demo Coding vs Intergenomic Seqs"
- DOW is for "Demo Human or Worm"
- DRE represents "Drosophila Enhancers Stark"

[5]https://huggingface.co/InstaDeepAI/agro-nucleotide-transformer-1b

Genome Type	Dataset		Examples
RNA	SNMD	Input Sequence	G A G T A ... T T G A G
		True Label	0 0 1 0 0 ... 0 0 1 0 0
		Prediction	0 0 0 0 0 ... 0 0 1 0 0
	SNMR	Input Sequence	T A C G A ... C T G A T
		True Label	T A C A A ... G T A A T
		Prediction	T A C A A ... C T G A T
	mRNA	Input Sequence	G G ... A C
		True Label	[0.1,0.3,0.2] [0.8,0.4,0.1] ... [0.9,0.4,0.3] [0.5,0.2,0.6]
		Prediction	[0.1,0.3,0.2] [0.8,0.4,0.1] ... [0.9,0.4,0.3] [0.5,0.2,0.6]
	bpRNA	Input Sequence	G G C G A ... C U U U U
		True Label	(((· · ... · ·)))
		Prediction	((((· ... ·))))
	Archive2	Input Sequence	A G U A G ... U U U G C U
		True Label	(((· · ... · ·)))
		Prediction	(((· · ... · ·)))
	RNAStralign	Input Sequence	A G U A G ... U U U G C U
		True Label	(((· · ... · ·)))
		Prediction	(((· · ... · ·)))
	Rfam	Input Sequence	A G U A G ... U U U G C U
		True Label	(((· · ... · ·)))
		Prediction	(((· · ... · ·)))

Table 7: The virtual input and output examples in RGB. The "..." represents the sequences that are omitted for better presentation and the red color indicates the wrong prediction in classification tasks. In the mRNA dataset, all single nucleotides have three values to predict. Note that "T" and "U" can be regarded as the same symbol in RNA sequences and depend on different datasets.

Task	# of datasets	Task Type	Total # of examples	# of classes	Metric	Sequence length
Polyadenylation	6	Sequence classification	738,918	2	macro F1	400
Splice site	2	Sequence classification	4,920,835	2	macro F1	398
LncRNA	2	Sequence classification	58,062	6	macro F1	101 – 6000
Promoter strength	2	Sequence regression	147,966	—	RMSE	170
Terminator strength	2	Sequence regression	106,818	—	RMSE	170
Chromatin accessibility	7	Multi-label classification	5,149,696	9 – 19	macro F1	1,000
Gene expression	6	Multi-variable regression	206,358	—	RMSE	6,000
Enhancer region	1	Sequence classification	18,893	2	macro F1	1,000

Table 8: The genomic tasks in the Plant Genomic Benchmark. This table briefly enumerates each task by name, the number of datasets available, the type of classification or regression analysis required, the range of sequence lengths, and the total number of samples in each dataset. Please find the dataset details of PGB in Agro-NT.

- HCE is short for "Human Enhancers Cohn"
- HEE denotes "Human Enhancers Ensembl"
- HRE abbreviates "Human Ensembl Regulatory"
- HNP shortens "Human Nontata Promoters"
- HOR is an abbreviation for "Human Ocr Ensembl"
- DME simplifies "Dummy Mouse Enhancers Ensembl"

The brief statistics for each dataset included in the GUE benchmark are displayed in Table 9. Similar to GUE, we run the evaluation on a subset of GB, where for each task we randomly select at most 10k samples from the original splits, e.g., training, testing and validation (if any) sets.

The experimental results presented in Table 12 demonstrate that **MP-RNA**-186M consistently achieves competitive performance across a diverse array of genomic tasks. Notably, **MP-RNA**-186M excels in the Human Ensembl Regulatory (HRE) task with an F1 score of 95.66, outperforming other models like DNABERT-2 and HyenaDNA in this specific benchmark. Additionally, **MP-RNA**-186M shows robust results in tasks involving enhancer predictions (HEE) and non-TATA promoters (HNP), underscoring its versatility and effective-

Task	Metric	Datasets	Training	Validation	Testing
Core Promoter Detection	macro F1	tata	4904	613	613
		notata	42452	5307	5307
		all	47356	5920	5920
Promoter Detection	macro F1	tata	4904	613	613
		notata	42452	5307	5307
		all	47356	5920	5920
Transcription Factor Prediction (Human)	macro F1	wgEncodeEH000552	32378	1000	1000
		wgEncodeEH000606	30672	1000	1000
		wgEncodeEH001546	19000	1000	1000
		wgEncodeEH001776	27497	1000	1000
		wgEncodeEH002829	19000	1000	1000
Splice Site Prediction	macro F1	reconstructed	36496	4562	4562
Transcription Factor Prediction (Mouse)	macro F1	Ch12Nrf2\iggrab	6478	810	810
		Ch12Zrf384hpa004051\iggrab	5395	674	674
		MelJun\iggrab	2620	328	328
		MelMafkDm2p5dStd	1904	239	239
		MelNelf\iggrab	15064	1883	1883
Epigenetic Marks Prediction	macro F1	H3	11971	1497	1497
		H3K14ac	26438	3305	3305
		H3K36me3	29704	3488	3488
		H3K4me1	25341	3168	3168
		H3K4me2	24545	3069	3069
		H3K4me3	29439	3680	3680
		H3K79me3	23069	2884	2884
		H3K9ac	22224	2779	2779
		H4	11679	1461	1461
		H4ac	27275	3410	3410
Covid Variant Classification	macro F1	Covid	77669	7000	7000
Enhancer Promoter Interaction	macro F1	GM12878	10000	2000	2000
		HeLa-S3	10000	2000	2000
		HUVEC	10000	2000	2000
		IMR90	10000	2000	2000
		K562	10000	2000	2000
		NHEK	10000	2000	2000
Species Classification	macro F1	fungi	8000	1000	1000
		virus	4000	500	500

Table 9: Statistics of tasks in the GUE, these details can be found in Section B.2. from Zhang et al. (2023).

ness in processing complex genomic sequences. These findings highlight the advanced capabilities of **MP-RNA**-186M in handling intricate genomic data, contributing significantly to the field of genomic research.

E Ablation Experiments

E.1 Modeling Efficiency

Table 13 presents the experimental results comparing three tokenization methods: SNT, BPE, and 3-mers (k=3), across different RNA datasets (mRNA, SNMD, SNMR, Archive2, bpRNA, and RNAStralign). The table reports the average sequence length (Len.), GPU memory consumption (Mem.), and training time per epoch (T./E.) for each method. The results demonstrate that although the average lengths for BPE and 3-mers are reduced by approximately threefold, the GPU memory usage and training time remain almost the same, indicating that the length of the modeling task is tied to the number of nucleotide labels, which necessitates padding the tokenized inputs to match the label length.

F The 1KP Initiative

The 1000 Plant Transcriptomes Initiative (1KP) was a comprehensive effort aimed at exploring genetic diversity across the green plant kingdom

Model	Model Performance (macro F1 Score)						
	Yeast EMP	Mouse TF-M	Virus CVC	Human TF-H	Human PD	Human CPD	Human SSP
DNABERT-2	75.85	**86.23**	58.23	81.80	90.17	82.57	85.21
HyenaDNA	73.08	73.44	27.59	77.62	91.19	84.31	83.34
NT-V2	74.93	78.10	27.49	79.12	90.87	84.70	84.13
SpliceBERT	77.66	84.97	47.17	**82.77**	**92.24**	83.96	**93.81**
3UTRBERT	71.89	71.46	34.84	74.85	82.37	**90.51**	81.95
MP-RNA (ours)	**78.51**	84.72	**64.41**	81.73	90.04	85.22	90.39

Table 10: The performance on GUE for **MP-RNA** and baseline FMs, where the results are reimplemented based on our evaluation protocol. The performance for each task is the average macro F1 score in all sub-datasets.

Task	# of Sequences	# of Classes	Class Ratio	Median Length	Standard Deviation
DME	1210	2	1.0	2381	984.4
DEM	100000	2	1.0	200	0.0
DOW	100000	2	1.0	200	0.0
DRE	6914	2	1.0	2142	285.5
HCE	27791	2	1.0	500	0.0
HEE	154842	2	1.0	269	122.6
HRE	289061	3	1.2	401	184.3
HNP	36131	2	1.2	251	0.0
HOR	174756	2	1.0	315	108.1

Table 11: The brief statistics of datasets reported in the genomic benchmark (Grešová et al., 2023).

Model	DEM F1	DOW F1	DRE F1	DME F1	HCE F1	HEE F1	HRE F1	HNP F1	HOR F1
DNABERT-2	92.67	95.17	43.77	77.21	75.58	80.66	78.14	85.80	68.03
HyenaDNA	88.21	94.13	70.11	76.44	70.38	79.58	96.33	85.99	67.03
NT-V2	91.66	94.32	78.20	81.72	71.98	79.85	93.30	85.30	68.53
SpliceBERT	94.72	96.42	72.29	74.70	73.50	79.60	95.23	89.57	68.89
3UTRBERT	89.50	90.22	74.35	80.14	70.23	76.33	98.47	82.49	66.78
MP-RNA-186M	94.16	93.49	77.17	80.34	73.51	82.23	95.66	87.87	68.97

Table 12: Performance of **MP-RNA** and baseline FMs across different tasks in the genomic benchmarks (GB), where the results are reimplemented based on our evaluation protocol. The performance (macro F1) for each task is the average macro F1 score in all sub-datasets.

Tokenization	mRNA			SNMD			SNMR			Archive2			bpRNA			RNAStralign		
	Len.	Mem.	T./E.	Len.	Mem.	T./E.	Len.	Mem.	T./E.	Len.	Mem.	T./E.	Len.	Mem.	T./E.	Len.	Mem.	T./E.
SNT	109	5459	76	202	7589	110	202	7589	110	142	19043	98	139	18913	458	114	19047	764
BPE	48	5459	76	42	7589	110	42	7589	110	29	19043	98	29	18913	458	27	19047	764
3-mers	38	5459	76	69	7589	110	69	7589	110	46	19043	98	46	18913	458	38	19047	764

Table 13: Comparison of tokenization methods (SNT, BPE, and 3-mers) on different RNA datasets. We report the average sequence length (Len.), GPU memory occupation (Mem., in MB), and training time per epoch (T./E., in seconds) for each dataset. The experiment uses the MP-RNA-186M model and evaluates on RGB datasets.

(Viridiplantae), sequencing the RNA from 1124 (1342 in other version) samples that represent over 1000 species, encompassing all major taxa within Viridiplantae. This includes streptophyte and chlorophyte green algae, bryophytes, ferns, angiosperms, and gymnosperms. The initiative's final or capstone publication presents three major analyses: inferring species trees, identifying whole genome duplications, and detecting gene family expansions. These findings are particularly valuable for plant and evolutionary scientists interested in specific gene families, whether their focus is across the entire green plant tree of life or within more narrowly defined lineages.

The sampling strategy for the 1KP was global and collaborative, with samples sourced from a wide range of environments including wild field collections, greenhouses, botanical gardens, laboratory specimens, and algal culture collections. The initiative prioritized the collection of live growing cells, such as young leaves, flowers, or shoots, to ensure a high abundance of expressed genes, though many samples also came from roots and other tissues. RNA extraction was performed using well-established protocols or commercial kits, facilitating the comprehensive analysis of transcribed RNA across this diverse set of species. This monumental effort not only sheds light on plant genetic diversity but also provides a rich data resource for ongoing and future research in plant science and evolutionary biology.

Ethics Statement

In this research, we utilized the open OneKP dataset, which does not contain human-related privacy. We ensure that such data is not exploited without fair compensation and acknowledgment of the source communities. The pretraining sequences are plant-based genomic data that involve potential harm to ecological systems, we do not permit the use of our model out of expectation, such as developing malicious bio-software or designing harmful RNA structures. The models and findings should support, not undermine, the conservation of plant species and their habitats. We adhere to principles of transparency and open science, using datasets that are publicly available and providing clear documentation of our methodologies and findings.

Overall, in conducting this research, we have committed to ethical scientific practices that respect biodiversity and aim to contribute positively to the field of genomic research. We encourage ongoing dialogue around the ethical use of plant RNA sequences and support initiatives that promote the sharing of benefits arising from such research with all stakeholders.