RNAGenesis: Foundation Model for Enhanced RNA Sequence Generation and Structural Insights

Zaixi Zhang*[1], Linlin Chao*[2], Ruofan Jin*†[1,2,3], Yikun Zhang*†[2,4], Guowei Zhou*[2], Yujie Yang[2], Yukang Yang[1], Kaixuan Huang[1], Qirong Yang[2], Ziyao Xu[2], Xiaoming Zhang[2,✉], Le Cong[5,✉], and Mengdi Wang[1,✉]

[1]Princeton University, NJ, USA
[2]BioMap Research, Beijing, China
[3]Zhejiang University, Zhejiang, China
[4]Peking University, Beijing, China
[5]Stanford University, California, USA
*Equal Contribution, † Work done during internship at BioMap
✉zhangxiaoming@biomap.com, congle@stanford.edu, and mengdiw@princeton.edu

ABSTRACT

RNA molecule plays an essential role in a wide range of biological processes. Gaining a deeper understanding of their functions can significantly advance our knowledge of life's mechanisms and drive the development of drugs for various diseases. Recently, advances in RNA foundation models have enabled new approaches to RNA engineering, yet existing methods fall short in generating novel sequences with specific functions. Here, we introduce RNAGenesis, a foundation model that combines RNA sequence understanding and de novo design through latent diffusion. With a Bert-like Transformer encoder with Hybrid N-Gram tokenization for encoding, a Query Transformer for latent space compression, and an autoregressive decoder for sequence generation, RNAGenesis reconstructs RNA sequences from learned representations. Specifically for the generation, a score-based denoising diffusion model is trained to capture the latent distribution of RNA sequences. RNAGenesis outperforms current methods in RNA sequence understanding, achieving the best results in 9 of 13 benchmarks (especially in RNA structure prediction), and further excels in designing natural-like aptamers and optimized CRISPR sgRNAs with desirable properties. Our work establishes RNAGenesis as a powerful tool for RNA-based therapeutics and biotechnology.

Introduction

RNA plays a pivotal role in a wide array of biological processes, including protein synthesis, enzymatic activities, and gene regulations. Serving as a key component in central dogma and the intermediary between DNA and proteins, RNA encompasses various types including messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA)[1]. Moreover, an expanding repertoire of non-coding RNAs (ncRNAs) has expanded our understanding of RNA functions[2], like microRNA[3], small interfering RNA (siRNA)[4], and long non-coding RNA (lncRNA)[5], each contributing to critical cellular functions such as metabolism, transport, and signal transduction. Furthermore, specialized RNA molecules such as CRISPR single guide RNA (sgRNA)[6] and aptamers[7] illustrate RNA's versatility in gene editing and therapeutic applications. The synthesizability and programmability of RNA make it an attractive candidate for advancements in therapeutic development and drug design[8,9]. As a cornerstone of molecular biology and biomedicine, understanding the multifaceted roles of RNA is essential for unraveling the complexities of cellular processes and the molecular underpinnings of diseases, thereby highlighting the significance of developing comprehensive RNA foundation models to advance research and clinical applications.

Recently, there have been a plethora of RNA Foundation Models pre-trained on vast RNA sequence data to learn evolutionary information for comprehensive downstream tasks[10–16]. Most of them take encoder-only transformer frameworks and leverage the Masked Language Modeling (MLM) as the pertaining strategy. For example, RNA-FM[11] and RiNALMo[15] are pioneering Bert-like Foundation models trained on RNACentral datasets and show remarkable generalization capability on downstream tasks of RNA families not encountered in the training set. RNAErnie[17] enhances pretraining by incorporating RNA motifs as biological priors and tokenizes RNA types (for example, miRNA, lnRNA) as stop words for type-guided fine-tuning. AIDO.RNA[16] scales these advancements by pushing the limits of model size (1.6B), setting new benchmarks on a comprehensive set of RNA tasks. There are also some specialized RNA Foundation models focusing on certain RNA types. For instance, UTR-LM[12] specializes in modeling untranslated regions (UTRs), uncovering critical regulatory motifs that influence gene expression and translation initiation. CodonBERT[18] uniquely focuses on modeling RNA sequences at the codon level, enabling a better understanding of

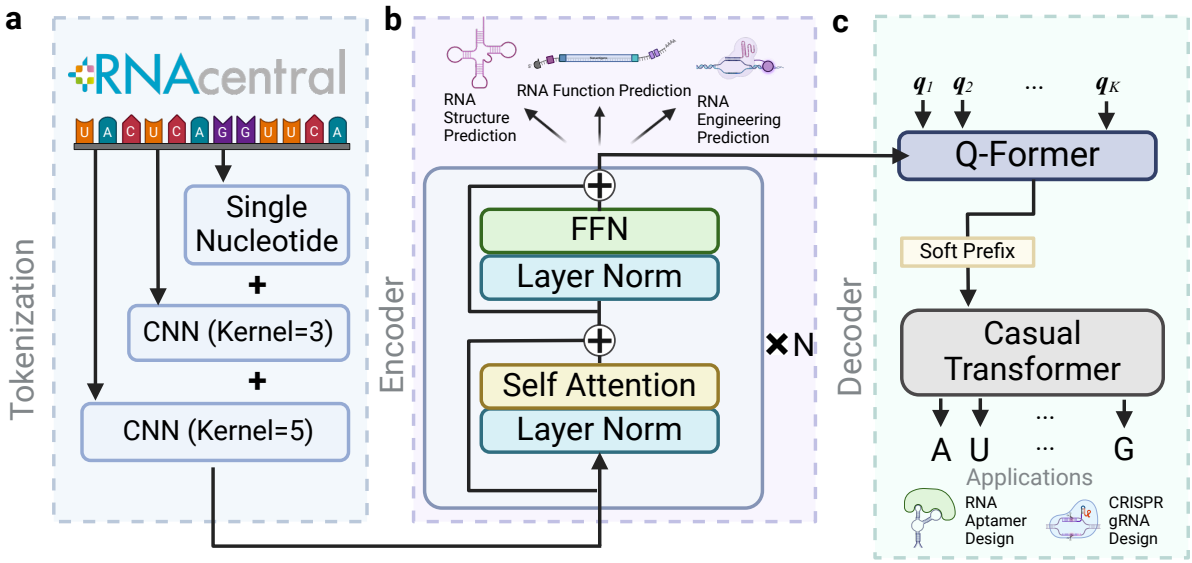


Figure 1. Overview of RNAGenesis. The components of RNAGenesis consist of Hybrid N-Gram Tokenization with single nucleotide representation and 1D CNNs (different kernel sizes) (**a**), Transformer layers as Encoder backbone (**b**), and Q-Former+Casual Transformer as Decoder for RNA sequence generation (**c**). The learned representation from Encoder can be used in downstream tasks including RNA Structure Prediction, RNA Function Prediction, and RNA Engineering Prediction. The Decoder can generate RNA sequences for Aptamer and CRISPR sgRNA design.

codon usage bias and its impact on translation efficiency.

However, most conventional methods lack the capability for sequence generation, thereby limiting the exploration and design of novel RNA molecules with specific functionalities. The varying length of RNA sequences as well as the complex motif semantics pose great challenges to sequence generation. In this paper, we propose RNAGenesis that bridges RNA sequence understanding and *de novo* sequence design with latent diffusion[19] (Figure. 1). The Encoder is a Bert-like Transformer with Hybrid N-Gram tokenization to well capture multi-granularity context information. Then, a Query Transformer[20] is used to compress the representations from the Encoder into a set of fixed-length latent vectors. The autoregressive decoder can reconstruct RNA sequences from these latent variables and we train a train a score-based denoising diffusion model to capture the distribution of RNAs in latent space. Extensive experiments show that RNAGenesis not only achieves superior performance on RNA sequence understanding tasks (best in 9 of 13), but also excels at designing natural-like RNA aptamers and CRISPR sgRNA.

Benchmark Results

Following BEACON[21], we evaluate RNAGenesis on three categories of tasks (Structure Prediction, RNA Function Prediction, and Engineering Prediction) and compare it with two kinds of baseline models (naive supervised models and pretrained RNA Foundation models).

The naive supervised models include: CNN[22], ResNet[23], and LSTM[23]. Following the design choices outlined in BEACON[21], the CNN consists of two layers, ResNet includes eight residual blocks, and the LSTM is configured with three Bi-LSTM layers. These models have approximately 5.4M, 11M, and 26.7M parameters, respectively. The RNA Foundation models include RNA-FM[11], RNABERT[10], RNA-MSM[24], SpliceBERT[13], 3UTRBERT[14], and UTR-LM[12]. For the Secondary Structure Prediction task, we further benchmark with RiNALMo[15] and AIDO.RNA[16]. These models vary significantly in size, ranging from 0.48M to 1.6B parameters, and are pre-trained on diverse RNA data sources such as ncRNA, pre-mRNA, mRNA-3'UTR, and mRNA-5'UTR. To ensure consistency, all models are fine-tuned using identical settings.

In Table. 1, we show the results of RNAGenesis compared with baseline methods and observed that it achieved the best results on 9 out of 13 downstream tasks, demonstrating the superior advantage of our model design. In particular, RNAGenesis demonstrates impressive performance on RNA structure prediction tasks. For example, RNAGenesis achieves Top-L precision of 71.16 at contact map prediction tasks while the second-best only achieves a score of 61.20. As for the secondary structure prediction task, we further benchmarked with RiNALMo and AIDO.RNA, the latest and largest RNA Foundation models so far in Figure. 2. RNAGenesis also achieves comparable or better performance on Precision, Recall, and F1 values, showing the strong capability of RNAGenesis for structure prediction.

Table 1. Benchmark results of 13 RNA tasks from BEACON benchmark. Our RNAGenesis is compared with naive supervised methods (e.g., CNN) and pretrained RNA Foundation Models (e.g., RNA-FM). Three color scales of blue first, second, and third are used to denote best performance among all the methods. The baseline results are borrowed from BEACON[21]. Mean (%) is reported for each experiment.

Task	SSP	CMP	DMP	SSI	SPL	APA	NcRNA	Modif	MRL	VDP	PRS	CRI-On	CRI-Off
Metric	F1 (%)	P@L (%)	R^2 (%)	R^2 (%)	ACC@K (%)	R^2 (%)	ACC (%)	AUC (%)	R^2 (%)	MCRMSE↓	R^2 (%)	SC (%)	SC (%)
CNN	49.95	43.89	27.76	34.36	8.43	50.93	88.62	70.87	74.13	0.361	45.40	29.69	11.40
ResNet	57.26	59.59	30.26	37.74	21.15	56.45	88.33	71.03	74.34	0.349	55.21	28.55	11.50
LSTM	58.61	40.41	44.77	35.44	36.66	67.03	88.78	94.83	83.94	0.329	55.45	26.83	8.60
RNA-FM	68.50	47.56	51.45	42.36	34.84	70.32	96.81	94.98	79.47	0.347	55.98	31.62	2.49
RNABERT	57.27	45.21	48.19	31.62	0.18	57.66	68.95	82.82	29.79	0.378	54.60	29.77	4.27
RNA-MSM	57.98	57.26	37.49	39.22	38.33	70.40	84.85	94.89	83.48	0.330	56.94	34.92	3.85
Splice-H510	64.93	45.80	55.56	38.91	44.80	58.65	95.92	62.57	83.49	0.321	54.90	26.61	4.00
Splice-MS510	43.24	52.64	10.27	38.58	50.55	52.46	95.87	55.87	84.98	0.315	50.98	27.13	3.49
Splice-MS1024	68.26	47.32	55.89	39.22	48.52	60.03	96.05	53.45	67.15	0.313	57.72	27.59	5.00
UTR-LM-MRL	59.71	45.51	55.21	39.52	36.20	64.99	89.97	56.41	77.78	0.325	57.28	28.49	4.28
UTR-LM-TE&EL	59.57	60.32	54.94	40.15	37.35	72.09	81.33	59.70	82.50	0.319	53.37	32.49	2.91
UTRBERT-3mer	60.37	51.03	50.95	34.31	44.24	69.52	92.88	95.14	83.89	0.337	56.83	29.92	4.48
UTRBERT-4mer	59.41	44.91	47.77	33.22	42.04	72.71	94.32	95.10	82.90	0.341	56.43	23.20	3.11
UTRBERT-5mer	47.92	44.71	48.67	31.27	39.19	72.70	93.04	94.78	75.64	0.343	57.16	25.74	3.93
UTRBERT-6mer	38.56	51.56	50.02	29.93	38.58	71.17	93.12	95.08	83.60	0.340	57.14	28.60	4.90
BEACON-B	64.18	60.81	56.28	38.78	37.43	70.59	94.63	94.74	72.29	0.320	54.67	26.01	4.42
BEACON-B512	58.75	61.20	56.82	39.13	37.24	72.00	94.99	94.92	72.35	0.320	55.20	28.17	3.82
RNAGenesis	73.72	71.16	70.31	-	37.70	89.15	97.81	95.14	85.75	-	58.12	35.55	-

Structure Prediction

Accurate RNA structural predictions are crucial for understanding RNA's structural and functional dynamics, offering insights into functional regions and interaction sites that are vital for drug discovery and genetic research. Identifying nucleotide interactions and predicting inter-nucleotide distances help infer RNA tertiary structures, advancing our knowledge of RNA folding and function. Structural score imputation further enhances RNA research by supplementing missing comprehensive structural information To summarize, the considered Structure Prediction tasks include Secondary Structure Prediction (SSP), Contact Map Prediction (CMP), Distance Map Prediction (DMP), and Structural Score Imputation (SSI). The detailed experimental settings are listed below.

Secondary structure prediction (SSP) involves the identification of paired regions (stems) and unpaired regions (loops, bulges, junctions) within RNA molecules. The target matrix, denoted as $y \in \mathbb{R}^{l \times l}$, indicates the formation of base pairs between nucleotides, reflecting the RNA molecule's secondary structure. In this study, we utilize the bpRNA-1m database[25], which provides comprehensive annotations for over 100,000 single-molecule RNA structures. The evaluation metric employed is the F1 score, with the threshold set to 0.5 to maintain alignment with other papers.

Contact Map Prediction (CMP) aims to identify pairs of nucleotides in RNA that are in close proximity in their three-dimensional structures. Each nucleotide pair is assigned a binary label $y \in \{0, 1\}$, where 1 indicates that the nucleotides are in contact (i.e., within a distance threshold of 8 Å), and 0 otherwise. Following[26], we utilize a dataset derived from non-redundant RNA 3D structures documented by[27], and evaluate the predictions using the Top-L precision metric.

Distance Map Prediction (DMP) aims to estimate the physical distances between pairs of nucleotides within an RNA molecule. The target distance matrix $y \in \mathbb{R}^{l \times l}$ records the distance between every pair of nucleotides within the sequence. The same dataset used for the contact map prediction task is employed, with R^2 serving as the evaluation metric.

Structural Score Imputation (SSI) predicts missing structural information in RNA molecules by assigning an experimentally derived structural score $y \in \mathbb{R}$ to each nucleotide. The dataset[28] is based on icSHAPE sequencing data from the HEK293 cell line. During training, 30% of the nucleotide scores are randomly masked as null, while downsampling introduces missing values in 3,095 fragments within the test set. The performance of SSI is evaluated using the coefficient of determination (R^2).

RNA Function Prediction

Accurate RNA analyses are essential for advancing our understanding of gene expression, regulation, and function. The RNA function prediction tasks include Splice Site Prediction (SPL), APA Isoform Prediction (APA), Non-coding RNA Function

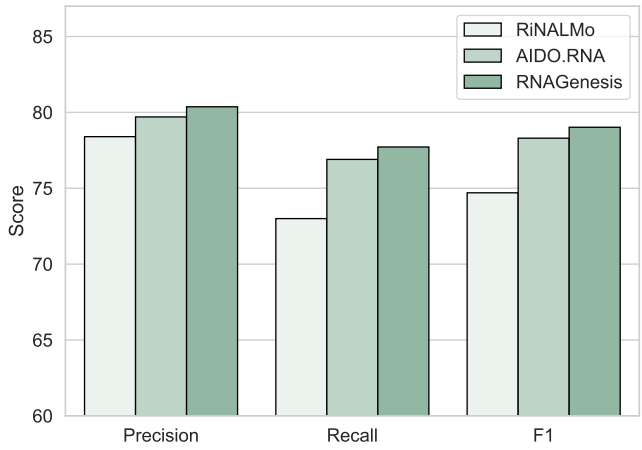


Figure 2. Benchmarking RNAGenesis with two state-of-the-art baselines RiNALMo and AIDO.RNA on Secondary Structure Prediction. Precision, Recall, and F1 scores are reported. The Recall and F1 values of RiNALMo and AIDO.RAN are borrowed from their original papers. The Precision values are then calculated based on Recall and F1.

Classification (ncRNA), Mean Ribosome Loading (MRL), and Modification Prediction (Modif). The details are described below.

Splice Site Prediction (SPL) classifies each base within a sequence into one of three categories: acceptor (a), donor (d), or neither (n), with the categorical label $y \in \{0, 1, 2\}$. The task utilizes Jaganathan's dataset[29] with Top-k accuracy serving as the evaluation metric.

APA Isoform Prediction (APA) predicts the usage ratio of the proximal polyadenylation site (PAS) in the 3' untranslated region (3' UTR) for each variant, recorded in the target $y \in \mathbb{R}$. We filter 228k sequences from over 3 million APA reporter gene data from Bogard's dataset[30], and this regression task assesses the proportion of proximal APA isoforms. The evaluation metric is the R^2 value.

Non-coding RNA Function Classification (ncRNA) involves the categorization of non-coding RNA molecules into distinct functional groups such as microRNAs (miRNAs), long non-coding RNAs (lncRNAs), and small interfering RNAs (siRNAs). Each molecule is assigned a categorical label $y \in \{0, 1, \ldots, 12\}$, which represents its functional classification. The dataset[31,32] utilized for this classification task integrates contributions from GENCODE, circBase, and Rfam, encompassing a broad range of ncRNA types. The evaluation metric for performance is the accuracy (ACC) at the sequence level.

Modification Prediction(Modif) aims to predict twelve commonly occurring types of RNA modifications from a given RNA sequence, with each modification represented by a categorical label $y \in \{0, 1, \ldots, 11\}$. We utilize Song's dataset, which consists of 20 epi-transcriptome profiles corresponding to 12 distinct types of RNA modifications, derived from 15 base-resolution technologies. This dataset includes over 300,000 modification sites, which are partitioned into training, validation, and test sets. The evaluation metric used in this task is AUC.

Mean Ribosome Loading (MRL) predicts the MRL value for a given sequence, with the target $y \in \mathbb{R}$ representing the level of mRNA translation activity into proteins. The data from Reid's dataset[33], which includes 91,519 5' UTR sequences and their variants, are utilized to compute the MRL for each sequence. The model's performance is evaluated using the R^2 value.

Engineering Prediction

Advancements in predictive models for vaccine stability, programmable RNA switches, and CRISPR technology are driving innovation in medicine and genetic engineering, enabling safer, more precise, and personalized therapeutic solutions. Here, we consider Engineering Prediction tasks including Vaccine Degradation Prediction (VDP), Programmable RNA Switches (PRS), CRISPR On-Target Prediction (CRI-On), and CRISPR Off-Target Prediction (CRI-Off). The detailed settings are listed below.

Vaccine Degradation Prediction (VDP) predicts the stability and shelf life of vaccines under various environmental conditions. For each nucleotide, three properties are recorded in the target vector $y \in \mathbb{R}^3$. We use data from the "Stanford OpenVaccine"[34] competition on Kaggle and the RNA design platform Eterna, which provides detailed measurements for 6,043 diverse RNA constructs. The model's performance is evaluated using the Mean Columnwise Root Mean Squared Error (MCRMSE) metric.

Programmable RNA Switches (PRS) involves the identification of synthetic RNA molecules capable of altering their conformation and function in response to specific signals. The target $y \in \mathbb{R}^3$ records the ON, OFF, and ON/OFF state activities

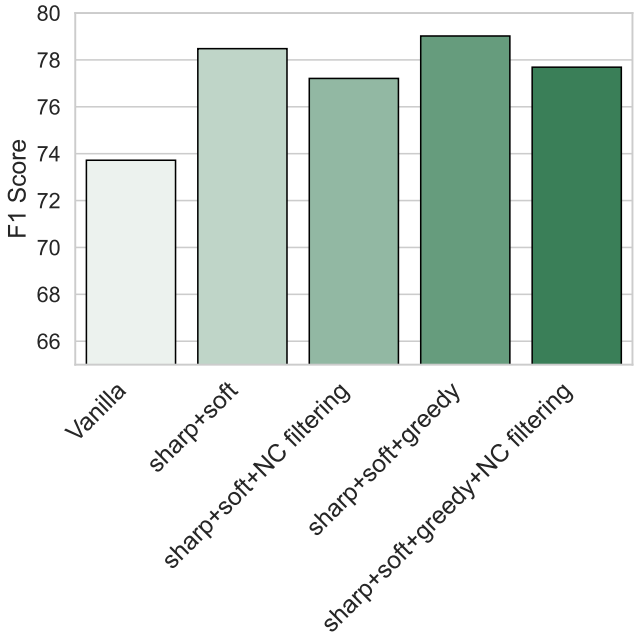


Figure 3. Ablation studies of RNAGenesis on Secondary Structure Prediction. The post-processing strategies include disallowing sharp hairpin loop (**sharp**), ignoring non-canonical base pairs (**NC filtering**), greedy approach for pairing (**greedy**), and considering a pairing to be correct if the predicted one differs by at most one nucleotide (**soft**). Besides, the classification threshold was tuned on the validation set.

given an RNA sequence. The dataset, analyzed by Angenent-Mari[35], consists of 91,534 toehold switches in vivo, spanning 23 viral genomes and 906 human transcription factors, with GFP signal intensity measurements indicating the ON and OFF state activity levels[36]. The effectiveness of these switches is evaluated using the R^2 metric.

CRISPR On-Target Prediction(CRI-On) evaluates the efficiency of single-guide RNAs (sgRNAs) directed by Cas proteins in gene editing at specific target sites. The knockout efficacy of each sgRNA is quantified and represented as the target $y \in \mathbb{R}$. The dataset[37] includes approximately 15,000 sgRNAs targeting 1,071 genes across four different cell lines, with performance assessed using the Weighted Spearman correlation coefficient.

CRISPR Off-Target Prediction (CRI-Off) predicts the likelihood and frequency of CRISPR-induced mutations at unintended genomic locations. The specificity of sgRNAs is quantified using a target $y \in \mathbb{R}$, representing the frequency of off-target cleavage events. The evaluation dataset[37] consists of data for approximately 160,000 potential off-target sites across 30 sgRNAs in various cell types. The model's performance is assessed using the Weighted Spearman correlation coefficient.

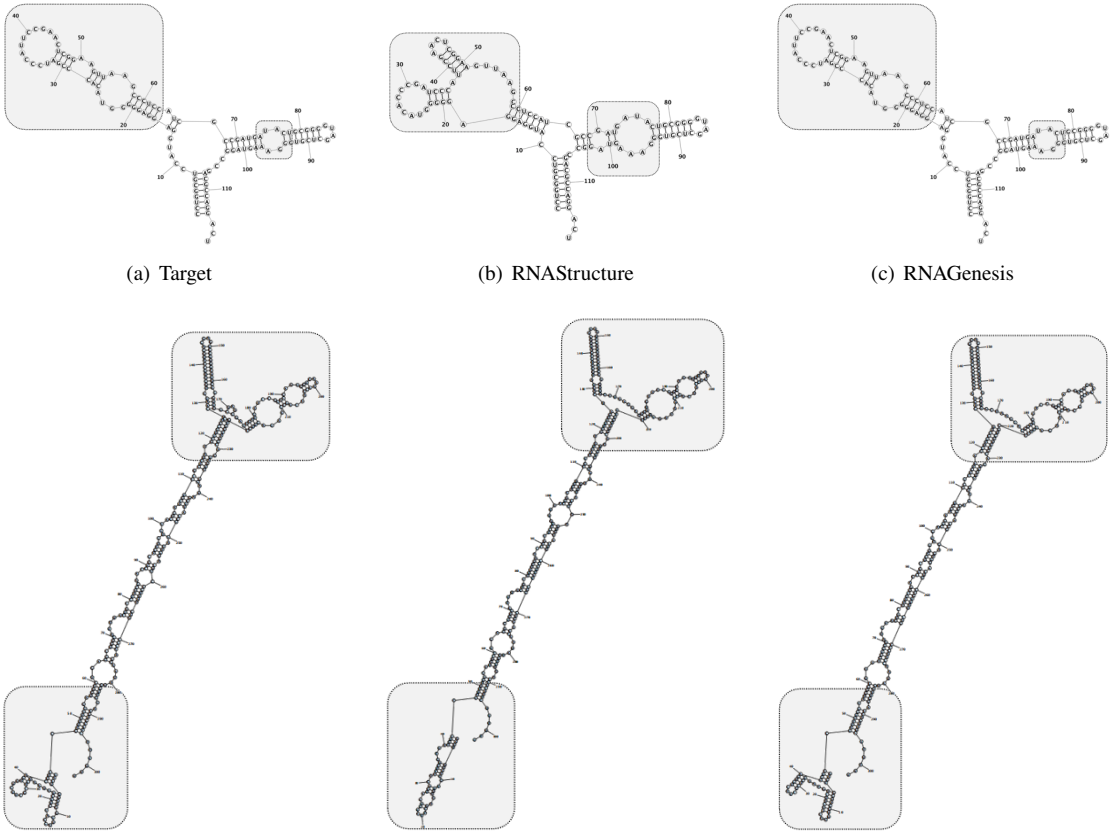
Further Exploration of RNAGenesis for Secondary Structure Prediction

Besides benchmark results, we here aim to provide further insights of secondary structure predictions. When evaluating prediction results, it is crucial to account for the inherent structural dynamics of RNA. Consequently, predictions that are "close enough" to the true pairings are treated as correct. Following the approach proposed by Mathews et al.[38], we therefore consider a pairing (i, j) to be correct if the predicted pair differs by at most one nucleotide in either index. Specifically, if (i, j) is the true pairing, then $(i \pm 1, j)$ and $(i, j \pm 1)$ are also counted as correct predictions (**soft**). To maintain alignment with Mathews et al.[38], we also determined the optimal threshold for F1 score calculation based on the results from the validation set. To convert base-pairing probabilities into a valid secondary structure, we implemented a simple greedy approach (**greedy**). We iteratively fix the nucleotide pair with the highest pairing probability and then exclude all clashing pairs from further iterations. Throughout this process, we ignore non-canonical base pairs (**NC filtering**) and disallow any pairing that would produce a "sharp" hairpin loop (i.e., the ith nucleotide cannot be paired with the jth nucleotide if $|i - j| < 4$) (**sharp**).

In Figure. 3, we show the results of ablation studies. We observe that sharp+soft+greedy leads to the best performance while ignoring non-canonical base pairs brings a drop in performance. This may be explained by the fact that non-canonical base pairs frequently contribute to RNA stability and folding, especially in complex motifs. By filtering them out, the model loses potentially informative interactions, resulting in predictions that deviate more strongly from the true conformation. Consequently, while the **sharp** and **soft** adjustments effectively capture local structural flexibility, discarding these non-canonical pairings

restricts the model's ability to identify important stabilizing contacts, thus lowering predictive accuracy.

In Figure. 4, we demonstrate the secondary structure prediction examples compared with RNAStructure[39], which relies on thermodynamic models. Compared to thermodynamic methods, our model RNAGenesis can achieve more accurate predictions. At the same time, we have found that our model can predict very well on short sequences, and there is room for improvement on long sequences, which we will further optimize in the future.



(a) Target (b) RNAStructure (c) RNAGenesis

(d) Target (e) RNAStructure (f) RNAGenesis

Figure 4. Example of secondary structure prediction. The examples show RNAGenesis can achieve nearly perfect prediction results on short sequences, and it also performs well on longer sequences, although the results are slightly inferior. In the future, we will further enhance the prediction capability for long sequences. The gray boxes highlight the RNA regions where RNAGenesis achieves better prediction than RNAStructure.

Aptamer Design

Aptamers are short RNA molecules that can bind to specific targets with high affinity and specificity, making them valuable tools in therapeutics, diagnostics, and biotechnology[40,41]. To evaluate the utility of RNAGenesis for in silico aptamer design, we assessed its capability to generate samples that align with key attributes of the real SELEX data from the HT-SELEX dataset A, as provided by AptaDiff[42]. Specifically, a latent diffusion model was trained while freezing RNAGenesis's encoder to generate RNA sequence embeddings. The generated aptamers were then evaluated using eight biological metrics: Levenshtein distance, 4-mer distance, G/C content, minimum free energy, secondary structure, Fréchet Inception Distance, DNA base composition ratios, and motif clustering analysis following previous works[19,42]. Additionally, latent space similarities were visualized using t-SNE maps.

We conducted a comparative analysis with RNA-FM and natural RNA sequences. Both RNAGenesis and RNA-FM generated 64,000 samples, and 13,802 samples (same as HT-SELEX dataset A) were randomly selected from each for evaluation. As illustrated in Figure. 5 and Figure. 6, the sequences generated by RNAGenesis exhibit higher similarity to natural sequences compared to those generated by RNA-FM. This advantage is evident across sequence-level metrics (e.g., Levenshtein distance, 4-mer distance, G/C content, DNA base composition, and Fréchet Inception Distance), structural-level metrics (e.g., minimum

free energy and secondary structure distances), and latent space organization visualized through t-SNE maps. Furthermore, motif clustering analysis using MEME[43] revealed closer alignment of RNAGenesis-generated sequences with authentic aptamers. As shown in Figure. 6, MEME's E-value scores indicate better motif clustering, with lower values reflecting improved clustering precision in specific categories.

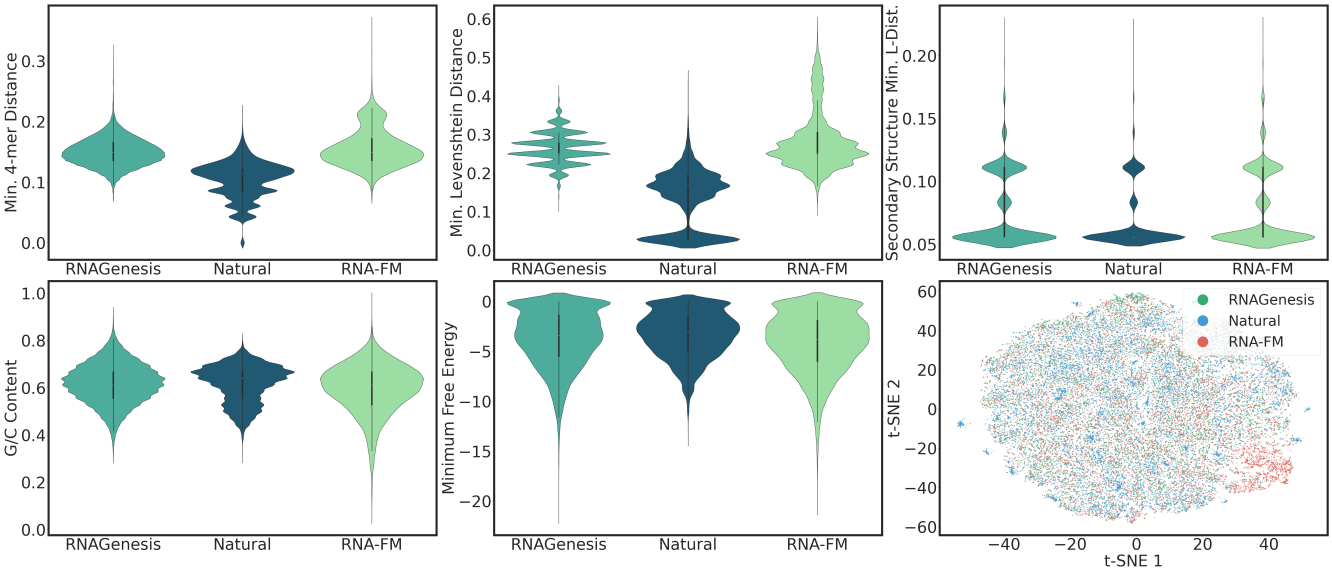


Figure 5. Comparison of generated aptamer sequences (RNAGenesis, RNA-FM, and natural aptamer sequences) based on the HT-SELEX dataset A, comprising 13,802 RNA samples. The evaluation metrics include minimum 4-mer distances, minimum sequence Levenshtein distances, minimum Levenshtein distances of RNA secondary structure, G/C content ratios, and minimum free energy, along with t-SNE visualizations of generated RNAs in latent embedding space.

CRISPR Optimization

Single guide RNAs (sgRNAs) are essential components of the CRISPR genome-editing system, where their design and optimization significantly impact the specificity and efficiency of target DNA cleavage, driving advancements in gene editing, therapeutics, and biotechnology[6,44]. In this section, we explore the potential of RNAGenesis for guiding the design and optimization of sgRNAs for CRISPR systems. Building upon our previous work with CRISPR-GPT[45], which demonstrated the ability to predict off-target effects and enhance sgRNA specificity, we aim to further refine these capabilities by integrating the powerful sequence generation and optimization features of RNAGenesis. Although we do not present experimental results in this version of preprint, we outline the approach for utilizing RNAGenesis in sgRNA design. Specifically, we leverage RNAGenesis' ability to understand RNA sequences and their structural context to design high-affinity sgRNAs that are optimized for minimal off-target cleavage. The model generates candidate sgRNAs by considering both the target sequence and surrounding genomic context, thereby improving on traditional methods that rely on heuristic rules. In future work, we will validate these predictions through experimental benchmarking, comparing the performance of RNAGenesis-designed sgRNAs against state-of-the-art CRISPR systems.

Discussion

The development of RNAGenesis marks a significant advancement in RNA sequence generation, offering a robust framework that integrates RNA sequence understanding with de novo design capabilities. By leveraging latent diffusion, our model overcomes key limitations of existing methods, enabling the design of functional RNA molecules with high specificity. Extensive benchmarks demonstrate its superior performance in both sequence understanding and sequence design, particularly in the generation of aptamers and CRISPR sgRNAs. These results underscore the potential of RNAGenesis to drive innovations in RNA-based therapeutics, biotechnology, and synthetic biology, providing a valuable tool for advancing RNA engineering across diverse applications.

While RNAGenesis achieves substantial progress in RNA sequence modeling and generation, there are inherent limitations that warrant attention. First, our model currently operates on RNA sequences in isolation, which may not fully leverage the broader context provided by RNA structure or functional annotations. Future work could explore multi-modal models that

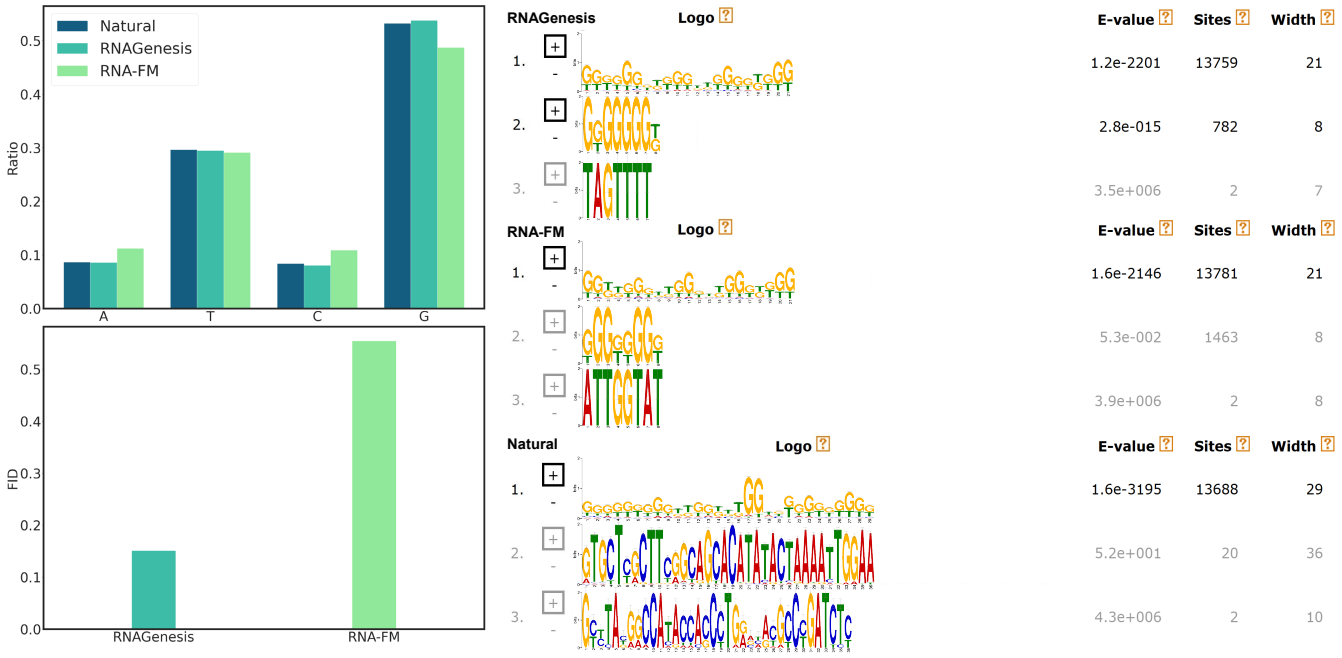


Figure 6. Comparison of generated aptamer sequences (RNAGenesis, RNA-FM, and natural aptamer sequences) based on Fréchet Inception Distance (FID), DNA base ratio, and motif clustering analysis by MEME suite.

integrate sequence information with structural data (e.g., secondary and tertiary structures) and functional annotations (e.g., gene expression or RNA modifications). Such multi-modal frameworks could enable a more comprehensive understanding of RNA function and its role in cellular processes.

Another potential avenue for improvement is the development of a unified model that can simultaneously handle both coding and non-coding RNAs. Currently, RNAGenesis is optimized for general RNA sequence tasks, but distinct models for coding versus non-coding RNAs may be required to capture the unique characteristics and functions of these classes of RNAs. A unified model that can differentiate between coding and non-coding sequences could streamline workflows and improve the accuracy of predictions across various RNA types.

Finally, looking further ahead, an exciting direction would be to extend RNAGenesis to a unified model that can handle RNA, DNA, and protein sequences in a single framework[46,47]. Such a unified model would enable researchers to investigate the intricate interactions between these fundamental biological molecules and provide insights into complex processes like gene expression regulation, protein-RNA interactions, and the central dogma[1].

Methods

Pretraining Datasets

We leverage the preprocessed dataset from RNACentral[48] for the pertaining, which totally has 42 million sequences. To reduce the impact of data redundancy on the model, we used MMSeq to cluster the sequences with a similarity threshold of 90%. After clustering, the total number of cluster centers was reduced to 6.4 million. Figure. 7 illustrates the distribution changes of RNA samples based on type before and after clustering. During model pretraining, we first randomly selected cluster centers and then sampled from the corresponding clusters to form a batch.

Model Architecture

Hybrid N-Gram Tokenization: A hybrid tokenization scheme is adopted including single nucleotide tokenization and N-Gram tokenization with CNN. RNA sequences are primarily composed of four nucleotide units A,U,C,G, and single nucleotide tokenization can represent the basic information. The N-Gram tokenization based on 1D CNN with different Kernel sizes can learn the higher level of interactions such as RNA motifs. The CNN transformation is unique in RNAGenesis compared with baseline models such as RiNALMo[15] and Uni-RNA[16]. The input into the Encoder is the sum of the hybrid tokenization representations. After tokenization, each nucleotide is transformed into a 1280-dimensional vector.

Encoder: The backbone of RNAGenesis is an Bert-style encoder-only Transformer that focuses on RNA sequence modeling and understanding. We denote the encoder by Enc which maps a sequence $\boldsymbol{x} = (\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^L)$ to a sequence of contextualized

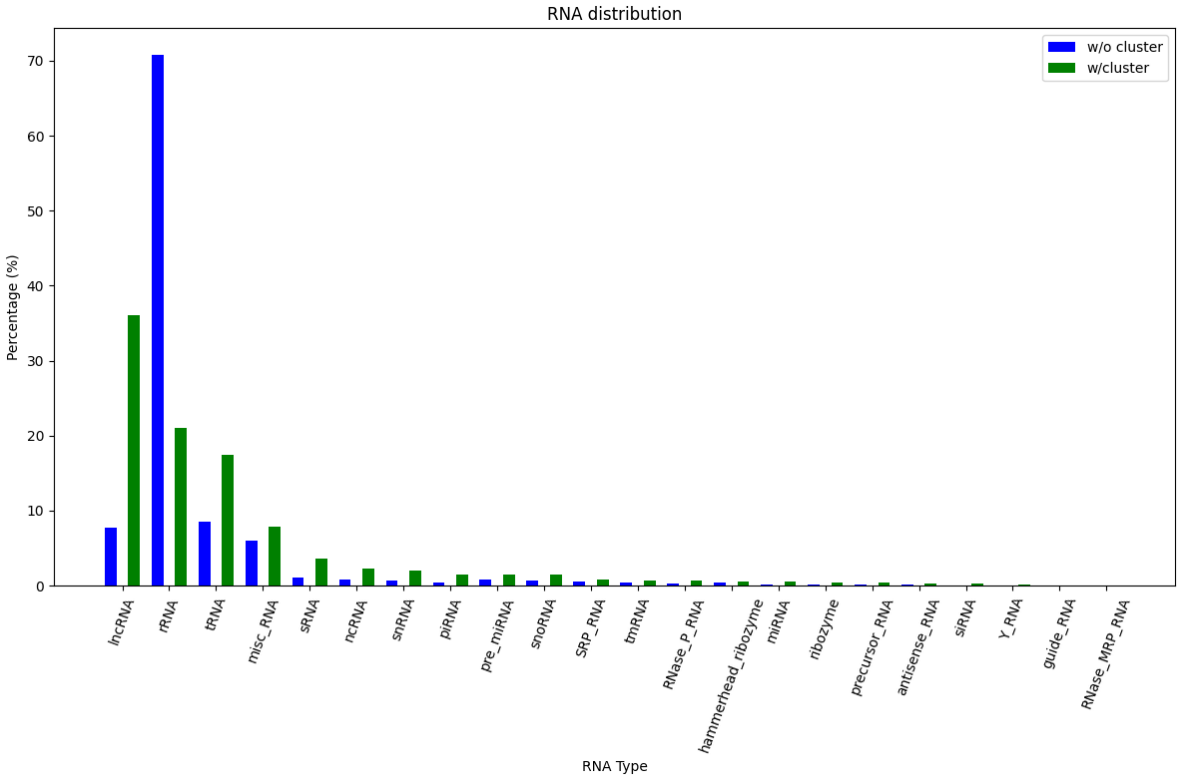


Figure 7. Data distribution comparison between clustered and non-clustered version.

token embeddings of the same length L. At its core lies the self-attention mechanism[49], which enables the capture of both local and global contextual information. RNAGenesis is composed of 32 Transformer blocks, each containing a multi-head attention mechanism and a feed-forward network (FFN). Positional encoding is provided by RoPE[50], effectively representing both relative and absolute positional cues. Each multi-head attention module has 20 attention heads. To enhance pre-training efficiency at this scale, we integrate IO-aware FlashAttention[51], a fast and memory-efficient exact attention mechanism. In the FFN, we employ two linear layers coupled with the SwiGLU activation function[52], which merges the benefits of the Swish activation function and gated linear units (GLU). Residual connections link the Transformer modules, and layer normalization[53] is applied inside each residual block to stabilize training and ensure well-behaved gradients at initialization.

Decoder: Besides encoding RNA sequences for understanding tasks, RNAGenesis has the unique capability of generating *de novo* RNA sequences. A Querying Transformer (QFormer)[20], denoted by $\text{QFormer}_\phi(\cdot)$, is used to bridge the encoder and the decoder. It takes K trainable query token embeddings $(\boldsymbol{q}^1, \ldots, \boldsymbol{q}^K)$ as input, conducts cross-attention with the embeddings from the Encoder, and summarizes the information into K latent vectors $z = (z^1, z^2, \ldots, z^K)$. K is a fixed number that does not depend on the input length L. Such a Q-Former preserves the key information for decoding while helping reduce the dimensionality by limiting K.

For decoding, we employ a standard causal Transformer[49], denoted by Dec_ψ. We first use a linear projection layer to convert the K latent vectors into a set of soft prompt embeddings with dimensions that match the decoder's input size. Conditioned on these soft prompts, the decoder then autoregressively produces RNA sequence. With the pretrainedd RNAGenesis encoder, the decoder including the Q-Former and the causal Transformer is trained to reconstruct the input RNA sequences with the following loss function:

$$\mathcal{L}_{\text{reconstruct}}(\psi, \phi; \boldsymbol{x}) = -\log p_{\text{Dec}_\psi}(\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^L \mid \text{QFormer}_\phi(\text{Enc}(\boldsymbol{x}))), \tag{1}$$

Taking the hybrid N-Gram tokenizer, the encoder backbone, and the decoder into consideration, the total model has 865.8M parameters.

Latent Diffusion: We leverage a continuous diffusion model[54] to model the distribution of the latent variables z, which generates samples from the target distribution by a series of denoising processes. Specifically, we adopt a denoising network to predict the added noise ϵ from the noisy input $(\sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon)$, where t denotes the time step, $\bar{\alpha}_t$ is the corresponding signal strength level, z_0 is the uncorrupted latent vector, and ϵ is the added Gaussian noise. For the denoising score network $_\theta$,

we utilize transformer-based denoising networks following[55]. We adopt the ϵ-prediction scheme from Denoising Diffusion Probabilistic Models (DDPM)[54] and employ the simplified training objective defined by

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{z}_0=\text{QFormer}_\phi(\text{Enc}(\mathbf{z}))}\mathbb{E}_{t,\epsilon}\|\epsilon - \theta(\sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2, \tag{2}$$

In our framework, the latent embeddings have fixed sizes, facilitating the training of diffusion models and allowing data to be treated as continuous, similar to images. Unlike other discrete diffusion models, our framework does not require truncating or padding sequences to a fixed length.

Training and Finetuning

We pretrained RNAGenesis using a masked language modeling (MLM) objective. In this setup, unlabeled RNA sequences are corrupted and then reconstructed by the model. Each nucleotide in the RNA sequence is treated as a single token. We randomly selected 30% of the tokens in the training sequence. Compared to a 15% mask ratio, our experiments demonstrate that the strategy with a 30% mask ratio is more effective. Among these, 80% are replaced with the special [MASK] token, 10% with a randomly chosen token from the vocabulary, and the remaining 10% are left unchanged. We used ordinary cross-entropy training loss function and our maximum sequence length is 1024. We use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$. The initial learning rate is set to 1e-5, which increases to 1e-4 during the warmup phase and then gradually decays thereafter.

After pre-trained, RNAGenesis's output embeddings can serve as a powerful sequence representation that has embedded structural and evolutionary information. Such a representation can be used as an enriched input to structural, functional, and engineering downstream tasks. We employed RNAGenesis for validation on 10 datasets provided by BEACON[21], achieving state-of-the-art results on 9 of them. The incorporation of the CNN module significantly enhanced the performance of RNAGenesis in structural prediction tasks.

Following BEACON[21], RNAGenesis incorporates three pipelines for different types of tasks.For sequence-level tasks, we do not use the CLS token for prediction; instead, we apply an average readout function to the representations. Then representations are processed through an MLP layer to derive the sequence-level predictions. For tasks that demand resolution at the nucleotide level, each nucleotide's representation is individually processed through an MLP to generate predictions specific to that nucleotide. To examine the relationships between nucleotides, a self-outer product of the nucleotide representations is calculated, resulting in a matrix that encapsulates pairwise interactions among nucleotides. This interaction matrix is subsequently processed through a simple ResNet to derive the final output. Except for the SSP task, which employed full fine-tuning, all other tasks utilized LoRA fine-tuning. For each task, we appropriately adjusted hyperparameters such as batch size, learning rate, and the LoRA ratio to optimize performance.

For aptamer design, we followed RNAdiffusion[19] and performed a comparative evaluation with RNA-FM. Specifically, we froze the pretrained encoder from RNAGenesis and trained the Q-former, modified from the ESM-2[47] model architecture, along with the decoder, which uses the ProGen2-small[56] architecture. The Q-former and decoder were trained from scratch on the ncRNA subset of the Ensembl database, comprising 1.1 million samples, over 2 epochs using two NVIDIA Tesla A100-SXM4-80GB GPUs. This training step encodes RNA sequences of varying lengths into fixed-size embeddings, with $K = 16$ and hidden dimension $D = 160$. Subsequently, a continuous diffusion model was trained using the same dataset for 2 epochs to model the distribution of the latent variables. After pretraining on the ncRNA dataset, the model was fine-tuned on the SELEX dataset A, using the same data-split strategy as AptaDiff[42]. The fine-tuning process was performed over 10 epochs. An identical pipeline was applied to RNA-FM to enable direct comparison.

Baselines

We surveyed and summarized the related RNA Foundation models in Table. 2.

Table 2. Related work of RNA foundation models.

FM	Architecture	Model size	Pre-training dataset	Tokenization	Pre-training objective
RNABERT[10]	encoder-only transformer	<10M	RNAcentral (76,237 human-derived small ncRNA sequences) & Rfam alignment	base	MLM(15%) + others
RNA-FM[11]	encoder-only transformer	100M	RNAcentral v19 (23M ncRNA sequences)	base	MLM(15%)
RNA-MSM[24]	MSA transformer	~200M	Rfam v14.7	base	MLM(20%)
Uni-RNA[16]	encoder-only transformer	<400M	RNAcentral & nt & Genome Warehouse (1B potential RNA sequences)	base	MLM
SpliceBERT[13]	encoder-only transformer	20M	2M pre-mRNA sequences from UCSC genome browser	base	MLM(15%)
CodonBERT[18]	encoder-only transformer	~100M	NCBI (10M mRNA coding sequences)	codon	MLM(15%) + others
3UTRBERT[14]	encoder-only transformer	~100M	86k 3'UTRs curated from GRCh38.p13, Release 40	k-mer	MLM(15%)
UTR-LM[12]	encoder-only transformer	<10M	Ensembl & eGFP & mCherry & Cao (700K 5' UTRs)	base	MLM + others
ATOM-1[57]	encoder-decoder transformer	unknown	chemical mapping data	base	unknown
RiNALMo[15]	encoder-only transformer	650M	RNAcentral & Rfam & nt & Ensembl (36M unique ncRNA sequences)	base	MLM
CaLM[58]	encoder-only transformer	86M	European Nucleotide Archive (9M cDNA sequences)	codon	MLM(25%)
RNAErnie[17]	encoder-only transformer	~100M	RNAcentral (23M ncRNA sequences)	base	MLM(15%) + others
AIDO.RNA[16]	encoder-only transformer	1.6B	RNAcentral v24.0 (42M ncRNA sequences)	base	MLM(15%)
RNAGenesis (ours)	encoder-decoder transformer	866M	RNAcentral v24.0 (42M ncRNA sequences)	Hybrid N-Gram	MLM(30%)

Data availability

This study's training and test data are available at Zenodo (https://zenodo.org/). The project website for RNAGenesis is at https://.

Code availability

The source code of this study is freely available at GitHub (https://github.com/biomap-research/RNAGenesis) to allow for replication of the results of this study.

References

1. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561–563 (1970).
2. Morris, K. V. & Mattick, J. S. The rise of regulatory rna. *Nature Reviews Genetics* **15**, 423–437 (2014).
3. Lu, T. X. & Rothenberg, M. E. Microrna. *Journal of allergy and clinical immunology* **141**, 1202–1207 (2018).
4. Hu, B. *et al.* Therapeutic sirna: state of the art. *Signal transduction and targeted therapy* **5**, 101 (2020).
5. Mercer, T. R., Dinger, M. E. & Mattick, J. S. Long non-coding rnas: insights into functions. *Nature reviews genetics* **10**, 155–159 (2009).
6. Cong, L. *et al.* Multiplex genome engineering using crispr/cas systems. *Science* **339**, 819–823 (2013).
7. Zhou, J., Bobbin, M. L., Burnett, J. C. & Rossi, J. J. Current progress of rna aptamer-based therapeutics. *Frontiers in genetics* **3**, 234 (2012).
8. Wang, F., Zuroske, T., Watts, J. K. *et al.* Rna therapeutics on the rise. *Nat Rev Drug Discov* **19**, 441–442 (2020).
9. Damase, T. R. *et al.* The limitless future of rna therapeutics. *Frontiers in bioengineering and biotechnology* **9**, 628137 (2021).
10. Akiyama, M. & Sakakibara, Y. Informative rna base embedding for rna structural alignment and clustering by deep representation learning. *NAR genomics and bioinformatics* **4**, lqac012 (2022).
11. Chen, J. *et al.* Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. *bioRxiv* 2022–08 (2022).
12. Chu, Y. *et al.* A 5 utr language model for decoding untranslated regions of mrna and function predictions. *Nature Machine Intelligence* 1–12 (2024).
13. Chen, K. *et al.* Self-supervised learning on millions of pre-mrna sequences improves sequence-based rna splicing prediction. *bioRxiv* (2023). URL https://www.biorxiv.org/content/early/2023/05/09/2023.01.31.526427. https://www.biorxiv.org/content/early/2023/05/09/2023.01.31.526427.full.pdf.
14. Yang, Y. *et al.* Deciphering 3' utr mediated gene regulation using interpretable deep representation learning. *bioRxiv* (2023). URL https://www.biorxiv.org/content/early/2023/09/12/2023.09.08.556883. https://www.biorxiv.org/content/early/2023/09/12/2023.09.08.556883.full.pdf.
15. Penić, R. J., Vlašić, T., Huber, R. G., Wan, Y. & Šikić, M. Rinalmo: General-purpose rna language models can generalize well on structure prediction tasks. *arXiv preprint arXiv:2403.00043* (2024).
16. Zou, S. *et al.* A large-scale foundation model for rna function and structure prediction. *bioRxiv* 2024–11 (2024).
17. Wang, N. *et al.* Multi-purpose rna language modelling with motif-aware pretraining and type-guided fine-tuning. *Nature Machine Intelligence* 1–10 (2024).
18. Li, S. *et al.* Codonbert: Large language models for mrna design and optimization. *bioRxiv* 2023–09 (2023).
19. Huang, K. *et al.* Latent diffusion models for controllable rna sequence generation. *arXiv preprint arXiv:2409.09828* (2024).
20. Li, J., Li, D., Savarese, S. & Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742 (PMLR, 2023).
21. Ren, Y. *et al.* Beacon: Benchmark for comprehensive rna tasks and language models. *NeurIPS* (2024).
22. Shanehsazzadeh, A., Belanger, D. & Dohan, D. Is transfer learning necessary for protein landscape prediction? *arXiv preprint arXiv:2011.03443* (2020).
23. Rao, R. *et al.* Evaluating protein transfer learning with tape. *Advances in neural information processing systems* **32** (2019).

24. Zhang, Y. *et al.* Multiple sequence alignment-based RNA language model and its application to structural inference. *Nucleic Acids Research* **52**, e3–e3 (2023). URL https://doi.org/10.1093/nar/gkad1031. https://academic.oup.com/nar/article-pdf/52/1/e3/55443207/gkad1031.pdf.

25. Danaee, P. *et al.* bprna: large-scale automated annotation and analysis of rna secondary structure. *Nucleic acids research* **46**, 5381–5394 (2018).

26. Sun, S., Wang, W., Peng, Z. & Yang, J. Rna inter-nucleotide 3d closeness prediction by deep residual neural networks. *Bioinformatics* **37**, 1093–1098 (2021).

27. Leontis, N. B. & Zirbel, C. L. Nonredundant 3d structure datasets for rna knowledge extraction and benchmarking. In *RNA 3D Structure Analysis and Prediction*, 281–298 (Springer, 2012).

28. Gong, J., Xu, K., Ma, Z., Lu, Z. J. & Zhang, Q. C. A deep learning method for recovering missing signals in transcriptome-wide rna structure profiles from probing experiments. *Nature Machine Intelligence* **3**, 995–1006 (2021).

29. Jaganathan, K. *et al.* Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548.e24 (2019).

30. Bogard, N., Linder, J., Rosenberg, A. B. & Seelig, G. A deep neural network for predicting and engineering alternative polyadenylation. *Cell* **178**, 91–106 (2019).

31. Amin, N., McGrath, A. & Chen, Y.-P. P. Evaluation of deep learning in non-coding RNA classification. *Nature Machine Intelligence* **1**, 246–256 (2019).

32. Fiannaca, A., La Rosa, M., La Paglia, L., Rizzo, R. & Urso, A. nrc: non-coding rna classifier based on structural features. *BioData mining* **10**, 1–18 (2017).

33. Sample, P. J. *et al.* Human 5' UTR design and variant effect prediction from a massively parallel translation assay. *Nat. Biotechnol.* **37**, 803–809 (2019).

34. Wayment-Steele, H. K. *et al.* Deep learning models for predicting rna degradation via dual crowdsourcing. *Nature Machine Intelligence* **4**, 1174–1184 (2022).

35. Angenent-Mari, N. M., Garruss, A. S., Soenksen, L. R., Church, G. & Collins, J. J. A deep learning approach to programmable rna switches. *Nature Communications* **11**, 5057 (2020). URL https://www.osti.gov/biblio/1816589.

36. Angenent-Mari, N. M., Garruss, A. S., Soenksen, L. R., Church, G. & Collins, J. J. A deep learning approach to programmable rna switches. *Nature Communications* **11**, 5057 (2020).

37. Chuai, G. *et al.* Deepcrispr: optimized crispr guide rna design by deep learning. *Genome biology* **19**, 1–18 (2018).

38. Mathews, D. H. How to benchmark rna secondary structure prediction accuracy. *Methods* **162**, 60–67 (2019).

39. Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**, 129 (2010).

40. Aquino-Jarquin, G. & Toscano-Garibay, J. D. Rna aptamer evolution: two decades of selection. *International journal of molecular sciences* **12**, 9155–9171 (2011).

41. Zhu, Y., Zhu, L., Wang, X. & Jin, H. RNA-based therapeutics: an overview and prospectus. *Cell Death Dis.* **13**, 644 (2022).

42. Wang, Z. *et al.* Aptadiff: de novo design and optimization of aptamers based on diffusion models. *Briefings in Bioinformatics* **25**, bbae517 (2024). URL https://doi.org/10.1093/bib/bbae517. https://academic.oup.com/bib/article-pdf/25/6/bbae517/59923079/bbae517.pdf.

43. Bailey, T. L., Elkan, C. *et al.* Fitting a mixture model by expectation maximization to discover motifs in bipolymers (1994).

44. Workman, R. E. *et al.* A natural single-guide rna repurposes cas9 to autoregulate crispr-cas expression. *Cell* **184**, 675–688 (2021).

45. Huang, K. *et al.* Crispr-gpt: An llm agent for automated design of gene-editing experiments. *arXiv preprint arXiv:2404.18021* (2024).

46. Nguyen, E. *et al.* Sequence modeling and design from molecular to genome scale with evo. *Science* **386**, eado9336 (2024).

47. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

48. Rnacentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic acids research* **49**, D212–D220 (2021).

49. Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

50. Su, J. *et al.* Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* **568**, 127063 (2024).

51. Dao, T., Fu, D., Ermon, S., Rudra, A. & Ré, C. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems* **35**, 16344–16359 (2022).

52. Shazeer, N. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202* (2020).

53. Ba, J. L. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

54. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020).

55. Li, X., Thickstun, J., Gulrajani, I., Liang, P. S. & Hashimoto, T. B. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems* **35**, 4328–4343 (2022).

56. Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N. & Madani, A. Progen2: exploring the boundaries of protein language models. *Cell systems* **14**, 968–978 (2023).

57. Boyd, N. *et al.* Atom-1: A foundation model for rna structure and function built on chemical mapping data. *bioRxiv* 2023–12 (2023).

58. Outeiral, C. & Deane, C. M. Codon language embeddings provide strong signals for use in protein engineering. *Nature Machine Intelligence* **6**, 170–179 (2024).

Acknowledgements

Author contributions statement

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Xiaoming Zhang, Le Cong, and Mengdi Wang.

Supplementary Information

Supplementary Notes

Supplementary Figures