RNA-FrameFlow for *de novo* 3D RNA Backbone Design

Rishabh Anand [*1] **Chaitanya K. Joshi** [*2] **Alex Morehead** [3]
Arian R. Jamasb [4] **Charles C. Harris** [2] **Simon V. Mathis** [2] **Kieran Didi** [2]
Bryan Hooi [1] **Pietro Liò** [2]

Open-source code: www.github.com/rish-16/rna-backbone-design

Abstract

We introduce RNA-FrameFlow, the first generative model for 3D RNA backbone design. We build upon $SE(3)$ flow matching for protein backbone generation and establish protocols for data preparation and evaluation to address unique challenges posed by RNA modeling. We formulate RNA structures as a set of rigid-body frames and associated loss functions which account for larger, more conformationally flexible RNA backbones (13 atoms per nucleotide) vs. proteins (4 atoms per residue). To tackle the lack of diversity in 3D RNA datasets, we explore training with structural clustering and cropping augmentations. Additionally, we define a suite of evaluation metrics to measure whether the generated RNA structures are globally self-consistent (via inverse folding followed by forward folding) and locally recover RNA-specific structural descriptors. The most performant version of RNA-FrameFlow generates locally realistic RNA backbones of 40-150 nucleotides, over 40% of which pass our validity criteria as measured by a self-consistency TM-score ≥ 0.45, at which two RNAs have the same global fold.

1. Introduction

Designing RNA structures. Proteins, and the diverse structures they can adopt, drive essential biological functions in cells. Deep learning has led to breakthroughs in structural modeling and design of proteins (Jumper et al., 2021; Dauparas et al., 2022; Watson et al., 2023), driven by the abundance of 3D data from the Protein Data Bank (PDB). Concurrently, there has been a surge of interest in *Ribonucleic Acids* (RNA) and RNA-based therapeutics for gene editing, gene silencing, and vaccines (Doudna and Charpentier, 2014; Metkar et al., 2024). RNAs play a dual role as carriers of genetic information coding for proteins (mRNAs) as well as performing functions driven by their tertiary structural interactions (riboswitches and ribozymes). While there is growing interest in designing structured RNAs for a range of applications in biotechnology and medicine (Mulhbacher et al., 2010; Damase et al., 2021), the current toolkit for 3D RNA design uses classical algorithms and heuristics to assemble RNA motifs as building blocks (Han et al., 2017; Yesselman et al., 2019). However, hand-crafted heuristics are not always broadly applicable across multiple tasks and rigid motifs may not fully capture the conformational dynamics that govern RNA functionality (Ganser et al., 2019; Li et al., 2023a). This presents an opportunity for deep generative models to learn data-driven design principles from existing 3D RNA structures.

What makes deep learning for RNA hard? The primary challenge is the paucity of raw 3D RNA structural data, manifesting as an absence of ML-ready datasets for model development (Joshi et al., 2023). Protein structure is primarily driven by hydrogen bonding along the backbone, and current geometric deep learning models incorporate this inductive bias through backbone frames to represent residues (Jumper et al., 2021). RNA structure, however, is often more conformationally flexible and driven by base pairing interactions across strands as well as base stacking between rings of adjacent nucleotides (Vicens and Kieft, 2022), all of which can only be learnt implicitly at present[1].

Additionally, RNA nucleotides, the equivalent of amino acids in proteins, include significantly more atoms as part of the backbone (13 compared to 4) which necessitates a generalization of backbone frames where the placement of most atoms needs to be parameterized by torsion angles. These complexities have contributed to relatively poor per-

[*] Equal contribution [1] National University of Singapore, Singapore [2] University of Cambridge, UK [3] University of Missouri, USA [4] Prescient Design, Genentech, Roche. Correspondence to: Rishabh Anand <mail.rishabh.anand@gmail.com>, Chaitanya K. Joshi <ckj24@cam.ac.uk>.

Published at the AI for Science workshop at the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1] See Eric Westhof's talk contrasting RNA and protein structure.

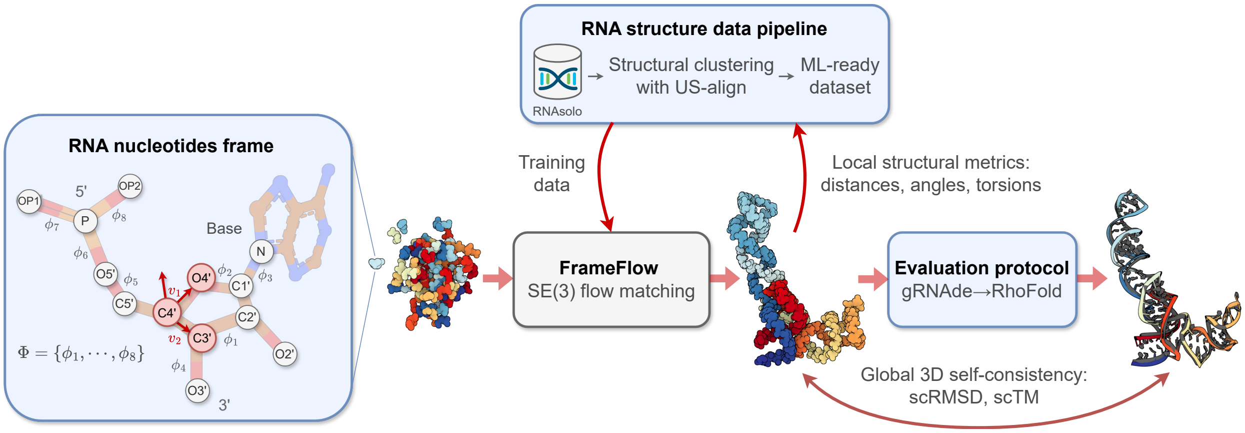

Figure 1. **The RNA-FrameFlow pipeline for 3D backbone generation.** Our implementation establishes RNA-specific protocols for data preparation and evaluation for FrameFlow (Yim et al., 2023a). (1) Each nucleotide in the RNA backbone is converted into a *frame* to parameterize the placement of $C4'$ by a translation, $C3' - C4' - O4'$ by a rotation, and the rest of the atoms via 8 torsion angles Φ. (2) We train generative models on all RNA structures of length 40-150 nucleotides from RNAsolo (Adamczyk et al., 2022). We also explore training with structural clustering and cropping augmentations to tackle the lack of diversity in 3D RNA datasets. (3) We introduce evaluation metrics to measure the recovery of local structural descriptors and global self-consistency of designed structures via inverse-folding with gRNAde (Joshi et al., 2023) followed by forward-folding with RhoFold (Shen et al., 2022).

formance of deep learning for RNA structure prediction compared to proteins (Kretsch et al., 2023; Abramson et al., 2024). Additionally, structure prediction models cannot directly be used for designing or generating *novel* RNA structures with desired constraints, which our work aims to do.

Our contributions. We develop RNA-FrameFlow, the first generative model for 3D RNA backbone design, illustrated in Figure 1. We adapt FrameFlow (Yim et al., 2023a), an $SE(3)$ equivariant flow matching model for proteins to RNA. We introduce RNA-specific modifications to the data preparation and loss formulation, including representing RNA nucleotides as rigid-body frames that parameterize all 13 atoms. We also introduce an evaluation pipeline to benchmark RNA backbone design models' capabilities at recovering local and global structure. Our best model is trained on RNAs of lengths 40-150 from the PDB and can unconditionally sample locally plausible backbones with over 40% validity as measured by a self-consistency TM-score ≥ 0.45.

Through this study, we aimed to evaluate the extent to which generative models for proteins can be adapted for RNA. This brought up critical challenges and limitations of deep learning for RNA modelling, such as a lack of explicit representations of the physical interactions that drive RNA structure as well as biases in 3D RNA datasets, which we have made preliminary efforts towards addressing. Together with our engineering contributions, we hope this work will stimulate future research in generative models for RNA design.

2. Method

We are concerned with building a generative model that unconditionally outputs all-atom RNA backbone samples. For a target sequence length of N_{res} nucleotides, we aim to generate a real-valued tensor $\mathbf{X}$ of shape $N_{\text{res}} \times 13 \times 3$ representing 3D atomic coordinates for each of the 13 backbone atoms per nucleotide.

2.1. Representing RNA Backbones

This work introduces a frame analog for nucleic acids that deals with the underlying complexity of working with their backbones. Unlike protein residues with just 4 atoms in the backbone, nucleic acid residues contain 12 atoms along the backbone. As shown in Figure 1, we use the $C4'$, $C3'$, and $O4'$ atoms as the reference frame as done by Morehead et al. (2023). All other backbone atoms are associated with 8 torsions $\Phi = \{\phi_1 \to \phi_8\}, \phi_i \in SO(2)$ that are predicted post-hoc after frame generation; these atoms are $C1', C2', N9$ (or $N1$), $O3', O5', P, OP1$, and $OP2$.

The Gram-Schmidt process is used on v_1, v_2 defined by the vectors along the $C4' - O4'$ and $C4' - C3'$ bonds; the $C5'$ plays the role of C_β and is imputed after the frames have been created. The 8 torsions, in order, are $C3' - C2'$, $O4' - C1'$, $C1' - N9$, $C3' - O3'$, $C5' - O5'$, $O5' - P$, $P - OP1$, and $P - OP2$. During sampling, *Adenine* (A) is the default nucleic acid base whose idealized geometry is obtained from Gelbin et al. (1996). Geometric imputation is not tractable given the complexity of torsion angles. Given the torsion angles, we autoregressively place non-frame atoms in order of the torsions Φ in Figure 1, constructing the final all-atom RNA backbone structure.

Choice of RNA frame. We choose $C3'$, $C4'$, and $O4'$ as they spatially shift the least in naturally occurring RNA (Harvey and Prabhakaran, 1986). The non-frame backbones - such as the remaining atoms in the ribose sugar ring ($C1'$, $C2'$) and the farther away Phosphorous (P) and two Oxygens ($OP1, OP2$) - are parameterized by torsion angles to account for their relative conformational flexibility. *Ring puckering* refers to the planar rotation of the ribose sugar ring about the $C4' - C5'$ bond. It affects how the RNA interacts with interaction partners to form complexes (e.g., protein-RNA) with high binding affinities (Clay et al., 2017). To reduce the sensitivity of our generative model to highly mobile atoms, we settle on the current frame composition.

Input. Given a set of 3D coordinates, a simultaneous rotation and translation (r, x) forms an orientation-preserving rigid-body transformation of the coordinates. The set of all such transformations in 3D is the Special Euclidean group $SE(3)$, which composes the group of 3D rotations $SO(3)$ and 3D translations in $\mathbb{R}^3$.

We can represent an RNA frame $T = (r, x)$ as a translation $x \in \mathbb{R}^3$ from the global origin to place $C4'$ and a rotation $r \in SO(3)$ to orient $C3' - C4' - O4'$. Compared to working with raw 3D coordinates for each backbone atom, using the frame representation entails performing flow matching on the space of $SE(3)$. This is an inductive bias to reduce the degrees of freedom the generative model needs to learn. Instead of predicting 13 correlated 3D coordinates independently (39 quantities) for each nucleotide, we instead predict one 3D coordinate (of $C4'$) and one 3×3 rotation matrix (12 quantities). We follow Chen and Lipman (2024) and Yim et al. (2023a)'s framework for flow matching on $SE(3)$, which we summarise subsequently.

Overview. Flow matching generates or learns how to place and orient a set of N frames $\mathbf{T} = \{T^{(n)}\}_{n=1}^N$, where $T^{(n)} = (r^{(n)}, x^{(n)})$, to form an RNA backbone of length N. To do so, we initialize frames at random in 3D space at time $t = 0$, and train a denoiser or flow model to iteratively refine the location and orientation of each frame for a specified number of steps until time $t = 1$.

Suppose $p_0(T_0)$ and $p_1(T_1)$ are the marginal distributions of randomly oriented and ground truth frames from our dataset of RNA structures, respectively. Suppose a non-unique time-dependent vector field u_t leads to an ODE between the two distributions p_0 and p_1, i.e., assume there is a way to map from noisy samples to the corresponding true samples. This solution forms a ground truth *probability path* p_t between the two distributions at time $t \in [0, 1]$, which we can use to transform samples from noise to the true distribution. The *continuity equation* $\frac{\partial p}{\partial t} = -\nabla \cdot (p_t u_t)$ relates the vector field u_t to the evolution of the probability path p_t.

Given a noisy frame T_0 sampled from $p_0(T_0)$ and the corresponding ground truth frame T_1 sampled from $p_1(T_1)$, we construct a *flow* T_t by following the probability path p_t between T_0 and T_1 for any time step t sampled from $\mathcal{U}(0, 1)$. As shown by Chen and Lipman (2024) for the $SE(3)$ group (and other manifolds), the shortest path between the two states T_0 and T_1 can be used to define an interpolation:

$$T_t = \exp_{T_0}(t \cdot \log_{T_0}(T_1)). \tag{1}$$

Here, $\exp(\cdot)$ and $\log(\cdot)$ are the *exponential* and *logarithmic* maps that enable moving (taking random walks) on curved manifolds such as the $SE(3)$ group. As we can decompose a frame $T = (r, x)$ into separate rotation and translation terms, we can obtain closed-form interpolations for the group of rotations $SO(3)$ and translations $\mathbb{R}^3$. This gives us two independent flows:

$$\text{Translations:} \quad x_t = t x_1 + (1 - t) x_0 \ , \tag{2}$$

$$\text{Rotations:} \quad r_t = \exp_{r_0}(t \cdot \log_{r_0}(r_1)) \ . \tag{3}$$

The random translation x_0 is sampled from a zero-centered Gaussian distribution $\mathcal{N}(0, \mathbf{I})$ in $\mathbb{R}^3$, and the random rotation r_0 is sampled from $\mathcal{U}(SO(3))$, a generalization of the uniform distribution for the group of rotations, $SO(3)$. For an RNA backbone consisting of a set of N frames $\mathbf{T} = \{\ T^{(n)}\}_{n=1}^N$, we can define the interpolation for each frame in parallel via the aforementioned procedure.

Training. During training, we would like to learn a parameterized vector field $v_\theta(\mathbf{T}_t, t)$, a deep neural network with parameters θ, which takes as input the intermediate frames $\mathbf{T}_t$ at time t sampled from $\mathcal{U}(0, 1)$, and predicts the final frames $\hat{\mathbf{T}} = \{\hat{T}^{(n)}\}_{n=1}^N$, where $\hat{T}^{(n)} = (\hat{r}_t^{(n)}, \hat{x}_t^{(n)})$. The ground truth vector field u_t for mapping from the intermediate frames $\mathbf{T}_t$ to the ground truth frames $\mathbf{T}_1$ can also be decomposed into a ground truth rotation and translation for each frame $T^{(n)}$:

$$\text{Translations:} \quad u_t(x^{(n)} | x_0^{(n)}, x_1^{(n)}) = x_1^{(n)} \ , \tag{4}$$

$$\text{Rotations:} \quad u_t(r^{(n)} | r_0^{(n)}, r_1^{(n)}) = \log_{r_t^{(n)}}(r_1^{(n)}) \ . \tag{5}$$

To train the model v_θ, we compute separate losses for the predicted rotation $\hat{r}_t \in SO(3)$ and translation $\hat{x}_t \in \mathbb{R}^3$. The combined $SE(3)$ flow matching loss over N frames is as follows:

$$\mathcal{L}_{SE(3)} = \mathbb{E}_{t, p_0(\mathbf{T}_0), p_1(\mathbf{T}_1)} \Bigg[\frac{1}{(1-t)^2} \sum_{n=1}^{N} \bigg\{ \left\| \hat{x}_t^{(n)} - x_1^{(n)} \right\|^2_{\mathbb{R}^3} + \left\| \log_{r_t^{(n)}}(\hat{r}_1^{(n)}) - \log_{r_t^{(n)}}(r_1^{(n)}) \right\|^2_{SO(3)} \bigg\} \Bigg]. \tag{6}$$

The architecture of the flow model v_θ is similar to the structure module from AlphaFold2 comprising Invariant Point

Attention layers interleaved with standard Transformer encoder layers, following Yim et al. (2023a;b). We use an MLP head to predict torsion angles Φ.

Auxiliary losses. The inclusion of auxiliary loss terms to the objective in Equation (6) can be seen as a form of adding domain knowledge into the training process (Yim et al., 2023b). We include 3 additional losses that operate on the all-atom structure inferred from the predicted frames, weighted by tunable coefficients to modulate their contribution to the total loss:

$$\mathcal{L}_{\text{tot}} = \mathcal{L}_{SE(3)} \, + \, \mathcal{L}_{\text{bb}} \, + \, \mathcal{L}_{\text{dist}} \, + \, \mathcal{L}_{\text{tors}} \, . \tag{7}$$

Suppose $S = \{C4', C3', O4'\}$ is the set of frame atoms[2] and the sequence length is N. We summarise the auxiliary losses subsequently.

- **Coordinate MSE** $\mathcal{L}_{\text{bb}}$: A direct all-atom MSE is computed between generated and ground truth coordinates. Here, $a, \hat{a}$ are the ground truth and predicted atomic coordinates for the frame atoms:

$$\mathcal{L}_{\text{bb}} = \frac{1}{|S|N} \sum_{n=1}^{N} \sum_{a \in S} \|a^{(n)} - \hat{a}^{(n)}\|^2. \tag{8}$$

- **Distogram loss** $\mathcal{L}_{\text{dist}}$: A distogram $D \in \mathbb{R}^{NS \times NS}$ containing all-to-all coordinate differences between the atoms in an RNA structure is computed. Let $D_{ab}^{(nm)} = \|a^{(n)} - b^{(m)}\|$ be the elements of the distogram for the ground truth structure. Here, atom a belongs to nucleotide n and atom b to nucleotide m. Given the corresponding predicted distogram $\hat{D}_{ab}^{(nm)}$, we compute another difference between the tensors:

$$\mathcal{L}_{\text{dist}} = \frac{1}{(|S|N)^2 - N} \sum_{\substack{n,m=1 \\ n \neq m}}^{N} \sum_{a,b \in S} \|D_{ab}^{(nm)} - \hat{D}_{ab}^{(nm)}\|^2. \tag{9}$$

- **Torsional loss** $\mathcal{L}_{\text{tors}}$: An angular loss between the 8 predicted torsions by the auxiliary MLP head and the angles from the ground truth all-atom structure. Suppose $\phi \in \Phi_n$ and $\hat{\phi} \in \hat{\Phi}_n$ are the ground truth and predicted torsion angles for residue n, we compute:

$$\mathcal{L}_{\text{tors}} = \frac{1}{8N} \sum_{n=1}^{N} \sum_{\phi \in \Phi_n} \Big(\|\phi - \hat{\phi}\|^2 \Big). \tag{10}$$

Sampling. To generate or unconditionally sample an RNA backbone of length N, we initialize a random point cloud of frames. We use our trained flow model v_θ within an ODE solver to iteratively transform the noisy frames into a realistic RNA backbone. For each nucleotide, we begin with a noisy frame $T_0 = (r_0, x_0)$ at time step $t = 0$, and integrate to $t = 1$ using the Euler method for a specified number of steps N_T, with step size $\Delta t = 1/N_T$. At each step t, the flow model v_θ predicts updates for the frames via a rotation $\hat{r}_1$ and translation $\hat{x}_1$:

$$\begin{aligned} \text{Translations:} \quad x_{t+\Delta t} &= x_t + \Delta t \cdot (\hat{x}_1 - x_t) \, , & (11) \\ \text{Rotations:} \quad r_{t+\Delta t} &= \exp_{r_t}(\, c \, \Delta t \cdot \log_{r_t}(\hat{r}_1)) \, , & (12) \end{aligned}$$

where $c = 10$ is a tunable hyperparameter governing the exponential sampling schedule for rotations.

Conditional generation. The unconditional sampling strategy described above aims to generate realistic RNA backbone structures sampled from the training distribution. However using generative models in real-world design tasks entails *conditional* generation based on specified design constraints or requirements (Ingraham et al., 2022; Watson et al., 2023), which we are currently exploring. For example, unconditional models can leverage inference-time guidance strategies (Wu et al., 2024), be fine-tuned conditionally (Denker et al., 2024) or in an amortized fashion for motif-scaffolding (Didi et al., 2023). For sequence conditioning and structure prediction, we can incorporate embeddings from language models (Penic et al., 2024; He et al., 2024).

2.2. Architecture

Following Yim et al. (2023a;b), we use the structure module from AlphaFold2 (Jumper et al., 2021) comprising Invariant Point Attention (IPA). We also use an auxiliary MLP head to predict torsion angles Φ. We provide hyperparameters, objective functions, and additional experimental setup details in Appendix A.1.

3. Experiments

3D RNA structure dataset. RNAsolo (Adamczyk et al., 2022) is a recent dataset of RNA 3D structures extracted from isolated RNAs, protein-RNA complexes, and DNA-RNA hybrids from the Protein Data Bank (as of January 5, 2024). The dataset contains 14,366 structures at resolution ≤ 4 Å (1 Å = 0.1nm). We select sequences of lengths between 40 and 150 nucleotides (5,319 in total) as we envisioned this size range containing structured RNAs of interest for design tasks.

Evaluation metrics. We evaluate our models for unconditional RNA backbone generation, analogous to recent work in protein design (Yim et al., 2023b;a; Bose et al., 2023; Lin and AlQuraishi, 2023). We generate 50 backbones for target lengths sampled between 40 and 150 at intervals of 10. We then compute the following indicators of quality for these

[2] In Appendix B.1, we show how including all backbone atoms better accounts for larger RNA nucleotides and improves validity of generated samples.

backbones:

- **Validity** ($\texttt{scTM} \geq 0.45$): We inverse fold each generated backbone using gRNAde (Joshi et al., 2023) and pass $N_{\text{seq}} = 8$ generated sequences into RhoFold (Shen et al., 2022). We then compute the self-consistency TM-score (scTM) between the predicted RhoFold structure and our backbone at the $C4'$ level. We say a backbone is *valid* if $\texttt{scTM} \geq 0.45$; this threshold corresponds to roughly the same fold between two RNA strands (Zhang et al., 2022). We expand on this framework in Figure 5.
- **Diversity**: Among the valid samples, we compute the number of unique structural clusters formed using qTMclust (Zhang et al., 2022) and take the ratio to the total number of samples. Two structures are considered similar if their TM-score ≥ 0.45. This metric shows how much each generated sample varies from others across various sequence lengths.
- **Novelty**: Among the valid samples, we use US-align (Zhang et al., 2022) at the $C4'$ level to compute how structurally dissimilar the generated backbones are from the training distribution. For a set of samples for a given sequence length, we compute the TM-score between all pairs of generated backbones and training samples, and for each generated backbone, we assign the highest TM-score. We call the average across this set, pdbTM.
- **Local structural measurements**: We measure the similarity between bond distances, bond angles, and dihedral angles from the set of generated samples and the training set. To do so, we compute histograms for each of the local structural metrics and use 1D Earth Mover's distance to measure the similarity between generated and training distributions.

Hyperparameters. We use 6 IPA blocks in our flow model, with an additional 3-layer torsion predictor MLP that takes in node embeddings from the IPA module. Our final model contains 16.8M trainable parameters. We use AdamW optimizer with learning rate 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$. We train for $120K$ gradient update steps on four NVIDIA GeForce RTX 3090 GPUs for about 18 hours with a batch size $B = 28$. Each batch contains samples of the same sequence length to avoid padding. Further hyperparameters are listed in Appendix A.1.

Training. Our filtered training dataset with sequences of lengths between 40 and 150 consists of 5,319 samples. For the denoiser, we use 6 IPA blocks and an additional torsion predictor head, the latter being a 3-layer MLP that takes in node embeddings from the IPA module to predict 8 torsion angles. Our final model contains 16.8M trainable parameters. We use the Adam optimizer with a learning rate of 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$. We train for $120K$ gradient update steps on four NVIDIA GeForce RTX 3090 GPUs for 15 hours with a batch size of $B = 20$. Each batch comprises padded samples from randomly selected structural clusters across sequence lengths.

Model	% Valid ↑	Diversity ↑	Novelty ↓
$N_T = 10$	16.7	**0.62**	0.70
$N_T = 50$	**41.0**	0.61	**0.54**
$N_T = 100$	20.0	0.61	0.69
$N_T = 500$	20.0	0.57	0.67
MMDiff	0.0	-	-

Table 1. **RNA backbone generation results**. The best performing model uses $N_T = 50$ timesteps for denoising.

4. Results

4.1. Global Evaluation of Generated RNA Backbones

We begin by analyzing RNA-FrameFlow's samples using the aforementioned evaluation metrics. For validity, we report percentage of samples with $\texttt{scTM} \geq 0.45$; for diversity, we report the ratio of unique structural clusters to total **valid** samples; and for novelty, we report the highest average pdbTM to a match from the PDB. For each sequence length between 40 and 150, at intervals of 10, we generate 50 backbones. Table 1 reports these metrics across different variants for the number of denoising steps N_T. We compare our model to protein-RNA-DNA complex co-design model MMDiff (Morehead et al., 2023), which is a diffusion model. As the original version of MMDiff was trained on shorted RNA sequences, we retrain it on our training set. Additionally, we inverse-folded MMDiff's backbones using gRNAde.

We identify $N_T = 50$ as the best-performing model that balances validity, diversity, and novelty; furthermore, it takes 4.74 seconds (averaged over 5 runs) to sample a backbone of length 100, as opposed to 27.3 seconds for MMDiff with 100 diffusion steps. We note that increasing N_T does not improve validity despite allowing the model to perform more updates to atomic coordinate placements. Our model also outperforms MMDiff. On manual inspection, samples from MMDiff had significant chain breaks and disconnected floating strands; see Appendix C.1.

4.2. Local Evaluation with Structural Measurements

For our best-performing model with number of timesteps $N_T = 50$, we plot histograms of bond distance, bond angles, and dihedral angles in Figure 2. We include the Earth Mover's distance (EMD) between measurements from the training and generated distributions as an indicator of local realism (using 30 bins for each quantity). An ideal

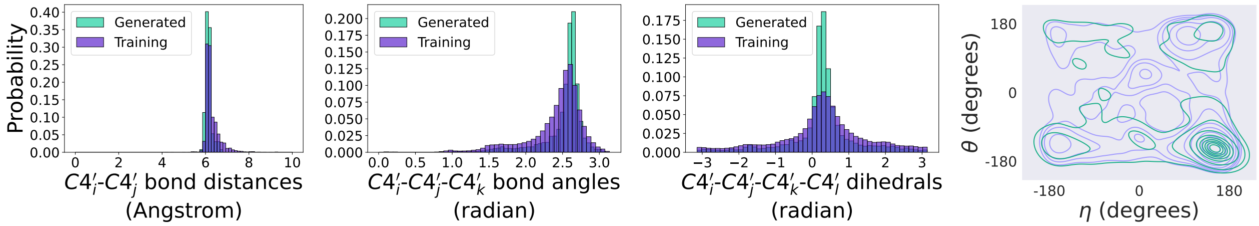

Figure 2. **Local structural metrics** from 600 generated backbone samples, compared to random Gaussian point cloud as a sanity check. Our model can recapitulate local structural descriptors. **(Subplots 1-3)** Histograms of inter-nucleotide bond distances, bond angles between nucleotide triplets, and torsion angles between every four nucleotides. **(Subplot 4)**: RNA-centric Ramachandran plot of structures from the training set (purple) and generated backbones (green).

Model	EMD ↓ (dist)	EMD ↓ (angles)	EMD ↓ (torsions)
RNA-FrameFlow	**0.17**	**0.11**	**2.36**
MMDiff (original)	1.38	0.43	3.06
MMDiff (RNAsolo)	0.39	0.21	3.23
Gaussian noise	29.00	6.35	4.37

Table 2. **Local structural metrics.** Earth Mover's Distance for local structural measurements compared to ground truth measurements from RNAsolo. Our model ($N_T = 50$) shows improved recapitulation of local structural descriptors compared to baselines.

generative model will score an EMD close to 0.0 (i.e. consistent with the training set comprising naturally occurring RNA). In Table 2, we observe EMD values from our best-performing model's backbones being significantly closer to 0.0 compared to MMDiff and random Gaussian all-atom point clouds (akin to an untrained model), which serve as sanity checks. We include histograms for MMDiff in Appendix C.1.

We also show RNA Ramachandran angle plots for generated samples and the training distribution in Figure 2. Keating et al. (2011) introduced $\eta - \theta$ plots, similar to Ramachandran angle plots for proteins, that track the separate dihedral angles formed by $\{C4'_i, P_{i+1}, C4'_{i+1}, P_{i+2}\}$ and $\{P_i, C4'_i, P_{i+1}, C4'_{i+1}\}$ respectively, for each nucleotide i along the chain. We observe that the dehedral angle distribution from RNA-FrameFlow closely recapitulates the distribution of naturally occuring RNA structures from the training set.

4.3. Generation Quality Across Sequence Lengths

We next investigate how sequence length affects the global realism of generated samples (measured by `scTM`). Figure 3 (Left) shows the performance of RNA-FrameFlow for different sequence lengths. We observe our model generates samples with high `scTM` for specific sequence lengths like 50, 60, 70, and 120 while generating poorer quality structures for other lengths. We believe the fluctuation of TM-scores may be due to certain lengths being over-represented in the training distribution. We can also partially attribute this to the inherent length bias of RhoFold; see Appendix A.2. With a better structure predictor, we expect an increase in valid samples that meet the 0.45 TM-score threshold.

We also analyze the novelty of our generated samples (measured by `pdbTM`) in Figure 3 (Middle). We are particularly interested in samples that lie in the right half with high `scTM` and low `pdbTM`, which means that the designs are highly likely to fold back into the sampled backbone but are structurally dissimilar to any RNAs in the training set. It is worth noting that our training set has high structural similarity among samples: running `qTMclust` on our training dataset revealed only 342 unique clusters from 5,319 samples, which indicates that the model does not encounter a diverse set of samples during training. This contributes to many generated samples from our model looking similar to samples from the training distribution. We include two such examples in Figure 3 (Right). Both generated RNAs yield relatively high `pdbTM` scores and look similar to their respective closest matching chain from the training set: a tRNA at length 70 and a 5S ribosomal RNA at length 120, respectively. We include comparative results on validity and novelty for MMDiff in Appendix C.1, finding that MMDiff does not generate any samples that pass the validity criteria.

4.4. Data Preparation Protocols

Due to the overrepresentation of RNA strands of certain lengths (mostly corresponding to tRNA or 5S ribosomal RNA) in our training set, our models generate close likenesses for those lengths that achieve high self-consistency but are not novel folds. To avoid this memorized recapitulation and promote increased diversity among samples, we sought to develop data preparation protocols to balance RNA folds across sequence lengths.

- **Structural clustering:** We cluster our training set using `qTMclust`. When creating a training batch, we sample random clusters and from each, a random structure. This ensures a batch does not comprise solely of samples for a single sequence length or is dominated

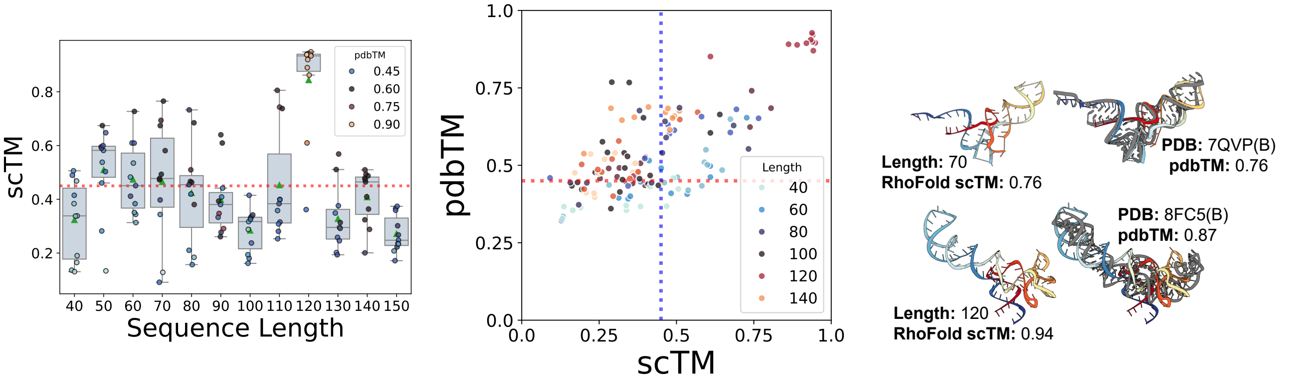

Figure 3. **Validity and novelty of generated backbones.** **(Left)** `scTM` of backbones of lengths 40-150 with the mean and spread of `scTM` for each length; we select the top 10 structures with the best validation scores per length. **(Middle)** Scatter plot of self-consistency TM-score (`scTM`) and novelty (`pdbTM`) across lengths. Vertical and horizontal dotted lines represent TM-score thresholds of 0.45. **(Right)** Selected samples with high pdbTM scores (colored) with the closest, aligned match from the PDB (gray). Our model generates valid backbones for certain sequence lengths and tends to recapitulate the most frequent folds in the PDB (e.g., tRNAs, small rRNAs).

MODEL	% VALID ↑	DIVERSITY ↑	NOVELTY ↓
BASE	**41.0**	0.62	0.54
+ CLUST	12.0	**0.88**	0.49
+ CROP	11.0	0.85	**0.47**

Table 3. **Impact of data preparation strategies**. Increasing the diversity of the training dataset using a combination of strategies improves diversity and novelty of generated structures but leads to fewer designs passing the validity threshold.

by over-represented folds. There are only 342 structural clusters for the 5,319 samples within sequence lengths 40-150, highlighting the lack of diversity in RNA structural data.

- **Cropping augmentation:** We expand our training set by cropping longer RNA strands beyond length 150 by sampling a random crop length in $[40, 150]$ and extracting a contiguous segment from the larger chains. As cropped RNA are not standalone molecules and serve only to augment the dataset, we consider a randomly chosen 20% of the training set size to balance uncropped and cropped samples; this gives 1,063 extra cropped samples.

We train identical models on these data splits for 120K gradient steps, with evaluation results reported in Table 3 showing improved diversity and novelty in the generated samples, at the cost of reduced validity. For structural clustering, each batch comprises padded samples up to a maximum length of 150 from randomly selected structural clusters across sequence lengths. See Appendix C.2 for full results for the two alternate data preparation protocols.

5. Limitations and Discussions

Altogether, our experiments demonstrate that the $SE(3)$ flow matching framework is sufficiently expressive for learning the distribution of 3D RNA structure and generating realistic RNA backbones similar to well-represented RNA folds in the PDB. Select examples are shown in Figure 6. We have also identified notable limitations and avenues for future work, which we highlight below.

Physical violations. While well-trained models usually generate realistic RNA backbones, we do observe some physical violations: generated backbones sometimes have chains that are either too close by or directly clash with one another, are highly coiled, have excessive loops and unrealistically intertwined helices, or have chain breaks. We highlight these limitations in Figure 4. RNA tertiary structure folding is driven by *base pairing* and *base stacking* (Vicens and Kieft, 2022) which influence the formation of helices, loops, and other tertiary motifs. Base pairing refers to nucleotides along adjacent chains forming hydrogen bonds, while base stacking involves interactions between rings of adjacent nucleotide bases along a chain. To our knowledge, all current deep learning models operate on individual nucleotides, only implicitly learning base pairing and stacking. Developing explicit representations of these interactions as part of the architecture may further minimize physical violations and provide stronger inductive biases to learn complex tertiary RNA motifs.

Generalisation and novelty. We observed that the best designs from our models (as measured by `scTM` score) are sampled at lengths 70-80 and 120-130, and often have closely matching structures in the PDB (high TM-scores). This suggests that models can recapitulate well-represented RNA folds in their training distribution (e.g., both tRNAs at length 70-90 and small 5S ribosomal RNAs at length 120

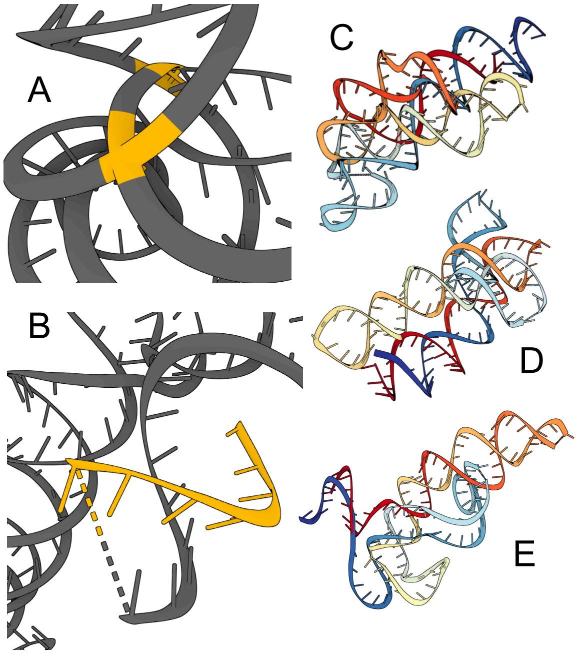

Figure 4. **Physical violations in generated samples.** (A) Steric clashes between generated chains (highlighted in yellow). (B) Chain breaks and stray strands (highlighted in yellow). (C)-(E) Excessive loops and intertwined helices.

are very frequent). However, self-consistency metrics were relatively poorer for less frequent lengths, suggesting that models are currently not designing novel folds.

We would also like to note that the models we use for structure prediction and inverse folding may be similarly biased to perform well for certain sequence lengths, leading to the overall pipeline being reliable for commonly occurring lengths and unreliable for less frequent ones (see Appendix A.2 for an analysis on RhoFold). We evaluated preliminary strategies for structural clustering and cropping augmentations during training, which improved the novelty of designed structures but led to fewer designs passing the validity filter. Overall, the relative scarcity of RNA structural data compared to proteins necessitates greater care in preparing data pipelines for scaling up training and/or incorporating inductive biases into generative models, which we hope to continue exploring.

6. Related Work

Here, we summarize recent developments in deep learning for 3D RNA modeling and design. Recent RNA structure prediction tools include RhoFold (Shen et al., 2022), RoseTTAFold2NA (Baek et al., 2022), DRFold (Li et al., 2023b), and AlphaFold3 (Abramson et al., 2024), each with varying performance that is yet to match the current state-of-the-art for proteins. However, structure prediction tools are not directly capable of designing new structures, which this work aims to address by adapting an $SE(3)$ flow matching framework for proteins (Yim et al., 2023a). MMDiff (Morehead et al., 2023), a diffusion model for protein-nucleic acid complex generation, is also capable of designing RNA structures. Our evaluation shows that our flow matching model significantly outperforms both the original and RNA-only versions of MMDiff.

Joshi et al. (2023) introduce gRNAde for 3D RNA inverse folding, a closely related task of designing new sequences conditioned on backbone structures. We use gRNAde followed by RhoFold in our evaluation pipeline to forward fold designed backbones and measure structural self-consistency. Independently and concurrent to our work, Nori and Jin (2024) propose RNAFlow, which uses gRNAde combined with RoseTTAFold2NA as a denoiser in the flow matching setup to design RNA sequences conditioned on protein structures. Our work tackles *de novo* 3D RNA backbone generation, an orthogonal RNA design task.

7. Conclusion

We introduce RNA-FrameFlow, a generative model for 3D RNA backbone design. Our evaluations show our model can design locally realistic and moderately novel backbones of length 40 – 150 nucleotides. We achieve a validity score of 41.0% and relatively strong diversity and novelty scores compared to diffusion model baselines and ablated variants. While generative models can successfully recapitulate well-represented RNA folds (e.g., tRNAs, small rRNAs), the lack of diversity in the training data may hinder broad generalization. We are actively exploring improved data preparation strategies combined with inductive biases that explicitly incorporate physical interactions that drive RNA structure. We hope RNA-FrameFlow and the associated evaluation framework serve as foundations for the community to explore 3D RNA design, towards developing conditional generative models for real-world design scenarios.

Acknowledgements

We thank Jason Yim and Emile Mathieu for helpful discussions. CKJ was supported by the A*STAR Singapore National Science Scholarship (PhD). AM was supported by a U.S. NSF grant (DBI2308699) and two U.S. NIH grants (R01GM093123 and R01GM146340). SVM was supported by the UKRI Centre for Doctoral Training in Application of Artificial Intelligence to the study of Environmental Risks (EP/S022961/1). This research was partially supported by a Cambridge Dawn Supercomputer Pioneer Project compute grant.

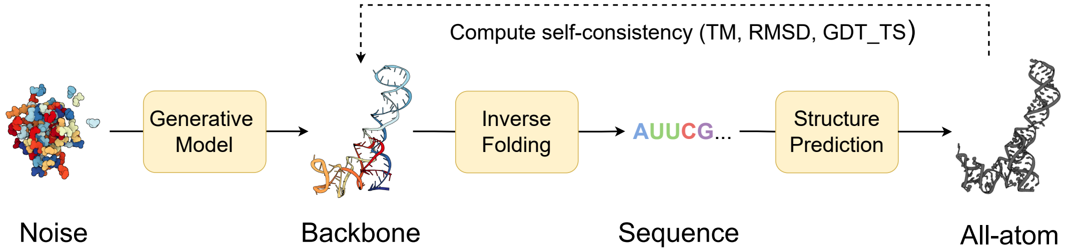

Figure 5. **Structural self-consistency evaluation.** We sample a backbone from our model and pass it through an inverse folding model (gRNAde) to obtain $N_{\text{seq}} = 8$ sequences. Each sequence is fed into a structure prediction model (RhoFold) to get the predicted all-atom backbone. Self-consistency between a predicted backbone and the generated sample is measured with TM-score (we also report RMSD and GDT_TS). For a given generated sample, we thus have $N_{\text{seq}} = 8$ TM-scores of which we take the maximum as its `scTM` score.

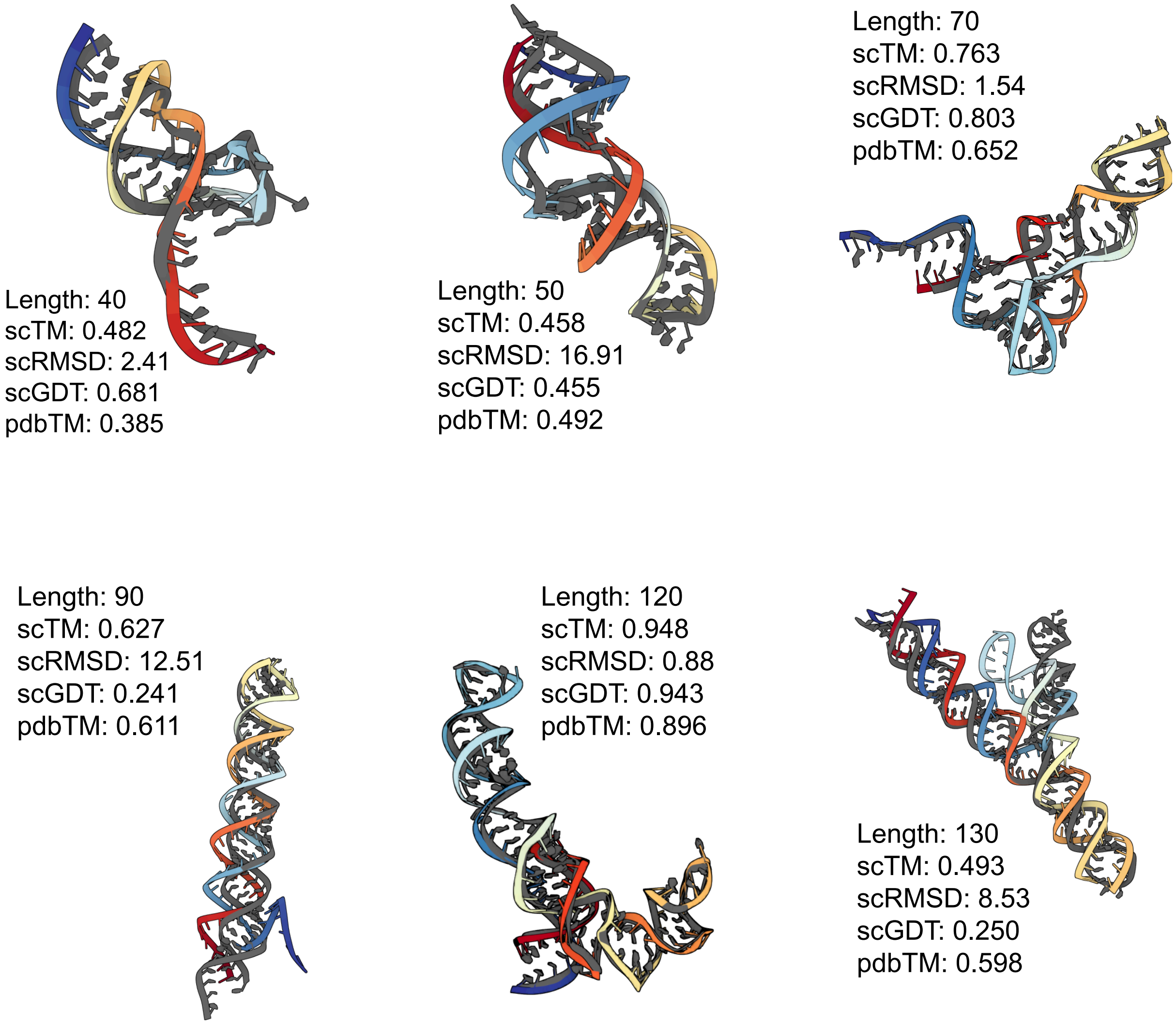

Figure 6. Some generated RNA backbones (colored) of varying lengths aligned with their RhoFold predicted structure (gray).

References

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zidek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021.

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*, 2022.

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*, 2023.

Jennifer A Doudna and Emmanuelle Charpentier. The new frontier of genome engineering with crispr-cas9. *Science*, 2014.

Mihir Metkar, Christopher S Pepin, and Melissa J Moore. Tailor made: the art of therapeutic mrna design. *Nature Reviews Drug Discovery*, 23(1):67–83, 2024.

Jerome Mulhbacher, Patrick St-Pierre, and Daniel A Lafontaine. Therapeutic applications of ribozymes and riboswitches. *Current opinion in pharmacology*, 2010.

Tulsi Ram Damase, Roman Sukhovershin, Christian Boada, Francesca Taraballi, Roderic I Pettigrew, and John P Cooke. The limitless future of rna therapeutics. *Frontiers in bioengineering and biotechnology*, 9:628137, 2021.

Dongran Han, Xiaodong Qi, Cameron Myhrvold, Bei Wang, Mingjie Dai, Shuoxing Jiang, Maxwell Bates, Yan Liu, Byoungkwon An, Fei Zhang, et al. Single-stranded dna and rna origami. *Science*, 2017.

Joseph D Yesselman, Daniel Eiler, Erik D Carlson, Michael R Gotrik, Anne E d'Aquino, Alexandra N Ooms, Wipapat Kladwang, Paul D Carlson, Xuesong Shi, David A Costantino, et al. Computational design of three-dimensional rna structure and function. *Nature nanotechnology*, 2019.

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

Yueyi Li, Anibal Arce, Tyler Lucci, Rebecca A Rasmussen, and Julius B Lucks. Dynamic rna synthetic biology: new principles, practices and potential. *RNA biology*, 2023a.

Jason Yim, Andrew Campbell, Andrew Y. K. Foong, Michael Gastegger, José Jiménez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S. Veeling, Regina Barzilay, Tommi Jaakkola, and Frank Noé. Fast protein backbone generation with se(3) flow matching, 2023a.

Bartosz Adamczyk, Maciej Antczak, and Marta Szachniuk. RNAsolo: a repository of cleaned PDB-derived RNA 3D structures. *Bioinformatics*, 2022.

Chaitanya K. Joshi, Arian R. Jamasb, Ramon Viñas, Charles Harris, Simon Mathis, Alex Morehead, Rishabh Anand, and Pietro Liò. grnade: Geometric deep learning for 3d rna inverse design, 2023.

Tao Shen, Zhihang Hu, Zhangzhi Peng, Jiayang Chen, Peng Xiong, Liang Hong, Liangzhen Zheng, Yixuan Wang, Irwin King, Sheng Wang, Siqi Sun, and Yu Li. E2efold-3d: End-to-end deep learning method for accurate de novo rna 3d structure prediction, 2022.

Quentin Vicens and Jeffrey S Kieft. Thoughts on how to think (and talk) about rna structure. *Proceedings of the National Academy of Sciences*, 2022.

Rachael C Kretsch, Ebbe S Andersen, Janusz M Bujnicki, Wah Chiu, Rhiju Das, Bingnan Luo, Benoît Masquida, Ewan KS McRae, Griffin M Schroeder, Zhaoming Su, et al. Rna target highlights in casp15: Evaluation of predicted models by structure providers. *Proteins: Structure, Function, and Bioinformatics*, 2023.

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

Alex Morehead, Jeffrey Ruffolo, Aadyot Bhatnagar, and Ali Madani. Towards joint sequence-structure generation of nucleic acid and protein complexes with se(3)-discrete diffusion, 2023.

Anke Gelbin, Bohdan Schneider, Lester Clowney, Shu-Hsin Hsieh, Wilma K Olson, and Helen M Berman. Geometric parameters in nucleic acids: sugar and phosphate constituents. *Journal of the American Chemical Society*, 118 (3):519–529, 1996.

Stephen C Harvey and M Prabhakaran. Ribose puckering: structure, dynamics, energetics, and the pseudorotation cycle. *Journal of the American Chemical Society*, 108 (20):6128–6136, 1986.

Mary C Clay, Laura R Ganser, Dawn K Merriman, and Hashim M Al-Hashimi. Resolving sugar puckers in rna excited states exposes slow modes of repuckering dynamics. *Nucleic acids research*, 45(14):e134–e134, 2017.

Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries, 2024.

Jason Yim, Brian L. Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. Se(3) diffusion model with application to protein backbone generation, 2023b.

John Ingraham, Max Baranov, Zak Costello, Vincent Frappier, Ahmed Ismail, Shan Tie, Wujie Wang, Vincent Xue, Fritz Obermeyer, Andrew Beam, et al. Illuminating protein space with a programmable generative model. *bioRxiv*, pages 2022–12, 2022.

Luhuan Wu, Brian Trippe, Christian Naesseth, David Blei, and John P Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Alexander Denker, Francisco Vargas, Shreyas Padhy, Kieran Didi, Simon Mathis, Vincent Dutordoir, Riccardo Barbano, Emile Mathieu, Urszula Julia Komorowska, and Pietro Lio. Deft: Efficient finetuning of conditional diffusion models by learning the generalised h-transform. *arXiv preprint arXiv:2406.01781*, 2024.

Kieran Didi, Francisco Vargas, Simon Mathis, Vincent Dutordoir, Emile Mathieu, Urszula Julia Komorowska, and Pietro Lio. A framework for conditional diffusion modelling with applications in motif scaffolding for protein design. In *NeurIPS 2023 Machine Learning for Structural Biology Workshop*, 2023.

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

Shujun He, Rui Huang, Jill Townley, Rachael C Kretsch, Thomas G Karagianes, David BT Cox, Hamish Blair, Dmitry Penzar, Valeriy Vyaltsev, Elizaveta Aristova, et al. Ribonanza: deep learning of rna structure through dual crowdsourcing. *bioRxiv*, 2024.

Avishek Joey Bose, Tara Akhound-Sadegh, Kilian Fatras, Guillaume Huguet, Jarrid Rector-Brooks, Cheng-Hao Liu, Andrei Cristian Nica, Maksym Korablyov, Michael Bronstein, and Alexander Tong. Se(3)-stochastic flow matching for protein backbone generation, 2023.

Yeqing Lin and Mohammed AlQuraishi. Generating novel, designable, and diverse protein structures by equivariantly diffusing oriented residue clouds, 2023.

Chengxin Zhang, Morgan Shine, Anna Marie Pyle, and Yang Zhang. Us-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nature methods*, 2022.

Kevin S Keating, Elisabeth L Humphris, and Anna Marie Pyle. A new way to see rna. *Quarterly reviews of biophysics*, 44(4):433–466, 2011.

Minkyung Baek, Ryan McHugh, Ivan Anishchenko, David Baker, and Frank DiMaio. Accurate prediction of nucleic acid and protein-nucleic acid complexes using rosettafoldna. *bioRxiv*, 2022.

Yang Li, Chengxin Zhang, Chenjie Feng, Robin Pearce, P Lydia Freddolino, and Yang Zhang. Integrating end-to-end learning with deep geometrical potentials for ab initio rna structure prediction. *Nature Communications*, 2023b.

Divya Nori and Wengong Jin. Rnaflow: Rna structure & sequence co-design via inverse folding-based flow matching. In *ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design*, 2024.

A. Additional Experimental Details

A.1. Architectural Details

We list hyperparameters used for our denoiser model in Table 4 below:

Category	Hyperparameter	Value
Invariant Point Attention (IPA)	Atom embedding dimension D_h	256
	Hidden dimension D_z	128
	Number of blocks	6
	Query and key points	8
	Number of heads	8
	Key points	12
Transformer	Number of heads	4
	Number of layers	2
Torsion Prediction MLP	Input dimension	256
	Hidden dimension	128
Schedule	Translations (training)	linear
	Rotations (training)	linear
	Translations (sampling)	linear
	Rotations (sampling)	exponential
	Number of denoising steps N_T	$[10, \mathbf{50}, 100, 500]$

Table 4. Hyperparameters for the baseline model.

A.2. RhoFold Length Bias

We investigate the performance of RhoFold on the RNAsolo training dataset used for our generative model. Figure 7 shows sequence length bias where RhoFold predicts structures with extremely low RMSDs for sequence lengths (like 70, 100, and 120) while predicting poor structures for other lengths with larger RMSDs. The performance across lengths is disparate (like AlphaFold2) and may influence what is considered valid. Furthermore, RhoFold is not optimized for *de novo* designed RNA, only naturally occurring RNA. To compensate for bias, we resort to a *ranking* instead of thresholding done by (Yim et al., 2023b;a) when measuring validity.

B. Ablations

B.1. Composition of Backbone Coordinate Loss

We also analyze how changing the composition of atoms considered in the inter-atom losses affects performance. We increase the number of atoms being supervised in the $\mathcal{L}_{\text{bb}}$ loss described above. Aside from the frame comprising $C3'$, $C4'$, and $O4'$, we try two settings with 3 and 7 additional non-frame atoms included in the loss. For the 3 non-frame atoms, we choose $C1'$, P, and $O3'$, and for the 7 non-frame atoms, we choose a superset $C1'$, P, $O3'$, $C5'$, $OP1$, $OP2$, and $N1/N9$. We posit the additional supervision may increase the local structural realism, which may further improve validity, as shown in Table 5.

FRAME COMPOSITION IN $\mathcal{L}_{\text{BB}}$	% VALIDITY ↑	DIVERSITY ↑	NOVELTY ↓
FRAME ONLY (BASELINE)	41.0	**0.62**	**0.54**
FRAME AND 3 NON-FRAME	45.0	0.28	0.79
FRAME AND 7 NON-FRAME	**46.7**	0.35	0.85

Table 5. Ablating composition of backbone loss $\mathcal{L}_{\text{bb}}$. Supervising more non-frame atoms improves validity but worsens diversity and novelty. Best per-column result is **bolded**.

We indeed observe increasing validity as we increase the frame complexity in the auxiliary backbone loss. The minute RMSD contributions from disordered fragments of the RNA may be minimal, accounting for greater likeness to the RhoFold

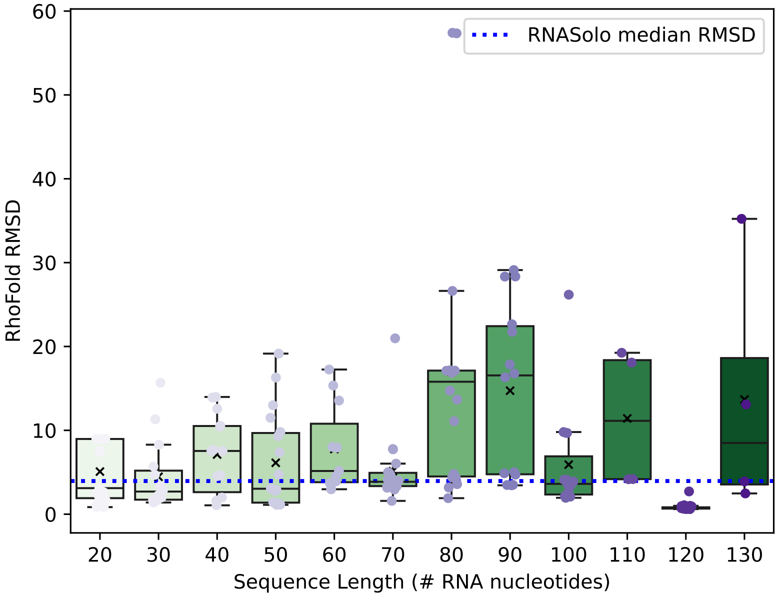

Figure 7. **RhoFold length bias**. RhoFold has a strong bias for certain sequence lengths over others. This affects its efficacy when used to compute the 3D self-consistency of generated backbones. The blue dotted line represents the median RMSD of RhoFold predictions to the samples from RNAsolo. To minimize the influence of this length bias, we use TM-score for self-consistency because it does not penalize flexible regions as much as RMSD.

predicted structures, scoring relatively higher scTM scores. However, the original frame-only baseline model has better diversity and novelty which we attribute to high local variation in atomic placements. This variation causes two generated structures for the same sequence length to look very different at an all-atom resolution.

B.2. Composition of Auxiliary Loss

We ablate the inclusion of different auxiliary loss terms that guide our $SE(3)$ flow matching setup; results are in Table 6. Although, there is an increase in EMD for bond distances as we remove distance-based losses like backbone coordinate loss $\mathcal{L}_{\text{bb}}$ and all-to-all pairwise distance loss ($\mathcal{L}_{\text{dist}}$). However, we also observe the model still learns realistic distributions despite removing different loss terms, indicating that each loss makes up for the absence of the other. Moreover, the best model still uses all losses with any removal causing a drop in validity.

$\mathcal{L}_{\text{BB}}$	$\mathcal{L}_{\text{DIST}}$	$\mathcal{L}_{SO(3)}$	EMD (DISTANCE) ↓	EMD (ANGLES) ↓	EMD (TORSIONS) ↓	% VALIDITY ↑
✓	✓	✓	0.17	0.11	2.36	41.0
✓		✓	0.18	0.14	3.85	35.0
✓	✓		0.23	0.11	3.72	13.3
	✓	✓	0.18	0.18	3.59	16.7

Table 6. Ablations of loss terms on Earth Mover's Distance scores for structural measurements compared to ground truth measurements from RNAsolo. The first row corresponds to the baseline model. Distance-based losses like the backbone coordinate loss ($\mathcal{L}_{\text{bb}}$) and all-to-all pairwise distance loss ($\mathcal{L}_{\text{dist}}$) are necessary to learn geometric properties like bond distances adequately.

Further inspecting the samples from the models without each loss term reveals structural deformities at the all-atom level. Figure 8 shows such artifacts resulting from not enforcing geometric constraints through explicit losses.

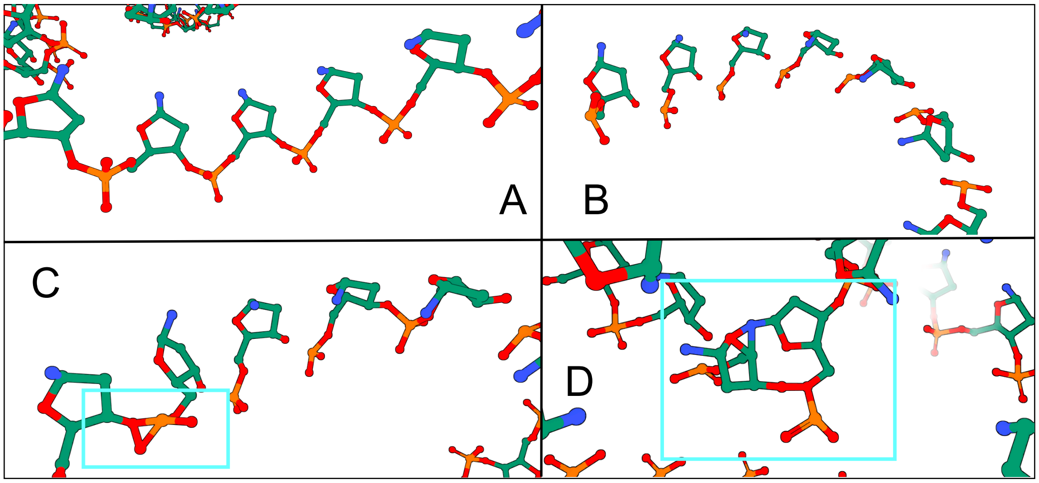

Figure 8. Not including auxiliary losses causes structural issues in generated RNAs. **(A)** RNA backbone from our baseline model with expected adherence to bonding between nucleotides. **(B)** Not including the rotation loss $\mathcal{L}_{SO(3)}$ causes nucleotides to have random orientations, preventing them from connecting contiguously. **(C)** Not including the backbone atom loss $\mathcal{L}_{\text{bb}}$ causes intra-residue atoms to be placed too close to one another resulting in bonds that should not exist. **(D)** Not including the all-to-all pairwise distance loss $\mathcal{L}_{\text{dist}}$ causes deformations and fusing between adjacent frames, and unrealistic nucleotide placements, especially along helices and loops.

C. Additional Results

C.1. Evaluation of MMDiff Samples

Here, we document global and local metrics from samples generated by MMDiff. MMDiff has a validity score of 0.0% as all the samples have a poor `scTM` score below the 0.45 threshold to the RhoFold predicted backbones. Even though none of the samples are valid, we show the average `pdbTM` scores for the samples, which are trivially low as there are no structures from the PDB that match them due to poor quality.

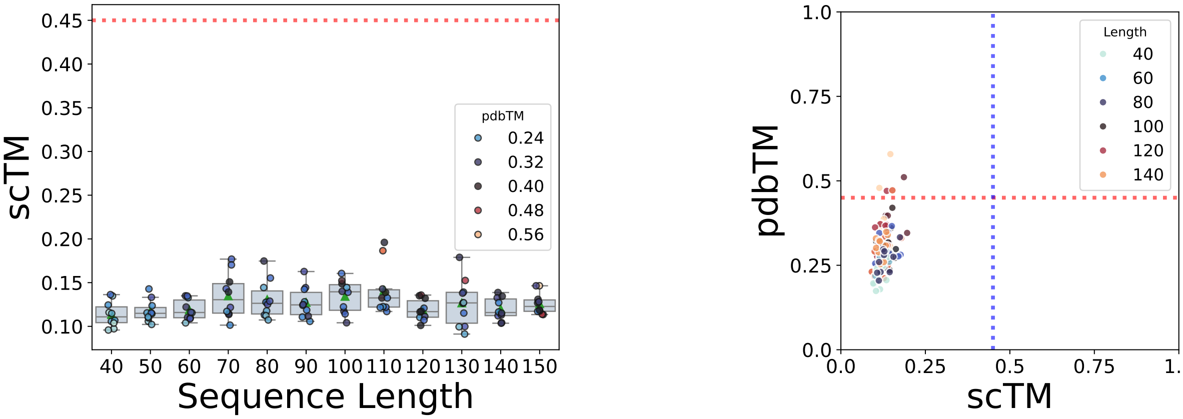

Figure 9. **Validity and novelty of retrained MMDiff's generated backbones. (Left)** `scTM` of backbones of lengths 40-150 with the mean and spread of `scTM` for each length. **(Middle)** Scatter plot of self-consistency TM-score (`scTM`) and novelty (`pdbTM`) across lengths. Vertical and horizontal dotted lines represent TM-score thresholds of 0.45. Overall, MMDiff retrained on RNAsolo does not generate realistic RNA structures.

While MMDiff's samples locally resemble RNA structures given realistic, manual inspection reveals multiple chain breaks and disconnected floating strands, resulting in 0.0% validity. In 10 **(Subplot 1)**, we see inter-residue $C4'$ distances slightly varying, causing the chain breaks. Furthermore, the Ramachandran plot in Figure 10 **(Subplot 4)** reveals a more complex angular distribution than found in RNAsolo, which may be a consequence of excessively folded regions or substructures that may have folded in on themselves.

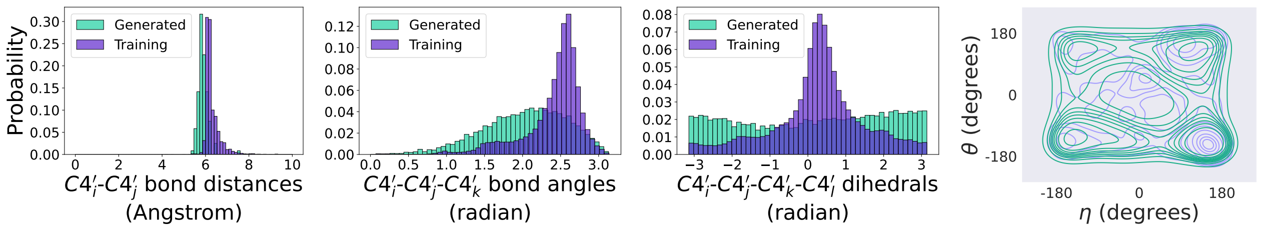

Figure 10. **Structural measurements** from samples generated by MMDiff. **(Subplots 1-3)** Left: histogram of inter-nucleotide bond distances in Angstrom. Middle: histogram of bond angles between nucleotide triplets. Right: histogram of torsion (dihedral) angles between every four nucleotides. **(Subplot 4)**: RNA-centric Ramachandran plot of structures from the training set (purple) and MMDiff's generated backbones (green).

C.2. Evaluation of Data Preparation Strategies

We include global evaluation metrics for the two data preparation strategies presented in the main text, namely structural clustering and cropping augmentation.

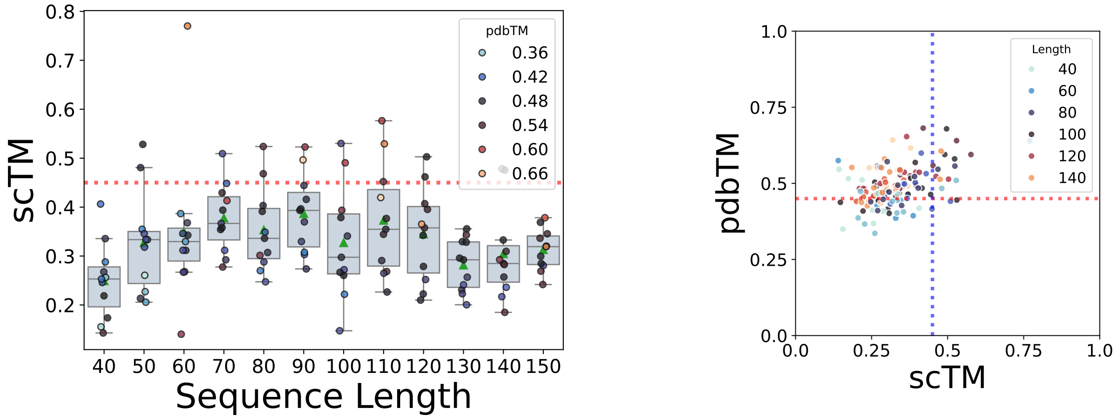

Figure 11. Validity and novelty of top-10 generated backbones from the model trained with only structural clustering. **(Left)** `scTM` of backbones of lengths 40-150 with the mean and spread of `scTM` for each length. **(Middle)** Scatter plot of self-consistency TM-score (`scTM`) and novelty (`pdbTM`) across lengths. Vertical and horizontal dotted lines represent TM-score thresholds of 0.45.

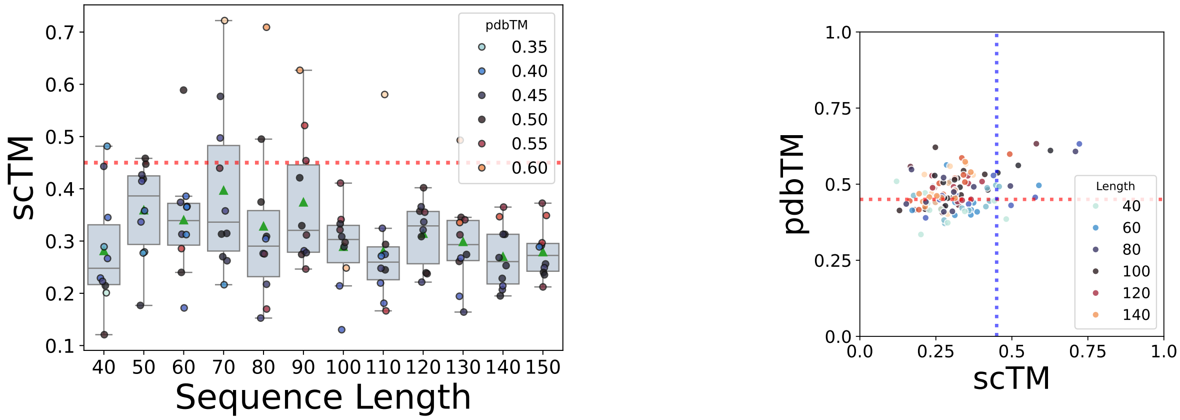

Figure 12. Validity and novelty of top-10 generated backbones from the model trained with structural clustering and cropping. **(Left)** `scTM` of backbones of lengths 40-150 with the mean and spread of `scTM` for each length. **(Middle)** Scatter plot of self-consistency TM-score (`scTM`) and novelty (`pdbTM`) across lengths. Vertical and horizontal dotted lines represent TM-score thresholds of 0.45.