RiboGen：RNA 序列和结构协同生成的等变多流方法

Dana Rubin
MIT CSAIL
MIT Media Lab, Molecular Machines
danaru@mit.edu

Allan dos Santos Costa
Center for Bits and Atoms
MIT Media Lab, Molecular Machines
allanc@mit.edu

Manvitha Ponnapati
Center for Bits and Atoms
MIT Media Lab, Molecular Machines

Joseph Jacobson
Center for Bits and Atoms
MIT Media Lab, Molecular Machines

Abstract

核糖核酸（RNA）在生物系统中扮演着基本角色，从携带遗传信息到执行酶功能。理解和设计 RNA 可以实现新的治疗应用和生物技术革新。为了提升 RNA 的设计，在本文中我们介绍了 RiboGen，这是第一个能够同时生成 RNA 序列和全原子 3D 结构的深度学习模型。RiboGen 利用了标准流匹配与离散流匹配在多模态数据表示中的应用。RiboGen 基于欧几里得等变神经网络来高效处理和学习三维几何形状。我们的实验表明，RiboGen 可以高效生成化学上合理且自洽的 RNA 样本，这表明序列和结构的同时生成是建模 RNA 的一种有竞争力的方法。

1 介绍

核糖核酸（RNA）是一种位于现代生物学和生命起源交汇处的基本生物分子。RNA 证明是一种多功能的分子，在其复杂的三维结构 (Fire et al., 1998) 中，它在信使功能 (Crick, 1970)、催化功能 (Altman & Guerrier-Takada, 1983)、调控以及各种生物过程中扮演着关键角色。虽然传统的计算方法在解码 RNA 结构和促进 RNA 设计方面存在局限性，但深度学习作为准确预测 RNA 结构、增强 RNA 工程并解锁其功能性作用新见解的强大方法应运而生。现有的深度学习模型通常独立地从序列预测 RNA 结构或为目标结构设计序列。然而，在基于深度学习的 RNA 建模中，同时生成序列和结构的能力仍然是一个很大程度上未被探索的研究领域。这种联合生成能力可以以新颖的方式探索序列-结构景观。为了填补这一空白，本文介绍了 RiboGen 用于 RNA 的全原子结构与序列的联合生成。RiboGen 基于 Multiflow(Campbell et al., 2024) 模型，该模型利用 Flow Matching(Lipman et al., 2022; Liu et al., 2022) 和 Discrete Flow(Gat et al., 2024; Campbell et al., 2024) 进行生成。我们训练了一个大型模型并评估了其化学有效性和自洽性。我们的结果展示了 RiboGen 在生成方面的功能，并强调联合生成模型是 RNA 建模的有前景的研究方向。

1.1 相关工作

之前在深度学习用于 RNA 设计方面的工作已经在计算方法预测 RNA 结构 (Shen et al., 2024; Abramson et al., 2024)、RNA-RNA 相互作用以及设计新型 RNA 序列方面取得了显著进展。最近关于 RNA 生成建模的进展集中在通过去噪扩散概率模型（DDPM）(Ho et al.,

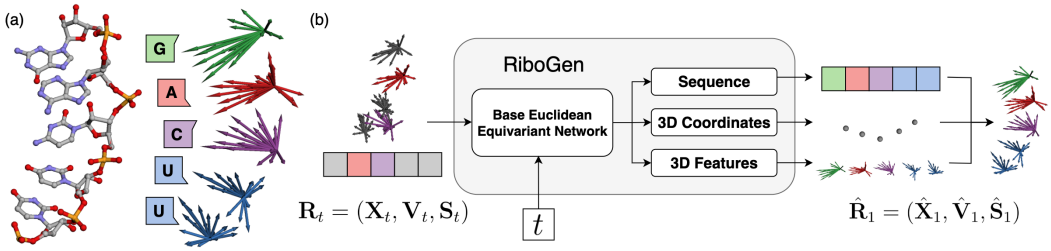

图 1: RNA **序列和结构协同生成** (a) 传统的分子结构展示了带有原子和键的核苷酸。右侧演示了每个核苷酸（G、A、C、U）如何作为离散序列元素（彩色方框）以及与其相关的三维点云表示（彩色方向特征）来展示，这些特征围绕着 $C3'$ 原子中心。(b) RiboGen 模型架构：该模型接受带有噪声的序列和几何特征 $\boldsymbol{R}_t$ 输入及一个时间参数 t，通过基础网络处理它们，并同时预测三个组成部分：RNA 序列、中心坐标以及三维特征。这些组件结合生成最终的 RNA 结构预测 $\hat{\boldsymbol{R}}_1$。

2020) 或流匹配 (Lipman et al., 2022) 来生成序列和结构。MMDiff(Morehead et al., 2023a) 使用离散 DDPM 共同生成 RNA、DNA 和蛋白质的序列和结构。我们的方法则采用流匹配及其离散变体 (Campbell et al., 2024)。RNA-FrameFlow (Anand et al., 2024) 通过刚体框架表示 RNA，并使用流匹配生成 3D 骨架，利用逆折叠模型 gRNAde(Joshi et al., 2023) 获得序列。相比之下，尽管类似地使用流匹配进行 3D 生成，我们的方法还建模了 RNA 生成的离散序列成分。RNAFlowNori & Jin (2024) 使用基于蛋白质结构和序列条件化的 GNN 来生成 RNA 序列，然后通过 RoseTTAFold2NABaek (2024) 预测骨架结构；该方法另外将蛋白质结构作为输入进行条件化。我们的方法则专注于孤立的 RNA，学习无条件直接序列-结构生成。最近将 Multiflow(Campbell et al., 2024) 框架应用于蛋白质序列-结构设计展示了联合生成的强大能力。我们的方法基于这些来自蛋白质设计的见解，并通过使 RNA 序列和全原子结构的联合生成成为可能来将其适应到 RNA 设计领域。

2 方法

2.1 RNA 表示方法

我们用一个序列和一组三维几何特征的气体 $\boldsymbol{R} = (\boldsymbol{S}, \boldsymbol{X}, \boldsymbol{V})$（图 1.a）来表示一个 RNA 分子，其中：

- $\boldsymbol{S} \in \mathcal{S}^N$ 是长度为 N 的序列，由标准核苷酸 $\mathcal{S} = \{A, C, G, U\}$ 组成。
- $\boldsymbol{X} \in \mathbb{R}^{N\times 3}$ 包含每个核苷酸的 C3' 原子的 3D 坐标，选定为**参考中心**。
- $\boldsymbol{V} \in \mathbb{R}^{N\times 24\times 3}$ 是编码每个核苷酸相对于其 C3' 中心最多 24 个重原子相对位置的几何特征，在规范排序中。这种表示方法涵盖了糖-磷酸主链原子和碱基原子，允许完整重建 RNA 结构。具有少于 24 个重原子的核苷酸对应的 $\boldsymbol{V}$ 通道用零填充。

生成三个组件后，预测向量 $\boldsymbol{V}$ 添加到预测中心 $\boldsymbol{X}$，并利用序列 $\boldsymbol{S}$ 对核苷酸和原子类型进行标注，以完整重建三维原子坐标。该表示同时编码了每个核苷酸的化学身份和几何形状，同时保持旋转和平移等方差性，这对于通过欧几里得等变神经网络进行下游学习至关重要。

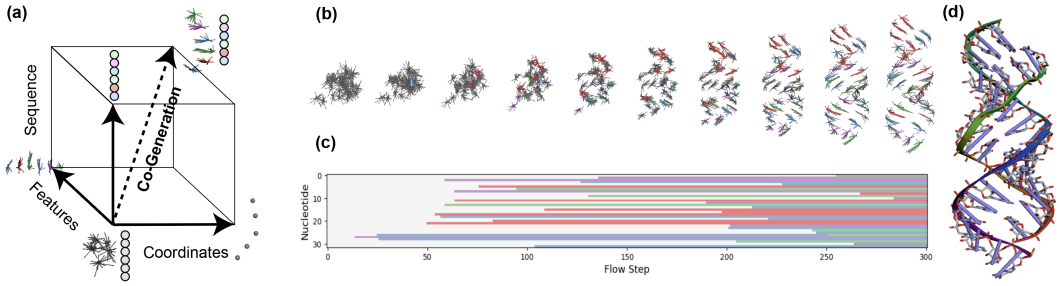

图 2: **多流用于 RNA 序列、主链和原子结构** (a) 我们多流方法的示意图，展示了三个维度-序列、坐标和特征。(b) 多个时间步长上的 RNA 结构生成可视化。(c) 模型中用于序列预测的离散流匹配可视化，每种颜色代表不同的核苷酸。(d) 最终产品，一个完整的生成 RNA 分子。

2.2 流匹配

为了建模 3D RNA 坐标 $\boldsymbol{X}$ 和特征 $\boldsymbol{V}$ 的分布，我们使用流匹配 (Lipman et al., 2022; Liu et al., 2022; Albergo et al., 2023)。流匹配通过学习一个条件速度场 $\hat{v}_t^\theta(\boldsymbol{X}_t) \approx v_t(\boldsymbol{X}_t|\boldsymbol{X}_1)$ 将来自先验分布 $\boldsymbol{X}_0 \sim \rho_0 = \mathcal{N}$ 的样本变换为目标数据分布 $\boldsymbol{X}_1 \sim \rho_1 = \rho_D$，从而参数化时间 t 上的条件概率路径 $\rho_t(\boldsymbol{X}_t|\boldsymbol{X}_1)$。为了学习这种传输，我们使用流匹配的标准形式通过线性插值及其相关速度获得 $\boldsymbol{X}_1$ 的噪声版本：

$$\boldsymbol{X}_t = (1-t)\boldsymbol{X}_0 + t\boldsymbol{X}_1 \tag{1}$$

$$v_t(\boldsymbol{X}_t|\boldsymbol{X}_1) = \boldsymbol{X}_1 - \boldsymbol{X}_0 \tag{2}$$

我们构建模型以从其噪声对应物 $\boldsymbol{X}_t$ 中重构目标 $\hat{\boldsymbol{X}}^\theta_{1|\boldsymbol{X}_t} \approx \boldsymbol{X}_1$。我们遵循重新参数化 (Jing et al., 2024; Pooladian et al., 2023) 并通过以下方式获得学习到的条件速度：

$$v_t^\theta(\boldsymbol{X}_t) = \frac{1}{(1-t)}\left(\hat{\boldsymbol{X}}^\theta_{1|\boldsymbol{X}_t} - \boldsymbol{X}_t\right) \approx v_t(\boldsymbol{X}_t|\boldsymbol{X}_1) \tag{3}$$

然后，我们通过对 $\boldsymbol{X}_1 = \boldsymbol{X}_0 + \int_0^1 v_t^\theta(\boldsymbol{X}_t)dt$ 进行积分来采样我们的学习模型，其中 $\boldsymbol{X}_0 \sim \mathcal{N}$。

2.3 离散流匹配

虽然流动匹配的标准形式对连续数据有效，但它不适合分类领域。因此，在建模 RNA 序列时，我们采用离散流动匹配的扩展框架 (Gat et al., 2024; Campbell et al., 2024)。在此设置下，序列数据 $\boldsymbol{S} \in \mathcal{S}^N$ 是基于词汇表 $\mathcal{S}$ 描述的。我们通过描述该分类空间上的概率向量的速度场来参数化离散流：

$$\boldsymbol{S}_t \sim \text{Cat}((1-t)\delta_{\boldsymbol{S}_0} + t\delta_{\boldsymbol{S}_1}) \tag{4}$$

$$v_t(\boldsymbol{S}_t|\boldsymbol{S}_1) = \delta_{\boldsymbol{S}_1} - \delta_{\boldsymbol{S}_0} \tag{5}$$

其中，$\delta_{\boldsymbol{S}\in\mathbb{R}^{N\times|\mathcal{X}|}}$ 是 $\boldsymbol{S}$ 的狄拉克 delta 表示，$\text{Cat}(\cdot)$ 表示分类分布。我们学习一个模型来预测概率向量 $p^\theta_{1|\boldsymbol{S}_t} \approx \delta_{\boldsymbol{S}_1}$。通过类似于公式 3 的重新参数化，我们得到近似条件速度如下：

$$v_t^\theta(\boldsymbol{S}_t) = \frac{1}{(1-t)}\left(\hat{p}^\theta_{1|\boldsymbol{S}_t} - \delta_{\boldsymbol{S}_t}\right) \approx v_t(\boldsymbol{S}_t|\boldsymbol{S}_1) \tag{6}$$

2.4 多流

为了生成完整的 RNA 表示，我们使用 Multiflow(Campbell et al., 2024) 并训练一个神经网络来学习多模态速度场 $d\boldsymbol{R}_t = (d\boldsymbol{S}_t, d\boldsymbol{X}_t, d\boldsymbol{V}_t)$，给定联合噪声数据 $\boldsymbol{R}_t$ 和时间 t（图 1.b）。我们采用标准的流匹配用于 $\boldsymbol{X}$ 和 $\boldsymbol{V}$ 及其离散对应部分用于 $\boldsymbol{S}$。这种分解使模型能够分别捕获序列、主链和原子位置的分布，允许无条件生成或基于特定结构或序列约束的有条件生成（图 2.a），例如结构预测或逆折叠。

2.5 架构与训练

我们使用欧几里得等变神经网络 (Geiger & Smidt, 2022) 来处理我们的 RNA 表示。为了处理 $\boldsymbol{R}$ 的不同模态，我们的模型由一个基础网络组成，该网络输入到每个数据组件的 3 个头部：序列、坐标和三维特征。序列头预测概率向量 $\hat{p}_{\boldsymbol{R}_1} \in \mathbb{R}^{N \times |\mathcal{S}|}$，而坐标和特征头则预测等变变量 $\hat{\boldsymbol{X}}_1$ 和 $\hat{\boldsymbol{V}}_1$。我们训练模型以重构原始结构 $(\boldsymbol{X}, \boldsymbol{V})$ 和序列 $\boldsymbol{S}$：

$$\mathcal{L} = \mathcal{L}_{\text{struct}} + \mathcal{L}_{\text{seq}} = \text{Mean}(\|V(\boldsymbol{X}_1, \boldsymbol{V}_1) - V(\hat{\boldsymbol{X}}_1, \hat{\boldsymbol{V}}_1)\|^2) + \text{CrossEntropy}(\hat{p}_{\boldsymbol{S}_1}, \delta_{\boldsymbol{S}_1}) \quad (7)$$

其中 $V(\boldsymbol{X}, \boldsymbol{V}) \in \mathbb{R}^{N_A \times N_A \times 3}$ 是系统中每个原子之间的 3D 向量图（大小为 N_A）。

3 结果

我们利用 RNASolo 数据集 (Adamczyk et al., 2022)，该数据集由蛋白质数据库 (PDB)(Berman et al., 2000) 提取的单个 RNA 结构组成，来训练我们的模型。

遵循 (Anand et al., 2024)，我们将完整数据集过滤至分辨率 < 4Å，序列长度在 40 到 150 之间，总数据集大小为 6090 个数据点。

该数据集表现出显著的长度不平衡，这最初导致了有偏差的模型性能。为了解决这个问题，我们实施了一种长度平衡采样，以确保 RNA 序列在训练中的均匀表示，如附录 A.1 中所述。

我们使用批量大小为 64，在 4 个 GPU 上训练 120k 步。

我们模型的流程在 300 个时间步长上进行训练。

为了评估我们的模型，我们遵循 (Anand et al., 2024)，并从序列长度 40-150 的每个序列长度中采样 50 个 RNA 结构，步长为 10。

3.1 化学有效性

为了评估我们生成结构的化学有效性，我们分析了定义 RNA 主链和碱基构象的关键几何参数。使用 MDAnalysis(Gowers et al., 2016) (Michaud-Agrawal et al., 2011) 我们计算了所有二面角以及核糖皱褶的伪角度。图 3 显示了训练数据（代表实验确定结构）和 50 个随机选择的 RiboGen 结构中二面角的分布，每个序列长度各选 5 个。结果表明，对于大多数角度，RiboGen 生成的结构捕捉到了二面角分布以及训练集的一般趋势，尽管存在一些差异。我们的 RiboGen 结构在 Alpha 角上显示出更宽泛的分布，这表明在这个特定扭转处与实验数据有轻微偏离。总体而言，这些结果显示 RiboGen 成功学习了 RNA 分子的几何约束。

3.2 自洽性

为了评估我们生成的 RNA 的质量和生物学合理性，我们采用了一种自我一致性验证过程。对于每个生成的 RNA 分子，我们提取其序列并使用 Boltz-1(Wohlwend et al., 2024) 获得

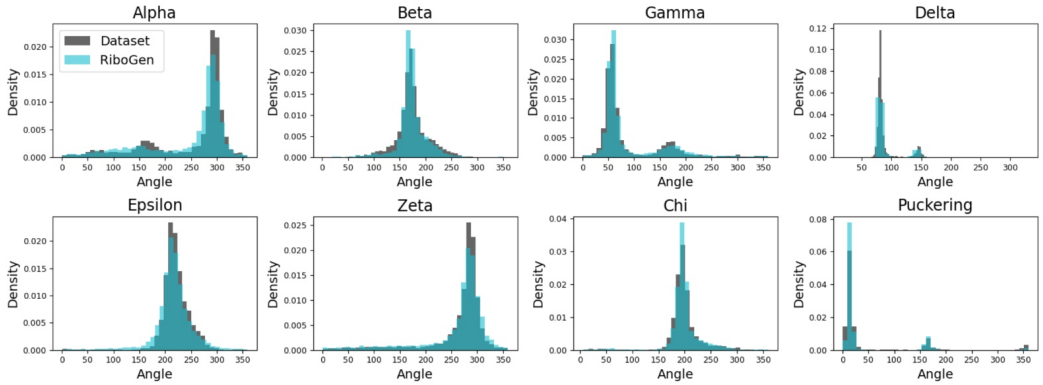

图 3: **核糖生成化学分析**训练数据集与 RiboGen 生成的 50 个随机样本（每个长度 5 个）在所有长度上的关键 RNA 几何参数分布比较。分析的参数包括 alpha、beta、gamma、chi 二面角以及核糖皱褶相位，这些都是 RNA 主链和化学有效性的强指标。

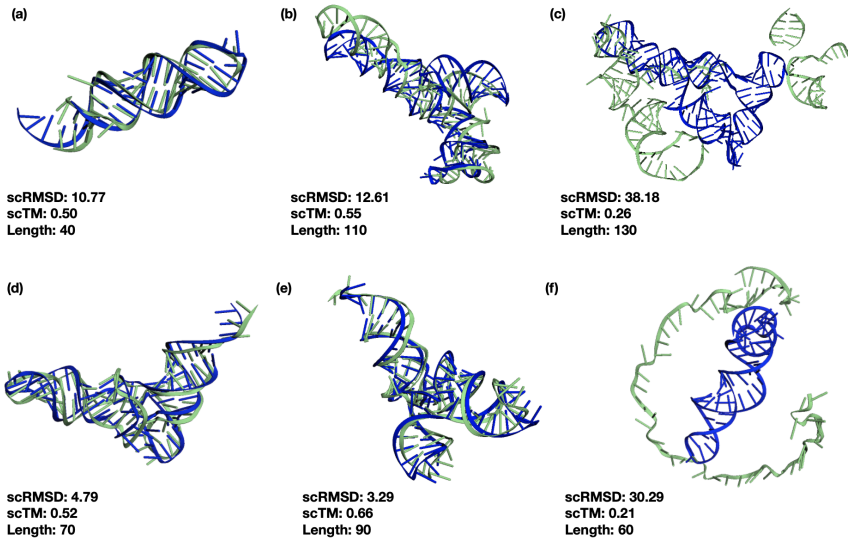

图 4: RiboGen **联合序列结构生成与** Boltz **结构对齐的自洽可视化** RiboGen 生成的 RNA 结构（绿色）与从 RiboGen 相应共生成序列中得出的 Boltz 结构预测（蓝色）对齐。不同序列长度的六个示例展示了不同程度的结构一致。值得注意的是，在某些情况下 (c, f)，RiboGen 生成了断裂或未折叠的结构，这表明在长序列或结构复杂的序列采样过程中存在失败模式。

参考结构。我们通过两种互补的度量标准来量化我们的生成结构与 Boltz 结构之间的相似性：均方根偏差（RMSD）和模板建模得分（TM-score）。我们在每种序列长度的所有 50 个样本中计算了这些指标。根据 (Anand et al., 2024)，在图 5(a) 和 (b)，我们报告了每个长度下表现最好的 10 个样本的结果。我们的预测 scTM 分数在 5(b) 中显示出比 Anand et al. (2024) 更低的方差，表明不同长度之间有更好的一致性和泛化能力。在图 5(c) 中，我们观察到 RiboGen 在大多数超过 70 个核苷酸的序列长度上实现了比 RNA-FrameFlow 更高的中位 TM 分数。这可能突显了协同生成的优势，特别是对于较长的 RNA。

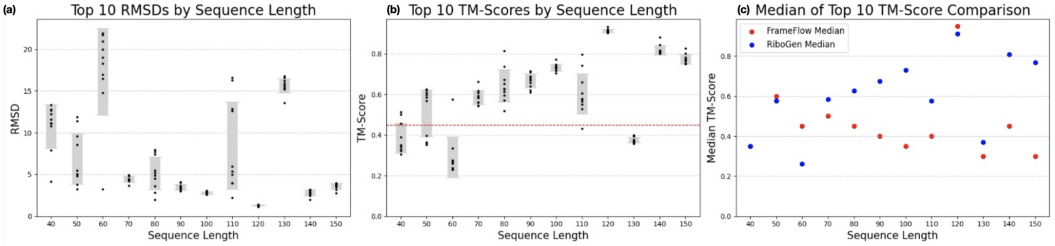

图 5: **自洽性评价**：(a) 我们生成的结构与 Boltz-1 预测结果之间的均方根偏差 (RMSD) 和 (b) 结构分数 (TM-score)，针对不同长度的序列（40-150 个核苷酸），展示每个长度前 10 名生成的结构。结构分数范围从 0 到 1，数值越高表示结构一致性越好，而较低的均方根偏差值则表明更好的结构相似性。(c) 前 10 名生成结构的中位数结构分数：该图比较了 RiboGen 和 FrameFlow 在各种 RNA 序列长度上的中位数，并展示了 RiboGen 对于 70-150 核苷酸之间的 RNA 序列实现了更高的结构分数，除了 120 核苷酸序列具有相似的中位数。FrameFlow 表现出了对于较短序列相当的性能，但随着序列长度增加显示出结构准确性的下降。

3.3 结构评估

为了评估 RiboGen 在生成有效 RNA 结构序列对方面的性能，我们使用了 (Anand et al., 2024) 提出的评估套件中的指标。虽然 RNA-FrameFlow 仅专注于结构生成，但 RiboGen 同时生成序列及其对应的结构。因此，我们没有使用 gRNAde(Joshi et al., 2025) 序列和相应的 RhoFold(Shen et al., 2024) 结构来计算 TM 得分，而是使用 Boltz-1 折叠共同生成的序列，并在与生成的结构对齐后计算 TM 得分。虽然 RNA-FrameFlow 在每次为每个骨架调用 gRNAde 八次时采用额外的 8 次逆向折叠过程，但 RiboGen 在一个采样过程中一次性共生成了序列和结构。这突显了 RiboGen 具有简化且更高效的 RNA 设计流程的潜力，避免了昂贵的事后序列推断需求。根据 (Anand et al., 2024)，将带有 TM-score ≥ 0.45 的样本视为有效。为了评估多样性，我们测量了在有效样本中唯一的 qTM 簇的数量，并通过总的有效样本数量进行归一化。尽管 RiboGen 同时采样 RNA 序列和结构，但在主链有效性及多样性方面其表现与 RNA-FrameFlow 相当。我们的结果显示出了具有竞争力的指标和高效的抽样，验证了 RiboGen 作为联合生成 RNA 结构序列的一个有前景的基础模型。

Model	Sampling Steps N_T	% Validity ↑	Diversity ↑	Time (s) ↓
RiboGen	100	27.17	0.604	1.18
	200	32.17	0.553	4.50
	300	34.17	0.585	9.06
RNA-FrameFlow *	10	16.7	0.62	–
	50	41.0	0.61	4.74
	100	20.0	0.61	–
MMDiff *	100	0	–	27.30

表 1: **无条件 RNA 结构生成模型的性能比较**此表展示了 RiboGen 在不同流采样时间步长（N_T）上的性能，以及其他模型的对比。指标包括结构有效性、通过 qTM 聚类衡量的多样性以及每生成一次所需的计算成本（秒）。* RNA-FrameFlow 和 MMDiff(Morehead et al., 2023b) 方法的结果来自 (Anand et al., 2024)。

4 结论

在这篇论文中，我们介绍了 RiboGen，这是首个通过学习单一多模态流场来同时生成 RNA 序列及其相应的全原子三维结构的生成模型。我们的方法利用 Flow Matching 处理连续结构组件，并在 Multiflow 框架内使用离散 Flow Matching 进行序列生成。我们证明了 RiboGen 能够生成化学上合理的 RNA 结构，这从二面角和核糖皱褶等关键几何参数的分布中得到了证实。我们的模型在自一致性评估（scTM 分数）方面优于以前的方法，并且适用于一系列不同的序列长度，特别是较长的 RNAs。早期结果显示，RiboGen 的生成提供了一个竞争性和高效的 RNA 设计工作流程，表明序列-结构共生成是 RNA 建模的一个强有力方法。随着 RNA 设计领域在治疗和生物技术应用中的重要性不断增加，我们相信像 RiboGen 这样的生成模型将成为探索和工程化 RNA 的重要工具。

致谢

此项研究得到了十一基金会、比特与原子中心以及麻省理工学院媒体实验室联盟的支持。他们的支持对于这项工作的开展至关重要。

参考文献

J. Abramson, J. Adler, J. Dunger, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. Nature, 2024. doi: https://doi.org/10.1038/s41586-024-07487-w.

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

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. arXiv preprint arXiv:2303.08797, 2023.

Sidney Altman and Cecile Guerrier-Takada. The rna moiety of ribonuclease p is the catalytic subunit of the enzyme. Cell, 35(3):849–857, 1983. doi: 10.1016/0092-8674(83)90117-4.

Rishabh Anand, Chaitanya K. Joshi, Alex Morehead, Arian R. Jamasb, Charles Harris, Simon V. Mathis, Kieran Didi, Bryan Hooi, and Pietro Liò. Rna-frameflow: Flow matching for de novo 3d rna backbone design. arXiv preprint, 2024. URL https://doi.org/10.48550/arXiv.2406.13839.

Minkyung Baek. Towards the prediction of general biomolecular interactions with ai. Nature Methods, 21:1382–1383, 2024. URL https://www.nature.com/articles/s41592.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. Nucleic acids research, 28(1):235–242, 2000.

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. arXiv preprint arXiv:2402.04997, 2024.

Francis Crick. Central dogma of molecular biology. Nature, 227:561–563, 1970. doi: 10.1038/227561a0.

Andrew Fire, SiQun Xu, Mary K. Montgomery, Steven A. Kostas, Samuel E. Driver, and Craig C. Mello. Potent and specific genetic interference by double-stranded rna in caenorhabditis elegans. Nature, 391(6669):806–811, 1998. doi: 10.1038/35888.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T. Q. Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching, 2024. URL https://arxiv.org/abs/2407.15595.

Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks. arXiv preprint arXiv:2207.09453, 2022. doi: 10.48550/arXiv.2207.09453. URL https://doi.org/10.48550/arXiv.2207.09453. draft.

R. J. Gowers, M. Linke, J. Barnoud, T. J. E. Reddy, M. N. Melo, S. L. Seyler, D. L. Dotson, J. Domanski, S. Buchoux, I. M. Kenney, and O. Beckstein. MDAnalysis: A Python package for the rapid analysis of molecular dynamics simulations. In S. Benthall and S. Rostrup (eds.), Proceedings of the 15th Python in Science Conference, pp. 98–105, Austin, TX, 2016. SciPy. doi: 10.25080/Majora-629e541a-00e.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. arXiv preprint arXiv:2006.11239, 2020. doi: 10.48550/arXiv.2006.11239. URL https://doi.org/10.48550/arXiv.2006.11239.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. arXiv preprint arXiv:2402.04845, 2024.

Chaitanya K. Joshi, Arian R. Jamasb, Ramon Viñas, Charles Harris, Simon V. Mathis, Alex Morehead, Rishabh Anand, and Pietro Liò. gRNAde: Geometric Deep Learning for 3D RNA Inverse Design. arXiv preprint arXiv:2305.14749, 2023. URL https://doi.org/10.48550/arXiv.2305.14749.

Chaitanya K. Joshi, Arian R. Jamasb, Ramon Viñas, Charles Harris, Simon V. Mathis, Alex Morehead, Rishabh Anand, and Pietro Liò. grnade: Geometric deep learning for 3d rna inverse design, 2025. URL https://arxiv.org/abs/2305.14749.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. arXiv preprint arXiv:2210.02747, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003, 2022.

N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, and O. Beckstein. MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. J. Comput. Chem., 32:2319–2327, 2011. doi: 10.1002/jcc.21787.

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, 2023a. URL https://arxiv.org/abs/2401.06151.

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. arXiv preprint arXiv:2401.06151, 2023b. URL https://doi.org/10.48550/arXiv.2401.06151. Presented at NeurIPS 2023 MLSB Workshop.

Divya Nori and Wengong Jin. Rnaflow: Rna structure & sequence design via inverse folding-based flow matching. arXiv preprint arXiv:2405.18768, 2024.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings, 2023. URL https://arxiv.org/abs/2304.14772.

Tian Shen, Zhen Hu, Shuxin Sun, et al. Accurate rna 3d structure prediction using a language model-based deep learning approach. Nature Methods, 21:2287–2298, 2024. doi: 10.1038/s41592-024-02487-0. URL https://doi.org/10.1038/s41592-024-02487-0.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, and Regina Barzilay. Boltz-1: Democratizing biomolecular interaction modeling. bioRxiv, 2024. doi: 10.1101/2024.11.19.624167. URL https://doi.org/10.1101/2024.11.19.624167. Preprint.

A 附录

A.1 数据分布与数据平衡

在 RNA 序列分析中，某些类型的 RNA（如 tRNA、rRNA）在不同的数据集中过度表示。如图 6 所示，我们的原始数据集表现出显著的长度不平衡，在特定长度范围（70-79 和 120-129 核苷酸）内有明显的峰值。这种不平衡导致模型性能偏差，其中预测精度在过度表示的长度范围内显著较高，但在较少表示的长度范围内则较差。为了解决这个问题，我们在数据集类中实现了一个长度平衡抽样方法，在训练过程中确保在整个长度范围内均匀表示 RNA 序列而不改变原始数据集。我们的算法将 RNA 数据划分为长度桶，每个桶的范围是 10；40-49、50-59 等。（最后一个桶除外，它包含 140-150）在训练期间，随机选择一个长度桶，并从该桶中随机抽取一个数据点。这种方法动态地平衡了训练过程中的数据集，允许模型查看所有可用的数据，同时防止过度表示的长度支配训练过程。这种平衡技术显著提高了模型在整个长度范围内的性能，特别是在先前较少表示的序列上。

Algorithm 1 长度平衡的 RNA 序列采样

Require: 数据集 D 包含长度为 $L \in [40, 150]$ 的 RNA 序列
Ensure: 均匀采样所有长度范围的序列

预处理：
    将序列分组到桶 B 中，按长度范围（40–49, 50–59, ..., 140–150）
for each sequence $s \in D$ do
    确定长度 l 的 s
    将 s 分配到桶中 $B[\lfloor l/10 \rfloor \times 10]$
end for
采样
从可用的桶中均匀随机选择目标长度范围 t
返回一个随机选中的序列 $B[t]$

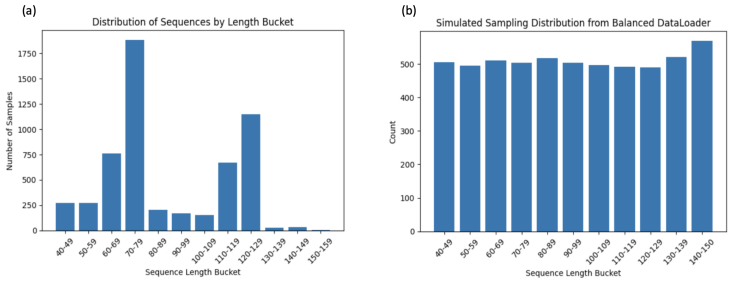

图 6: **训练数据集中 RNA 序列长度的分布，在平衡前和平衡后，按桶划分：** (a) 训练数据集中 RNA 序列的原始分布，按每 10 个核苷酸划分的长度区间分类。该分布显示出显著的不平衡，70-79 核苷酸和 120-129 核苷酸区间有明显的峰值，这可能对应于过度代表的 tRNA 和 rRNA 类别。相比之下，在 80-109 范围内的序列以及超过 130 个核苷酸的序列数量显著不足，某些区间样本数少于 200。(b) 我们在训练分布上实现的平衡采样方法使得所有长度区间的采样均匀，每个区间大约有 500 个序列用于训练。这种均匀分布确保了模型在整个 40 至 150 核苷酸长度范围内接收到了等量的 RNA 序列暴露。