bioRxiv, 2025, pp. 1–10

doi: N/A
Upload Date: 25 February, 2025
Preprint

PREPRINT

EvoFlow-RNA: Generating and Representing non-coding RNA with a Language Model

Sawan Patel,[1,*] Fred Zhangzhi Peng,[2,*] Keith Fraser,[3,4,5] Pranam Chatterjee[2,6,7,†] and Sherwood Yao[1,†]

[1]Atom Bioworks, [2]Department of Biomedical Engineering, Duke University, [3]Department of Biological Sciences, Rensselaer Polytechnic Institute, [4]Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, [5]Future of Computing Institute, Rensselaer Polytechnic Institute, [6]Department of Computer Science, Duke University and [7]Department of Biostatistics anad Bioinformatics, Duke University

[†]Corresponding authors. s.yao@atombioworks.com, pranam.chatterjee@duke.edu

Abstract

RNA plays a critical role across numerous biological functions. Recent advances in language modeling show promise with representing RNA, but the possibility of large-scale RNA design and optimization not been fully explored. We propose **EvoFlow-RNA**, a bidirectional non-coding RNA language model leveraging a masked discrete diffusion model (MDM) formulation for both generative modeling and representation learning. EvoFlow-RNA bridges the gap between RNA sequence representation and design. It outperforms leading RNA models on six BEACON tasks critical to understanding RNA function, such as secondary structure prediction. For unconditional generation, it synthesizes diverse RNA sequences with native-like biophysical properties. Furthermore, EvoFlow-RNA can optimize aptamer sequences while preserving binding recognition sites. Our results demonstrate the effectiveness of EvoFlow-RNA in RNA modeling, highlighting the capability and potential of masked discrete diffusion for both recapitulating and enhancing existing RNAs.

Key words: RNA, diffusion, design, optimization

Introduction

Deep learning has revolutionized biomolecular design, serving as a foundational technology that drives innovation in biotechnology and medicine. Recent breakthroughs such as RFDiffusion [Watson et al., 2023], ProteinMPNN [Dauparas et al., 2022], and RosettaFold [Baek et al., 2021] exemplify the potential of computational approaches in protein representation with unprecedented *in-silico* to experimental success in generating functional proteins. However, the majority of deep learning efforts have been concentrated on protein modeling and design. By comparison, *de novo* RNA design —a field with immense biomedical and synthetic biology potential —has received relatively little attention. RNA molecules play crucial roles in gene regulation[Phizicky and Hopper, 2010, Olsen and Woese, 1993, Suzuki, 2021, Cai et al., 2009], catalysis [Dana et al., 2017], and information storage, making them key targets for applications such as RNA-based therapeutics, vaccine development, and programmable genetic circuits. The ability to design and optimize RNA molecules has opened new frontiers in therapeutic applications, including RNA-based drug delivery (e.g. in gene therapies) and immune modulation [Coller and Ignatova, 2024, Hu et al., 2020, Kanasty et al., 2013, Morrow et al., 2024]. Synthetic RNA constructs, such as RNA aptamers, have shown significant potential in precision medicine, offering high-affinity target binding for disease treatment and biomolecular recognition [Ng et al., 2006, Zhou et al., 2012]. As RNA-based technologies continue to evolve, their diverse functional repertoire underscores the need for advanced computational models tailored to RNA sequence and structure design.

Recent efforts in RNA modeling have begun integrating structural information to enhance generative and predictive capabilities. One such approach is a diffusion-based model that designs fluorescing RNA aptamers with minimal sequence homology to known examples while maintaining functional integrity [Wong et al., 2024]. Despite their effectiveness, structure-based approaches are largely constrained by the limited number of crystal structures in the PDB. Autoregressive language models trained explicitly for sequence generation have also demonstrated advantages in RNA design, achieving high fidelity in generating sequences with biologically meaningful properties [Zhang et al., 2024b]. However, these models are inherently constrained by their left-to-right generation mechanism, which limits their ability to fully capture bidirectional dependencies essential for RNA structure-function relationships. On the other hand, BERT-style masked language models have gained traction for RNA representation learning,

*Contributed equally.

© The Author 2025. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

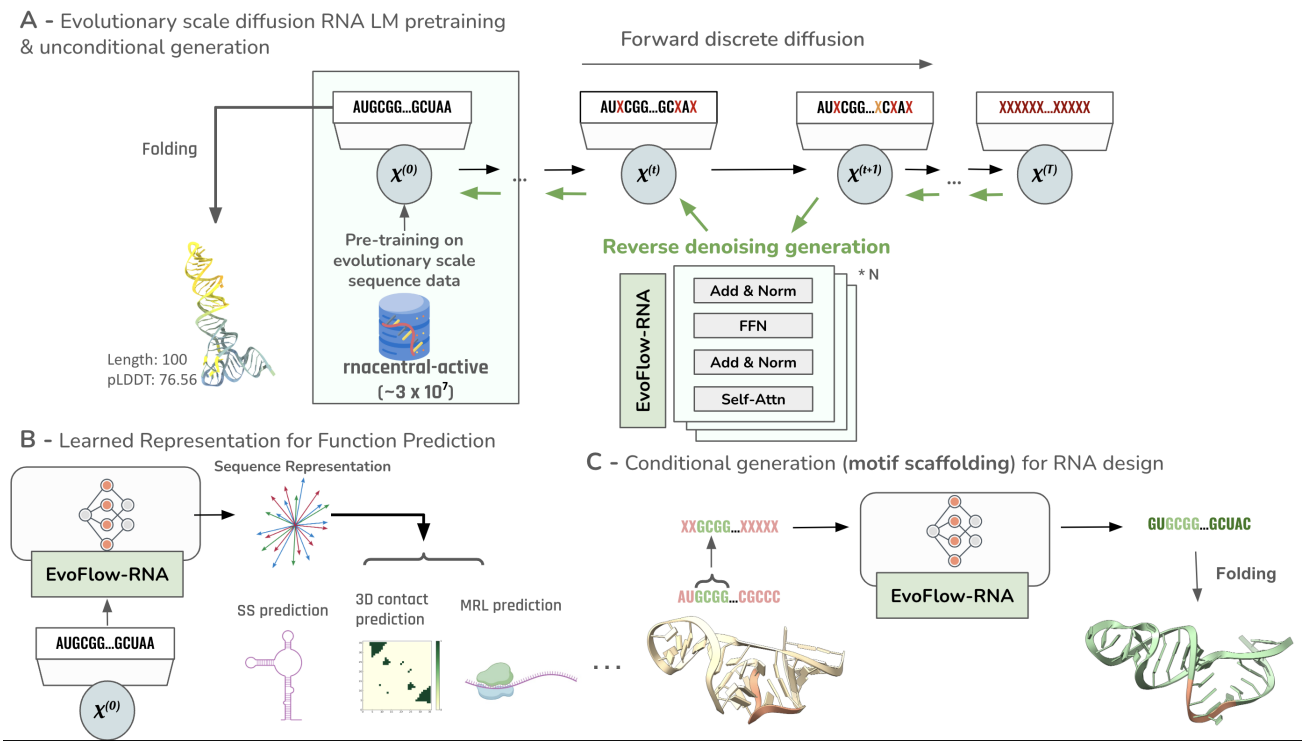

Fig. 1. General framework of EvoFlow-RNA. A) Pre-training, modeling and unconditional generation. B) Non-coding RNA representation learning for structure, functional and engineering design tasks. C) Conditional generation of RNAs via motif scaffolding.

excelling in tasks such as functional annotation and mutation effect prediction [Penić et al., 2024, Chen et al., 2022, Zhang et al., 2024a, Chen et al., 2024, Akiyama and Sakakibara, 2022]. By pretraining on large-scale non-coding RNA datasets, these models learn meaningful embeddings that can be effectively transferred to downstream tasks, including RNA-protein interaction prediction, structural stability assessment, and classification of regulatory elements. While powerful for function prediction, these models are not explicitly designed for sequence generation, limiting their utility in RNA design applications.

Masked discrete diffusion models, also known as masked diffusion models (MDMs), have emerged as a promising alternative to autoregressive models for the generative modeling of discrete data while also maintaining the bidirectional context formulation of BERT-like models. In MDMs, the forward process involves progressively masking elements of the input data, effectively adding noise by replacing original tokens with a mask token. The reverse process then aims to reconstruct the original data by predicting the masked tokens at each step, thereby learning the data distribution through iterative denoising. This approach has been shown to be effective in various applications, including language modeling and biological sequence generation [Shi et al., 2024, He et al., 2022, Wang et al., 2024, Peng et al., 2025]. In biological contexts, MDMs have comparable performance with leading BERT-based models in representation learning tasks [Hayes et al., 2025] in addition to improved generation quality compared to autoregressive language models [Peng et al., 2025, Sahoo et al., 2024].

In this work, we propose EvoFlow-RNA (Figure 1), an RNA language modeling with bidirectional generative capabilities and representation learning. EvoFlow-RNA is an MDM, providing a unified framework for generation and representation by combining bidirectional attention and discrete flow learning. We evaluate EvoFlow-RNA in three key areas: representation learning, unconditional RNA generation, and RNA design (scaffolding).

In representation learning, EvoFlow-RNA is benchmarked on 13 downstream tasks from the BEACON suite [Ren et al., 2024], outperforming leading pretrained RNA models and supervised baselines on six tasks. It excels in structural prediction, particularly secondary structure and contact map prediction, demonstrating the effectiveness of masked diffusion representations. For unconditional RNA generation, EvoFlow-RNA generates de novo non-coding RNA sequences with biophysical properties closely aligned to native sequences, outperforming RiNALMo models. Across various lengths, it maintains consistency in key metrics such as pLDDT (foldability), GC content, minimum free energy, and sequence entropy, producing diverse structures, including tRNA-like motifs. In RNA design (scaffolding), EvoFlow-RNA's bidirectional generation optimizes aptamer sequences while preserving key binding motifs. It successfully redesigns aptamers binding to HIV-1 Rev peptide and AMP, maintaining functional elements with motif-specific RMSDs below 1Å while exhibiting novel global conformations. As follows, we show that EvoFlow-RNA maintains well-formed representations of non-coding RNAs in its latent space while having an unparalleled ability for *de-novo* generation.

Background

In this section, we describe the key components of the Discrete Flow Model (DFM) that underlies EvoFlow-RNA. We first introduce the notation and problem formulation, then describe the construction of the generative flow using continuous-time Markov chains (CTMCs). We detail our method formulation with special emphasis on the mask formulation, and finally explain the sampling procedure and training objective.

Notation

Let $x \in \{1, \ldots, S\}^D$ denote a discrete sequence with D dimensions, where each element takes one of S possible states. For ease of exposition, we assume $D = 1$; the extension to higher dimensions follows by appropriate factorization.

We define $p_0(x)$ as the initial (noise) distribution, $p_{\text{data}}(x)$ as the target data distribution, and $p_t(x)$ as the time-dependent marginal distribution for $t \in [0, 1]$ that interpolates between p_0 at $t = 0$ and p_{data} at $t = 1$. The dynamics of $p_t(x)$ are generated by a time-dependent rate matrix $R_t \in \mathbb{R}^{S \times S}$, whose off-diagonal entries are non-negative. In the context of continuous-time Markov chains (CTMCs), the transition probability for an infinitesimal time step Δt is given by

$$p_{t+\Delta t}(j \mid x) = \delta\{x, j\} + R_t(x, j)\Delta t, \tag{1}$$

where $\delta\{x, j\}$ is the Kronecker delta, and $R_t(x, x)$ is defined so that the row sums are equal to 1. We also introduce a conditional noise flow $p_{t|1}(x_t \mid x_1)$ that interpolates from a prescribed noise distribution $p_{\text{noise}}(x)$ at $t = 0$ to a data point x_1 at $t = 1$.

Problem Formulation

Our objective is to design a generative model that transforms noise samples into samples from the data distribution. More precisely, we seek a *generative flow* $\{p_t(x)\}_{t\in[0,1]}$ satisfying

$$p_0(x) = p_{\text{noise}}(x) \quad \text{and} \quad p_1(x) = p_{\text{data}}(x).$$

The evolution of the marginal $p_t(x)$ is governed by the Kolmogorov forward (or continuity) equation:

$$\partial_t p_t(x) = \sum_{j \neq x} R_t(j, x) p_t(j) - \sum_{j \neq x} R_t(x, j) p_t(x). \tag{2}$$

In our approach, the generative flow is constructed via an expectation over conditional flows:

$$p_t(x_t) = \mathbb{E}_{x_1 \sim p_{\text{data}}}\left[p_{t|1}(x_t \mid x_1)\right], \tag{3}$$

which decomposes the global evolution into simpler, datapoint-conditional processes.

Method Formulation

A key component of our approach is the explicit construction of the conditional flow $p_{t|1}(x_t \mid x_1)$. We consider two variants: a uniform corruption flow and a mask-based flow. In this work, we focus on the mask formulation, which has been found effective for discrete data.

Mask Formulation

The mask formulation defines the conditional flow as

$$p^{\text{mask}}_{t|1}(x_t \mid x_1) = \text{Cat}\Big(t\,\delta\{x_1, x_t\} + (1 - t)\,\delta\{M, x_t\}\Big), \tag{4}$$

where M is an artificially introduced mask state. This construction satisfies the following limits: at $t = 0$, $p^{\text{mask}}_{t|1}(x_t \mid x_1) = \delta\{M, x_t\}$, so that all probability mass is on the mask state, and at $t = 1$, $p^{\text{mask}}_{t|1}(x_t \mid x_1) = \delta\{x_1, x_t\}$, thereby recovering the true data point. Using (4), the overall generative flow in (3) becomes

$$p_t(x_t) = \mathbb{E}_{x_1 \sim p_{\text{data}}}\left[p^{\text{mask}}_{t|1}(x_t \mid x_1)\right]. \tag{5}$$

In order to sample from $p_t(x_t)$, we require a rate matrix $R_t(x_t, j)$ that generates this flow. One strategy is to first derive a conditional rate matrix $R_t(x_t, j \mid x_1)$ that produces the desired $p_{t|1}(x_t \mid x_1)$ and then aggregate it via the denoising distribution:

$$R_t(x_t, j) = \mathbb{E}_{x_1 \sim p(x_1|x_t)}\left[R_t(x_t, j \mid x_1)\right], \tag{6}$$

with

$$p(x_1 \mid x_t) = \frac{p_{t|1}(x_t \mid x_1) p_{\text{data}}(x_1)}{p_t(x_t)}.$$

This formulation decouples the design of the conditional dynamics (which can be chosen in closed form) from the unconditional generative process.

Sampling

To sample from the data distribution, we simulate the CTMC defined by the rate matrix $R_t(x_t, j)$. Starting from an initial sample $x_0 \sim p_0(x)$, we discretize the time interval $[0, 1]$ with a step size Δt and update the state according to

$$x_{t+\Delta t} \sim \text{Cat}\Big(\delta\{x_t, x_{t+\Delta t}\} + R_t(x_t, x_{t+\Delta t})\Delta t\Big). \tag{7}$$

This process is repeated iteratively until $t = 1$, at which point the final state is a sample from $p_{\text{data}}(x)$.

Training

The parameters of our model are learned by training a neural network $p^{\theta}_{1|t}(x_1 \mid x_t)$ to approximate the denoising distribution $p(x_1 \mid x_t)$. The training objective is a standard cross-entropy loss:

$$\mathcal{L}_{\text{ce}} = -\mathbb{E}_{\substack{x_1 \sim p_{\text{data}} \\ t \sim \mathcal{U}(0,1) \\ x_t \sim p_{t|1}(x_t|x_1)}}\left[\log p^{\theta}_{1|t}(x_1 \mid x_t)\right]. \tag{8}$$

Here, t is sampled uniformly from $[0, 1]$ and x_t is generated using the conditional flow (e.g., via the mask formulation in (4)). Importantly, this loss does not depend on the choice of the conditional rate matrix $R_t(x_t, j \mid x_1)$, which allows flexibility in designing the sampling dynamics.

Methods

Dataset

RNAcentral is a comprehensive database that consolidates non-coding RNA (ncRNA) sequences from over 40 expert databases, providing a unified platform for ncRNA data. As of its latest release, RNAcentral encompasses more than 30 million ncRNA sequences spanning a diverse range of organisms and RNA types. We take our pretraining data from RNACentral [Consortium, 2021], which has a processed dataset of over 30 million non-coding RNAs. We made no further preprocessing adjustments.

Tokenization

Each nucleotide is treated as a single token, following a character-level tokenization scheme specifically designed for RNA sequences. During tokenization, all uracil (U) bases

are replaced with thymine (T) to maintain consistency across sequence representations, resulting in a primary vocabulary of four nucleotides: "A", "C", "T", and "G". Additionally, we include IUPAC degenerate nucleotide codes such as "I", "R", "Y", "K", "M", "S", "W", "B", "D", "H", "V", "N", and "-" to account for sequence ambiguities commonly found in biological datasets. To align with standard masked language modeling (MLM) practices, we incorporate the special tokens [CLS], [EOS], [MASK], and [PAD], which respectively denote sequence classification, end-of-sequence markers, masked tokens for pretraining, and padding for fixed-length sequences. EvoFlow-RNA processes sequences of up to 1024 tokens, dynamically truncating or padding them as needed. Random sequence cropping is applied during training to ensure the model observes diverse sequence contexts across epochs, which enhances its ability to generalize across RNA families.

Encoder Architecture

EvoFlow-RNA is a BERT-style encoder-only Transformer that extends the RiNALMo-150M model [Penić et al., 2024] and is trained on a masked diffusion modeling objective while leveraging the same tokenization scheme and architectural framework. Given an input sequence $x = (x_1, x_2, ..., x_L)$ of length L, the model maps it to a sequence of context-aware token embeddings of shape $(L \times M_{\text{emb}})$. The architecture consists of 33 Transformer blocks, where each block contains a multi-head self-attention (MHSA) module with 20 attention heads, a feed-forward network (FFN) using the SwiGLU activation function [Shazeer, 2020], and Layer Normalization (LayerNorm) within residual connections to stabilize training dynamics. EvoFlow-RNA adopts Rotary Positional Embeddings (RoPE) [Su et al., 2024] to encode both absolute and relative positional information, thereby improving its ability to model long-range dependencies in RNA sequences. To enhance computational efficiency, we employ FlashAttention-2 [Dao, 2024], a memory-optimized exact attention mechanism that significantly reduces peak activation memory without compromising precision, enabling training on long RNA sequences with high batch throughput.

EvoFlow-RNA is initialized from a pretrained RiNALMo-150M checkpoint and continues training using a masked diffusion objective with a token corruption strategy where $t \sim \mathcal{U}(0, 1)$ of the input sequence is masked, namely, replaced by [MASK]. This approach ensures that the model retains the strong representation learning capacity of RiNALMo while being effectively adapted for diffusion-based sequence generation. Compared to GELU-based architectures, the incorporation of SwiGLU activation improves gradient stability, while RoPE enables better positional encoding compared to conventional absolute position embeddings. Notably, we observe significant memory savings with FlashAttention-2 while maintaining model convergence, making EvoFlow-RNA computationally efficient for large-scale RNA sequence modeling.

Diffusion Training

We take two-stage training as described in Manocha et al. [2021], which involves the initial training of a masked language model (MLM) followed by continuing that training with a masked diffusion modeling (MDM) objective. We truncate all RNAs to 1022 tokens and have no further data processing. Our model is trained for 100K updates with a batch size of 320K. We additionally utilize a batch sampler to construct batches using sequences of similar lengths. Practically, we did not implement the first-stage training. Instead, we employed RiNALMo, an RNA language model that is trained on MLM loss. We finetune RiNALMo using the MDM objective with the weighted cross-entropy loss where the weight is $\frac{1}{mask_ratio}$, and the $mask_ratio$ is uniformly sampled from 0 to 1 and clamped to $[\frac{1}{500}, 1]$.

Sampling

We follow the P2 sampling scheme from Peng et al. [2025], specifically self-planning. In P2, the denoiser makes predictions and the planner selects positions to update (unmask to the prediction or remask an existing token). In self-planning, the planner is the denoiser itself where we leverage the denoiser's predicted probabilities to score positions and unmask the mask tokens that have the top predicted probabilities and remask the unmasked tokens that have the lowest. The overall sampling scheme is depicted in algorithm 1.

Algorithm 1 Sampling Scheme.

Initialize: $t \leftarrow 0$, $x_0 \leftarrow (M, \ldots, M)$, planner G^θ, denoiser D^θ, scheduler K
for $t = 1 : L$ **do**
  Plan:
  Sample $y \sim D^\theta(x_t)$
  UpdatePos $\leftarrow \text{TopPos}_{\kappa(t)}\big(\tilde{G}^\theta_\cdot(y, x_t)\big)$
  Denoise:
  for $j \in$ UpdatePos **do**
    if $[x_t]_j = M$ **then**
      $[x_t]_j \leftarrow y_j$
    end if
  end for
  for $j \notin$ UpdatePos **do**
    if $[x_t]_j \neq M$ **then**
      $[x_t]_j \leftarrow M$
    end if
  end for
end for
return x_L

Results

Representation Benchmarking

We evaluate EvoFlow-RNA representations using the BEACON benchmark suite [Ren et al., 2024]. Accordingly, the model is fine-tuned and assessed for performance across 13 related tasks spanning three categories: structure, function and engineering. We compare our model's performance against leading pretrained large-language models (LLMs) and supervised models. The supervised models include a convolutional neural network (CNN), ResNET, and LSTM. The pretrained LLMs include RNA-FM [Chen et al., 2022], RNA-BERT [Akiyama and Sakakibara, 2022], RNA-MSM [Zhang et al., 2024a], SpliceBERT [Chen et al., 2024], 3UTRBert [Yang et al., 2024], and UTR-LM [Chu et al., 2024]. These models vary widely in size, up to 27M parameters for the supervised models and up to 100M parameters for the pretrained LLMs. They are also pretrained on different data sources, including but not limited to ncRNA, pre-mRNA, mRNA-3'UTR, and mRNA-5'UTR. For EvoFlow-RNA, we utilize the exact configurations listed in [Ren et al., 2024] for each task to ensure consistency. The only exceptions are secondary structure prediction, contact map prediction and distance map prediction,

Table 1. BEACON Representation Learning. Benchmarking across 13 downstream RNA-related tasks across structure, function and engineering. EvoFlow-RNA is compared with both naive supervised methods and pretrained RNA language models. Baseline results for all other models are taken from BEACON [Ren et al., 2024].

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	-
EvoFlow-RNA	74.51	74.05	38.77	42.52	32.66	72.61	86.46	71.06	86.28	0.47	55.03	41.48	5.34

where we were not able to train with a batch size of 32 due to hardware limitations. We instead trained with a batch size of 2, with all other hyperparameters held constant.

In Table 1, we show the performance of EvoFlow-RNA against all models evaluated against the BEACON benchmark, pulled across BEACON [Ren et al., 2024] and RNAGenesis [Zhang et al., 2024b]. Evidently, we observe that EvoFlow-RNA achieves state-of-the-art performance on 6 of the 13 evaluation tasks, reflecting the quality of masked diffusion model representations in the context of ncRNA. Most notably, EvoFlow-RNA performs well at the structure-related tasks, such as secondary-structure prediction (SSP) and contact map prediction (CMP). The latter requires an understanding of the three-dimensional representation of an RNA, showing that this property can be learned to a degree from sequence. Interestingly, though the model performs well at contact map prediction, it does not perform as well at distance map prediction (DMP), another task requiring a precise understanding of RNA tertiary structure. Still, we maintain that EvoFlow-RNA possesses state-of-the-art structural representations of non-coding RNAs, in addition to its knack for assessing the functional capabilities of non-coding RNAs such as single-guide RNAS for CRISPR on-target gene editing (CRI-On) and predicting the mean ribosomal load of a given mRNA sequence (MRL).

ncRNA De-novo Generation Evaluation

Unconditional generation

Synthetic non-coding RNAs have shown to be viable tools in therapeutics, diagnostics and biotechnology. These can encompass engineered tRNAs, codon-optimized mRNAs or siRNA therapeutics. Moreover, RNA aptamers also have a long history of utility, being short RNA molecules that can bind to specific biomolecules with high affinity, selectivity and specificity. As such, we evaluated EvoFlow-RNA's ability to design non-coding RNAs that align with key properties computed from the ncRNA training dataset. In doing so, this would allow for the generation of a diverse library of potential RNA aptamers, each exhibiting unique structure-stability profiles, for use as aptamers. In theory, this would also diminish the need for multiple rounds of SELEX as is the case in prototypical aptamer design.

To evaluate the model's generative capability, we compared EvoFlow-RNA with RiNALMo-150M, RiNALMo-650M, and a generated aptamers from the diffusion model latent space using RDM sampling. Lastly, we also featured native ncRNAs as a baseline for comparing key properties of the generated sequences. To facilitate comparisons, we computed pLDDT via RhoFold as a proxy for foldability, GC content, minimum free energy via ViennaRNA [Lorenz et al., 2011], and sequence entropy. Libraries of size 50, 100, 200 and 400 were generated to faciliate this study. Per Figure 2, we report that EvoFlow-RNA designed

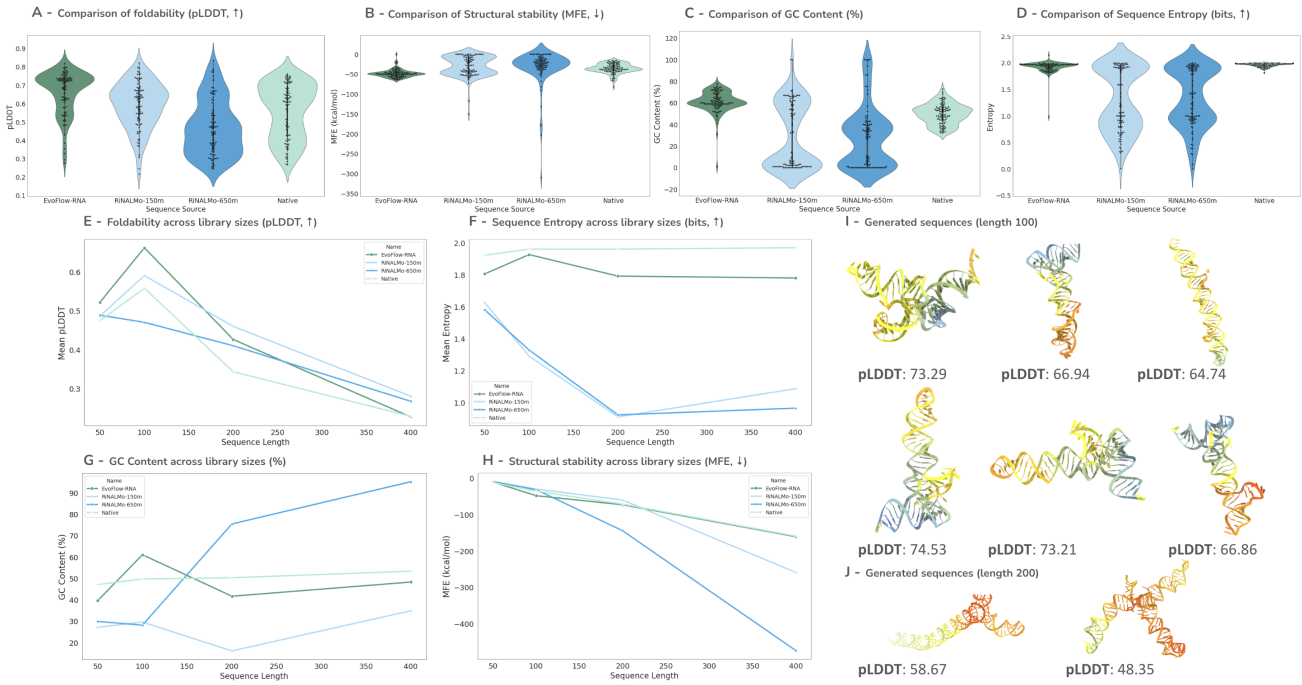

Fig. 2. Unconditional generation evaluation with P2 (RDM, $\eta = 1.0$). We use RhoFold as the folding model to generate predicted structures and calculate the pLDDT score for all sampled sequences. Additionally, the minimum free energy (MFE) is computed by ViennaRNA. GC content and sequence entropy are computed as the percent of bases within the sequence that are either guanine or cytosine and the Shannon entropy, respectively. (A-D) Metrics for sequences of length 100. E) Average pLDDT, F) MFE, G) GC content, and H) sequence entropy computed for libraries of size 100 as a function of library sequence length. De-novo, unconditionally-generated sequences of I) length 100 and J) 200. Bases are colored according to their pLDDT and the AlphaFold2 color scheme. Additional structures can be found in Figure A2

ncRNAs exhibit similar biophysical profiles as native sequences than either RiNALMo base model. This is evident across all metrics, most notably GC content, MFE and sequence entropy. The EvoFlow-RNA pLDDT distribution is relatively higher than the native sequence library, which may be a byproduct of oversampling particular sequences with higher foldability. More extensive results across library sizes are shown in Figure A1. Most notably, we show that the consistency in alignment between EvoFlow-RNA and native ncRNAs holds across all evaluated sequence lengths, regardless of the metric. All models converge to low average pLDDT values as the sequence length increases, which is expected as a product of the folding model. De-novo designs also exhibit a variety of tertiary structures, with several exhibiting tRNA-like structures (Figures **??**I, **??**).

To assess sampling quality, we performed an ablation using various P2-derived strategies using three generation sources: EvoFlow-RNA (denoted as MDM), RiNALMo-150M and RiNALMo-650M in Table 2. We compare four sampling cases: ancestral (P2, $\eta = 0$), RDM (P2, $\eta = 1$), P2 with a calibrated η, and P2 with a dedicated BERT planner (RiNALMo-150M). It is apparent that training an MDM overall yields higher quality *de-novo* ncRNAs over a pretrained BERT model such as RiNALMo, regardless of model size. Notably, ncRNAs generated from EvoFlow-RNA are comparable to a set of subsampled native ncRNA sequences, as per the objective metrics. The only exception, as also shown above, is pLDDT where sequences generated synthetically tend to exhibit higher values.

Moreover, we also demonstrate that P2 sampling with a supervisory planner yields sequences with 'best' performance across all metrics, though we note the additional deviation from native sequences in the training dataset.

Conditional generation

A noteworthy ability of masked diffusion models, as opposed to their autoregressive counterparts, is the ability to inform generation using bidirectional context. This property is, in theory, useful for building scaffolds around key binding motifs. In this task, we fix a key motif subsequence within an experimentally validated aptamer with a crystal structure and optimize the rest of its sequence using a generative model. This task is particularly suited for masked diffusion models as bases generated upstream of the fixed motif can be generated in the context of the fixed motif, which autoregressive models are not capable of doing.

To test this capability, we investigate two known RNA aptamers from the protein data bank: an RNA aptamer binding selectively to the HIV-1 Rev peptide (1ULL, length 35) from Ye et al. [1996] and an RNA aptamer binding selectively to the AMP molecule (1RAW, length 36) from Dieckmann et al. [1996]. Within their respective depositions, we identified the key recognition sites of each aptamer which participate in the binding interaction with the intended target. The subsequences corresponding to these recognition sites were fixed (bases 3:12, 21:23, 25:28 and 30:33 for 1ULL and bases 7-11 for 1RAW). The remainder of the sequence was masked and infilled using EvoFlow-RNA. Generated aptamers from these scaffolds were of the same length as the original.

Using this procedure, we designed a library of model-optimized aptamers. The result is shown in Figure 3, where we feature two optimized, conditionally-generated aptamers for

Table 2. RNA Sequence Generation Benchmark. The "Native" row represents subsampled natural RNA sequences. "MDM" refers to a pretrained 150M Masked Diffusion Model trained on RNACentral [Consortium, 2021].

Sequence Source	pLDDT (↑)	MFE (kcal/mol) (↓)	Entropy (↑)	GC Content (%) (↑)
Native	48.26	-35.83	**1.96**	49.64
RiNALMo-150M	59.01	-30.12	1.29	29.50
RiNALMo-650M	46.99	-31.90	1.33	28.06
EvoFlow-RNA + Ancestral	68.12	-48.46	1.93	60.84
EvoFlow-RNA + RDM	67.35	-47.54	1.89	59.42
EvoFlow-RNA + P2 (self-planning)	**69.41**	-48.21	1.89	59.84
EvoFlow-RNA + P2 + RiNALMo-150M Planner	**73.28**	**-51.91**	1.86	**65.47**

each original aptamer. All optimized aptamers maintain similar MFE and GC content profiles as the original aptamer. Notably, the optimized aptamers for HIV-1 Rev peptide align fairly well with the original aptamers, with motif-specific RMSDs of approximately 1 A and whole-structure RMSDs of approximately 4 A, following a least-squares fit. The common motif RMSDs for the optimized AMP aptamers and the original are below 1 A, highlighting that the key recognition elements do not deviate substantially from their original locations using a scaffolding approach. However, we do note that the whole-structure RMSDs following a least-squares fit are well above 10 A. This indicates that the optimized aptamers have a globally unique structure in comparison to the originals, though they maintain similarities at the local-level, and particularly at binding sites. Additional scaffolds for the HIV-1 Rev peptide system can be found in Figure A3.

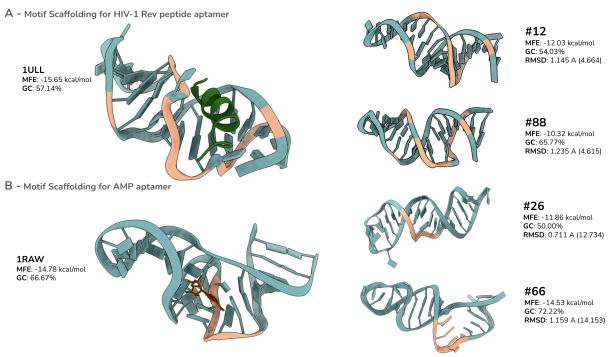

Fig. 3. Conditionally generating scaffolds for RNA aptamers. A) 1ULL (Left), an RNA aptamer/HIV-1 rev peptide complex, is taken as a simpler study with several bases contributing to the binding recognition site (salmon) on the aptamer. (Right) Two conditionally-generated scaffold RNA aptamers sharing the binding recognition site (salmon) of the original aptamer. B) 1RAW (Left), an RNA aptamer/AMP molecule complex, is a more complicated study with fewer bases contributing to the binding recognition site (salmon) on the aptamer. (Right) Two conditionally-generated scaffold RNA aptamers sharing the binding recognition site (salmon) of the original aptamer.

Conclusion

We introduced **EvoFlow-RNA**, a bidirectional RNA language model based on masked discrete diffusion, unifying representation learning and sequence generation. Trained on 30M non-coding RNA sequences, it outperforms leading models on six BEACON tasks, excels in secondary structure prediction, and generates diverse, native-like RNA sequences. For RNA design, EvoFlow-RNA optimizes aptamer scaffolds while preserving functional motifs with sub-1Å RMSD. Our results demonstrate the effectiveness of masked diffusion for RNA modeling, providing a scalable alternative to existing methods.

Despite these results, we observe a few key limitations that could be items for future work. First, we observe that successful performance on representation learning tasks was not seen across the board, with EvoFlow-RNA performing below expectation on mRNA-related tasks. Though this is expected, given the model was trained entirely on ncRNA, its subpar performance on distance map prediction (DMP) is peculiar given its excellent ability on the other two structure-related tasks. Second, though we observe that EvoFlow-RNA-generated libraries most closely mirror native ncRNAs across metrics for sequence composition and biophysical properties, the pLDDT distributions do not always align closely. Often, EvoFlow-RNA generates sequences with pLDDT's above those computed for native sequences. We note that this may be a consequence of the folding model. Lastly, we acknowledge that experimental validation for any generated RNAs, particularly the scaffolds, has yet to be done. The real-world applications using this model largely concern its generative capability, which can only be conclusively assessed in the lab. Our motif scaffolding experiments demonstrate a possible application of EvoFlow-RNA with a straightforward validation method. If such results hold true experimentally, the possibility of non-coding RNA optimization can seamlessly be applied to the growing field of tRNA therapeutics. We leave these items for future works.

Competing interests

No competing interest is declared.

Author contributions statement

S.P. proposed the idea and conducted all experiments described in the paper. F.Z.P. outlined the mathematical framework and proposed the sampling algorithm experiments. S.P., F.Z.P. and K.F. jointly wrote the manuscript, with all others contributing edits. P.C. and S.Y. supervised the project.

References

Manato Akiyama and Yasubumi Sakakibara. Informative rna base embedding for rna structural alignment and clustering by deep representation learning. *NAR genomics and bioinformatics*, 4(1):lqac012, 2022.

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

Yimei Cai, Xiaomin Yu, Songnian Hu, and Jun Yu. A brief review on the mechanisms of mirna regulation. *Genomics, Proteomics and Bioinformatics*, 7(4):147–154, 2009.

Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, et al. Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. *arXiv preprint arXiv:2204.00300*, 2022.

Ken Chen, Yue Zhou, Maolin Ding, Yu Wang, Zhixiang Ren, and Yuedong Yang. Self-supervised learning on millions of primary rna sequences from 72 vertebrates improves sequence-based rna splicing prediction. *Briefings in Bioinformatics*, 25(3):bbae163, 2024.

Yanyi Chu, Dan Yu, Yupeng Li, Kaixuan Huang, Yue Shen, Le Cong, Jason Zhang, and Mengdi Wang. A 5 utr language model for decoding untranslated regions of mrna and function predictions. *Nature Machine Intelligence*, 6(4):449–460, 2024.

Jeff Coller and Zoya Ignatova. trna therapeutics for genetic diseases. *Nature Reviews Drug Discovery*, 23(2):108–125, 2024.

RNACentral Consortium. Rnacentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic acids research*, 49(D1):D212–D220, 2021.

Hassan Dana, Ghanbar Mahmoodi Chalbatani, Habibollah Mahmoodzadeh, Rezvan Karimloo, Omid Rezaiean, Amirreza Moradzadeh, Narges Mehmandoost, Fateme Moazzen, Ali Mazraeh, Vahid Marmari, et al. Molecular mechanisms and biological functions of sirna. *International journal of biomedical science: IJBS*, 13(2):48, 2017.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

THORSTEN Dieckmann, ELLEN Suzuki, GERALD K Nakamura, and JULI Feigon. Solution structure of an atp-binding rna aptamer reveals a novel fold. *Rna*, 2(7):628–640, 1996.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, page eads0018, 2025.

Zhengfu He, Tianxiang Sun, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. Diffusionbert: Improving generative masked language models with diffusion models. *arXiv preprint arXiv:2211.15029*, 2022.

Bo Hu, Liping Zhong, Yuhua Weng, Ling Peng, Yuanyu Huang, Yongxiang Zhao, and Xing-Jie Liang. Therapeutic sirna: state of the art. *Signal transduction and targeted therapy*, 5(1): 101, 2020.

Rosemary Kanasty, Joseph Robert Dorkin, Arturo Vegas, and Daniel Anderson. Delivery materials for sirna therapeutics. *Nature materials*, 12(11):967–977, 2013.

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

Pranay Manocha, Anurag Kumar, Buye Xu, Anjali Menon, Israel D Gebru, Vamsi K Ithapu, and Paul Calamia. Dplm: A deep perceptual spatial-audio localization metric. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 6–10. IEEE, 2021.

Alyssa Kramer Morrow, Ashley Thornal, Elise Duboscq Flynn, Emily Hoelzli, Meimei Shan, Gorkem Garipler, Rory Kirchner, Aniketh Janardhan Reddy, Sophia Tabchouri, Ankit Gupta, et al. Ml-driven design of 3'utrs for mrna stability. *bioRxiv*, pages 2024–10, 2024.

Eugene WM Ng, David T Shima, Perry Calias, Emmett T Cunningham Jr, David R Guyer, and Anthony P Adamis. Pegaptanib, a targeted anti-vegf aptamer for ocular vascular disease. *Nature reviews drug discovery*, 5(2):123–132, 2006.

Gary J Olsen and Carl R Woese. Ribosomal rna: a key to phylogeny. *The FASEB journal*, 7(1):113–123, 1993.

Fred Zhangzhi Peng, Zachary Bezemek, Sawan Patel, Sherwood Yao, Jarrid Rector-Brooks, Alexander Tong, and Pranam Chatterjee. Path planning for masked diffusion model sampling, 2025. URL `https://arxiv.org/abs/2502.03540`.

Rafael Josip Penić, Tin Vlašić, Roland G Huber, Yue Wan, and Mile Šikić. Rinalmo: General-purpose rna language models can generalize well on structure prediction tasks. *arXiv preprint arXiv:2403.00043*, 2024.

Eric M Phizicky and Anita K Hopper. trna biology charges to the front. *Genes & development*, 24(17):1832–1860, 2010.

Yuchen Ren, Zhiyuan Chen, Lifeng Qiao, Hongtai Jing, Yuchen Cai, Sheng Xu, Peng Ye, Xinzhu Ma, Siqi Sun, Hongliang Yan, et al. Beacon: Benchmark for comprehensive rna tasks and language models. *arXiv preprint arXiv:2406.10391*, 2024.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *arXiv preprint arXiv:2406.07524*, 2024.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K Titsias. Simplified and generalized masked diffusion for discrete data. *arXiv preprint arXiv:2406.04329*, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Tsutomu Suzuki. The expanding world of trna modifications and their disease relevance. *Nature Reviews Molecular Cell Biology*, 22(6):375–392, 2021.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Dplm-2: A multimodal diffusion protein language model. *arXiv preprint arXiv:2410.13782*, 2024.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De

novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Felix Wong, Dongchen He, Aarti Krishnan, Liang Hong, Alexander Z Wang, Jiuming Wang, Zhihang Hu, Satotaka Omori, Alicia Li, Jiahua Rao, et al. Deep generative design of rna aptamers using structural predictions. *Nature Computational Science*, pages 1–11, 2024.

Yuning Yang, Gen Li, Kuan Pang, Wuxinhao Cao, Zhaolei Zhang, and Xiangtao Li. Deciphering 3'utr mediated gene regulation using interpretable deep representation learning. *Advanced Science*, 11(39):2407013, 2024.

Xiaomei Ye, Andrey Gorin, Andrew D Ellington, and Dinshaw J Patel. Deep penetration of an α-helix into a widened rna major groove in the hiv-1 rev peptide–rna aptamer complex. *Nature structural biology*, 3(12):1026–1033, 1996.

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

Zaixi Zhang, Linlin Chao, RuoFan Jin, Yikun Zhang, Guowei Zhou, Yujie Yang, Yukang Yang, Kaixuan Huang, Qirong Yang, Ziyao Xu, et al. Rnagenesis: Foundation model for enhanced rna sequence generation and structural insights. *bioRxiv*, pages 2024–12, 2024b.

Jiehua Zhou, Maggie L Bobbin, John C Burnett, and John J Rossi. Current progress of rna aptamer-based therapeutics. *Frontiers in genetics*, 3:234, 2012.

Supplementary Figures

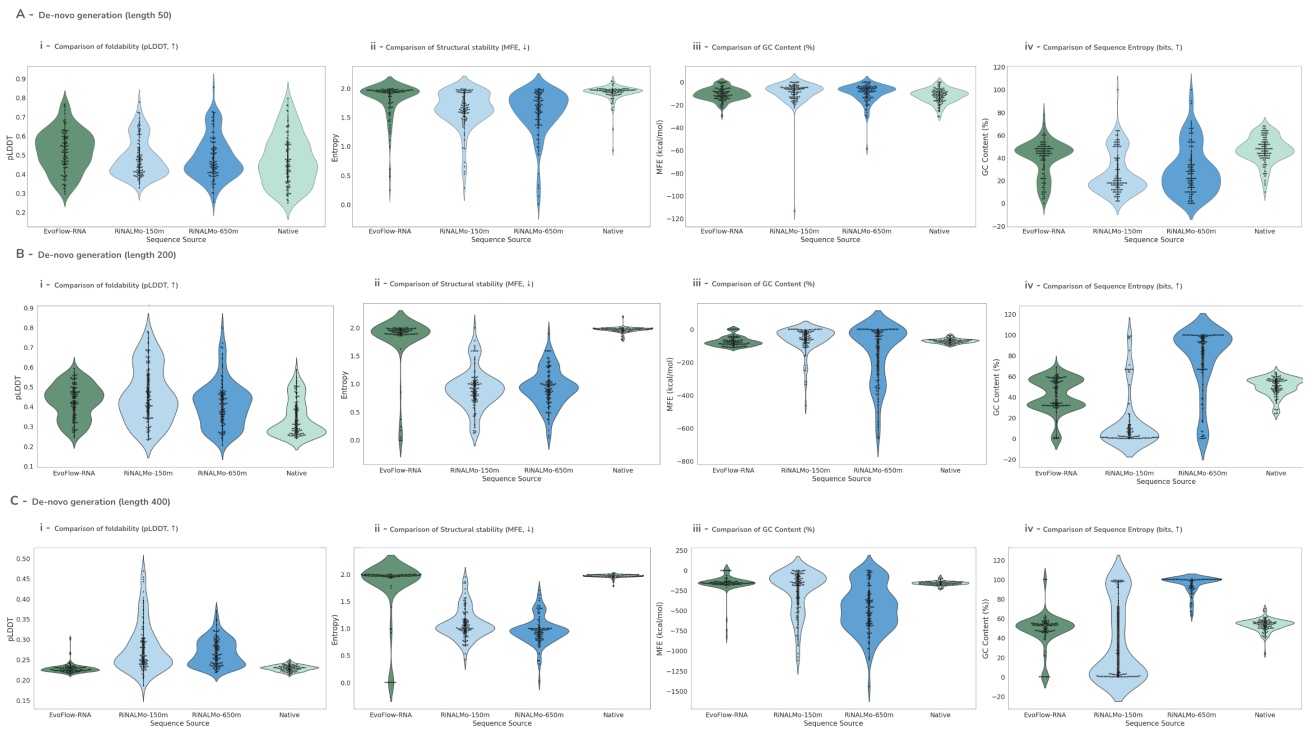

Fig. A1. Unconditional generation evaluation with P2 (RDM, $\eta = 1.0$) for libraries of length 50, 200 and 400. Metrics shown for libraries of length A) 50, B) 200, and C) 400.

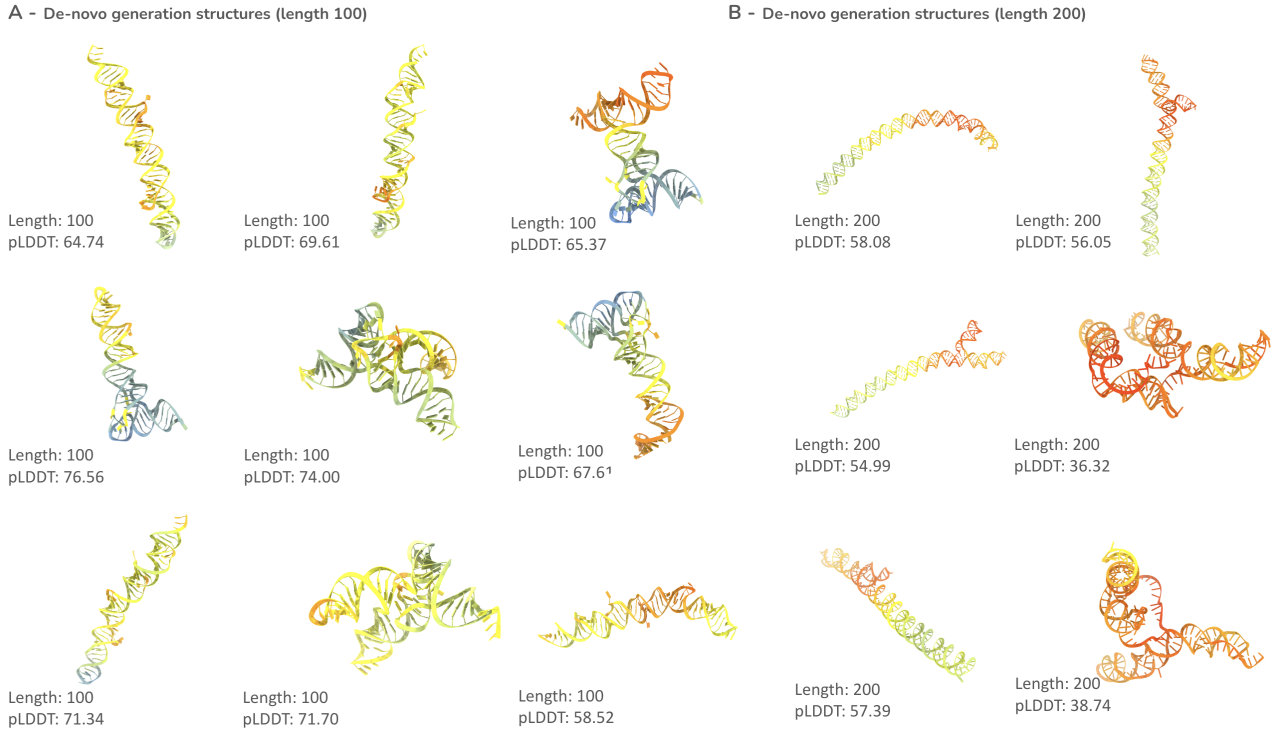

Fig. A2. Additional structures for non-coding RNAs generated unconditionally by EvoFlow-RNA of length A) 100 and B) 200.

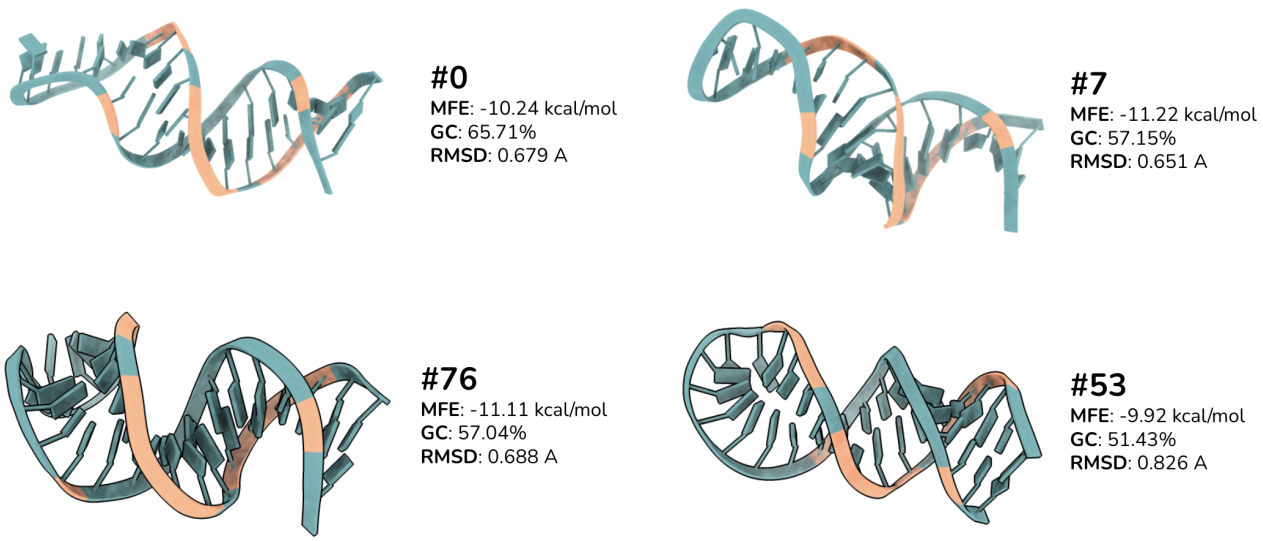

Fig. A3. Additional scaffolds generated from the 1ULL HIV-1 Rev peptide binding aptamer.