中国科学技术大学

本科毕业论文

基于扩散语言模型的分子生成方法研究

作者姓名：	武宇星
学　　号：	PB21020651
专　　业：	计算机科学与技术
导师姓名：	何向南 教授
完成时间：	2025 年 5 月 12 日

何向南

摘 要

分子生成是药物发现和材料科学等领域的关键问题，旨在通过计算方法设计特定性质的新型分子。尽管现有方法取得进展，但在处理非自回归的复杂语法分子表示（SELFIES）、有效利用预训练模型知识及平衡生成分子的有效性、新颖性和多样性方面仍存挑战。针对这些问题，本文提出了一种基于扩散语言模型的 SELFIES 分子生成方法（DMLM）。该方法将离散扩散模型的迭代细化能力与预训练掩码语言模型 MOLGEN 的强大上下文理解能力相结合，借鉴 DiffusionBERT 在文本生成领域的成功经验，采用带离散吸收态的正向扩散过程、感知符号信息量的纺锤形噪声调度及时间无关解码策略进行反向去噪。实验结果表明：在 MOSES 数据集上，DMLM 在生成分子的有效性、唯一性（@10k）、新颖性及分子稳定性方面均达到 100%，表现达 SOTA 水平，验证了其生成分子的有效性。然而，在结构更复杂的 Natural Product 数据集上，DMLM 在化学特性分布相似性指标上表现不佳，显示了其处理高度多样和复杂化学结构时的局限性。本研究为结合预训练语言模型与离散扩散模型进行高质量 SELFIES 分子生成提供了新思路。

关键词：分子生成；扩散模型；掩码语言模型；深度学习

ABSTRACT

Molecule generation is a pivotal issue in interdisciplinary fields such as drug discovery and materials science, aiming to design novel molecules with specific properties computationally. Despite advancements, existing methods still face challenges in handling molecular representations with complex grammars like SELFIES, effectively leveraging the prior knowledge of large-scale pre-trained models, and balancing the chemical validity, structural novelty, and diversity of generated molecules. To address these issues, this paper proposes a Diffusion Molecular Language Model (DMLM）for SELFIES molecule generation. The core of this method lies in combining the iterative refinement capability of discrete diffusion models with the strong contextual understanding of MOLGEN, a pre-trained masked language model optimized for SELFIES representation. Drawing insights from DiffusionBERT, the DMLM employs a forward diffusion process with a discrete absorbing state (`[MASK]` token), a spindle noise schedule that perceives and differentially handles the information content of SELFIES tokens, and a time-agnostic decoding strategy for reverse denoising, which does not require explicit timestep input. Experimental validation on the public benchmark dataset MOSES and the more structurally complex Natural Product dataset shows that: on the MOSES dataset, DMLM achieves 100% in validity, uniqueness, and novelty, as well as 100% in molecular stability, reaching state-of-the-art performance and demonstrating its effectiveness in generating drug-like molecules. However, in experiments on the Natural Product dataset, DMLM performed poorly on metrics such as the similarity of chemical property distributions, revealing its limitations in handling molecules with high diversity and complex chemical structures. This research provides a novel approach and implementation for high-quality, high-validity SELFIES molecule generation by integrating pre-trained language models with discrete diffusion models.

Key Words: Molecule Generation; Diffusion Models; Masked Language Models; Deep Learning

目 录

第一章　绪论

第一节 研究背景与意义

分子设计与发现是化学、材料科学以及药物研发等多个关键领域的基石。所谓“化学空间”，指的是所有可能的分子和化合物在遵循一组给定的构造原理和边界条件下所跨越的属性空间。据估计，其中具有潜在药理活性（即药用价值）的分子数量级可达10^{60}[1]，而整个化学空间的分子数量更是难以估量。这种巨大的搜索空间使得通过传统的穷举实验来发现具有特定性质的新型分子几乎是不可能的。传统的分子发现方法，如高通量筛选（High-Throughput Screening, HTS），通过针对性的实验批量筛选已有的样本库，其中的化合物成本高昂，且此方法的本质仍是枚举式的。因此，开发高效的计算方法，利用机器学习等人工智能技术来智能地探索化学空间、预测分子性质并生成具有期望功能的新分子结构，已成为推动相关领域发展的迫切需求，具有重大的科学意义和广阔的应用前景。

近年来，随着人工智能，特别是深度学习和生成模型技术的飞速发展，从头分子生成（De Novo Molecule Generation）已成为一个备受瞩目的研究方向。这种方法旨在自动创建针对特定分子特征的新化学结构，它利用现有有效分子的知识来设计具有独特结构特征的新型分子。与传统方法 HTS 相比，AI 驱动的 De Novo Molecule Generation 减少了对广泛实验验证的依赖，有望显著加速新材料和新药物的发现进程，降低研发成本，并为解决特定化学挑战（如设计全新的具有特定结合亲和力、高药理活性或低毒性的分子）提供创新的解决方案。然而，尽管取得了显著进展，当前的分子生成模型仍面临诸多挑战。有关分子生成的报告指出[2]，这些挑战包括如何确保生成分子的化学有效性、结构多样性、新颖性，以及如何针对特定性质进行高效优化。此外，如何有效地表示分子结构（包括 1D 字符串、2D 图和 3D 几何结构）并将其与日益强大的深度学习模型相结合，以及如何弥合不同表示方法和生成策略之间的鸿沟，仍然是该领域持续探索的核心问题。本研究旨在这一快速发展的交叉领域贡献一份力量，专注于利用先进的生成模型和分子表示方法来提升分子生成的效率和质量。

第二节 相关工作

本节将回顾分子生成领域的相关研究工作。首先，本文将从分子的物理本质出发，系统梳理从三维到一维的各类分子表示方法及其代表性生成模型；随后，本文将总结主流的分子生成模型范式。

分子的计算机表示是所有基于机器学习的分子设计任务的起点。根据其描述的维度和抽象层次，主流表示方法可归纳为三维几何表示、二维图表示和一维字符串表示。

1. 3D 几何表示（3D Molecular Geometry）：分子的三维空间构象直接决定了其物理化学性质和生物活性，因此，直接描述原子三维坐标的几何表示法能够提供最丰富、最真实的信息。这种表示对于理解分子间相互作用和基于结构的药物设计（Structure-Based Drug Design, SBDD）[3]至关重要。然而，其挑战在于模型必须处理旋转和平移的对称性问题（即 E(3)等变性），且计算成本高昂。

   近年来，随着几何深度学习的发展，直接在三维空间中生成分子的方法取得了显著进展。其中，等变扩散模型（Equivariant Diffusion Models, EDM）[4]是一个里程碑式的工作。该工作直接对原子的三维坐标和类别特征（如原子类型）进行扩散和去噪。该模型的核心是一个等变图神经网络，它确保了在去噪过程的每一步，无论分子如何旋转或平移，模型的输出都能保持相应的变换，从而在生成几何上合理的分子构象方面表现出色，显著优于早期的三维生成方法。

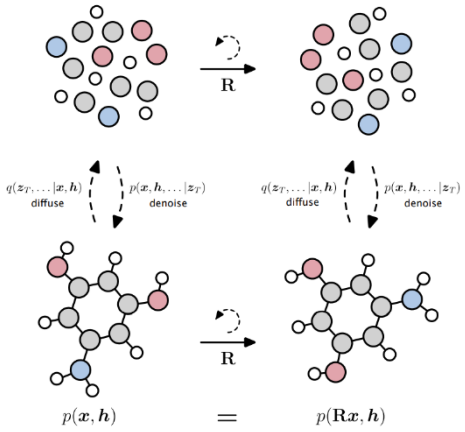

图 1 等变扩散模型

2. 2D 图表示（2D Graph Representation）：为了在保留关键结构信息和降低计算

复杂度之间取得平衡，研究者们广泛采用二维图表示，将分子抽象为原子（节点）和化学键（边）构成的图。这种表示直观地捕捉了分子的拓扑连接关系，是当前分子性质预测和图生成模型的主流。这类模型的挑战在于如何有效定义图的生成步骤（例如，一次性生成邻接矩阵，还是逐步添加节点和边）。

在逐步生成图方面，许多工作采用图神经网络（Graph Neural Networks, GNNs）[5]来指导生成过程。例如 Bongini 等人提出的 MG²N²（Molecular Generative Graph Neural Networks）[6]就是一个代表。该模型将图生成分解为一系列决策步骤：在每个步骤中，模型通过 GNN 模块决定是否添加一个新节点（原子）、新节点的类型，以及新节点与图中已有节点的连接方式（化学键类型）。通过在每一步利用 GNN 充分捕捉当前已生成子图的全部信息，MG²N²能够生成化学结构合理的分子，展现了 GNN 在序列化图生成任务中的潜力。

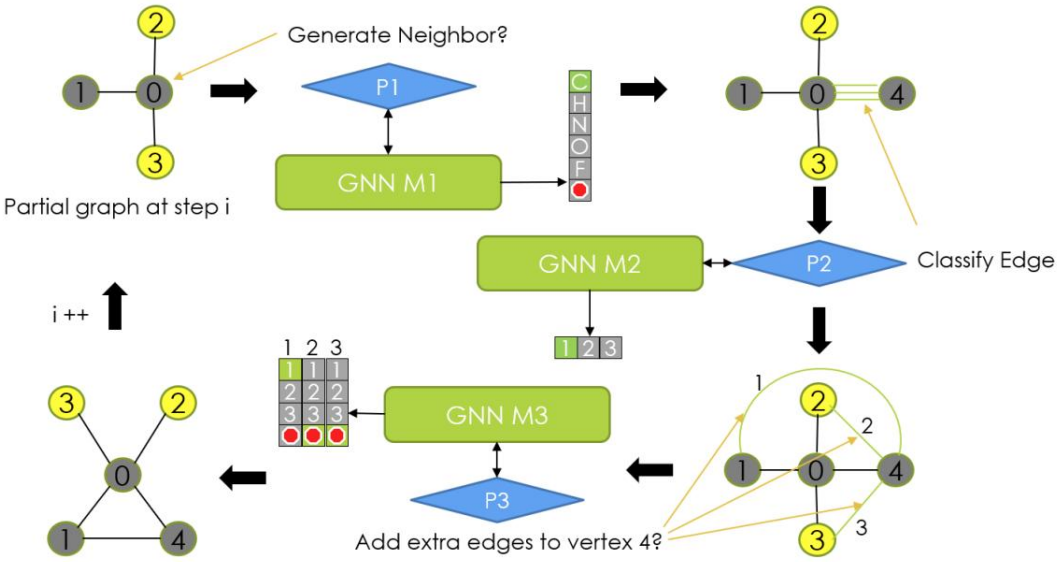

图 2 图神经网络分子生成过程

3. 一维字符串表示（1D String Representation）：将分子结构编码为线性字符序列，是与自然语言处理（Natural Language Processing, NLP）领域中成熟的序列模型（如 Transformer[7]）结合最紧密的方式，具有处理便捷、易于与大规模语言模型集成且计算成本较低的优势。

（1）SMILES（Simplified Molecular Input Line Entry System）[8]：SMILES 因其简洁性，在早期和当前的分子生成研究中都得到了广泛应用。一个代表性的研究方向是利用自编码器（Autoencoder，AE）[9]及其变体如变分自编码器

（Variational Autoencoder，VAE）[10]来学习 SMILES 字符串的连续隐空间表示。Blaschke 等人在其工作中[11]，系统地研究了多种自编码器架构用于 SMILES 表示的分子生成。该工作通过将大量分子编码到低维连续的隐空间，然后从该空间中采样并解码回 SMILES 字符串，成功生成了与已知药物结构类似的新分子，并证明了该隐空间保留了化学相似性原理，可以用于性质导向的分子优化。

然而，这项工作也清晰地暴露了所有基于 SMILES 生成方法所面临的共同且核心的缺陷：SMILES 语法的脆弱性与生成有效性瓶颈。Blaschke 等人在其研究中发现，即使是经过优化的模型，在从隐空间解码生成 SMILES 字符串时，也仅有约78%的产出是化学上有效的。这意味着有超过五分之一的计算资源被浪费在生成无意义的、不符合化学语法规则的字符串上（例如括号不匹配、环编号未闭合等）。这一固有的“低效率”问题，源于 SMILES 表示法本身，它并没有内在机制来保证任意字符组合的化学合理性。当生成任务越复杂，或者目标分子结构越新颖时，这个问题往往会变得更加严重。

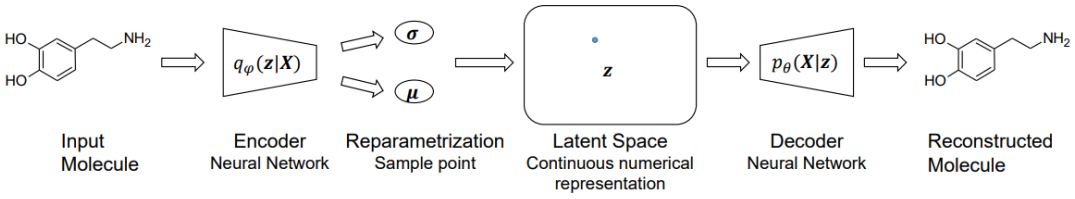

图 3 编码器生成 SMILES 字符串流程

（2）SELFIES（Self-Referencing Embedded Strings）[12]：为了从根本上解决 SMILES 的有效性难题，SELFIES 应运而生。它基于一套严格的上下文无关文法，通过将原子、键、环、分支等化学概念映射为独立的、语法上自洽的符号，确保了任何由其字母表符号组成的序列都能 100%被解码为一个有效的分子图结构。

综上所述，考虑到一维字符串表示与强大的序列模型天然适配且节省计算资源，而基于 SMILES 的生成方法尽管验证了利用深度学习模型探索化学空间的可行性，但其固有的有效性问题限制了其实际应用效率。因此，本研究选择 SELFIES 作为核心的分子表示方法，旨在彻底规避 SMILES 的语法陷阱，将模型的学习能力完全集中在捕捉和生成有意义的化学结构与语义上。

同时，SELFIES 与同样由离散符号组成的文本存在差异，文本具有前后的顺序依赖关系，因此自回归模型在文本生成领域有着优异表现，而 SELFIES 表示

的是一个分子，并没有严格的前后依赖关系。

针对上述背景和挑战，本文聚焦于离散分子序列的生成，提出了一种基于扩散语言模型的SELFIES分子生成方法DMLM（Diffusion Molecule Language Model，分子语言扩散模型），以捕捉SELFIES符号之间复杂的长程依赖关系，并深入理解序列的句法和语义结构。

第二章　预备知识

第一节 扩散模型

扩散模型，也称作去噪扩散概率模型（denoising diffusion probabilistic models）最早由Sohl-Dickstein等人[13]提出，Ho等人[14]在其基础上将训练与采样流程系统化，使之成为图像生成领域的重要模型。扩散模型本质上是一类潜变量模型，由正向扩散过程和反向扩散过程两部分组成，用以在像素空间或隐空间中逐步破坏并重建数据分布。

一、正向扩散过程

给定真实样本$x_0 \sim q(x_0)$，在分布正向过程里通过一个马尔可夫链 $x_1, \dots, x_T$逐步注入高斯噪声，公式为

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{(1-\beta_t)}\, x_{t-1}, \beta_t \mathbf{I}\right) \tag{2.1}$$

其中 $\{\beta_t\}_{t=1}^{T} \subset (0,1)$ 为噪声调度表，决定每一步加入噪声的强度。随着 t 递增，样本逐渐退化为各向同性高斯分布 $\mathcal{N}(0,1)$。当足够小且 T 足够大时，整个扩散序列在理论上可近似一个连续扩散过程。

二、反向扩散过程

由于正向过程是马尔可夫链，在 β_t 很小的前提下，其反向转移概率 $q(x_{t-1} \mid x_t)$同样近似高斯分布。训练时可以用参数化网络 p_θ 去拟合这一逆变换：

$$p_\theta(x_{t-1} \mid x_t, t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta(x_t, t)\right) \tag{2.2}$$

其中$\mu_\theta(\cdot)$, $\sum_\theta(\cdot)$可由U-Net[15]，Transformer等架构实现，用于预测每一步重建的均值与方差。

三、训练目标

为了最大化对数似然 $\log p_\theta(x_0)$，通常最小化其变分下界（Variational Lower Bound），将其作为损失函数进行训练：

$$\mathcal{L}_{vlb} = E_q\left[\mathrm{D_{KL}}\left(q(x_T \mid x_0) \,||\, p_\theta(x_T)\right)\right]$$

$$+E_q\left[\sum_{t=2}^{T} D_{KL}\left(q(x_{t-1} \mid x_t,\, x_0) \,|\, p_\theta(x_{t-1} \mid x_t,\, t)\right)\right] - \log p_\theta(x_0 \mid x_1) \quad (2.3)$$

其中 $E_q[\cdot]$ 表示对联合分布 $q(x_{0:T})$ 的期望；第一个 KL 项负责匹配终态先验，第二个 KL 项度量每一逆向步的匹配误差，最后一项为重建误差。

第二节 离散扩散模型

连续域扩散模型在图像、语音等实值数据上表现卓越，但在文本或分子 SELFIES 这类离散符号序列中，直接注入高斯噪声既不合理，也会导致梯度断裂。为此，Austin 等人[17]及后续工作将扩散思想推广到离散域：用类别转移噪声逐步随机化符号，并在反向过程中学习去噪概率分布。

一、正向扩散过程

设序列长度为 L，每个时间步向量 $x_t \in \{0,1\}^{L\times K}$ 为 one-hot 堆栈（K为为符号类别数；对自然语言 $K = |\mathrm{V}|$。定义转移矩阵 $Q_t \in [0,1]^{K\times K}$，其中

$$[Q_t]_{i,j} = q\left(x_t^l = j \mid x_{t-1}^l = i\right), \quad \forall\, l \in \{1,\dots,L\},\, i,j \in \{1,\dots,K\} \quad (2.4)$$

即对序列中每个符号独立施加同一随机映射。则一步正向扩散为

$$q(x_t \mid x_{t-1}) = \mathrm{Cat}(x_t;\, p = x_{t-1}Q_t) \quad (2.5)$$

其中 $\mathrm{Cat}(\cdot)$ 表示按行独立的多类别分布；$x_{t-1}Q_t$ 可视为对原 one-hot 张量乘以转移矩阵后的类别概率。

二、后验闭式解与联合转移矩阵

设累计转移矩阵$\overline{Q}_t = Q_1Q_2\cdots Q_t$，则在给定原始样本$x_0$的条件下，有

$$q(x_t \mid x_0) = \mathrm{Cat}\left(x_t; p = x_0\overline{Q}_t\right) \quad (2.6)$$

利用贝叶斯规则可得到一步精确后验：

$$q(x_{t-1} \mid x_t, x_0) = \mathrm{Cat}\left(x_{t-1}; p = \frac{x_tQ_t^{\top} \odot x_0\overline{Q}_{t-1}}{x_0\overline{Q}_t x_t^{\top}}\right) \quad (2.7)$$

其中“$\odot$”为元素乘，分母按行归一化，确保每个符号概率和为 1。这一闭式解为后续变分推导提供了可行的 KL 目标。

三、反向扩散建模与训练目标

与连续情形的式（2.2）类比，定义可学习去噪分布

$$p_\theta(x_{t-1} \mid x_t, t) = \mathrm{Cat}\big(x_{t-1}; \pi_\theta(x_t, t)\big) \tag{2.8}$$

其中网络输出$\pi_\theta(\cdot) \in \Delta^{K-1}$为每个符号的类别概率向量，可由 Transformer 或 掩码语言模型 （Masked Language Model， MLM）实现。训练时仍最小化式（2.3）的变分下界， 让模型输出逼近真后验分布 q.

第三节 掩码语言模型

传统的语言模型通常是单向的（从左到右或从右到左），这限制了模型在预训练阶段捕获上下文信息的能力。例如，在从左到右的模型中，每个词的表示只能依赖于其左侧的词。这种单向性对于需要理解完整上下文的任务而言是次优的。为了克服这一限制并预训练出能够同时利用左右两侧上下文信息的深度双向表示， Devlin 等人在其 BERT[18] （Bidirectional Encoder Representations from Transformers） 模型中引入了掩码语言模型 （Masked Language Model， MLM） 的预训练目标。

一、核心思想

MLM 的核心思想借鉴了心理语言学中的 Cloze 任务。其核心做法是：在输入序列中随机遮盖一定比例的词元（token），然后训练模型去预测这些被遮盖词元的原始身份。具体来说，对于一个输入序列：

1. 随机遮盖：选择序列中一定百分比（例如 BERT 中为 15%）的词元进行处理。
2. 处理方式：对于被选中的词元，并非总是替换为特殊的“[MASK]”标记。BERT 采用了一种策略来减轻预训练和微调阶段之间的不匹配问题（因为“[MASK]”标记在微调阶段通常不存在）：
   （1） 80%的概率：用“[MASK]”标记替换该词元。
   （2） 10%的概率：用一个随机选择的其他词元替换该词元。
   （3） 10%的概率：保持该词元不变。
3. 预测目标：模型的目标是基于未被遮盖的其他词元（即其左右上下文）来预测被处理词元的原始词元 ID。这通常通过在对应位置 Transformer 输出之上加一个 Softmax 分类层来实现，输出词汇表大小的概率分布。通过这种方式，模型被迫学习每个词元在特定上下文中的分布式表示，因为它无法预知哪些词元会被要求预测，也无法预知哪些词元是原始的或被随机替换的。这种机制使得表示能够融合来自左右两侧的上下文信息，从而训练出深度的双向

Transformer 编码器。

二、模型架构与输入表示

BERT 的模型架构是一个多层的双向 Transformer 编码器，基于 Vaswani 等人提出的原始 Transformer 结构。每个 Transformer 层都包含多头自注意力机制（multi-head self-attention）和前馈神经网络。双向性体现在自注意力机制允许每个词元关注到序列中所有其他词元（包括其左侧和右侧的词元）。

为了处理不同的下游任务，BERT 的输入表示能够将单个句子或句子对统一表示为一个词元序列。其输入表示由三部分加和构成：

1. 词元嵌入（Token Embeddings）：表示词元本身的语义信息，通常使用 WordPiece 等子词切分方法。
2. 片段嵌入（Segment Embeddings）：用于区分句子对中的不同句子（例如句子 A 和句子 B）。对于单一句子输入，则只有一种片段嵌入。
3. 位置嵌入（Position Embeddings）：为模型提供词元在序列中的位置信息，因为 Transformer 本身不包含序列顺序的概念。

第四节 分子生成评价指标

评估分子生成模型的性能至关重要，这需要一套全面的指标来衡量生成分子的质量、多样性、新颖性以及与真实分子分布的相似性。Polykovskiy 等人在其提出的 MOSES[19]（Molecular Sets）基准平台中，系统地整理并应用了一系列评价指标。本节将主要依据 MOSES 框架介绍常用的分子生成评价指标。通常，这些指标的计算会涉及一个生成的分子集合 G 和一个参考的测试集 R（通常包含真实的，符合特定化学或物理性质要求的分子）。除非特别说明，所有指标均在通过有效性检查的生成分子上计算。

一、基础指标

1. 有效性（Validity，Valid）：指生成的分子串（如 SMILES 或 SELFIES）能够被化学信息学工具（如 RDKit）成功解析并被确认为化学上合理的分子结构的比例。这是最基本的指标，衡量模型是否学习到了基本的化学规则，如原子价态、成键规则等。对于使用 SELFIES 这种保证语法有效性的表示法，此指标主要验证从 SELFIES 到分子图的转换以及更深层次的化学合理性。
2. 唯一性（Uniqueness@k, Unique@k）：在生成的前k个有效分子中，不重复的

分子的比例。通常会报告$k = 1{,}000$和$k = 10{,}000$时的结果。该指标衡量了模型是否倾向于生成少量重复的“典型”分子，即是否存在模式坍塌（mode collapse）的问题。高唯一性表明模型能够产生多样化的输出。

3. 新颖性（Novelty）：生成的有效分子中，不存在于训练集中的分子的比例。该指标衡量了模型是否具有泛化能力并能产生新的化学结构，而不是简单地记忆和复现训练样本。低新颖性可能表示模型过拟合。

二、基于化学特征的相似性指标

1. 片段相似性（Fragment Similarity，Frag）：比较生成分子集 G 和参考集 R 中化学片段的分布相似性。通常使用余弦相似度计算两个集合中各片段出现频率向量的相似度。

$$\mathrm{Frag}(G,R) = \frac{\sum_{f\in F}\left(c_f(G)\cdot c_f(R)\right)}{\sqrt{\sum_{f\in F} c_f^2(G)}\sqrt{\sum_{f\in F} c_f^2(R)}} \tag{2.9}$$

其中 $c_f(A)$ 是片段 f 在集合 A 中出现的次数， F 是在 G 或 R 中出现的所有片段的集合。该指标评估生成分子在子结构层面是否与真实分子相似。如果某些重要片段在生成分子中过多或过少，该指标会降低。

2. 骨架相似性（Scaffold Similarity，Scaf）：与片段相似性类似，但比较的是 Bemis-Murcko 骨架的频率分布。Bemis-Murcko 骨架包含分子所有的环结构以及连接这些环的链状片段。计算方式同片段相似性，只是将片段替换为骨架。该指标评估生成分子在核心骨架结构上是否与参考集相似。这对于药物发现等领域尤为重要，因为骨架通常决定了分子的核心药效团或理化性质。
3. 与最近邻分子的相似性（Similarity to Nearest Neighbor，SNN）：计算生成集G中每个分子与其在参考集R中最相似分子之间的平均 Tanimoto 相似度（基于分子指纹，如 Morgan 指纹）。

$$\mathrm{SNN}(G,R) = \frac{1}{|G|}\sum_{m_G\in G}\max_{m_R\in R} T(m_G,m_R) \tag{2.10}$$

该指标可以看作是一种“精确率”的衡量。如果生成分子与真实分子集合的流形相距较远，SNN 会较低。

4. 内部多样性（Internal Diversity，IntDiv）：评估生成分子集G内部的化学多样性。计算方式为集合内所有分子对之间的平均 Tanimoto 相似度的补数。

$$\mathrm{IntDiv_p}(G) = 1 - \left(\frac{1}{|G|^2} \sum_{m_1, m_2 \in G} T(m_1, m_2)^p \right)^{\frac{1}{p}} \tag{2.11}$$

检测模式坍塌的另一个重要指标。高内部多样性表示生成的分子结构各不相同，覆盖了更广泛的化学空间。

第三章　基于扩散语言模型的分子生成方法

第一节 引言

本章详细阐述了本文提出的用于分子生成的新型方法。该方法的核心思想是将离散扩散模型（Discrete Diffusion Model）的强大生成能力与掩码语言模型（Masked Language Model, MLM）在序列数据理解方面的优势相结合，专门针对SELFIES（Self-Referencing Embedded Strings）分子表示进行优化。在模型的框架中，一个预训练的 MLM 将扮演去噪器的关键角色，在扩散模型的反向过程中逐步从噪声中恢复出结构合理且具有化学意义的 SELFIES 字符串。

第二节 问题定义与建模思路

分子生成任务可以形式化地定义为：给定一个由大量已知分子构成的训练数据集所隐式定义的目标分子分布 $p(x_0)$，目标是学习一个生成模型 $p_\theta(x_0)$，该模型能够从一个简单的先验分布（如噪声分布）中采样，并生成新的、化学有效的、结构多样化的分子表示（在本工作中为 SELFIES 字符串）。这些生成的 SELFIES 字符串随后可以被解码为具有特定化学结构和潜在期望性质的分子。

一、SELFIES 表示的选择

本文选择 SELFIES 作为分子的线性表示，主要基于其核心优势：100%的语法有效性。与 SMILES 等传统表示法相比，任何 SELFIES 符号的组合都能保证对应一个化学上可解析的分子图结构，从而从根本上避免了生成大量无效分子的问题。SELFIES 的这种鲁棒性使其非常适合作为离散生成模型的目标。

二、离散扩散模型的选择

考虑到 SELFIES 本质上是离散的符号序列，本文选择离散扩散模型作为基础

生成框架。如第二章所述，离散扩散模型通过一个正向过程逐步向数据中引入噪声（例如，将符号替换为特殊标记或随机符号），并通过一个反向过程学习从噪声中恢复原始数据。这种逐步去噪的机制为生成复杂离散结构提供了强大的建模能力。

三、掩码语言模型（MLM）作为去噪核心的动机

在离散扩散模型的反向（去噪）过程中，核心任务是在给定部分损坏的序列 x_t 和当前时间步 t 的条件下，预测更接近原始序列的 x_{t-1}（或直接预测原始序列 x_0）。而预训练的掩码语言模型 （MLM）非常适合承担这一角色：

1. 处理 SELFIES 的非序列依赖：SELFIES 符号之间的相互作用和约束往往不是严格的从左到右的顺序。MLM，特别是基于 Transformer 的 MLM，其核心的双向注意力机制能够捕获序列中任意两个符号之间的依赖关系，无论它们在序列中的距离如何，也无论其顺序如何。这对于理解和重构 SELFIES 这种具有复杂内部语法的表示至关重要。
2. 学习 SELFIES 的语法与语义：通过在大量 SELFIES 数据上进行预训练，MLM 能够学习到 SELFIES 语言的内在语法规则以及一定的化学结构语义。这种学到的知识对于在去噪过程中做出合理的符号预测至关重要。
3. 利用 MOLGEN[20]的预训练知识：本文选择 MOLGEN 作为 MLM 的基础。MOLGEN 是一个专门为分子生成任务设计的预训练模型，它通过重构超过一亿个分子 SELFIES 来内化结构和语法知识。将其作为去噪器，可以期望利用其强大的分子理解能力来指导扩散模型的生成过程，从而可能加速收敛并提升生成分子的质量。

第三节 模型架构

本文的模型由两个主要过程组成：一个固定的正向扩散过程和一个可学习的反向去噪过程。正向过程从一个干净的 SELFIES 分子串 x_0 开始，通过 T 个离散的时间步逐步引入噪声，最终得到一个完全噪声化的序列 x_T（即全由“[MASK]”符号组成的序列）。反向过程从 x_T 开始，模型学习逐步去除噪声，迭代地从 x_t 预测 x_{t-1}，最终生成一个干净的 SELFIES 串 x_0。这个去噪步骤由预训练的的掩码语言模型 MOLGEN 通过 $p_\theta(x_{t-1}|x_t,t)$ 来参数化。相比正常的离散扩散模型，

本工作对模型框架的主要改进如下：

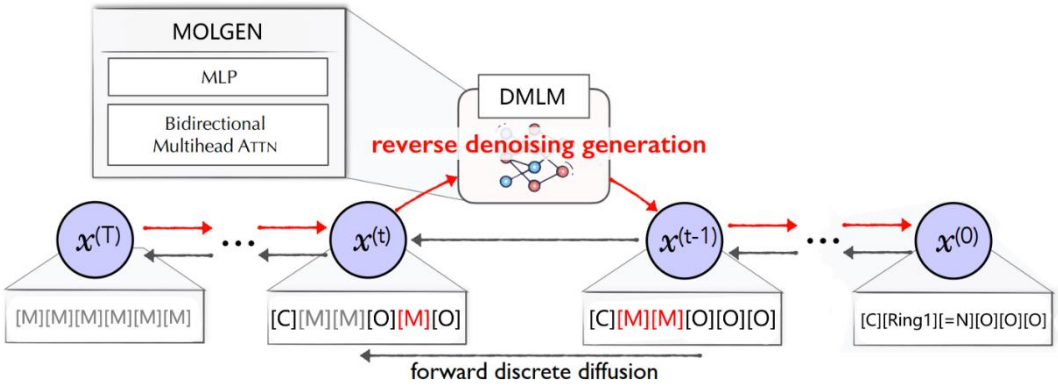

图 4 模型框架

一、正向过程：带离散吸收态的扩散

在为离散的 SELFIES 符号序列设计扩散过程时，面临的首要问题是如何有效地引入噪声。一种常见的离散噪声方式是将符号替换为其他随机符号，然而考虑到本文的去噪网络将基于预训练的掩码语言模型 MOLGEN，其预训练任务通常是“完形填空”，即预测被特殊[MASK]符号遮盖的词元。

为了使扩散模型的去噪任务与 MLM 的预训练目标尽可能对齐，从而更有效地利用 MOLGEN 的预训练知识，本文在马尔可夫扩散过程中引入了一个离散的吸收态，即词汇表中除去 SELFIES 词元额外的“[MASK]”符号。在每个时间步 t，序列中的每一个 SELFIES 符号 x_{t-1}^{i}（其中i为序列中的位置索引）要么保持不变，要么以一定的概率β_t转换为“[MASK]”符号。形式上，这一步的转移矩阵 $[Q_t]_{ij} = q\left(x_t^{(l)} = j \mid x_{t-1}^{(l)} = i\right)$ 定义如下：

$$[Q_t]_{ij} = \begin{cases} 1 & if\ i = j = [\mathrm{M}], \\ \beta_t & if\ j = [\mathrm{M}], i \neq [\mathrm{M}], \\ 1 - \beta_t & if\ i = j \neq [\mathrm{M}], \end{cases} \tag{3.1}$$

其中“[M]”代表“[MASK]”符号。经过 T 步扩散后，原始序列 x_0 将以高概率收敛到一个所有符号均为“[MASK]”的序列 x_T。对于任意一个原始符号 x_0^i，经过 t步扩散后，其条件概率分布 $q(x_t^i|x_0^i)$ 可以表示为：

$$q(x_t^i|x_0^i) = \begin{cases} \bar{\alpha}_t & if\ x_t^i = x_0^i, \\ 1 - \bar{\alpha}_t & if\ x_t^i = [\mathrm{M}], \end{cases} \tag{3.2}$$

其中$\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$，$x_t^i$ 表示时间步 t 序列中的第 i 个词元。

二、纺锤形噪声调度

在定义了如何引入噪声、之后，接下来的关键问题是如何确定噪声调度，即

每一步的噪声引入概率 β_t（或等价地 $\overline{\alpha_t}$），对扩散模型的性能有显著影响。传统调度（如 $\beta_t = (T-t+1)^{-1}$）通常平等对待序列中的所有符号位置。然而，SELFIES 符号与自然语言词元类似，其在序列中承载的信息量是不同的。这样的调度方案存在以下问题：（1）未明确量化每步加入的噪声量；（2）假设所有符号携带相同信息量，忽略了差异；（3）与去噪语言模型“先易后难”的生成特性相悖。

为了更精细化地控制噪声引入过程，本文采用了 DiffusionBERT[21]中提出的纺锤形噪声调度。纺锤形噪声调度的核心思想是根据每个 SELFIES 符号的信息量（以熵 $H(x_0^i)$ 衡量）来差异化地引入噪声，并力求在整个扩散过程中均匀地分布“被损坏的信息总量”。其目标是使得信息量较大的符号在正向过程中更早被掩码，相应地，在反向生成过程中更晚被恢复。具体地，第 i 个符号在时间步 t 保持未被掩码的概率 $\bar{\alpha}_t^i$ 由以下公式确定：

$$\bar{\alpha}_t^i = 1 - \frac{t}{T} - S(t) \cdot \tilde{H}(x_0^i) \tag{3.3}$$

$$S(t) = \lambda \sin\left(\frac{t\pi}{T}\right) \tag{3.4}$$

$$\tilde{H}(x_0^i) = 1 - \frac{\sum_{j=1}^{n} H(x_0^j)}{nH(x_0^i)} \tag{3.5}$$

其中 T 是总的扩散步数，$S(t)$ 是一个正弦形状的函数，用于控制信息量 $\tilde{H}(x_0^i)$ 在不同时间步 t 的影响强度，λ 是一个超参数。$S(0) = S(T) = 0$ 确保了在初始和最终状态信息的完全保留或完全损坏。$\tilde{H}(x_0^i)$ 是对原始符号 x_0^i 信息熵的归一化度量，使得信息量越大的符号，其值也越大。这种设计使得前向扩散过程成为一个非马尔可夫过程，因为在时间步 t 将符号 x_{t-1}^i 转换为 x_t^i 的概率 β_t^i 间接依赖于原始符号 x_0^i 的信息量。

三、基于 MOLGEN 的去噪网络与时间无关解码

反向过程是扩散模型的核心学习部分，其目标是训练一个网络 p_θ 从噪声序列 x_t 预测更早时间步的序列 x_{t-1}，或者更直接地预测原始的、未损坏的 SELFIES 序列 x_0。考虑到 SELFIES 序列的复杂性和捕捉其长程依赖的需求，我们选择强大的预训练掩码语言模型 MOLGEN 作为去噪网络的基础。在反向去噪过程中，其核心组件是基于 MOLGEN 的掩码语言模型。然而，标准扩散模型在进行反向去噪 $p_\theta(x_{t-1}|x_t, t)$ 时，通常需要将当前的时间步 t 作为显式条件输入给去噪网络，因为不同时间步的噪声水平和去噪难度差异巨大。但 MOLGEN 在其原始预训练阶

段并未包含此类时间步条件。直接引入时间步嵌入需要对 MOLGEN 的架构进行修改，其预训练学到的强大表示能力也可能会因为不熟悉的时间步输入而未能充分利用。

为了充分利用 MOLGEN 的预训练能力并简化模型适配，本工作沿袭了 DiffusionBERT 中被证明有效的“时间无关解码”策略。在此策略下，时间步 t 不作为显式输入提供给 MOLGEN 网络。取而代之的是，模型被期望从输入的噪声序列 x_t 本身（例如，通过序列中“[MASK]”符号的数量、分布或整体的损坏程度）隐式地推断出当前的噪声水平或等效的时间阶段。因此，去噪网络的形式简化为 $p_\theta(x_0|x_t)$。这种设计不仅简化了模型的输入，也使得预训练的 MLM（如 MOLGEN，在其原始预训练中通常不处理显式的时间步输入）更容易适配到扩散框架中。

将预训练好 MOLGEN 模型作为去噪网络的基础。在训练整个扩散模型时，MOLGEN 的参数会针对从不同噪声水平的 SELFIES 序列中恢复原始结构这一特定去噪任务进行微调。通过这种方式，本工作旨在充分利用 MOLGEN 已学习到的丰富的分子结构知识和 SELFIES 语法规则，来指导和增强扩散模型的生成能力。

第四节 训练过程

一、训练目标

训练的目标是让基 MOLGEN 的去噪网络 p_θ 能够准确地从噪声样本 x_t 中恢复出原始的 SELFIES 序列 x_0。遵循离散扩散模型的一般框架，并考虑到我们的去噪网络旨在预测原始符号，损失函数可以简化为一个交叉熵损失，用于最小化模型预测的符号分布与真实原始符号之间的差异。对于给定的原始 SELFIES 序列 x_0，时间步 t，以及通过正向过程 $q(x_t|x_0)$ 得到的噪声序列 x_t，损失函数 $\mathcal{L}$ 定义为：

$$\mathcal{L} = E_{t\sim U[1,T],x_0\sim q_{\text{data}},x_t\sim q(x_t|x_0)}\left[\sum_{i=1}^{L} -\log p_\theta\left(x_0^i \mid x_t, t\right)\right] \tag{3.6}$$

其中 L 是序列长度，x_0^i 是原始序列中第 i 个位置的符号，$p_\theta(x_0^i|x_t,t)$ 是模型在给定 x_t 和（隐式）时间 t 的条件下，预测第 i 个位置为 x_0^i 的概率。在实践中，这个损失通常只计算那些在 x_t 中被掩码或改变的位置。

二、训练算法

训练过程遵循以下步骤：

1. 从训练数据集中随机采样一批完整的 SELFIES 分子串 x_0。
2. 随机均匀采样一个时间步 t（从 1 到 T）。
3. 根据 3.3.2 节描述的纺锤形噪声调度和正向扩散过程 $q(x_t|x_0)$，生成噪声样本 x_t（即根据 $\bar{\alpha}_t^i$ 的概率决定 x_0^i 是否被替换为“[MASK]”）。
4. 将噪声样本 x_t 输入到基于 MOLGEN 的去噪网络 p_θ。
5. 模型 p_θ 输出对原始 SELFIES 符号 x_0 的预测。
6. 根据预测结果和真实的 x_0 计算损失函数 $\mathcal{L}$。
7. 使用反向传播算法计算损失相对于模型参数 θ 的梯度。
8. 使用优化器（AdamW）更新模型参数 θ。
9. 重复以上步骤，直至模型在完成预设的训练迭代次数。

第五节 分子生成过程

训练完成后，模型可以用于生成新的 SELFIES 分子串。生成过程是一个迭代的去噪过程，从一个完全噪声化的状态开始：

1. 起始状态（x_T）：生成一个长度为 L（可以是固定的或动态确定的）的序列，其中所有符号都被初始化为`[MASK]`符号。这对应于扩散过程的最大时间步 T。
2. 迭代去噪：从 $t = T$ 开始，逐步迭代到 $t = 1$：
   （1） 将当前的噪声样本x_t输入到训练好的基于 MOLGEN 的去噪网络 p_θ。
   （2） 对于序列中的每个位置，模型输出一个关于下一个状态 x_{t-1}^i 的 SELFIES 符号的概率分布。
   （3） 从这些概率分布中为每个位置采样一个符号，形成 x_{t-1}。根据概率分布进行随机抽样，并引入温度系数 T 来控制生成的多样性。
1. 终止与输出：当 $t = 1$ 完成最后一步去噪后，得到的序列 x_0 即为一个生成的 SELFIES 分子串。

第四章　实验结果

第一节 实验设定

一、数据集与预处理

本文共使用了三个数据集进行实验：

1. ZINC250K：该数据集包含了 250,000 个分子的子集，主要用于化学语言模型的基准测试和分子属性回归任务的训练。本文使用该数据集进行了模型训练的初步尝试，成功完成了分子生成，验证了实验思路的正确性。
2. MOSES[19]：该数据集的训练集包含约 160 万个分子，测试集约 17.6 万个分子，其中剔除了含有带电原子或 C、N、S、O、F、Cl、Br、H 以外元素的分子，以及包含超过 8 元环的分子。本文使用该数据集完成了模型的完整训练并在基础指标上对模型进行评估。
3. Natural product[22]：从天然产物活性与物种来源数据库（Natural Product Activity & Species Source Database）中获取的 30,926 个化合物。在这些化合物中，任意选择了 30,126 个用于训练，并保留了 800 个用于测试。本文使用该数据集训练模型并进行化学性质指标的评估。

对于以上数据集，需要预先进行词元的频率统计，用于控制纺锤型的噪声调度。

二、实验环境与参数设定

本文的实验环境配置具体如下表所示：

表 4.1 实验环境

Python	Pytorch	datasets	nltk	numpy
3.8	2.21+cuda121	2.21.0	3.8.1	1.24.4
torchmetrics	transformers	rdkit	molsets	selfies
1.4.1	4.21.1	2022.3.5	1.0	2.2.0

本文的模型训练过程参数设定如下表所示：

表 4.2 模型训练参数

Learning Rate	Batch Size	Epoch	Time Steps

5e-5	256	100	2048

本文的模型采样生成过程参数设定如下表所示：

表 4.3 模型采样参数

Time Steps	Temperature	Topk	Step Size
512	0.2	30	2

本文模型根据以上配置及参数在 4 张 NVIDIA A40 上进行训练，所有实验结果可以通过 https://github.com/MikeWYX/Diffusion-Molecular-Language-Model 的代码以及以上参数进行复现。

第二节 MOSES 实验结果

本实验采用 MOSES 数据集对模型进行训练，并使用 CharRNN[22]，VAE[10]，JTN-VAE[23]，LatentGAN[24]等近年来分子生成的模型作为基线模型，参照 MOSES（https://github.com/molecularsets/moses）的流程进行训练采样及评估，具体的评估结果如下所示：

表 4.4 不同模型在 MOSES 数据集上的分子生成表现

Model	Valid	Unique@10k	Novelty
CharRNN	0.975	1.0	0.842
VAE	0.977	1.0	0.695
JTN-VAE	1.0	1.0	0.913
LatentGAN	0.897	1.0	0.950
MOLGEN	1.0	1.0	1.0
DMLM	1.0	1.0	1.0

在这些基础指标上，本文的模型 DMLM 均达到了 SOTA（state of the art）水平，鉴于作为模型框架中去噪网络的 MOLGEN 模型也为 SOTA 水平，本文在“atom stable”和“mol stable”两个指标上继续评估。

表 4.5 与基线模型的进一步比较

Model	atom stable	mol stable
MOLGEN	1.0	0.994
DMLM	1.0	1.0

如上表所示，本文模型生成的分子在稳定性上有着进一步提升，结合表 4.4，表明本文模型非常适合在不发生过拟合的情况下发现新化学结构并探索未知化学空间。

第三节 Natural-product 实验结果

本实验采用 natural-product 数据集对模型进行训练，使用 MOLGEN 所提供的 CharRNN，VAE，JTN-VAE，LatentGAN 等基线模型的结果加以比较，具体评估结果如下所示：

表 4.4 不同模型在 natural product 数据集上的分子生成表现

Model	Frag	Scaf	SNN	IntDiv	FCD
CharRNN	0.277	0.884	0.532	0.601	45.53
VAE	0.884	0.456	0.395	0.872	4.32
JTN-VAE	0.880	0.501	0.375	0.874	12.03
LatentGAN	0.277	0.088	0.532	0.601	45.53
MOLGEN	0.999	0.841	0.815	0.888	0.65
DMLM	0.779	0.005	0.392	0.714	35.65

如上表所示，本文的模型在捕捉数据集的化学性质分布上的表现不尽如人意，分析其原因，可能包括以下几点：

1. 天然产物的复杂性：天然产物分子结构复杂、骨架多样，且通常包含较多的手性中心和稠环体系。这对生成模型的学习能力提出了极高的要求。
2. 预训练与微调的领域差异：MOLGEN 本身在更广泛的化学空间（包括大量合成化合物）上进行了预训练。虽然 MOLGEN 在直接应用于天然产物生成时表现优异，但当将其作为去噪网络嵌入到 DMLM 框架中，并针对天然产物数据集进行微调时，扩散过程的特性（如多步迭代去噪、噪声调度等）可能未能很好地适应天然产物独特的结构特征。

3. 数据量限制：虽然 natural product 数据集包含数万个天然产物（10^4），但相对于 MOSES 等合成化合物数据集（10^6），其规模仍然较小，这可能限制了 DMLM 从头学习捕捉天然产物复杂分布的能力，尤其是在扩散模型的框架下。

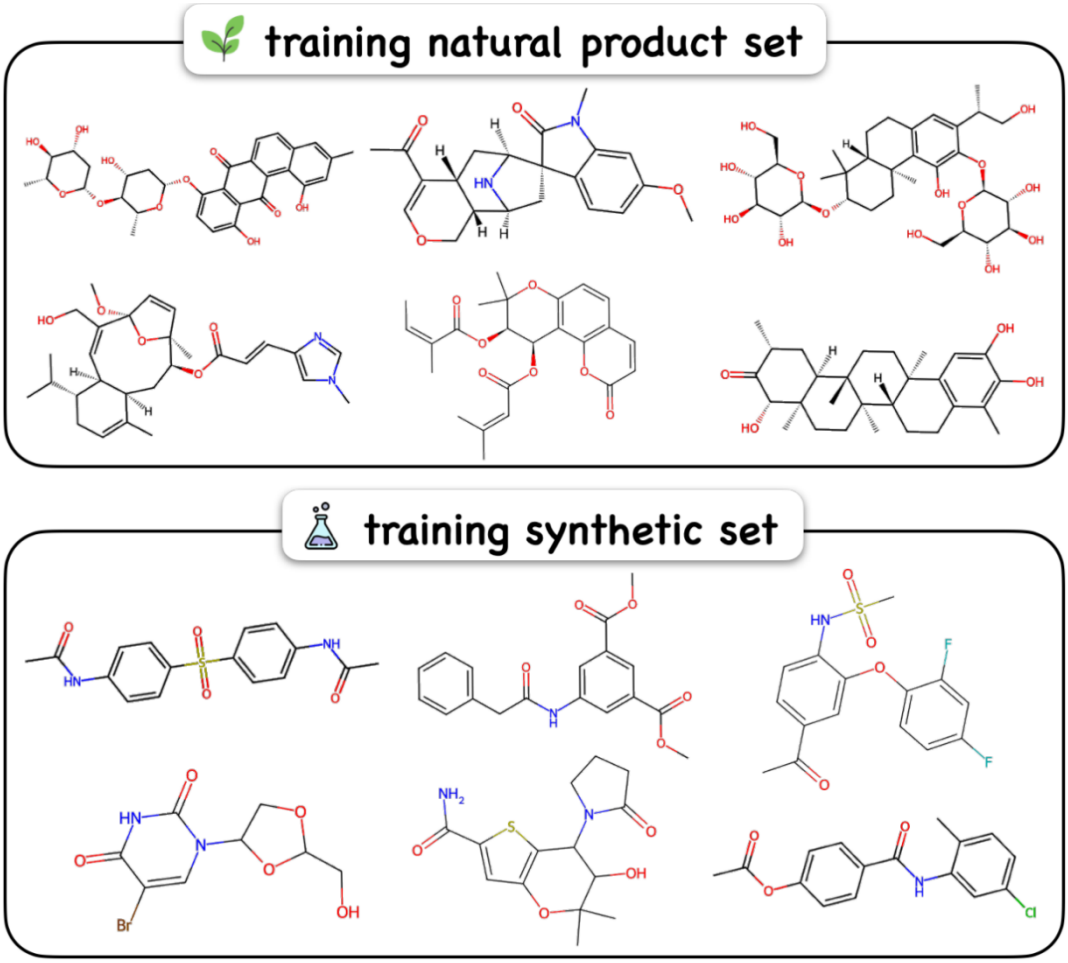

图 5 MOSES 和 Natural 训练集分子对比

第五章 总结与展望

第一节 结果讨论

本文针对分子生成这一化学信息学和药物发现领域的关键问题，提出了一种基于扩散语言模型的 SELFIES 分子生成方法（DMLM）。该方法的核心创新在于将强大的预训练掩码语言模型（MOLGEN）作为核心去噪网络，嵌入到离散扩散模型的框架中，并针对 SELFIES 这种 100%有效的分子表示进行优化。

主要研究工作和贡献概括如下：

1. 模型框架设计：本文构建了一个基于吸收态的离散扩散过程，用于 SELFIES 的逐步噪声化和去噪生成。借鉴 DiffusionBERT 的成功经验，引入了能够感知符号信息量的纺锤形噪声调度策略，以及简化模型设计并有助于利用预训练模型的时间无关解码机制。
2. 预训练知识的融合：通过采用专为 SELFIES 预训练的 MOLGEN 模型作为去噪网络，并对其进行微调，旨在 MOLGEN 学习到的丰富分子结构知识和 SELFIES 语法规则有效地迁移到分子生成任务中。
3. 实验验证与分析：在广泛使用的 MOSES 基准数据集上，DMLM 在生成分子的有效性、唯一性、新颖性以及原子/分子稳定性等指标上均达到 SOTA 水平，展示了其在生成分子方面的强大能力。然而，在结构更为复杂的天然产物数据集上，DMLM 在捕捉化学特性分布方面的表现不佳，揭示了模型在处理高度多样化和复杂化学结构时的局限性。

总的来说，本研究探索了扩散模型与预训练语言模型在分子生成领域的结合潜力，特别是在处理 SELFIES 表示方面提供了一种新的思路。实验结果表明，该方法在分子生成任务中具有一定竞争力，同时也指出了其在更复杂化学空间中面临的挑战。

第二节 不足和未来工作

尽管本文提出的 DMLM 模型在某些方面取得了令人鼓舞的结果，但仍存在一些不足之处，并为未来的研究提供了若干方向：

一、不足之处

1. 对复杂化学空间的适应性：如实验结果所示，DMLM 在处理结构高度复杂和多样化的天然产物数据集时表现不佳。这表明模型当前的架构和训练策略可能难以完全捕捉这类分子的独特分布特征。
2. 计算成本：扩散模型通常需要较多的迭代步数进行采样生成，相对于一些一次性生成或自回归步数较少的模型，其计算开销可能仍然较大，尤其是在生成大量分子时。
3. 超参数敏感性：扩散模型的性能往往对噪声调度、总扩散步数、采样参数（如温度、top-k）等超参数较为敏感，需要细致的调优。纺锤形噪声调度中的参数 λ 以及信息熵的计算方式也可能需要针对不同数据集进行适配。

4. 可控生成能力有限：当前模型主要进行无条件分子生成，缺乏直接生成具有特定化学或生物学性质的分子的能力。

二、未来工作展望

1. 提升对复杂化学空间的建模能力：探索更具适应性的噪声调度机制，研究如何将化学结构先验知识（如官能团信息、骨架约束）更有效地融入到扩散过程或去噪网络中。
2. 条件分子生成与性质导向：引入分类器引导（Classifier Guidance）：可以训练一个独立的离散属性分类器（例如，预测分子是否具有某种期望的药理活性、低毒性或高合成可及性）。在反向扩散的每一步，利用该分类器的梯度信息（或其在离散空间中的等价形式）来引导去噪网络的输出，使其倾向于生成具有目标属性的 SELFIES。具体而言，可以在采样步骤 $p_\theta(x_{t-1}|x_t)$ 中，结合分类器 $p_\varphi(y|x_{t-1})$的对数概率的梯度，从而调整最终采样的符号概率。这种方式借鉴了连续扩散模型中 classifier guidance[25]和 classifier-free guidance[26]的思想，并将其适配到离散的 SELFIES 生成框架中。
3. 多模态与多目标优化：结合分子的其他模态信息（如三维结构、文本描述）进行多模态分子生成。将扩散模型与多目标优化算法（如强化学习、遗传算法）相结合，以同时优化多个分子性质。

通过对上述方向的探索，本文期望能够进一步提升基于扩散语言模型的分子生成方法的性能、可控性和实用性，为新药研发和新材料设计等领域贡献更强大的计算工具。

参 考 文 献

[1] Polishchuk P G, Madzhidov T I, Varnek A. Estimation of the size of drug-like chemical space based on GDB-17 data[J]. *Journal of computer-aided molecular design*, 2013, 27: 675-679.

[2] Du Y, Fu T, Sun J, et al. Molgensurvey: A systematic survey in machine learning models for molecule design[J]. *arXiv preprint arXiv:2203.14500*, 2022.

[3] Anderson A C. The process of structure-based drug design[J]. *Chemistry & biology*, 2003, 10(9): 787-797.

[4] Hoogeboom E, Satorras V G, Vignac C, et al. Equivariant diffusion for molecule generation in 3d[C]//*International conference on machine learning*. PMLR, 2022: 8867-8887.

[5] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. *arXiv preprint arXiv:1609.02907*, 2016.

[6] You J, Liu B, Ying Z, et al. Graph convolutional policy network for goal-directed molecular graph generation[J]. *Advances in neural information processing systems*, 2018, 31.

[7] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. *Advances in neural information processing systems*, 2017, 30.

[8] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules[J]. *Journal of chemical information and computer sciences*, 1988, 28(1): 31-36.

[9] Ng A. Sparse autoencoder[J]. CS294A *Lecture notes*, 2011, 72(2011): 1-19.

[10] Kingma D P, Welling M. An introduction to variational autoencoders[J]. *Foundations and Trends® in Machine Learning*, 2019, 12(4): 307-392.

[11] Blaschke T, Olivecrona M, Engkvist O, et al. Application of generative autoencoder in de novo molecular design[J]. *Molecular informatics*, 2018, 37(1-2): 1700123.

[12] Krenn M, Häse F, Nigam A K, et al. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation[J]. *Machine Learning: Science and Technology*, 2020, 1(4): 045024.

[13] Sohl-Dickstein J, Weiss E, Maheswaranathan N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//*International conference on machine learning*. pmlr, 2015: 2256-2265.

[14] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. *Advances in neural information processing systems*, 2020, 33: 6840-6851.

[15] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//*Medical image computing and computer-assisted intervention–*

MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer international publishing, 2015: 234-241.

[16] Austin J, Johnson D D, Ho J, et al. Structured denoising diffusion models in discrete state-spaces[J]. *Advances in neural information processing systems*, 2021, 34: 17981-17993.

[17] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//*Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019: 4171-4186.

[18] Polykovskiy D, Zhebrak A, Sanchez-Lengeling B, et al. Molecular sets (MOSES): a benchmarking platform for molecular generation models[J]. *Frontiers in pharmacology*, 2020, 11: 565644.

[19] Fang Y, Zhang N, Chen Z, et al. Domain-agnostic molecular generation with chemical feedback[J]. *arXiv preprint arXiv:2301.11259*, 2023.

[20] He Z, Sun T, Wang K, et al. Diffusionbert: Improving generative masked language models with diffusion models[J]. *arXiv preprint arXiv:2211.15029*, 2022.

[21] Zhao H, Yang Y, Wang S, et al. NPASS database update 2023: quantitative natural product activity and species source database for biomedical research[J]. *Nucleic Acids Research*, 2023, 51(D1): D621-D628.

[22] Segler M H S, Kogej T, Tyrchan C, et al. Generating focused molecule libraries for drug discovery with recurrent neural networks[J]. *ACS central science*, 2018, 4(1): 120-131.

[23] Jin W, Barzilay R, Jaakkola T. Junction tree variational autoencoder for molecular graph generation[C]//*International conference on machine learning*. PMLR, 2018: 2323-2332.

[24] Prykhodko O, Johansson S V, Kotsias P C, et al. A de novo molecular generation method using latent vector based generative adversarial network[J]. *Journal of cheminformatics*, 2019, 11: 1-13.

[25] Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis[J]. *Advances in neural information processing systems*, 2021, 34: 8780-8794.

[26] Ho J, Salimans T. Classifier-free diffusion guidance[J]. *arXiv preprint arXiv:2207.12598*, 2022.

致　谢

我在中国科大读书的四年时间转瞬即逝，在这四年的本科生涯里学会了很多，在此我要对指导过我的老师，帮助过我的同学，照顾过我的朋友说一声谢谢。

首先感谢我的导师何向南老师和罗晏宸师兄对我的悉心教诲，激发了我对科学研究的兴趣和对科研工作的责任感。此论文是在导师的指导下完成的，在论文的结构、撰写和修改等方面都提出了宝贵的意见，也提供了很多的帮助。导师认真严谨的科研态度，敏锐的科学洞察力，独到的见解使我受益良多，也会深刻影响我以后的工作和生活。罗晏宸师兄帮助我确定了研究方向，并在实验中为我提供了许多思路。在此表达我最诚挚的谢意！

2025 年 5 月