DiffRNAFold: Generating RNA Tertiary Structures with Latent Space Diffusion

Mihir Bafna
Georgia Institute of Technology
mbafna@gatech.edu

Vikranth Keerthipati
Georgia Institute of Technology
vkeerthipati@gatech.edu

Subhash Kanaparthi
Georgia Institute of Technology
skanaparthi3@gatech.edu

Ruochi Zhang
Broad Institute of MIT and Harvard
zhangruo@broadinstitute.org

Abstract

RNA molecules provide an exciting frontier for novel therapeutics. Accurate determination of RNA structure could accelerate development of therapeutics through an improved understanding of function. However, the extremely large conformation space has kept the RNA 3D structure space largely unresolved. Using recent advances in generative modeling, we propose DiffRNAFold, a latent space diffusion model for RNA tertiary structure design. Our preliminary results suggest that DiffRNAFold generated molecules are similar in 3D space to true RNA molecules, providing an important first step towards accurate structure and function prediction *in vivo*.

1 Introduction

1.1 Why RNA?

RNA tertiary structure design is vital for drug discovery and therapeutics.The rapid development of mRNA vaccines during the COVID pandemic highlights the significance of RNA-based therapies. Over 400 RNA-targeting drug programs are underway[26], targeting various diseases. Understanding RNA function and 3D structure is essential for effective therapeutics. However, determining RNA structure has been denoted a grand challenge and even stated to be more difficult than protein structure prediction [25]. The reason for this is simply the flexibility of RNA molecules. While protein molecules, with three torsional angles at each residue, generate enough diversity to make structure prediction difficult, RNA molecules have seven torsional angles at each nucleotide [17]. Due to this large conformational sample space, traditional Monte-Carlo approaches that aim to randomly sample and choose the molecules with lowest free energy, often fail to converge in reasonable time. To overcome this issue, and partly due to the recent success of protein structure prediction with AlphaFold [12], deep learning based methods have been proposed [18, 22]. These methods have shown promising results in structure prediction. With DiffRNAfold, we propose a framework that takes this one step forward with RNA structure generation and design.

1.2 Why Diffusion?

Score based generative modeling and diffusion denoising models [24, 23, 13, 9, 8] are architectures that iteratively add noise to the input samples following the diffusion stochastic differential equation until the sample represents pure noise. The model then seeks to learn the incremental reverse diffusion (denoising) steps and reconstruct the input. After training, the denoising diffusion part of

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

the model can be use to construct high quality samples from pure noise. These models have had major success in the computer vision domain beating GANs in image and 3D shape synthesis [6, 30] and also with the advent of stable-diffusion [19] that utilizes latent space diffusion for high quality text-to-image generation. More recently, diffusion has achieved great results in the computational chemistry domain, specifically for molecular docking, small molecule generation, and even protein structure generation/dynamics [4, 10, 11, 3, 28, 27]. Due to the recent positive results of diffusion models for chemical structure generation, we find it a fitting model for RNA structure generation as well. However, as RNA molecules are much larger than small molecule drugs, we seek to use a latent space diffusion model, where we first encode the molecule's into latent representations before diffusing and denoising. This architecture allows for conditional generation (based on linear sequence of nucleotides) analogous to the aforementioned text-to-image model.

2 Methods

2.1 DiffRNAFold Architecture

DiffRNAFold (see Figure 1) consists of three major parts: (a) a graph autoencoder, (b) the latent space diffusion denoising layers, and c) an optional language model for conditional input. At a high level, the pipeline is as follows. The autoencoder takes RNA features and points (X) as input (Section 2.2), and embeds them into a robust latent space that contextualizes the RNA molecule and simultaneously learns how to decode the latent vector back into the RNA point cloud. Next, to enable high quality generation of RNA molecules, we diffuse (step (b)) on the latent vector by adding Gaussian noise incrementally. The denoising layers then learn to reconstruct the original latent vector from noise. If conditional input (step (c)) is provided, then the linear sequence of an RNA, embedded by a language model, is concatenated with the noisy vector before denoising. This “conditional" generation guides the denoising layers to reconstruct a latent that describes the structure and condition simulateneously. All of these working parts are detailed below.

2.1.1 Graph Autoencoder

See Figure 1a. We use Graph Neural Networks (GNNs). GNN layers essentially involve a series of message passing and aggregation steps. We can think of this process as a function $Z = f(X, A)$, where the graph's vertex features X and adjacency matrix A are used to transmit messages among neighboring vertices. We specifically used graph convolutional neural networks (GCNs) [15]. These layers can be stacked similar to traditional convolutional neural networks. Furthermore, we utilize stacked graph convolutional layers, incorporating the following message-passing rule([15]) :

$$Z^{(l+1)} = \sigma\left(\tilde{D}^{\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}Z^{(l)}W^{(l)}\right),\ Z^{(0)} = X \tag{1}$$

At GCN layer 0, $Z^{(0)}$ is the initial input node features X. The graph's input adjacency matrix is symmetrically normalized shown in (1). Note that $\tilde{A} = A + I_n$ and $\tilde{D}$ is the degree matrix of $\tilde{A}$. At each layer l, there is a learnable weight parameter $W^{(l)}$. Finally, the representations are passed through the sigmoid $\sigma(\cdot)$ nonlinearity.

Following the canonical autoencoder structure [21], we define a GNN encoder $\mathcal{E}(\cdot)$ and decoders $\mathcal{D}_1(\cdot), \mathcal{D}_2(\cdot)$. Incorporating the GCN layers, the encoder ($Z = \mathcal{E}(X, A)$) takes as input the molecule points and features X and uses the edges A in the message passing scheme defined in (1), resulting in a refined latent representation Z.

Z is then used as input to both decoders where one reconstructs the RNA atomic point cloud P' via Multilayer Perceptrons (MLP) and the other reconstructs the adjacency matrix A' via inner product.

$$P' = D_1(Z) = MLP(Z) \qquad A' = D_2(Z) = \sigma(ZZ^T) \tag{2}$$

To measure the Graph Autoencoder's reconstruction capabilities, we incorporate two methods of loss. First, we use Chamfer Distance (CD) as the loss between the ground truth atomic coordinates P and reconstructed coordinates P'. This loss is standard in point cloud reconstruction tasks [30] and is formally described as such,

$$\mathcal{L}_{CD} = \sum_{x\in P} \min_{y\in P'} ||x-y||_2^2 + \sum_{y\in P'} \min_{x\in P} ||x-y||_2^2 \tag{3}$$

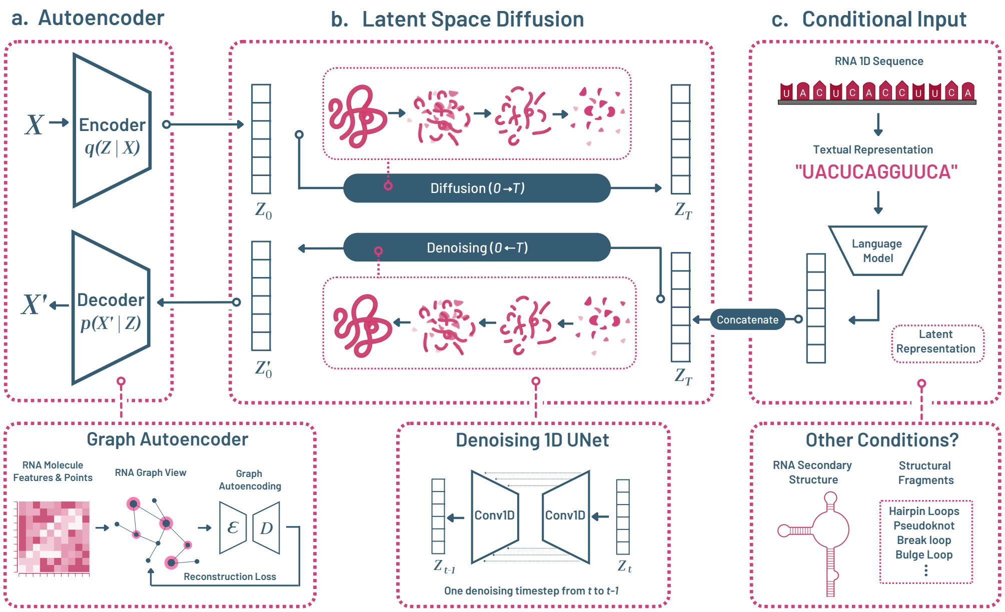


Figure 1: DiffRNAFold architecture visualization, inspired by [19]. (a) Graph Autoencoder to encode RNA molecule graphs into latent representations. (b) Latent Space Diffusion and Denoising layers to generate high quality latents. (c) Optional conditional input to guide the diffusion process.

Secondly, we use binary cross entropy (BCE) loss between the ground truth A edges and reconstructed edges A'. Both are summed up for the final loss of the GAE.

2.1.2 Denoising Diffusion Layers

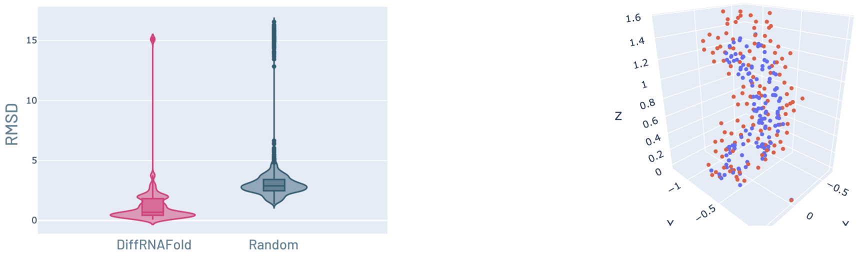
We utilize diffusion (see Figure 1b) on the latent representations of the RNA molecule from Section 2.1.1, to enable high quality latent vector generation (which can then be decoded into RNA molecules). In the forward process, the latent vector at timestep 0 ($\mathbf{z_0}$) is injected with noise over many iterations until at timestep T, $\mathbf{z_T}$ essentially represents only noise. $\mathbf{z_T}$ is then passed through the denoising layers where from each timestep t to $t-1$, the latent vector is denoised by means of a 1D UNet convolutional layer [20]. The model learns to reconstruct the original latent $\mathbf{z_0}'$, which can then be decoded into the original RNA molecule. During training, both the forward process of adding noise and the reverse process of denoising is utilized, and is optimized with a canonical loss function among the successful diffusion models. Using the reparameterization trick [14], it has been shown that predicting the original latent $\mathbf{z_0}$ is equivalent to predicting the source noise added at each timestep. Let $\epsilon_0 \sim \mathcal{N}(\epsilon; \mathbf{0}, \mathbf{I})$ be the added noise. We constructed a 1D UNet that predicts this noise, denoted using $\hat{\epsilon}_\theta(\mathbf{z_t}, t)$. Thus, according to [8, 16, 19], the loss at from the denoising diffusion layers can be generalized to matching the noise as such:

$$\mathcal{L}_{diff} = \mathbb{E}_{z,\epsilon \sim \mathcal{N}(0,1),t}\left[||\epsilon_0 - \epsilon_\theta(z_t, t)||_2^2\right] \tag{4}$$

Note that the forward process of injecting noise is not utilized during generation. Rather, sampling a vector of Gaussian noise and passing it through the denoising layers results in a high quality latent that is ready for decoding.

2.1.3 Language Model for Conditional Input

The graph autoencoder and the latent space diffusion model is already capable of generating RNA molecules. To guide the diffusion process, we utilized conditioning via concatenation of a representation of the condition and the random sample (see Figure 1c). Specifically, we conditioned on RNA sequences using pretrained embeddings from the RNABERT model [2] which uses a bidirectional transformer language model [5] on RNA linear sequences.



(a) (b)

Figure 2: (a) Distribution of pairwise euclidean distances between the centroid of each point cloud (diffrnafold vs. random) and the centroids of real RNA molecule points.(b) Sample reconstructed RNA point cloud (orange is generated | blue is real atomic points)

2.2 Data & Preprocessing

We obtained 2,500 molecules from the RNASolo database [1]. To overcome the large variance in RNA sizes (50-2,000 atoms), we selected the 240 RNA molecules with a size range of 100-140 atoms. The PDB files were then parsed into tensors using the coordinates and features, and each point cloud was then padded with zeros to the size of 140 atoms and normalized to fit on the unit sphere prior to a 85-5-10 (train, val, test) split. To construct our graph representation of each RNA molecule $\mathcal{G}(X, A)$, the atomic coordinates (P) along with basic molecular features were organized into the feature matrix $X \in \mathbb{R}^{n\times(3+f)}$, where $n = 140$ is the number of points and f is the number of molecular features. The adjacency matrix $A \in \mathbb{R}^{n\times n}$ represents the edges (bonds) between atoms and additional edges based on nearest-neighbor ($k = 5$) proximity. These molecular graphs were used as input.

3 Results

DiffRNAFold is a generative model that produces RNA-like structures, and thus it cannot be compared directly with other methods that predict RNA structure. To assess DiffRNAFold's generative capabilities, we designed an experiment to explore the 3D space occupied by DiffRNAFold's molecular point clouds, and their relationship to (a) real RNA molecules in our data-set, and (b) random molecular point clouds as baseline. To accomplish this, we sampled 100 RNA molecules from DiffRNAFold, retrieved 100 real RNA molecules from our dataset at random, and generated 100 random molecular point clouds as a baseline. We computed the distribution of pairwise Euclidean distances between the centroid of each DiffRNAFold point cloud to each real RNA point cloud (Figure 2a; pink). We repeated the computation for the distance between Random and real point cloud centroids (blue).

DiffRNAFold-to-Real distances (in pink) Euclidean distance values (median 0.673) were significantly closer to zero compared to the Random-to-Real centroid Euclidean distance (median 2.900; Rank-sum test p-value: 2.06e-16). Thus, DiffRNAFold's molecular point clouds are indeed much similar in 3D space to real RNA molecules—a crucial first step in determining RNA molecule validity. Additionally, while this first experiment indicates the overall 3D space in which DiffRNAFold generated molecules lie in, we also provide a small proof of concept. Figure 2b shows the atomistic point cloud of a real RNA molecule in our dataset (in orange) and the autoencoder reconstructed point cloud by DiffRNAFold (in blue). The remarkable visual similarity suggested that even with a small dataset, sufficient properties of RNA structures could be obtained.

4 Conclusion & Work in Progress

With DiffRNAFold, we have proposed the first latent space diffusion model for the generation of novel RNA tertiary structures. However, in parallel work, a latent diffusion model for other *non-RNA* molecules was proposed [29] further validating and motivating our strategy. Our preliminary results indicate a good starting point, but also point to exciting new directions. On the algorithmic side, we plan to develop a roto-translational equivariant graph autoencoder using [7] to obtain better latent representations. Secondly, we plan to incorporate a hierarchical diffusion method as many RNA 3D structures can be directly informed by their 2D motifs (hairpin loops, pseudoknots, etc.). Our work could also be improved through larger data collections, perhaps incorporating training on accurately simulated RNA samples, or breaking up larger RNA molecules into smaller functional domains. Lastly, we plan to incorporate more rigorous analysis of DiffRNAFold's chemical validity, especially in regards to the conditional generation. Overall, with this work, we hope to emphasize the importance of research on designing RNA molecules, and promoting its application to development of novel drug therapy.

References

[1] Bartosz Adamczyk, Maciej Antczak, and Marta Szachniuk. Rnasolo: a repository of cleaned pdb-derived rna 3d structures. *Bioinformatics*, 38(14):3668–3670, 2022.

[2] 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.

[3] Namrata Anand and Tudor Achim. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv preprint arXiv:2205.15019*, 2022.

[4] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[7] Mario Geiger, Tess Smidt, Alby M., Benjamin Kurt Miller, Wouter Boomsma, Bradley Dice, Kostiantyn Lapchevskyi, Maurice Weiler, Michał Tyszkiewicz, Simon Batzner, Dylan Madisetti, Martin Uhrin, Jes Frellsen, Nuri Jung, Sophia Sanborn, Mingjian Wen, Josh Rackers, Marcel Rød, and Michael Bailey. Euclidean neural networks: e3nn, April 2022.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[9] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. 2023.

[10] Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. *arXiv preprint arXiv:2206.01729*, 2022.

[11] Bowen Jing, Ezra Erives, Peter Pao-Huang, Gabriele Corso, Bonnie Berger, and Tommi Jaakkola. Eigenfold: Generative protein structure prediction with diffusion models. *arXiv preprint arXiv:2304.02198*, 2023.

[12] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[13] Diederik P Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. On density estimation with diffusion models. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[15] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[16] Calvin Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022.

[17] Xiujuan Ou, Yi Zhang, Yiduo Xiong, and Yi Xiao. Advances in rna 3d structure prediction. *Journal of Chemical Information and Modeling*, 62(23):5862–5874, 2022.

[18] Robin Pearce, Gilbert S Omenn, and Yang Zhang. De novo rna tertiary structure prediction at atomic resolution using geometric potentials from deep learning. *bioRxiv*, pages 2022–05, 2022.

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[21] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[22] Tao Shen, Zhihang Hu, Zhangzhi Peng, Jiayang Chen, Peng Xiong, Liang Hong, Liangzhen Zheng, Yixuan Wang, Irwin King, Sheng Wang, et al. E2efold-3d: End-to-end deep learning method for accurate de novo rna 3d structure prediction. *arXiv preprint arXiv:2207.01586*, 2022.

[23] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[24] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[25] Raphael JL Townshend, Stephan Eismann, Andrew M Watkins, Ramya Rangan, Maria Karelina, Rhiju Das, and Ron O Dror. Geometric deep learning of rna structure. *Science*, 373(6558):1047–1051, 2021.

[26] F. Wang, T. Zuroske, and J. K. Watts. RNA therapeutics on the rise. *Nat Rev Drug Discov*, 19(7):441–442, Jul 2020.

[27] 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*, pages 1–3, 2023.

[28] Kevin E Wu, Kevin K Yang, Rianne van den Berg, James Y Zou, Alex X Lu, and Ava P Amini. Protein structure generation via folding diffusion. *arXiv preprint arXiv:2209.15611*, 2022.

[29] Minkai Xu, Alexander Powers, Ron Dror, Stefano Ermon, and Jure Leskovec. Geometric latent diffusion models for 3d molecule generation. *arXiv preprint arXiv:2305.01140*, 2023.

[30] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. *arXiv preprint arXiv:2210.06978*, 2022.