Diffusion models in de novo drug design

Amira Alakhdar, Barnabas Poczos, Newell Washburn ALIA@ADOBE.COM

Carnegie Mellon University, Pittsburgh, PA, 15213, USA

Editor: A detailed contributor list can be found in the appendix of this paper.

Abstract

扩散模型已成为分子生成的强大工具，特别是在 3D 分子结构的上下文中。受非平衡统计物理的启发，这些模型能够生成具有特定属性或要求的 3D 分子结构，这对于药物发现至关重要。扩散模型特别成功地学习了 3D 分子几何的复杂概率分布及其对应的化学和物理属性，通过正向和反向扩散过程。这篇综述侧重于针对 3D 分子生成的扩散模型的技术实现。它比较了用于分子生成任务的各种扩散模型的性能、评估方法和实现细节。我们涵盖了原子和键的表示策略、反向扩散去噪网络的架构，以及生成稳定 3D 分子结构所面临的挑战。这篇综述还探讨了扩散模型在 *de novo* 药物设计和计算化学相关领域中的应用，例如基于结构的药物设计，包括目标特异性分子生成、分子对接和蛋白质-配体复合物的分子动力学。我们还讨论了在物理属性上的条件生成、构象生成和基于片段的药物设计。通过总结用于 3D 分子生成的最先进的扩散模型，这篇综述阐明了它们在推动药物发现中的作用及其当前的局限性。

Date: Oct 27, 2024

1 Introduction

生成模型在现代药物发现中与分子科学的集成越来越多，以帮助开发新的治疗药物，如小分子?、抗体?、基因治疗和 mRNA 疫苗?。小分子的生成以字符串形式已经得到充分建立?。然而，3D 分子结构的生成仍然滞后，主要是由于分子形状的复杂性和任何模型对 E(3) 和 SE(3) 等变性要求的限制，这些要求使得分子在旋转和位移下保持不变?。3D 几何形状是决定分子的电化学性质的主要因素，从而决定其药理学、药代动力学、代谢、毒性和免疫反应，因为它决定了分子与生物靶点口袋以及各种生物分子（如酶、受体、抗体等）的相互作用。因此，在 3D 空间中学习分子可以改善在结构基础药物设计应用中对化学空间的探测，以寻找可能的蛋白质和 DNA 配体。此外，它还可以通过对分子的特定结构或量子力学属性进行条件化，帮助推进材料发现。

历史上，计算分子设计或 *de novo* 药物设计（DNDD）最初是通过较简单的方法，如生长和进化算法? 进行的。然而，随着深度生成模型的进步，它们已经占据主导地位，并被用于生成 1D、2D，最近则是 3D 空间中的分子。为此，采用了多种深度学习（DL）架构，包括递归神经网络（RNN）、变分自编码器（VAE）、强化学习（RL）、生成对抗网络（GAN）、卷积神经网络（CNN）和图神经网络（GNN）。RNN 通常与基于文本的表示（如 SMILES 和 SELFIES）一起使用，以 “自回归” 方式预测表示分子的序列的下一个标记。RNN 还用于生成图网络中的分子图，如 GraphNet ?、MolRNN ? 和 MRNN ?，这些基于序列的模型通常与更复杂的架构（如 RL、VAE 和 CNN）结合使用?。此外，基于 VAE 的方法，如 GraphVAE ?、CGVAE ?、NVAE ?，基于 GAN 的模型，如 ORGAN ?，基于 GNN 的模型，如 GraphINVENT ?，以及基于流的模型，如 GraphNVP ?、GraphAF ?、MoFlow ? 和 GraphDF ? 等也得到了应用。

生成邻接矩阵的 2D 图和在某些情况下 3D 原子坐标。为在 3D 空间中生成分子，开发了几种特定的算法，包括自回归模型和基于图表示的模型，这些模型可以从变分自编码器（VAE）、等变正则化流或扩散过程中解码，并在 Bailiff 等人? 最近发表的综述中得到了很好的总结。其他综述则集中于截至 2022 年 5 月的这些模型的实现细节?；然而，鉴于扩散模型在生成 3D 分子结构方面的快速进展，许多具有前所未有结果的模型未被包含在该综述中。

受到非平衡统计物理的启发，扩散模型首次由 Sohl-Dickstein 等人? 在 2015 年引入，以通过正向和反向过程学习复杂的概率分布?。然而，它们在 2020 年因在图像生成中取得前所未有的结果而变得非常流行?，随后许多研究跟进以改进这些结果???，包括 Song 等人提出的基于评分的方法?。扩散模型也在图生成中变得流行，并被用于计算生物学和化学的几个应用，如构象生成、分子对接、蛋白质生成与建模、蛋白质-配体复合物结构预测、分子生成和分子动力学??。扩散模型首次用于 3D 分子生成的是 E(3) 等变扩散模型（EDM）?，他们使用最初为判别任务开发的 E(n) 等变图神经网络（EGNN）来去噪和学习分子结构分布?。随后，它们被广泛应用于 3D 分子生成任务，并结合 GNN 或基于变换器的模型进行编码和学习分子结构?。

在这篇综述中，我们旨在总结用于 3D 分子生成的扩散模型的技术实现方面，尤其是在药物发现中的具体应用，如 *de novo* 药物设计。文献中有几篇关于深度生成模型在 *de novo* 药物设计中的应用的综述，其中 Xie 等人? 最近发表的综述就是一个非常全面的例子。还有几篇关于扩散模型在生物信息学和计算生物学中的应用的报告??。然而，文献中缺乏深入探讨最近开发的扩散模型并比较它们的性能、评估方法和实现细节的综述。一项相关的调查覆盖了截至 2023 年 4 月的扩散模型在化学中的应用，包括药物发现?。他们还涵盖了抗体设计、蛋白质设计和材料设计应用以及该领域的新兴挑战。在这篇综述中，我们深入探讨了原子和键的表示与编码，以及这些模型如何旨在同步学习，以避免生成具有不一致原子和键结构的不稳定分子。我们涵盖了用于前向扩散的不同策略、用于反向扩散的架构，以及这些模型在药物发现过程中的各种应用。使用扩散模型的分子生成过程的概述见图 1。

2 扩散模型

扩散模型是概率生成模型，它向数据添加噪声以扭曲数据，然后反转该过程以生成样本。目前，扩散模型研究围绕三个主要公式展开：去噪扩散概率模型（DDPMs）、基于评分的生成模型（SGMs）和受随机微分方程启发的模型

（评分 SDEs）。对扩散模型的研究集中在改善这三种公式的多个方面，例如更快和更高效的采样、准确的似然和密度估计，以及处理具有特殊结构的数据（例如，置换不变性、流形结构、离散数据）?。用于分子生成和 DNDD 应用的最流行公式是 DDPM，但也使用了其他公式?。图 2 展示了应用于 3D 分子图的扩散过程的示意图。

2.1 去噪扩散概率模型（DDPMs）

去噪扩散概率模型（DDPM）?? 包含两个马尔可夫链：一个前向链将数据转化为标准高斯噪声，另一个反向链通过学习由深度神经网络参数化的去噪变换将噪声转化回数据。通过从相同分布中采样随机向量，可以生成新分子，随后使用反向链进行祖先采样。正式地，给定数据分布 $x_0 \sim q(x_0)$，前向马尔可夫过程将使用高斯扰动转移核 $q(x_t|x_{t-1})$ 逐步将数据分布 $q(x_0)$ 转化为可处理的先验分布 $q(x_T)$，其中 T 是时间步骤的数量，如下所示：

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I) \tag{1}$$

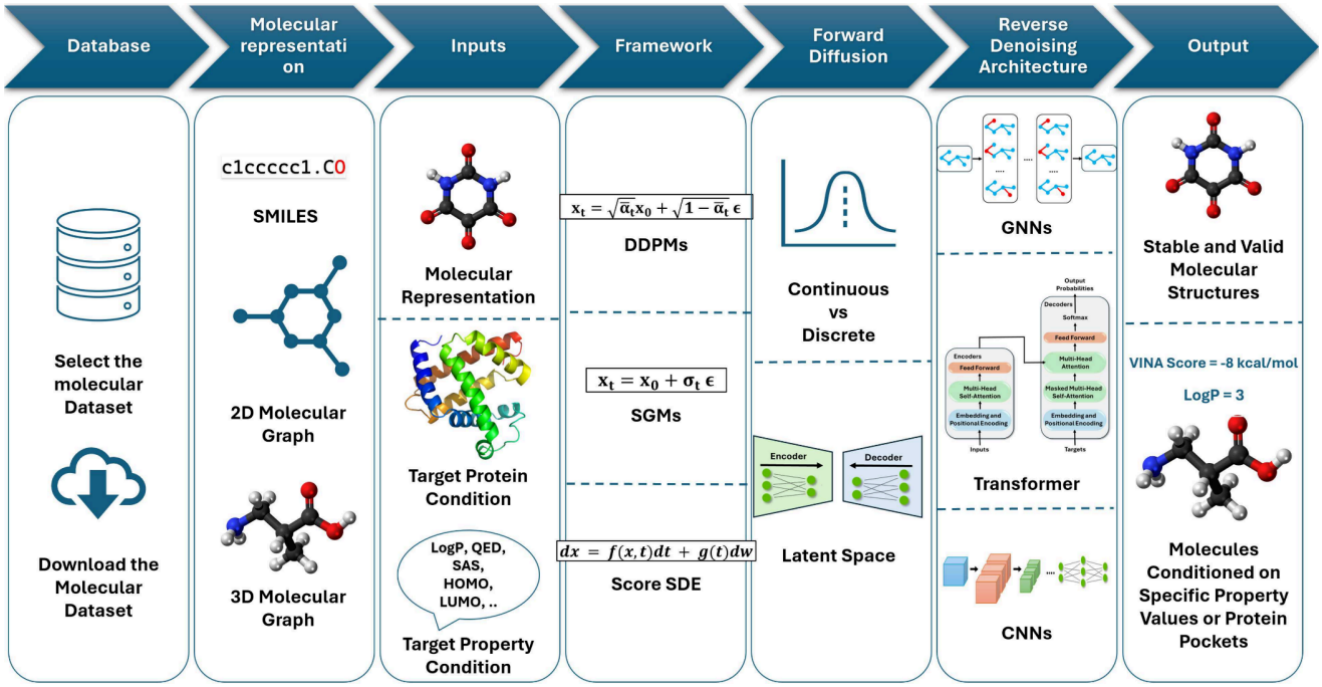

图 1: 使用扩散模型生成分子的过程概述。首先，获取相关数据集，以适当的分子表示表达分子，并确定扩散条件。接下来，选择扩散框架（DDPM、SGM、评分 SDE），并设计正向和反向扩散策略。去噪架构可以包括变换器、GNN、CNN 和混合架构。输出结果生成后，使用多个评估指标根据药物发现过程中的特定任务对生成的分子进行评估 ?。

该过程随后是一个可学习的反向去噪链，它反转时间步骤并逐渐学习检索数据。反向链采用先验分布 $p(x_T) \sim \mathcal{N}(0, I)$ 和可学习的转移核 $p_\theta(x_{t-1}|x_t)$，如下所示：

$$q(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t)x_0, \Sigma_\theta(x_t, t)) \tag{2}$$

其中 θ 表示模型参数，均值 $\mu_\theta(x_t, t)$ 和方差 $\Sigma_\theta(x_t, t)$ 由深度神经网络参数化。通过从高斯分布 $p(x_T)$ 中采样，然后从方程 3 中的核迭代采样，便可生成新样本，直到达到 $t = 1$。通过最小化前向分布 $q(x_0, x_1, \ldots, x_T)$ 与反向分布

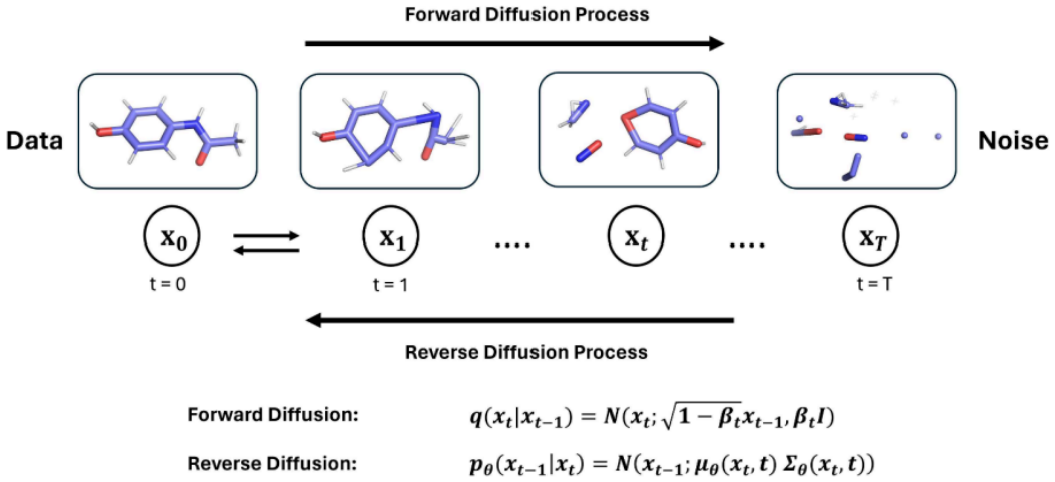

图 2: 应用于 3D 分子的扩散过程概述。在正向扩散过程中，通过从分布 $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$ 中采样，逐渐向分子添加噪声，其中 $\beta_t \in (0,1)$ 是模型训练前指定的超参数，I 是单位矩阵，$t \in \{1,2,\ldots,T\}$ 是时间步骤。为了生成分子，从标准正态噪声 x_T 开始，迭代地从分布 $p_\theta(x_{t-1}|x_t)$ 中抽样。这些分布由预训练的去噪神经网络学习。

$p_\theta(x_0, x_1, \ldots, x_T)$ 马尔可夫链之间的 Kullback-Leibler（KL）散度来学习神经网络的参数 θ，这等价于最大化数据 x_0 的对数似然的变分下界（VLB），如方程 4 所示。

$$\mathbb{E}[-\log p_\theta(x_0)] \leq KL(q(x_0, x_1, \ldots, x_T)||p_\theta(x_0, x_1, \ldots, x_T)) = -\mathbb{E}_{q(x_0,x_1,\ldots,x_T)}[\log p_\theta(x_0, x_1, \ldots, x_T)] + \text{const} = -\mathbb{E}_{q(x_0,x_1,\ldots,x_T)}[\tag{3}$$

Ho 等人**?** 提出了通过优化神经网络以预测高斯噪声来替代 $\mathcal{L}_{\text{VLB}}$ 损失，如方程 5 所示。

$$\mathbb{E}_{t\sim U[1,T], x_0\sim q(x_0), \epsilon\sim\mathcal{N}(0,I)}[\lambda(t)||\epsilon - \epsilon_\theta(x_t, t)||^2] \tag{4}$$

其中 $\lambda(t)$ 表示正权重函数，x_t 可以计算为 $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$，$\epsilon_\theta$ 是将被训练以估计高斯噪声的神经网络，$U[1,T]$ 是在集合 $\{1,2,\ldots,T\}$ 上的均匀分布。

DDPM 已在多种分子扩散模型中采用，如 EDM **?**、MiDi **?**、GCDM **?**、GeoDiff **?** 以及其他几种模型**???**。

2.2 基于评分的生成模型（SGMs）

（Stein）评分的概念（也称为评分或评分函数）是基于评分的生成模型的关键量。它被定义为对数概率密度的梯度 $\nabla_x \log p(x)$ **??**。基于评分的生成模型（SGMs），也称为噪声条件评分网络（NCSN）**?**，使用一系列增加的高斯噪声来扰动数据，然后训练一个基于噪声水平的深度神经网络模型来预测评分函数。

为了正式描述 SGMs，我们假设有一个数据分布 $q(x_0)$，以及一系列增加的噪声水平 $0 < \sigma_1 < \sigma_2 < \cdots < \sigma_T \cdots < \sigma_T$。数据将通过高斯噪声分布 $q(x_t|x_0) = \mathcal{N}(x_t; x_0, \sigma_t^2 I)$ 从 x_0 扰动到 x_t，以获得嘈杂数据密度 $q(x_1), q(x_2), \ldots, q(x_T)$，其中 $q(x_t) := \int q(x_t|x_0)q(x_0)dx_0$，并且评分函数 $\nabla_x \log q(x_t)$ 将通过一个噪声条件深度神经网络 $s_\theta(x,t)$ 进行估计。

因此，目标损失函数变为：

$$\mathbb{E}_{t\sim U[1,T],x_0\sim q(x_0),x_t\sim q(x_t|x_0)}[\lambda(t)\sigma_t^2||\nabla_x \log q(x_t) - s_\theta(x_t,t)||^2] \tag{5}$$

$$=\mathbb{E}_{t\sim U[1,T],x_0\sim q(x_0),x_t\sim q(x_t|x_0)}[\lambda(t)||\frac{x_t - x_0}{\sigma_t} - \sigma_t s_\theta(x_t,t)||^2] + \text{const} \tag{6}$$

类似于方程 5，$\lambda(t)$ 是一个正权重函数，x_t 可以计算为 $x_t = x_0 + \sigma_t\epsilon$。通过比较方程 5 和方程 6 中的损失函数，我们可以看到 DDPM 和 SGM 的训练目标是等价的。对于采样，从高斯噪声开始，SGMs 使用 $s_\theta(x_t,T), s_\theta(x_{T-1},T-1),\ldots,s_\theta(x_0,0)$ 的顺序链来生成新的数据实例。退火朗之万动力学（ALD）是 SGMs 中最常用的采样生成方法之一，但也研究了其他方法，如随机微分方程、常微分方程及其与 ALD 的组合 **?**。

2.3 随机微分方程（Score SDEs）

评分 SDEs **?** 是 DDPM 和 SGM 的扩展，以包括无限的时间步骤或噪声水平，涉及扰动和去噪过程，并且它们涉及求解随机微分方程（SDEs），在其中 SDE 用于噪声扰动和样本生成，而去噪是通过估计噪声数据分布的评分函数。基于评分的 SDE 扩散中的数据扰动由以下 SDE 控制：

$$dx = f(x,t)dt + g(t)dw \tag{7}$$

其中 $f(x,t)$ 是 SDE 扩散函数，$g(t)$ 是 SDE 漂移函数，w 定义为标准维纳过程或布朗运动。DDPM 和 SGM 都是此 SDE 在时间上的离散化，DDPM 的 SDE 可以表示为**?**：

$$dx = -\frac{1}{2}\beta(t)xdt + \sqrt{\beta(t)}dw \tag{8}$$

任何形式为方程 7 的扩散过程都可以通过求解以下 SDE 来反转**?**：

$$dx = [f(x,t) - g(t)\nabla_x \log q_t(x)]dt + g(t)dw \tag{9}$$

其中 w 表示反向时间扩散过程中的标准维纳过程，dt 是朝相反时间方向的微小负步长。反向 SDE 的解具有与正向 SDE 相同的边际密度，但时间方向相反。

在每个时间步骤 t 估计评分函数 $\nabla_x \log q(x)$ 后，反向时间 SDE（方程 9）被求解，并可以使用包括数值 SDE/ODE 求解器在内的数值方法进行采样 **????**，退火朗之万动力学 **?** 以及与这些方法的 MCMC 组合（预测-校正方法）**?**。与 SGMs 类似，时间依赖的评分神经网络 $s_\theta(x,t)$ 被训练以估计每个时间步骤的评分函数，使用方程 10 中的目标函数，该方程是对方程 6 在连续时间上的推广，

$$\mathbb{E}_{t\sim U[0,T],x_0\sim q(x_0),x_t\sim q(x_t|x_0)}[\lambda(t)\sigma_t^2||s_\theta(x_t,t) - \nabla_x \log q_t(x_t|x_0)||^2]. \tag{10}$$

这里 $U[0,T]$ 表示在 $[0,T]$ 上的均匀分布。类似于方程 5 和方程 6，$\lambda(t)$ 是一个正权重函数，θ 是神经网络的可训练参数。

基于评分的模型，包括 SGMs 和评分 SDEs，已在多个用于分子生成的扩散模型中得到应用，例如 DiffBridges **?**、GDSS **?** 和其他几种模型 **???**。

3 分子表示

多年来，有多种方式表示分子，例如库仑矩阵、分子指纹、键袋、国际化学标识符（InChI）、简化分子输入行系统（SMILES）和图 **?**。然而，扩散模型只广泛报道了与 SMILES、2D 和 3D 图相关的应用，并在分子图生成中取得

了前所未有的成果 **??**。因此，我们将重点讨论这些分子数据表示。此外，常用于使用扩散模型进行分子生成的数据集概述总结在表 1 中。

SMILES 是一种将分子结构转换为一维符号字符串的表示法。尽管基于序列的自回归模型（如 RNN）在基于文本的表示（包括 SMILES 和 SELFIES）中取得了成功，但扩散模型的研究探索了在更复杂和定制的应用中应用 SMILES。例如，DIFFMOL 模型 **?** 将扩散模型与变换器架构相结合，以标记 SMILES 并生成具有指定框架和属性的分子。

分子图表示已成为生成模型中广泛使用的表示法，其中分子结构表示为图 $G = (V, E)$，其中节点 $v_i \in V$ 表示原子，边 $(v_i, v_j) \in E$ 表示键。原子相互作用通常在 2D 及以上表示，在 2D 图中，键作为边表示，原子类型作为节点特征表示。在 3D 图中，分子在 3D 空间中表示，原子类型和 3D 位置都作为节点特征进行编码。在这种情况下，3D 生成所需的既有连通性也有坐标。能够显式学习 3D 原子坐标的扩散模型可以隐式建模完全连接图的 2D 和 1D 结构表示，如 EDM 模型 **?**。原子类型和键类型通过某种嵌入类型进行编码，其中每条边表示一个原子或键类型。

表 1: 使用扩散模型进行分子生成的常用数据集

数据集	**分子数量**	**数据**	**任务**
QM9 **?**	133,885	包含最多九个原子的有机分子及其 DFT 计算的量子化学属性。	无条件和有条件的 3D 分子生成。
GEOM-DRUG **?**	超过 450,000	37 万种分子的 3700 万个构象。	无条件和有条件的 3D 分子生成。
ZINC250k **?**	249,455	ZINC 数据库中具有生物活性数据的药物类似子集。	药物类似分子的无条件和有条件生成。
MOSES **?**	1,936,962	从 ZINC 清洁主链集合中过滤。	无条件和有条件的 2D 分子生成。
CrossDocked **?**	18,450 个复合物，2260 万个姿势	一组配体-蛋白质复合物，其中配体与几个相似的蛋白质靶标对接。	SBDD 任务：目标感知 3D 生成、分子对接等。
PDBbind **?**	23,496 个复合物	在 PDB 中具有实验测量结合亲和力的生物分子复合物，包括 19,443 个蛋白质-配体、2,852 个蛋白质-蛋白质、1,052 个蛋白质-核酸和 149 个核酸-配体复合物。	蛋白质-蛋白质对接和 SBDD 任务：目标感知 3D 生成、分子对接等。

4 扩散模型在分子图生成中的基本要求

E(3) 不变性和 SE(3) 等变性：当模型对分子的 3D 结构的平移、旋转和反射不变时，称该模型为 E(3) 等变。与此同时，作为一个群体，SE(3) 对旋转和平移不变，这意味着它们对手性敏感，从而改变分子的 3D 几何形状**?**。基于图的表示的主要优势在于它们可以通过与几何图神经网络（GNNs）结合，例如 EGNN **?** 和在 MiDi 模型中引入的 eFGNN **?**，轻松满足 E(3) 等变性要求，以实现旋转不变性，并通过将质心设置为零来对分子进行中心化，以实现平移不变性。

置换不变的图生成：图是置换不变的，这意味着它们在置换操作（如更改邻接矩阵中的行或列顺序，或与分子图表示 $G = (V, E)$ 中的原子对应的节点和与键对应的边的顺序）下保持不变。因此，置换不变性是任何基于图的生成（包括分子生成）的主要要求之一。扩散模型克服了置换不变图生成的问题，超越了显示对生成节点序列依赖的自回归模型**?**。

考虑离散性：图表示的另一个关键问题是原子类型和键类型特征是离散的，这使得使用高斯噪声扩散来训练去噪神经网络以学习分子结构的分布变得具有挑战性。将在第 5 节和第 6 节中讨论应用于分子图的各种正向和反向扩散方法。

捕捉基础数据分布：生成模型需要学习数据的基础分布，以生成逼真的样本，并准确生成在 2D 化学空间和 3D 空间中的构象上表示训练分布的变体。分子图 (2D 结构) 决定了原子和键类型的分布以及化合物中的骨架和功能基团，3D 构象则表示更复杂的分布，例如键角、二面角和立体异构的情况。

生成样本的保真度：确保生成分子的有效性可以有多个层面，包括整个分子图的连通性、化学稳定性，以及遵循已建立的分子结构规则，如原子价、基于元素在周期表中的位置的可能离子电荷、稳定的环结构，以及可以根据基础数据分布变化的可接受的角度和扭转应变水平。例如，一组药物类似分子的基础分布应该与在化学合成过程中出现的更具反应性的过渡态有不同的基础分布。在第 6 节中，我们将讨论如何使用不同的架构来提高生成分子的保真度。

5 正向扩散过程（离散与连续）

在正向扩散中，通过逐步注入噪声来破坏数据，直到其达到标准高斯噪声。由于原子和键类型的分类变量性质，它们在经过独热编码后被表示为分子图中的离散特征。假设我们有一个分类数组 h 表示类别 $c_1, \ldots, c_k$。在这种情况下，可以使用独热函数将其编码为数组 h^{onehot}，即 $h_{i,j}^{\text{onehot}} = 1$，然后在正向扩散中可以对数组 h^{onehot} 应用噪声。然而，将高斯噪声应用于这些特征并去噪图以维持生成数据中这些类别的相同分布可能会很棘手。

尽管如此，几种模型将高斯噪声应用于离散特征，包括 EDM **?**，它对独热编码向量应用连续扩散，使用预定义的噪声缩放调度 $q(h_t^{(l)}|h) = \mathcal{N}(z^{(l)}; a^{(l)}, \sigma_t^2 I)$，其中概率分布参数 p 被定义为与从 $-\frac{1}{2}$ 到 $+\frac{1}{2}$ 积分的正态分布成比例：

$$p(h_i^{(l)}) = C(h|p), \quad p \propto \int_{-\frac{1}{2}}^{+\frac{1}{2}} \mathcal{N}(u; z_i^{(l)}, \sigma_0) du \tag{11}$$

这里 $C(h|p)$ 是一个分类分布，而 p 被归一化以使其和为 1。这确保在 $h_0^{(l)}$ 的每一行中有一个类别是活动的，实际上，正态噪声分布的参数经过调优，以确保从反向扩散过程中采样的类别与原始活动的原子或键类别相匹配。

DiffMol **?** 遵循了类似的正向扩散过程。然而，他们在原始元素空间中添加了一个非类型，所有原子类型或键类型将逐渐被扰动到这个新类型中。他们称之为吸收类型，因为原子或键类型在正向扩散过程结束时逐渐被吸收到这个特定类型中。因此，这个扰动后的概率分布将成为反向过程的起始点。此外，正向扩散分为两个阶段，第一阶段专注于扰动键类型，第二阶段专注于原子类型和坐标。

其他模型，如 GDFMIDI **?**，应用具有可学习参数的高斯噪声来控制噪声的强度。类似地，JODO **?** 通过可学习的正弦位置嵌入表示噪声，然后将其用作去噪过程中的条件特征。许多研究更倾向于使用连续扩散，因为它提供无分类器引导，即不需要显式分类器来引导生成过程，从而导致更高效的训练和采样。连续扩散还可以提供可解释性，并基于先进的 ODE 求解器提供更快的扩散算法 **??**。

另一方面，其他研究表明，离散图扩散可以生成更高质量的样本，其分布更接近原始数据分布，这由真实和生成分布之间较低的最大均值差异（MMD）值所表明。MMD 是一种用于比较概率分布的距离度量，正式讨论在下面的评估指标部分。离散图扩散也在一些分子图模型中引入，例如 DiGress 作为基于 DG 的模型，以在扩散过程中更好地边际化节点和边类型的分布。简单来说，在 DiGress **?** 中，离散正向扩散被定义为对每个节点和边应用的分类扰动，使用转移概率矩阵。给定一个独热编码的节点（原子）矩阵 $X \in \mathbb{R}^{a \times c}$，其中 a 是节点（原子）的数量，c 是原子类别的数量，以及一个独热编码的边张量 $E \in \mathbb{R}^{a \times a \times b}$，其中 b 是边类型的数量，包括作为特定类型的无边。在这种情况下，转移概率可以由转移矩阵 $\mathbf{Q}_v$ 和 $\mathbf{Q}_e$ 定义，如下所示：

$$q(G^{t-1}|G^t) = (X^{t-1}\mathbf{Q}_v, E^{t-1}\mathbf{Q}_e) \tag{12}$$

通过从分类分布中对每个节点和边类型进行采样：

$$q(G^{t-1}) = (X^{t-1}\mathbf{Q}_v, E^{t-1}\mathbf{Q}_e) \quad \text{和} \quad q(G|G') = (X\mathbf{Q}_v, E\mathbf{Q}_e) \tag{13}$$

这里 $\mathbf{Q}_v = Q_v^1 \dots Q_v^t$，$\mathbf{Q}_e = Q_e^1 \dots Q_e^t$。相同的离散正向扩散过程也被后续模型 MiDi ? 使用。然而，它与连续扩散相结合，用于被高斯噪声扰动的 3D 坐标。此外，MiDi 遵循了一种自适应噪声调度，专门调整以使模型在去噪过程中先预测键类型和原子坐标，然后再预测原子类型 ?。

5.1 潜在图扩散

一些模型倾向于将数据嵌入连续空间以应用稳定扩散，例如 GEOLDM ?，它使用几何自编码器将分子结构转换为 3D 等变潜在特征，然后在潜在空间中应用稳定扩散。另一个名为 3M-Diffusion ? 的模型使用图编码器将分子图编码到与其对应文本描述对齐的连续空间（潜在空间），然后采用解码器根据给定的文本描述检索分子图。因此，该模型可以同时学习分子结构及其文本描述，并根据给定的文本描述生成新分子。潜在空间扩散的示意图如图 3 所示。

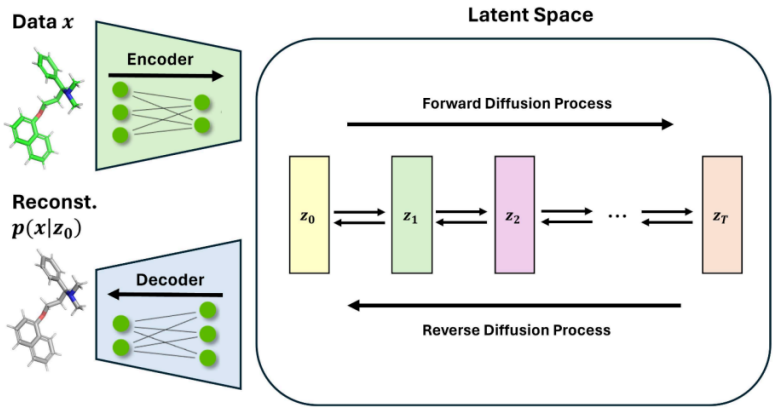

图 3: 潜在空间扩散过程概述。首先，分子被编码到一个连续的潜在空间，然后在潜在空间中应用稳定扩散。为了生成分子，首先从潜在空间中采样，然后使用解码器检索回原始离散空间 ?。

6 反向扩散去噪神经网络架构

反向扩散负责使用去噪神经网络架构学习数据的分布，该网络逐渐去除在正向扩散中添加的噪声。针对分子生成的去噪神经网络架构的几种架构已被研究，这些架构可以归入以下三类之一：变换器、GNN 和 CNN（图 4 和表 2），或它们的组合。EDM 还成功地通过将条件作为输入来生成基于属性 c 的分子，从而使 EGNN 能够学习。

6.1 图神经网络（GNNs）

图神经网络（GNNs）? 是一种专门设计用于处理图形结构数据的深度学习架构，通过节点之间传递消息，使每个节点能够理解其邻居和更广泛的网络。因此，它们可以在分子的图表示上操作，并已广泛应用于预测分子属性和深度生成分子模型。不同模型中也使用了几种变体的 GNN，如通过随机微分方程系统的图扩散（GDSS）?、E(n) 等变图

神经网络（EGNNs）?、几何完整感知机神经网络（GCPNET）? 和 ShapeMol ?。GNN 也被用于在 3M-Diffusion 模型 ? 中对分子的离散图结构进行编码，在该模型中，它们使用图同构网络（GINs）? 将分子映射到连续潜在空间。

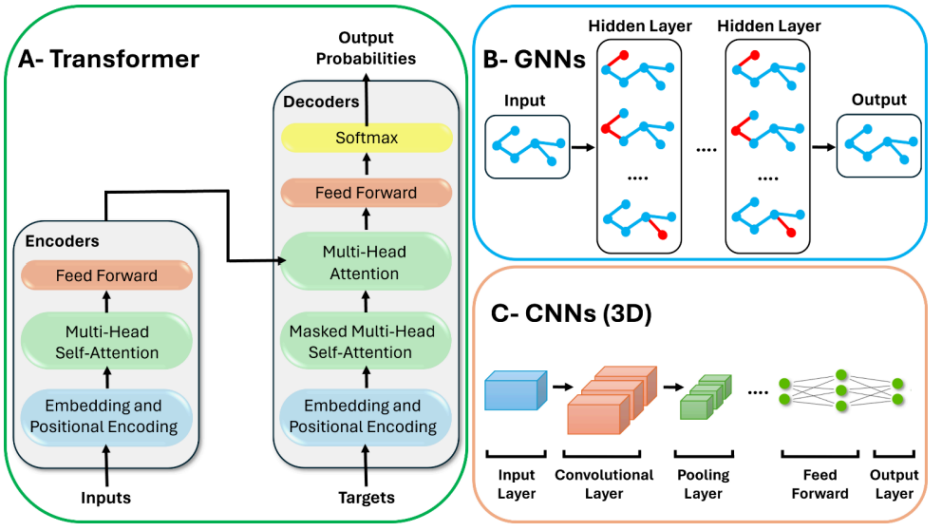

图 4: 反向扩散中常用的神经网络架构的简单示意图：A- A 变换器，B-GNN，C-CNN（在 3D 中）。

6.1.1 EGNNs

EGNNs 模型最初是为判别任务开发的，是在使用扩散模型进行分子生成时最流行的 GNN 架构 ?。它最早在 EDM 模型中用于通过消息传递生成 3D 分子，通过几个等变卷积层（EGCL）? 实现。在 EDM 模型中，EGNN 使用了一个完全连接的图 G，其中节点 $v_i \in V$，每个节点 v_i 具有坐标 x_i 作为其特征，因此共价键并未明确说明，而是从 3D 坐标中推断出来。共价键的隐式表示导致了几个问题，因为模型无法学习不同原子类型的价和控制键形成及分布的相邻原子之间的约束关系。在大药物类似分子的数据库中，这个问题表现得更为明显，例如 GEOM-DRUGs ?，其中最多有 181 个原子，平均有 44.4 个原子（24.9 个重原子），生成的分子相比于最多有 29 个原子（9 个重原子）和每个分子平均 18 个原子的 QM9 数据集 ? 更不稳定。EDM 还成功地通过将属性 c 作为输入添加条件，从而使 EGNN 能够学习。

已经进行了多次尝试以改善依赖相同 EGNN 的 EDM 模型的各个方面，并进行了修改。例如，EEGSDE ? 使用 SDE 框架，并根据分子 z 和属性 c 之间的一致性添加了能量函数的引导，这有助于捕捉两个变量之间的依赖关系，并生成具有期望属性的分子。DiffBridges ? 采用了类似的方法，通过基于 Lyapunov 函数方法设计物理相关的扩散桥来实现。受 AMBER 力场 ? 和从数据计算的分子几何统计（如键长、键角、扭转角等）的启发，向扩散桥引入了能量函数。DiffBridges 在 QM9 数据集上的表现优于 EDM。然而，导致不稳定分子的原子类型和键的不一致性问题仍然存在。

MolDiff ? 模型的开发旨在解决 EGNNs 基于反向扩散中的原子-键不一致性问题，因此实施了两阶段扩散模型。在第一阶段，键类型被扰动，同时保持原子类型固定，以便所有原子略微结合。然后，模型通过去除噪声并在正向扩散结束时从原子类型恢复键，反转该过程。在第二阶段，模型继续仅对原子添加噪声，并在反向过程中检索原子类型，同时固定键类型。训练权重和梯度根据各自的原子-键邻接矩阵进行调整，直到所有键标记为吸收类型。在第二阶段，原子坐标被扰动，直到它们达到无类型的先验分布。MolDiff 生成的有效分子比例高于 EDM；然而，它没有报告生成的 3D 结构和原子价的稳定性。

HierDiff ? 旨在生成更高质量的药物类似分子，并采用完全不同的策略来解决局部环境中的原子-键不一致性问题，使用分层粗粒度生成模型，其中每个节点编码一个片段。首先，一个普通的 EGNN 获取所有片段的嵌入，然后节点被组装以使用自下而上或自上而下的 EGNN 生成 3D 分子，这些 EGNN 预测这些片段之间的连接。粗粒度节点随后通过另外两个 EGNN 模块解码为细粒度片段：一个消息传递的自下而上 EGNN 和一个迭代精炼采样的自下

而上 EGNN 模块。类似地，GEOLDM 使用 EGNN 模块实现编码器、潜在扩散和解码器 ?。HierDiff 和 GEOLDM 通过使用 EGNN 将结构嵌入到较小的域中，避免了原子-键不一致性问题。

6.1.2 GCPNET 与 GCDM

GCPNET ? 是另一种 SE(3) 等变图神经网络，设计用于判别任务。它在 3D 分子图上的 3D 几何依赖预测任务（如蛋白质-配体结合亲和力和手性识别）中表现优越，特别是在直肠/左旋（RS）3D 分子数据集 ? 上。它被重新用作 GCDM 模型 ? 中的生成任务。与 EGNNs 类似，通过称为 GCPConv 的几何完整图卷积层进行消息传递；然而，这些层之前是一个几何完整感知机嵌入层 GCP，用于将输入节点和边特征编码为标量和值向量。GCPConv 层后面还有另一个 GCP，输出每个 GCP 的嵌入。尽管 GCPNET 相较于 EGNN 有多个优势，例如其几何完整性和手性感知的消息传递，但在 GCDM 模型 ? 中未能改善较大分子（如 GEOM-DRUGs 数据集 ?）的 3D 稳定性。

6.1.3 ShapeMol

ShapeMol 根据分子的形状生成 3D 分子，使用两个 GNN：一种不变图神经网络（INV-GNN）来预测原子特征，另一种等变图神经网络（EQ-GNN）来预测原子坐标 ?。

6.1.4 SimGen 模型的 MACE

MACE 模型 ? 是一种等变消息传递神经网络（MPNNs）架构，旨在创建计算上高效且准确的基于机器学习的力场。因此，它用于从 SPICE ? 数据集中提取特征，该数据集包含 100 万个原子数量在 3 到 100 之间的分子，用于零样本分子生成模型 SimGen ?。QM 力场特征与时变局部相似性核相结合，以定义一个可以生成任意大小分子的基于评分的扩散模型，并执行条件生成而不改变模型。

6.1.5 LigandDiff

LigandDiff ? 是一个条件扩散模型，旨在生成 3D 过渡金属配合物，采用消息传递神经网络（MPNNs）进行去噪，并结合几何向量感知机（GVPs）将分子嵌入到分子表示中。LigandDiff 在固定上下文下有条件地生成分子；例如，重原子的数量和所使用的过渡金属可以用作条件。

6.2 卷积神经网络（CNN）

CNN 是深度神经网络，特别擅长处理组织成网格状结构的数据，如图像。CNN 通常用于计算机视觉应用，包括目标检测、图像分类和图像识别 ?。

6.2.1 MDM 模型中的 SchNet

SchNet ? 是一种连续滤波 CNN，学习分子中原子的表示，类似于图像中的像素，其中原子根据原子类型进行嵌入，具有三个高交互块建模原子间相互作用。该模型是等变的和旋转不变的，因为它使用原子间距离在固定向量网络中对分子建模。SchNet ? 被适配到 DPM 模型 ? 中，以生成结构中的 3D 分子。后来的基于 SchNet 的模型与双等变评分神经网络对齐输出网络合并来自两个通道的原子特征，并输出最终的分子。MUFormer 在 QM9 数据集 ? 上成功生成了有效和稳定的 3D 几何结构和属性预测任务。

6.2.2 DiGress

DiGress 采用离散去噪架构来检索 2D 分子图作为分类原子和键特征。去噪架构基于 ? 的图变换网络，使用 FiLM 层 ? 结合边特征和全局特征。类似于 MUFormer ?，理论结构特征如循环和谱特征，以及分子特征如每个原子的当前价和整个分子的当前分子量，用于增强分子的表示以改善去噪过程。

6.2.3 MiDi

MiDi 还采用了一种图去噪变换器架构，联合预测分子的 2D 和 3D 图。MiDi 允许分子原子保持正式电荷并成为离子；例如，一个碳原子可以持有-1、0 和 1 的电荷。去噪架构逐步学习预测分子结构，其中原子坐标和键类型首先被预测，然后是原子类型和正式电荷。MiDi 的去噪架构在其更新块中集成了放松的 EGNN（rEGNN）层。MiDi 使用边、节点、成对和全局特征通过主邻域聚合（PNA）层 ? 聚合到节点表示中。

6.3 混合架构

将来自各种架构的元素（如变换器和 GNN）结合起来，可以利用每种方法的优势来增强去噪性能。一个好的例子是基于离散图结构的条件扩散模型（CDGS）?，其中为去噪架构采用了混合消息传递块（HMPB）。HMPB 包括两种变体的消息传递层：一种称为 GINE ? 的 GNN 层，用于离散数据类型以聚合局部邻居节点-边特征，另一种基于全连接的注意力变换器 ? 用于全局图特征学习。

与 ShapeMol ? 类似，Diff-Shape ? 提出了一种名为图控制网（GrCN）的预训练 GNN，以使扩散过程以 3D 分子形状为条件。图控制网受到 ControlNet ? 的启发，该网络使用数十亿图像进行预训练，以引导文本到图像的扩散模型。同样，预训练的图控制网（GrCN）与 MiDi ? 的无条件预训练模型结合，以生成受 3D 形状约束的分子。虽然 MiDi ? 包含变换器架构，而 GrCN 是基于 GNN 的，但该模型可以被视为一种混合架构。Diff-Shape ? 还可以通过以与目标蛋白质的高对接分数预先对接的配体的 3D 形状为条件进行结构基础药物设计。

7 基于结构的药物设计（SBDD）的分子生成

使用扩散模型的分子生成可以利用蛋白质靶点结合口袋作为生成条件。在 3D 生成的情况下，模型还可以生成分子在口袋内的姿势（图 5 和表 3）。DiffSBDD ? 和 DiffBP ? 是第一个引入基于 SBDD 的扩散模型的模型，它们都使用等变的 EGNN 来建模分子；然而，DiffSBDD ? 的研究后来扩展以探索更多案例。目前，它包括两种不同的方法。

在第一种方法中，使用固定的口袋表示作为 3D 分子生成的条件，而在第二种方法中，联合建模配体-口袋复合物的联合分布，无条件近似。它们还研究了交互策略，其中分子可以在概率转移步骤中与相同的口袋保持交互。Impainting 允许对口袋系统的任意部分进行掩蔽、替换或固定。因此，它可以与基于评分的函数结合，以进行条件 3D 分子生成，通过固定部分分子并执行部分分子生成或片段生成。这允许通过探索局部化学空间来优化候选药物分子（线索），同时保持 3D 分子结构的部分，以及应用药物设计中广泛使用的技术，如片段生成、骨架跳跃、片段生长和连接体设计。

TargetDiff ? 是另一个以目标为导向的分子生成模型，采用图注意力层进行参数化。它研究了子结构或片段生成和结合亲和力预测的案例。另一个结合亲和力预测工具是 Virna Scores ?，其基于 CrossDocked2020 数据集 ?，以及一些最先进的非扩散模型，如 Pocket2Mol ? 和 GraphBP ?。然而，就其他药物设计指标而言，它们的改进不那么明显，例如 SBDD 和 ego4D 评分。

另一个 SBDD 模型 DecoMPOT ? 旨在通过结合在支架下分解两个部分的领域来探索 DECOMPOT [102]，以便更有效地进行药物设计与口袋氨基酸残基形成相互作用，支架连接所有臂以形成完整的分子。DECOMPOPT ? 使用

关于臂和支架的结构先验，包括从参考分子估计的先验和从目标结合位点的子口袋提取的口袋先验，这些口袋先验是通过 AlphaSpace ?? 得到的。他们表明，口袋先验可以增强生成分子的结合力，与 TargetDiff ? 相比。DECOMPOPT ? 通过使用 R-组优化和支架跳跃等技术，对局部子结构进行条件控制应用迭代和可控优化。例如，该模型用于在保持与目标的结合亲和力的同时提高生成分子的药物相似性指标，例如 QED 和 SAS，通过对分子臂进行条件设置，使其与结合口袋形成相互作用 ?。

Binding Adaptive Diffusion Models (BindDM) ? 是另一个模型，通过从蛋白质-配体复合物图中提取每个时间步的基本结合亚复合图，使用可学习的结构策略，自适应生成分子。在此过程中，复合图和其子复合图的两个层次通过两个交叉层次交互节点相互作用：复合物到子复合物（C2S）和子复合物到复合物（S2C）。BindDM ? 的性能与 TargetDiff ? 和 DecompOPT ? 相似。KGDiff ? 还整合了蛋白质-配体结合亲和力的化学知识，以引导每一步的去噪过程。与 TargetDiff ? 相比，它在 CrossDocked2020 数据集 ? 上的平均 Vina 评分提高了 46.2%。IPTDiff ? 应用了与 BindDM ? 类似的策略，其中一个蛋白质-配体相互作用先验网络（IPNET）经过预训练，以从化学性质和 3D 结构中学习配体-蛋白质相互作用。然后，预训练的 IPNET 作为先验来增强生成的扩散过程。他们还引入了两种先验策略：先验转移，其中基于 IPNET 学习的蛋白质-分子相互作用来转移扩散过程，和先验条件设置，其中扩散过程以先前估计的蛋白质-配体复合物为条件。

PROMPTDIFF ? 也是一个目标感知生成模型，使用一组选择的配体提示——即具有理想特征（如与目标的高结合亲和力、药物相似性和可合成性）的化合物——来引导生成过程，以产生满足设计要求的相似分子。该模型采用几何蛋白-分子相互作用网络（PMINet），提取有关蛋白-配体对之间相互作用的信息作为嵌入，可以用作引导扩散过程的提示。该模型提出了两种引导生成过程的方法：自引导，其中在每个时间步骤从生成的配体中提取嵌入，和示例引导，其中从具有所需属性的示例配体中提取的嵌入由经过预训练的 PMINet 用于引导反向生成过程 ?。

表 2: Summary of the denoising architectures used in diffusion models for molecular generation, and the datasets used for training those models.

Denoising Architecture	Model	Denoising Model	Condition	Framework	Datasets
11*GNNs	EDM ?	EGNNs ?	Conditioned, unconditioned	DDPM	QM9 ?, GEOM-Drugs ?
	DiffBridges ?	EGNNs ?	Unconditioned	SMLD[a]	QM9 ?, GEOM-Drugs ?
	EEGSDE ?	EGNNs ?	Conditioned, unconditioned	Score SDE	QM9 ?, QM9 ?
	GEOLDM ?	EGNNs ?	Conditioned, unconditioned	DDPM	QM9 ?, GEOM-Drugs ?
	MolDiff ?	EGNNs ?	Unconditioned	DDPM	QM9 ?, QM9 ?
	HierDiff ?	EGNNs ?	Conditioned	DDPM	GEOM-Drugs ?, Cross-Docked2020 ?
	GCDM ?	GCPNET ?	Conditioned, unconditioned	DDPM	QM9 ?, GEOM-Drugs ?, RS ?
	ShapeMol ?	INV-GNNs, EQ-GNNs	Conditioned	DDPM	MOSES ?
	GDSS ?	EGNNs ?	Unconditioned	Score SDE	QM9 ?, ZINC250k ?
	LigandDiff ?	MMPNs, GVPs	Conditioned	DDPM	Subset of Cambridge Structural Database (CSD) ??
3*CNNs	MDM ?	Schnet ?	Conditioned, unconditioned	SMLD[a]	QM9 ?, GEOM-Drugs ?
	VoxMol ?	3D U-Net ?	Unconditioned	SGM	QM9 ?, QM9 ?
7*Transformer	DiGress ?	EGNNs ?	Conditioned, unconditioned	DDPM	QM9 ?, GuacaMol ?, MOSES ?
	MiDi ?	EGNNs ?	Conditioned	DDPM	QM9 ?, QM9 ?
	DIFFUMOL ?	EGNNs ?	Conditioned	DDPM	GuacaMol?, MOSES ?, ZINC250k ?
	JODO ?	DGT	Conditioned, unconditioned	Score SDE	QM9?, QM9?
	MUDiff ?	MUformer	Conditioned, unconditioned	DDPM	QM9?
	GFMDiff ?	DTN	Conditioned, unconditioned	DDPM	QM9?, GEOM-Drugs ?
6*Other	CDGS ?	HMPB (hybrid)	Unconditioned	Score SDE	QM9 ?, ZINC250k ?
	Diff-Shape ?	hybrid	Conditioned	DDPM	GEOM-Drugs ?, PDB-Bind dataset ?
	SimGen ?	hybrid	Conditioned	Score SDE	SPICE ?
	3M-Diffusion ?	EGNNs ?	Conditioned	Score SDE	ChEBI-20 ?, PubChem ?, PCDes ?, MoMu ?
	GSDM ?	hybrid	Unconditioned	Score SDE	QM9 ?, ZINC250k ?

[a] SMLD: score matching with Langevin dynamics, a sub category of SGMs where Langevin dynamics are used

PMDM ? 采用了一种新颖的方法，将分子生成条件设定在局部和全局口袋氨基酸相互作用上。该模型采用 EGNN 对口袋-配体结构建模，其中口袋几何作为条件，SchNet 用于生成条件蛋白语义信息编码，并将配体原子特征嵌入到相互作用表示中，随后通过交叉注意机制融合两者。

分子超出分布扩散（MOOD）? 模型是一个基于评分的扩散模型，旨在生成具有理想化学属性的分子，如结合亲和力、药物相似性和可合成性。该模型训练一个单独的网络来预测特定属性，然后将属性预测器的梯度用于引导反向扩散过程。超出分布的引导通过一个超参数 $\lambda \in (0, 1)$ 控制，可以根据生成的主要目标适应不同的大小。MOOD ?

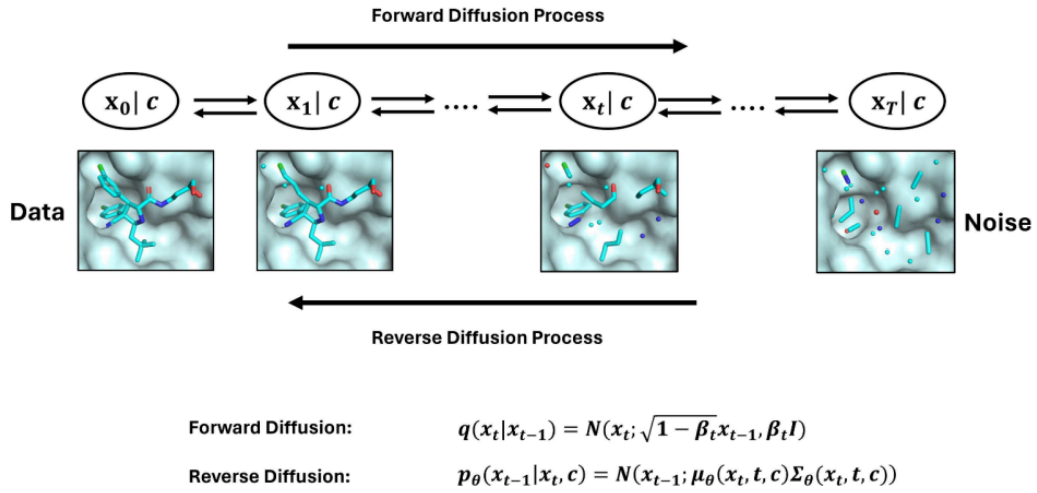

图 5: 使用扩散模型生成以蛋白质口袋为条件的 3D 分子。

能够生成同时满足多个约束的分子，这在生成化合物的高新颖前 5% 评分中得以体现（即前 5% 独特分子的平均值，QED $>$ 0.5 且 SA $<$ 5）。PILOT ? 也旨在通过重要性采样实现多目标条件。该模型将口袋条件与合成可及性（SA）引导相结合。首先，该模型在 ZINC 数据库中存在的 Enamine Real Diversity 子集上进行无条件预训练，然后在以口袋为条件的 CrossDocked2020 数据集 ? 上进行微调。最后，使用重要性采样作为推断，引导扩散过程朝向特定属性，如 SAS。

与 VoxMol ? 类似，一个基于分数的条件模型，VoxBind ? 被开发用于使用 3D 原子密度网格进行目标感知分子生成。VoxBind 应用了一个两步法，首先用 Langevin 动力学应用噪声配体，然后使用条件神经经验贝叶斯（NEB）去噪器 ?? 预测干净的分子。该模型在 CrossDocked2020 数据集 ? 上实现了成功的 Vina 评分，超过了 DiffSBDD ?、TargetDiff ? 和 DecompOPT ?。MolPainter ? 是另一个最近发布的模型，其中扩散过程由 3D 药效团引导，结合了与目标蛋白相互作用的重要特征，同时平衡了如结合亲和力和脂亲和接触等属性。

该模型使用了 MolDiff ? 进行具有片段的分子生成，但设计用于生成以各种子结构为条件的 3D 化合物。此外，一些模型被设计为使用基于片段的药物设计生成以各种子结构为条件的 3D 化合物，如 DiffLinker ?、FragDiff ? 和 AutoFragDiff ?，将在下一部分中描述。

8 其他应用

除了在 3D 空间生成分子，扩散模型还被开发用于药物设计过程中至关重要的其他应用，例如构象生成、分子对接、分子动力学以及基于片段的药物设计或连接器生成（图 6 和表 3）。

8.1 基于片段的药物设计和连接器设计

DiffLinker ? 是一个基于 EGNNs 的 3D 条件扩散模型，用于连接器生成。该模型应用了基于片段的药物设计的概念，它将一组不相连的片段作为条件并将其连接。连接器的附着位置及其大小，特别是所需的原子数量，可以由 DiffLinker 自动预测。与上述应用于连接器设计的 SBDD 模型（如 DiffSBDD ? 和 PMDM ?）类似，DiffLinker 也在连接器设计中使用了蛋白质口袋作为条件。

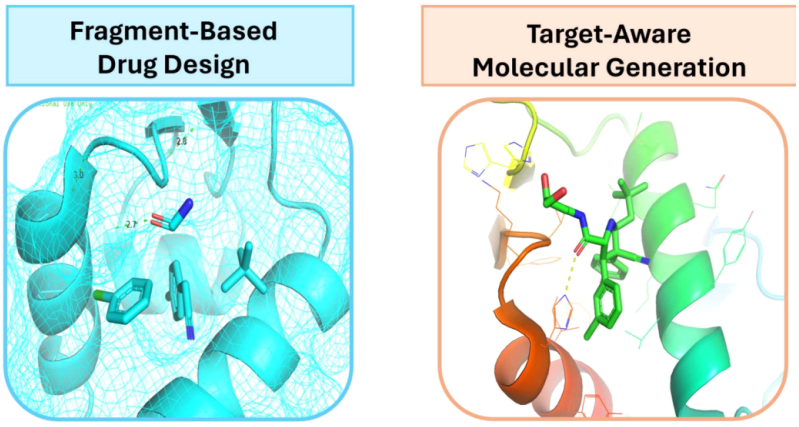

图 6: 扩散模型在药物发现过程中的一些应用。

FragDiff ? 是一个自回归模型，使用 E(3)-等变扩散生成模型逐片段地生成以蛋白质口袋为条件的 3D 分子。AutoFragDiff ? 采用类似的方法，但它使用几何向量感知器（GVPs）来预测以分子支架和蛋白质口袋为条件的分子片段。

选择性迭代潜变量精化（SILVR）? 模型生成以筛选出的与蛋白质结合位点的片段为条件的分子。首先，命中片段通过前向扩散过程，在每个时间步骤中添加高斯噪声，生成在不同噪声水平下的不同时间步骤的参考片段集。然后，该组噪声片段在反向扩散过程中用于条件化 EDM，以生成与参考片段相似的新样本。

8.2 构象生成

构象生成是药物设计中的一个关键步骤，它涉及预测分子可以采用的各种空间排列或构象。几个扩散模型被设计用于有效地对分子的构象空间进行采样。动态图得分匹配（DGSM）? 是一个基于得分的模型，旨在通过使用图神经网络（GNNs）建模局部和长程相互作用来预测平衡分子构象。分子图是使用全连接图构建的，具有截止距离，其中每个节点代表分子中的一个原子，并假设与构造截止距离所形成的球体内的所有原子相互作用。Geodiff ? 采用了一种称为图场网络（GFN）的等变卷积层来建模分子，其中每个原子被视为一个粒子，反向扩散过程训练一个马尔可夫链以生成分子的稳定构象。Geodiff...

超越了 DGSM ? 在 GEOM-DRUGs 数据集 ? 上的匹配和覆盖得分。扭转扩散 ? 通过仅在由扭转角定义的超圆环上操作，使用扭转玻尔兹曼生成器生成构象，并且扩散模型使用一种外在到内在的模型进行评分，该模型预测特定于分子的扭转分数（外在坐标），并以其在欧几里得空间中的构象的 3D 点云表示（内在坐标）作为输入。

SDEGen ? 旨在使用随机微分方程（SDE）模型对低能构象进行建模，该模型预测注入的高斯噪声。分子构象被表示为通过添加辅助键（如两跳边（1–3 角相互作用）和三跳边（1–4 二面角相互作用））扩展的分子图。多层感知机（MLPs）用于嵌入高斯噪声、原子间距离矩阵、高维空间和边缘信息并将它们加在一起，形成距离嵌入。然后，图同构网络（GINs）?，一种类型的 GNNs，将距离嵌入与原子特征嵌入结合，并通过更新距离嵌入执行反向去噪扩散过程，直到将其映射到高斯噪声的维度空间。SDEGen 在 QM9 ? 和 GEOM-DRUGs ? 数据集上与其他模型相比，报告了较高的覆盖率和匹配得分。

8.3 分子对接

分子对接旨在通过模拟两种分子之间的相互作用来预测配体在受体结合口袋中的最佳结合方向（姿态），然后根据这些相互作用对姿态进行评分。DiffDock ? 是一种基于 SDE 的得分生成模型（Score SDE），经过训练以预测配

体姿态。从随机姿态开始，反向扩散过程在平移、旋转和扭转角度上操作并采样多个姿态。另一种被称为置信模型的模型被训练来评分这些姿态，并根据估计最可能样本的置信得分对其进行排名。DiffDock 能够在 PDBBind 数据集 ? 上实现显著的准确性，检索出 38.2% 的配体姿态，其均方根偏差（RMSD）值 < 2 Å 和 63.2% 的配体，其 RMSD < 5 Å，超越了需要参数化的基于物理的评分函数和搜索算法（如 VINA ?）以及基于深度学习的方法（如 GNINA ?）的传统对接方法。DiffDock-PP ? 与 DiffDock ? 类似，但它执行刚性蛋白质-蛋白质对接。

PLANTAIN ? 是一种受物理启发的扩散模型，旨在通过最小化评分函数来预测和评分配体在蛋白质口袋中的姿态。该模型使用蛋白质结合口袋编码、2D 配体表示以及配体-配体和配体-受体的原子间距离，迭代地细化配体姿态并进行评分。PLANTAIN 能够预测 21.5% 的配体在蛋白质口袋中的姿态，其 RMSD 值 ≤ 2 Å，在 CrossDocked 数据集 ? 上同样超越了 VINA ? 和 GNINA ?。DiffEE ? 采用类似的方法，其中大规模蛋白质语言模型（PLM）对输入蛋白质序列进行编码，迭代过程更新 3D 分子图和采样的姿态。DiffEE 具有可比结果，

VINA ? 和 GNINA ? 在 PDBBind 基准 ? 上，针对 2 Å 和 5 Å 的 RMSD 截止值。PocketCFDM ? 使用结合口袋的数据增强技术，在配体分子周围创建多个人工结合口袋，以统计模拟实际蛋白-配体复合物中发现的非键合相互作用。PocketCFDM 的统计方法能够在 PDBBind 数据集 ? 上以 2 Å 的 RMSD 截止值实现 23.9% 的准确性，低于 DiffDock；然而，它在非键合相互作用、清晰的内涵和推理速度方面表现更好。

NeuralPLexer ? 是一种基于 SDE 的扩散模型，可以通过训练由等变结构去噪模块（ESDM）和粗粒度、自回归接触预测模型（CPM）组成的主网络，在原子分辨率下预测蛋白-配体复合物结构。给定输入的蛋白质语言模型（PLM）特征和结构模板，首先从输入的蛋白质序列中检索。接下来，使用输入配体的分子图表示和 PLM 及模板特征集合作为输入，NeuralPLexer 采样一组结合的蛋白-配体复合物确认。该模型在刚性盲蛋白-配体对接中能够超越 DiffDock，预测出高达 78% 的配体姿态，且 RMSD 值与参考 PDBBind2020 基准 ? 的差异为 < 2 Å。

神经欧拉旋转方程（NERE）? 是一种无监督的去噪得分匹配（DSM）基于能量的扩散模型，通过模拟配体和蛋白之间的原子间力和扭矩来预测旋转。该模型的对数似然被定义为配体-蛋白复合物的结合亲和力，并训练一个 SE(3) 等变旋转预测网络，其中力是关于原子坐标的能量函数的梯度。NERE 在 PDBBind 基准的晶体和对接结构上实现了结合亲和力与预测能量之间的皮尔逊相关值分别为 0.656 和 0.651 ?。

上述基于扩散的对接方法报告的性能与传统方法相比具有可比性或更优。然而，Yu 等人的研究 ? 认为，基于深度学习的模型在整个蛋白的口袋对接中更具优势，而传统方法在给定口袋上对接时表现更好。因此，当在对接之前执行口袋搜索时，与深度学习方法相比，传统方法在预定口袋中的对接效果更佳。此外，PoseBusters ? 发现，包括 DiffDock ? 在内的深度学习工具并未生成物理有效的姿态。然而，这些研究只涵盖了 DiffDock ?，而没有涉及后续的模型。随着基于扩散的模型的快速发展，超越传统模型在口袋搜索和给定口袋对接方面的潜力巨大。

8.4 分子动力学

分子动力学（MD）是一种计算建模技术，用于通过模拟原子和分子的物理运动来研究它们的时间依赖行为。它具有多种应用，例如理解药物分子与其目标蛋白之间的结合机制、蛋白质-蛋白质相互作用以及研究生物分子的动力学，例如蛋白质和核酸。DiffMD ? 是一种基于得分的扩散模型，旨在加速分子的 MD。DiffMD 采用等变几何变换器来计算分子构象的对数密度的梯度，其中原子的运动方向和速度由 3D 球面傅里叶-贝塞尔基函数表示。DiffMD 在 MD17 ? 和 $C_7O_2H_{10}$ 同分异构体数据集上超越了 EGNNs ? 等最新基准。另一种模型 ? 旨在使用去噪力场（DFF）作为基于得分的扩散模型来估计粗粒度（CG）力场。该模型成功地增强了各种蛋白质模拟的性能，适用于多达 56 个氨基酸系统。

表 3: Applications of diffusion models in drug design

Application	Model name	Condition/Guidance	Framework	Network architecture
13*Target-conditioned molecular generation	DiffSBDD ?	Protein pockets	DDPM	EGNNs ?
	DiffBP ?	Protein pockets	DDPM	EGNNs ?
	TargetDiff ?	Protein pockets	DDPM	EGNNs ?
	DECOMPOPT ?	Protein pockets	DDPM	EGNNs ?
	DECOMPDIFF ?	Protein pockets	DDPM	GNNs
	BindDM ?	Protein pockets	DDPM	Hybrid
	PROMPTDIFF ?	Ligand prompt set and target protein	DDPM	PMINet[a]
	IPDiff ?	Protein-ligand interaction prior	DDPM	Hybrid
	PMDM ?	Protein pockets	SGM	Hybrid (GNNs and GCN[b])
	MOOD ?	Chemical properties	Score SDE	Hybrid
	PILOT ?	Protein pockets and SAS	SGM	GNNs
	KGDiff ?	Protein-ligand binding affinity	DDPM	GNNs
	VoxBind ?	Protein pockets	SGM	3D U-Net ?
	MolSnapper ?	3D pharmacophores and Target protein	DDPM	EGNNs variant
4*Fragment-Based Drug Design and Linker Design	DiffLinker ?	Protein pockets	DDPM	EGNNs ?
	FragDiff ?	Protein pockets	DDPM	EGNNs ?
	AutoFragDiff ?	Target protein	DDPM	GVPs[c]
	SILVR ?	Hit fragments	DDPM	EGNNs ?
3*Conformations Generation	DGSM ?	Molecular graph	SGM	GNNs
	Geodiff ?	Molecular graph	DDPM	GFN[d]
	Torsional diffusion ?	Molecular graph	Score SDE	Hybrid
	SDEGen ?	Molecular graph	Score SDE	GNNs
8*Molecular Docking	DiffDock ?	Protein-ligand complex	Score SDE	GNNs variant (Hybrid)
	DiffDock-PP ?	Protein-protein complex	Score SDE	GNNs variant (Hybrid)
	PLANTAIN ?	Protein-ligand complex	SGM	GNNs variant (Hybrid)
	DiffEE ?	Protein sequence and ligand	DDPM	PLM and EGNNs ?
	PocketCFDM ?	Unconditioned	Score SDE	GNNs variant (Hybrid)
	NeuralPLexer ?	Protein-ligand complex	Score SDE	GNNs variant (Hybrid)
	NERE ?	Unconditioned	SGM	MPNNs[e]
2*Molecular Dynamics	DiffMD ?	Unconditioned	Score SDE	Transformer
	DFF ?	Unconditioned	Score SDE	Transformer

9 评估指标

评估用于分子生成的扩散模型可以包括根据任务不同而采用的各种策略。然而，生成分子的质量是最重要的方面，特别是在 3D 生成中，通常需要多层面的评估策略。有些评估指标可以用于任何分子生成，无论是 1D、2D 还是 3D，某些指标则专门针对 3D 分子图或药物样分子设计。本节将介绍在使用扩散模型进行分子生成时通常使用的评估指标。

9.1 分子生成的评估指标

有效性 - 有效分子的百分比，即在原子价和芳香环中键的一致性方面的化学正确性。该指标通常与 2D 分子图和 SMILES 一起使用，并通过 RDKit 进行评估。

新颖性 - 不包含在训练数据集中的分子百分比；该指标衡量模型生成数据集外分子的能力。

独特性 - 在一组生成的分子样本中，唯一且有效分子的百分比。

原子稳定性 - 具有正确原子价的原子百分比（通常用于 3D 分子生成）。

分子稳定性 - 所有原子均稳定的分子的百分比（通常用于 3D 分子生成）。

连通性/连接 - 连通分子的百分比，即具有单一连通分量的生成分子（通常用于分子图生成）。

9.2 相似性指标

距离度量在量化生成的分子数据分布与训练数据分布之间的相似性方面至关重要。在分子生成任务中，两种广泛使用的指标是最大均值差异（MMD）和杰森-香农（JS）散度。它们用于评估生成样本中键长、键角和二面角的分布与原始数据集的比较。

最大均值差异（MMD）- 一种度量两个概率分布之间距离的公式，通过将概率嵌入到再生核希尔伯特空间（RKHS）中来实现 **?**。给定两个分布 P 和 Q，以及相应的特征映射 $\phi(x)$ 和 $\psi(y)$，MMD 可以表示为：

$$\mathrm{MMD}^2(P,Q) = |\mathbb{E}_{x\sim P}[\phi(x)] - \mathbb{E}_{y\sim Q}[\psi(y)]|^2_{\mathcal{H}}$$

其中 $\mathbb{E}_{x\sim P}[\phi(x)]$ 和 $\mathbb{E}_{y\sim Q}[\psi(y)]$ 是从分布 P 和 Q 中抽取样本的特征映射的期望值，$\|\cdot\|_{\mathcal{H}}$ 是 RKHS 中的范数 **?**。

杰森-香农（JS）散度 - 一种度量两个概率分布之间相似性或差异性的指标，基于库尔巴克-莱布勒（KL）散度。给定两个分布 P 和 Q，它们之间的 JS 散度应为：

$$\mathrm{JS}(P||Q) = \frac{1}{2}\cdot\mathrm{KL}(P||M) + \frac{1}{2}\cdot\mathrm{KL}(Q||M)$$

其中 $M = \frac{1}{2}(P+Q)$ 是平均分布，KL 是库尔巴克-莱布勒散度，其形式为：

$$\mathrm{KL}(P||Q) = \int_{-\infty}^{\infty} P(x)\cdot\log\left(\frac{P(x)}{Q(x)}\right)\,dx$$

9.3 药物相似性指标

logP - 分配系数的对数。用于测量化合物在两种不相溶相之间的分布，通常是在有机相（通常是辛醇）和水相之间，常用于评估化合物的亲脂性，这是确定其吸收、分布、代谢和排泄（ADME）特性的一个重要方面。

QED - 药物相似性的定量估计。一个综合分数，用于评估化合物成为药物的可能性，它结合了多种落在已知药物范围内的主要分子属性。

SAS - 合成可及性分数。评估生成分子的合成可行性，计算为片段分数和复杂度惩罚的总和 **?**。

Lipinski 的五条规则测量分子是否满足药物相似性的五条规则，这是评估分子是否符合药物标准的一个宽松规则，分子应满足以下条件：最多五个氢键供体，最多十个氢键受体，分子量小于 500 Da，logP 值小于五。

9.4 基于属性的条件生成的评估指标

条件生成可以与任何属性进行。然而，通常与 QM9 数据集一起使用作为概念证明，其中属性分类网络，通常是 EGNN，在 QM9 数据集的一半上训练而模型则在另一半上训练，然后将属性值作为输入，要求采样具有该属性值的分子。输入属性值与生成分子的预测值之间的平均绝对误差（MAE）用于评估模型的条件生成能力。常见的 QM9 属性有：

- α - 立体极化率，以立方 Bohr 半径（Bohr^3）表示。
- μ - 偶极矩，以德拜（D）为单位。
- ϵ_{HOMO} - 最高占据分子轨道能量，以毫电子伏（meV）为单位。
- ϵ_{LUMO} - 最低未占据分子轨道能量，以毫电子伏（meV）为单位。
- $\Delta\epsilon$ - ϵ_{HOMO} 和 ϵ_{LUMO} 之间的差，以毫电子伏（meV）为单位。
- C_v - 298.15K 时的摩尔热容，以卡路里每开尔文每摩尔（$\mathrm{cal}\cdot\mathrm{mol}^{-1}\cdot\mathrm{K}^{-1}$）表示。

9.5 基于结构的生成评估指标

Vina 分数 - 一种评分函数，用于估计配体（小分子）与蛋白靶点之间在分子对接模拟中的结合亲和力，使用 Autodock vina **?**。

Vina Min - 与 Vina 分数相同的评分函数，但 Vina 平台在估计之前进行局部结构最小化 **?**。

Vina Dock - 与 Vina 分数相同的评分函数，但在评分之前重新对接分子，反映生成分子的最佳结合亲和力 **?**。

Vina 能量 - 使用 Autodock vina **?** 估计配体（小分子）与蛋白靶点之间结合亲和力的能量。

高亲和力百分比 - 当结合到靶蛋白时，具有低于参考（真实）分子 Vina 能量的化合物百分比。

多样性 - 通过计算每个目标口袋生成的所有分子之间的平均成对不相似性（1 - Tanimoto 相似性）来衡量生成分子的多样性。

9.6 构象生成评估指标

用于评估构象的评估指标基于均方根偏差（RMSD），可以通过使用 Kabsch 算法 **?** 计算两个重叠分子的原子化为公式计算如下：

$$\mathrm{RMSD}(\mathbf{C}, \hat{\mathbf{C}}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 + (z_i - \hat{z}_i)^2}$$

其中 $\hat{\mathbf{C}}$ 是预测的分子构象，具有一组 $i \in \{1, 2, \ldots, N\}$ 原子坐标 (x_i, y_i, z_i)，而 $\mathbf{C}$ 是参考分子构象，其对应的参考原子位置为 $(x_i^{\mathrm{ref}}, y_i^{\mathrm{ref}}, z_i^{\mathrm{ref}})$，$N$ 是原子数量。然后，可以计算如平均最小 RMSD（AMR）或匹配（MAT）和覆盖（COV）等指标，用于从 RMSD 值评估生成的分子构象的精度 **?**。

COV - 一组结构在另一组结构中被覆盖的百分比，其中覆盖表示两个构象之间的 RMSD 在指定阈值 δ 内。给定两组构象，S_g 作为生成集，S_r 作为参考集，召回的覆盖 $\mathrm{COV}-R$ 衡量模型找到所有参考构象的能力（召回或预测的参考构象的百分比），而精度的覆盖 $\mathrm{COV}-P$ 衡量模型生成的构象中有多少与参考集中的构象相关，公式如下：

$$\text{COV-R}(S_g, S_r) = \frac{1}{|S_r|}\sum_{\hat{\mathbf{C}} \in S_r} \mathbf{1}\{\mathbf{C} \in S_g | \mathrm{RMSD}(\mathbf{C}, \hat{\mathbf{C}}) \leq \delta, \hat{\mathbf{C}} \in S_r\}$$

MAT/AMR - 同一集合中的构象与另一集合中最近邻的平均 RMSD，即构象的平均最小 RMSD（AMR），较低的 MAT 分数反映出生成更真实的构象。召回匹配（MAT-R）和精度匹配（MAT-P）定义类似于覆盖指标 **?**。

$$\text{MAT-R}(S_g, S_r) = \frac{1}{|S_r|}\sum_{\hat{\mathbf{C}} \in S_r} \min_{\mathbf{C} \in S_g} \mathrm{RMSD}(\mathbf{C}, \hat{\mathbf{C}})$$

10 当前局限性

扩散模型在分子生成，特别是 3D 分子图生成方面取得了显著进展。然而，仍然存在一些挑战需要进一步推动这一领域的发展。

手性是 3D 生成模型当前的局限之一，因为大多数扩散模型是 E(3) 等变的，意味着它们对手性不敏感。此问题仅在 GCDM 模型 **?**、GCPNET **?** 的反向扩散神经网络设计中考虑过，类似地，在设计新模型或改进当前最先进模型时，手性也应予以考虑。

第二个局限性是分子生成模型普遍存在的，与药物发现和材料设计等不同应用中的指标与实际表现之间存在脱节。虽然合成可及性评分（SAS）**?** 和 Vina 对接分数 **?** 或结合亲和力等指标提供了良好的估算，但这并不保证该分子能在化合物库中找到，或能够通过当前合成化学方法合成，或者与其他候选分子相比，合成这些分子的成本效率。

此外，深度生成模型（包括扩散模型）的黑箱特性使得很难追溯为何扩散模型生成特定分子或构象。这种缺乏可解释性使得识别和解决对某些化学空间的潜在偏见或生成过程中的局限变得复杂。

第三个局限性与化学应用中的数据可用性相关。由于实验验证的成本较高，例如在药物设计和材料设计中的基于结构的应用，训练可用的数据仍然有限，并未涵盖生成分子的药代动力学和毒性等重要方面。数据也可能引入对已知空间的偏见，限制真正新颖和突破性结构的发现，尤其是当模型优化使用奖励这种偏见的指标时，例如最大均值差异（MMD）距离和其他相似性评分。缺乏数据是多个药物设计应用中普遍存在的问题，不仅限于分子生成。为了解决这一问题，Hsu 等人使用扩散模型生成特定于某些终点（如药代动力学性质、毒性和 hERG **?**）的合成数据。

对于 3D 生成而言，更具体的局限性是缺乏统一的基准；即使广泛使用的指标如 3D 稳定性，在每个模型中也有不同的定义。例如，一些模型（如 MiDi **?**）计算基于模型生成的邻接矩阵的共价键类型的稳定性，而其他模型（如 JODO **?**）则根据生成的 3D 坐标的距离阈值计算共价键，然后相应评估稳定原子的百分比。然而，连简单的稳定原子的定义也并不统一。例如，一些模型（如 MiDi **?**）认为带电原子是稳定的，因此碳原子可以具有 4 的价数（正式电荷（FC）= 0）、3（FC = 1, -1）或 5（FC = -1），而其他模型（如 JODO **?** 和 EDM **?**）仅在其正式电荷等于零时才认为原子是稳定的。此外，基于量子力学属性（如 QM9 极化率 (α) 或偶极矩 (μ) 等）的分子条件生成是基于 EGNN **?** 预测模型进行评估的，而 EGNN 模型本身与真实的密度泛函理论（DFT）计算之间存在误差。考虑到计算资源的进步，直接生成合理的分子样本并使用 DFT 计算评估生成的分子会更好。

扩散模型的另一个局限性是高训练和采样成本，特别是对于复杂数据如 3D 分子。这些计算模型在分子大小增加时会显著增加。然而，开发新模型架构和训练策略可以帮助实现更好的训练和采样。

PoseCheck ? 是一项基准研究，旨在评估针对 SBDD 设计的深度生成模型，报告了扩散模型在目标感知生成方面的另一个局限性，其中两个基于扩散的模型（DiffSBDD **?** 和 TargetDiff **?**）生成了一些与蛋白质口袋存在高立体冲突的无效分子，并建议对立体冲突进行惩罚以避免此问题。

11 结论

扩散模型在各种化学应用中具有巨大的潜力。早期的成功激励着对数据可用性、模型性能和与不同应用相关性等挑战的深入研究。通过解决这些局限性并积极探索新的方法和模型架构，扩散模型可以有效地应用于各种化学应用。考虑到可用的大型数据集以及几十年的深入研究，药物发现为扩散模型的进一步发展提供了理想的平台。

12 致谢

我，Amira，感谢 Sijie Fu、Calvin Gang、Chenghui Zhou 和 Euxhen Hasanaj 对扩散模型的有益讨论和评论，这些讨论促成了本综述的形成。同时，我特别感谢 Arav Agarwal 在数据收集方面的帮助以及他提供的宝贵反馈和意见。在撰写本文时，使用了 ChatGPT、Gemini 和 Quillbot 等 AI 工具来辅助处理一些有限的任务。