

密级：公开

中图分类号：TP391



浙江工商大学

硕士学位论文

论文题目：用于创新药物设计的分子

图扩散生成框架

作者姓名：徐璨

学科专业：统计学

研究方向：数理统计

指导教师：王伟刚

提交日期：2023年12月

**Dissertation Submitted to Zhejiang Gongshang University for
Master's Degree of Engineering**

A Diffusion-based Graph Generative Framework for *de novo* 3D
molecule generation

Author: Can Xu

Major: Statistics

Supervisor: Prof. Weigang Wang



Dec. 2023

School of Information and Electronic Engineering

Zhejiang Gongshang University

Hangzhou, 310018, P.R. China

摘要

去噪扩散模型在多个研究领域显示出巨大潜力。现有的基于扩散生成模型在全新的三维药物分子设计任务中面临两个主要挑战。由于分子中的大部分重原子能够通过单键与多个原子相连，使用成对距离来建模分子几何结构是不足够的。因此，第一个挑战是提出一个有效的神经网络作为去噪核，能够捕捉复杂的原子间关系并学习高质量的特征。此外，由于图的离散性质，当前主流的基于扩散的药物分子设计模型严重依赖预设的规则，并以间接的方式生成边。故第二个挑战是将分子生成的过程与扩散的学习过程相结合，有效地准确预测键的存在。本文认为扩散过程中分子构象的迭代更新方式与分子动力学一致，故提出一种名为几何辅助的分子扩散（Geometric-facilitated Molecular Diffusion / GFMDiff）的新型药物分子设计框架。针对第一个挑战，我们引入了一种双轨 Transformer 网络（Dual-track Transformer Network / DTN），以全面挖掘全局空间关系并学习高质量的表示，从而提高对特征和几何结构的准确预测能力。至于第二个挑战，我们设计了几何辅助损失（Geometric-facilitated Loss / GFLoss），该损失在训练期间干预键的形成，而不是直接将边嵌入隐变量空间。本文在多个基准数据集上开展了全面实验表明 GFMDiff 性能达到该领域时下最优的水平。

关键字： 扩散模型；分子生成；分子学习；图神经网络；几何神经网络

Abstract

Denoising diffusion models have shown great potential in multiple research areas. Existing diffusion-based generative methods on *de novo* 3D molecule generation face two major challenge. Since majority heavy atoms in molecules allow connections to multiple atoms through single bonds, modeling molecule geometries using pair-wise distance is insufficient. Therefore, the first one involves proposing an effective neural network as the denoising kernel that is capable to capture complex interatomic relationships and learn high-quality features. Due to the discrete nature of graphs, mainstream diffusion-based methods for molecules heavily rely on predefined rules and generate edges in a indirect manner. The second one involves accommodating molecule generation to the learning process of diffusion and accurately predicting the existence of bonds effectively. In our research, we view the iterative way of updating molecule conformations in diffusion process is consistent with molecular dynamics and introduce a novel molecule generation method named Geometric-Facilitated Molecular Diffusion (GFMDiff). For the first challenge, we introduce a Dual-track Transformer Network (DTN) to fully excevate global spatial relationships and learn high quality representations which contribute to accurate predictions of features and geometries. As for the second challenge, we design Geometric-facilitated Loss (GFLoss) which intervenes the formation of bonds during the training period, instead of directly embedding edges into the latent space. Comprehensive experiments on current benchmarks demonstrate the superiority of GFMDiff.

Keywords: **Diffusion models; Molecule generation; Molecular learning; Graph neural networks; Geometry neural Networks**

目 录

中文摘要	I
英文摘要	II
插图	V
表格	VI
1 几何促进的 3D 分子图生成 (GFMDiff)	1
1.1 双轨 Transformer 网络 (DTN)	1
1.2 几何信息促进的损失函数 (GFLoss)	5
1.3 扩散及去噪过程	7
1.4 目标函数与评价指标	8
2 实验结果及分析	11
2.1 药物分子设计	11
2.1.1 实验设置	11
2.1.2 在 GEOM-QM9 上的药物分子设计	13
2.1.3 在 GEOM-QM9 上有条件的药物分子设计	15
2.1.4 在 GEOM-Drugs 上的药物分子设计	17
2.2 分子性质预测	18
2.2.1 实验设置	18
2.2.2 在 GEOM-QM9 数据集上的分子性质预测	19
2.2.3 在 OC20 数据集上的分子性质预测	20
参考文献	22
附录	25
A 化学键长	25
B 超参数设置	26
攻读硕士学位期间取得的研究成果	27
致谢	28

独创性声明和论文使用授权说明	29
----------------------	----

插 图

1.1 GFMDiff 模型框架示意图.....	1
1.2 DTN 去噪内核结构示意图.....	2
1.3 GFLoss 损失函数.....	5
2.1 GFMDiff 在 GEOM-QM9 上药物分子设计样本示意	13
2.2 GFMDiff 在 GEOM-QM9 上有条件的药物分子设计样本示意	16
2.3 GFMDiff 在 GEOM-Drugs 上药物分子设计样本示意	18

表 格

2.1 GEOM-QM9 上药物分子设计结果对比	14
2.2 GEOM-QM9 上有条件的药物分子设计结果对比	15
2.3 GEOM-Drugs 上药物分子设计结果对比	17
2.4 GEOM-QM9 上药物分子性质预测结果对比	19
2.5 OC20 上分子性质预测结果对比	20
A.1 典型单键键长.....	25
A.2 典型双键键长.....	25
A.3 典型三键键长.....	25

1 几何促进的 3D 分子图生成 (GFMDiff)

在本节中，本文将着重介绍创新 3D 药物分子设计的模型框架，具体包括本文提出的 E(n) 等变去噪内核，几何信息促进的损失函数，扩散及去噪过程和优化目标。本文提出的创新 3D 药物分子设计整体框架 GFMDiff 由图 1.1 所示。

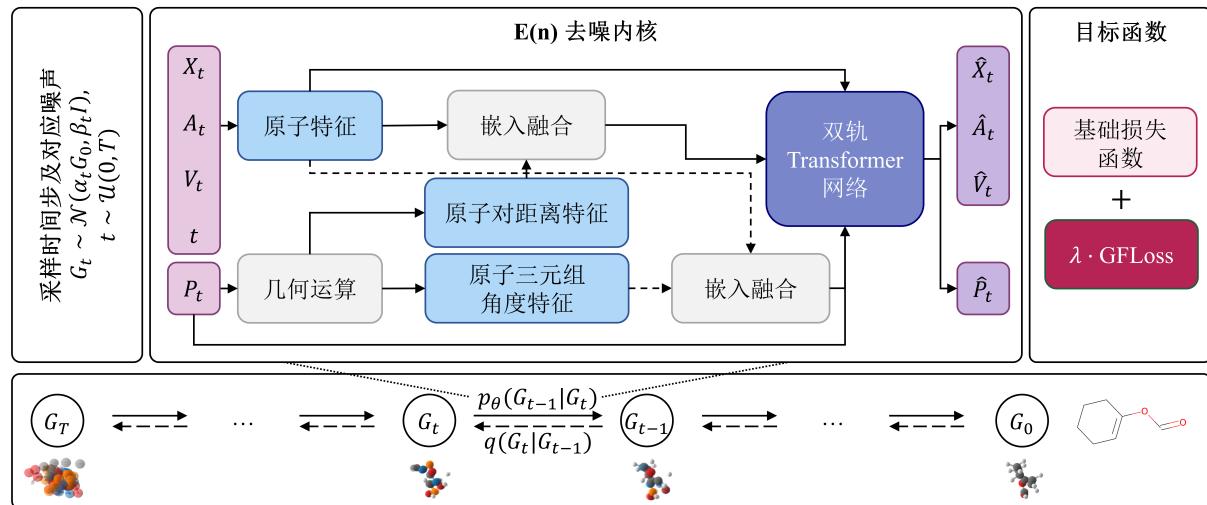


图 1.1 GFMDiff 模型框架示意图

对于每个训练样本，模型的输入为随机选取的时间步 t 和采样自对应时刻噪声分布的样本。通过几何计算及编码，得到的原子特征，原子对距离特征，三元组角度特征将被用于 E(n) 等变的去噪内核 DTN 中的分子学习，并得到更新后的样本在 $t - 1$ 时刻的分布，即对应的原子特征和位置信息。

1.1 双轨 Transformer 网络 (DTN)

在这个小节中，我们将详细介绍作为 GFMDiff 的 E(n) 等变去噪内核的双轨 Transformer 网络 (Dual-track Transformer Network / DTN)。DTN 被设计用于有效捕捉原子之间的关系和原子特征。由于三维分子几何具有旋转、平移、反射和排列等不变性质，使得去噪核满足这些性质是很重要的。本文所提出的 DTN 不仅是 E(n) 等变的，还能充分利用空间信息，进而预测高质量的原子及分子特征。图 1.2 为 DTN 的模型结构。

在我们提出的方法中，我们将具有总原子数 N 的输入分子视为 $G = (P, X, A, V)$ ，其中 $P = (p_1, p_2, \dots, p_N) \in \mathbb{R}^{N \times 3}$ 表示原子坐标， $X = (x_1, x_2, \dots, x_N) \in \mathbb{R}^{N \times n_f}$ 表示原子编号的独热编码， $A = (a_1, a_2, \dots, a_N) \in \mathbb{R}^N$ 表示原子编号， $V = (v_1, v_2, \dots, v_N) \in \mathbb{R}^N$ 表示原子化合价数。为了确保等变性，DTN 利用原子对距离信息和三元组角度信息捕捉几何信

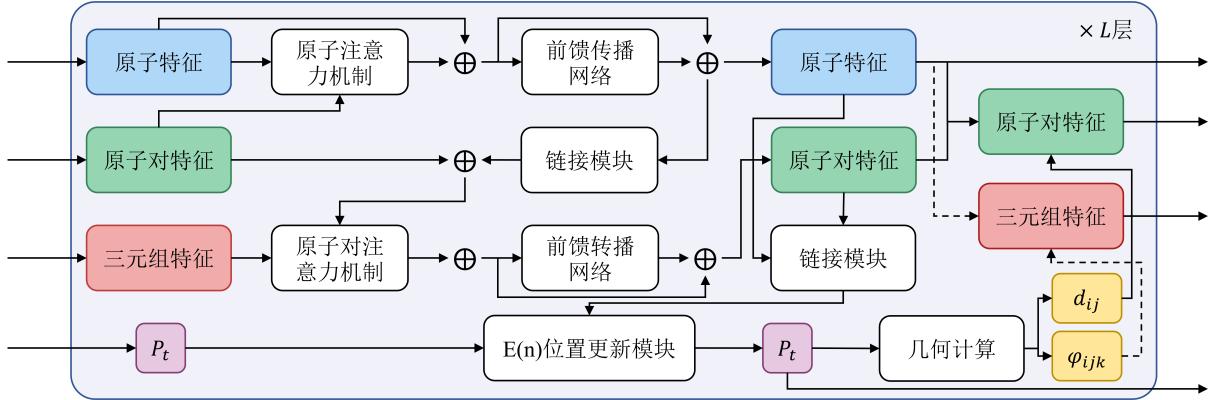


图 1.2 DTN 去噪内核结构示意图

息。原子 i 和 j 之间的欧几里得距离反映了原子间相互作用的强度，可以通过以下公式获得：

$$d_{ij} = \|p_i - p_j\|_2. \quad (1-1)$$

在分子中，除了氢原子，其他多数原子能够形成超过一个单键，故存在大量原子间的多体复杂关系。因此仅使用原子对距离是不足以充分提取空间几何信息。本文提出进一步使用以下公式计算原子三元组间的夹角：

$$\varphi_{ijk} = \arccos \left(\frac{(p_i - p_j) \times (p_i - p_k)}{\|p_i - p_j\|_2 \times \|p_i - p_k\|_2} \right). \quad (1-2)$$

经过上述几何计算，本文通过径向基神经网络（Radial Basis Function / RBF）编码得到可以被用作神经网络学习的原子对距离特征和三元组角度特征：

$$e_{ij} = \text{Linear}(\text{RBF}(d_{ij}), e_i, e_j), \quad (1-3)$$

$$e_{ijk} = \text{Linear}(\text{RBF}(\varphi_{ijk}), e_i, e_j, e_k), \quad (1-4)$$

其中 $e_i = \text{Embedding}(x_i, a_i, v_i)$ 是原子 i 的节点嵌入，由原子序数和化合价决定。这些特征随后被输入到 L 层的 DTN 中。RBF 函数是一种常用的径向基函数。在机器学习和模式识别领域中，RBF 函数常用于对数据进行特征编码和表征，以距离特征计算为例：

$$K(d_{ij}, \mu_k) = \exp \left(-\frac{\|d_{ij} - \mu_k\|^2}{2\sigma^2} \right), \quad (1-5)$$

$$\text{RBF}(d_{ij}) = \text{Linear}(K(d_{ij}, \mu_k)), \quad (1-6)$$

其中中心 μ_k 和自由参数 σ 为 RBF 函数参数。在距离特征计算时， μ 满足 $0\text{\AA} \leq \mu_k \leq 10\text{\AA}$ 并以 0.2\AA 的间隔切片， σ 满足 $\frac{1}{2\sigma^2} = 10\text{\AA}$ 。经过切片操作后， $K(d_{ij}, \mu_k)$ 的维度为 50，故

需要通过线性层将 RBF 特征映射到与原子对特征维度一致的高位特征空间内。在角度计算时， μ 满足 $0\text{\AA} \leq \mu_k \leq \pi\text{\AA}$ 并以 0.1\AA 的间隔切片， σ 满足 $\frac{1}{2\sigma^2} = 10\text{\AA}$ 。

DTN 的每一层由以下组件组成：原子对轨道、对-三元轨道和连接模块。原子对轨道模拟了原子之间的相互作用力对目标原子的影响，而对-三元轨道模型则模拟了潜在键角对边的影响。连接模块作为两个轨道之间的桥梁，将原子特征注入到成对特征中，以促进更好的表示学习。

原子对轨道涉及预测其他原子和原子之间相互作用力对目标原子的影响。该轨道以原子嵌入 e_i 和对嵌入 e_{ij} 作为输入：

$$e_i = \text{LayerNorm}(e_i), e_{ij} = \text{LayerNorm}(e_{ij}), \quad (1-7)$$

$$\mathbf{Q}_i = \text{Linear}(e_i), \mathbf{K}_i = \text{Linear}(e_i) + \text{Linear}(e_{ij}), \quad (1-8)$$

$$a_i = \text{Dropout} \left(\text{softmax} \frac{\mathbf{Q}_i \mathbf{K}_i^T}{\sqrt{d_h}} \right), \quad (1-9)$$

$$\mathbf{V}_i = \text{Linear}(e_{ij}) + \text{Linear}(e_i) + \text{Linear}(e_j), \quad (1-10)$$

$$\hat{e}_i = \text{Linear}(a_i \mathbf{V}_i^T), \quad (1-11)$$

其中 d_h 是头的数量。原子嵌入首先通过添加原子对轨道的预测来更新，然后传递给前馈网络。在每一层中，原子吸收其他原子和相应的原子对的聚合表示。

类似地，对-三元轨道预测了复杂几何亚结构对原子间相互作用力的影响。

$$e_{ij} = \text{LayerNorm}(e_{ij}), \quad (1-12)$$

$$e_{ijk} = \text{LayerNorm}(e_{ijk}), \quad (1-13)$$

$$\mathbf{Q}_{ij} = \text{Linear}(e_{ij}), \mathbf{K}_{ij} = \text{Linear}(e_{ij}) + \text{Linear}(e_{ijk}), \quad (1-14)$$

$$a_{ij} = \text{Dropout} \left(\text{softmax} \frac{\mathbf{Q}_{ij} \mathbf{K}_{ij}^T}{\sqrt{d_h}} \right), \quad (1-15)$$

$$\mathbf{V}_{ij} = \text{Linear}(e_{ij}) + \text{Linear}(e_{ijk}), \quad (1-16)$$

$$\hat{e}_{ij} = \text{Linear}(a_{ij} \mathbf{V}_{ij}^T). \quad (1-17)$$

值得注意的是，三元嵌入 e_{ijk} 在 Transformer 结构中不会得到更新，因为这会显著增加计算资源的需求。它们只会在原子坐标更新时得到更新。

连接模块的作用是将原子嵌入融合到对嵌入中。对于对嵌入 e_{ij} ，它同时吸收来自

连接模块的原子特征信息和来自对-三元轨道的局部空间信息。

$$e_{ij} = \text{LayerNorm}(e_{ij} + \text{Linear}(\text{Linear}(e_i) \otimes \text{Linear}(e_j))) \quad (1-18)$$

在更新坐标的方法上，我们遵循 EDM^[1] 和 MDM^[2] 的相关设计。

$$\hat{p}_i = p_i + \sum_{j \neq i} \frac{p_i - p_j}{d_{ij} + 1} \text{Linear}(\hat{e}_i, \hat{e}_j, d_{ij}^2, \hat{e}_{ij}) \quad (1-19)$$

由于在该坐标更新模块中，仅依赖于原子间相对距离更新原子坐标，故其严格遵循 E(n) 等变性要求。在原子坐标得到更新后，原子对和原子三元组的嵌入也将得到更新：

$$e_{ij} = \text{Linear}[\text{Linear}(\text{RBF}(\hat{d}_{ij}), \hat{e}_{ij}), \hat{e}_i, \hat{e}_j], \quad (1-20)$$

$$e_{ijk} = \text{Linear}(\text{RBF}(\hat{\varphi}_{ijk}), e_{ijk}). \quad (1-21)$$

更新后的原子对距离特征和三元组角度特征将被用作下一层 DTN 网络的输入，或是直接作为去噪内核的输出。运用该种方法能够避免通过神经网络直接更新三元组嵌入，故而在极大减小计算需求的同时，又能通过简单的线性层更新能够反映几何特征的三元组特征。在算法 1.1 中详细的介绍了每一层的 DTN 的计算流程。

算法 1.1: DTN 去噪内核伪代码

Input 原子特征 e_i , 原子对特征 e_{ij} , 三元组特征分子 e_{ijk} 和几何坐标 p_t ;

层归一化 e_i 和 e_{ij} ;

计算多头注意力概率 a_i 和值 \mathbf{V}_i ;

得到更新的原子特征残差 \hat{e}_i 并更新原子特征 e_i ;

通过连接模块将原子特征 e_i 融入原子对特征 e_{ij} ;

层归一化 e_{ij} 和 e_{ijk} ;

计算多头注意力概率 a_{ij} 和值 \mathbf{V}_{ij} ;

得到更新的原子特征残差 \hat{e}_{ij} 并更新原子特征 e_{ij} ;

通过坐标更新模块更新坐标 \hat{p}_t ;

进行几何计算，并对新的原子对特征 e_{ij} 和三元组特征 e_{ijk} 进行编码;

Return 新的原子特征 e_i , 原子对特征 e_{ij} , 三元组特征分子 e_{ijk} 和几何坐标 p_t .

1.2 几何信息促进的损失函数（GFLoss）

预测化学键的存在是分子图生成中的基本且不可或缺的任务。与以往的研究完全依赖于预设规则生成边不同，我们提出在训练过程中积极干预化学键的形成，通过设计一种精细的训练目标函数，命名为几何促进损失（Geometric-facilitated Loss / GFLoss）。这个损失函数的目的是引导模型生成既具有有效的拓扑结构，又具有稳定构象的分子。在药物分子设计中，我们认为原子的价是一种非常重要的辅助特征类型。因此，在上述提到的分子学习网络 DTN 中，原子的价被作为原子特征的一部分进行了整合。这使得模型能够学习和利用原子的价信息，具体计算流程和伪代码如图 1.3 和算法 1.2 所示。

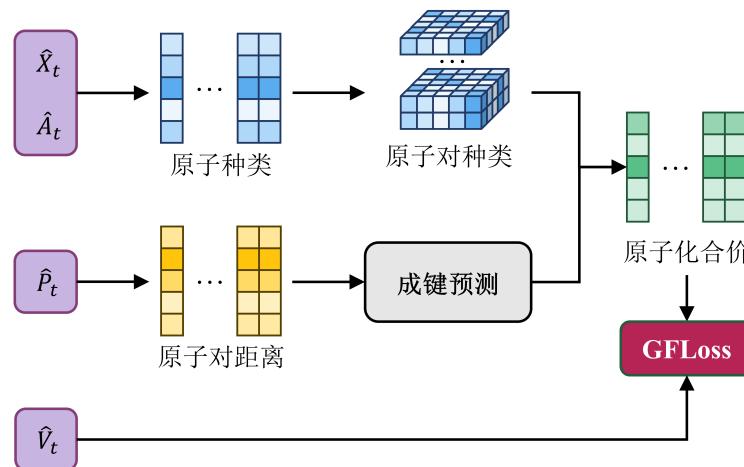


图 1.3 GFLoss 损失函数

根据预定义的规则，具有适当距离的原子对被认为是由化学键连接的。对于单键、双键或三键，某些原子之间存在典型距离。如果一对原子之间的距离在某个范围内，这两个原子被认为是由对应类型的键连接的。假设预定义的距离和边界为 $\mathbf{D} \in \mathbb{R}^{nf \times nf \times 3}$ 和 $\mathbf{M} \in \mathbb{R}^3$ ，其中 3 表示键的类型数量。基于 DTN 的输出 $\hat{G}_t = (\hat{P}_t, \hat{X}_t, \hat{A}_t, \hat{V}_t)$ ，我们首先使用 softmax 函数预测原子类型的概率：

$$\mathbf{p}_t(\hat{X}_{atom}) = \text{softmax}(\hat{X}_t) \in \mathbb{R}^{N \times nf}, \quad (1-22)$$

其中 \hat{X}_t 在此处表示维度为 nf 的独热编码格式的预测原子类型。原子对的类型概率为

$$\mathbf{p}_t(\hat{X}_{pair}) = \mathbf{p}_t(\hat{X}_{atom}) \cdot \mathbf{p}_t(\hat{X}_{atom}) \in \mathbb{R}^{N \times N \times nf \times nf}. \quad (1-23)$$

利用预测的原子坐标 \hat{P}_t , 可以得到原子对之间的距离矩阵 $\mathbf{d}_t \in \mathbb{R}^{N \times N}$, 并将其扩展为方便操作的 $\mathbb{R}^{N \times N \times nf \times nf \times 3}$ 。然后, 原子对之间的距离与典型的键距离之间的边界 \mathbf{m}_t 计算如下:

$$\mathbf{m}_t = \mathbf{d}_t - (\mathbf{D} + \mathbf{M}) \in \mathbb{R}^{N \times N \times nf \times nf \times 3}. \quad (1-24)$$

以原子 i 和 j 为例, 假设它们被认为是碳原子的概率大于零, 如果边界 $\mathbf{m}_t(i, j, C, C, :)$ 中的任何元素小于零, 则表示原子 i 和 j 之间存在一条键。键的具体类型由边界 $\mathbf{m}_t(i, j, C, C, :)$ 中最小值的索引确定。如果 $\arg \min(\mathbf{m}_t(i, j, C, C, :))$ 为 1, 则它们由一条单键连接。如果 $\arg \min(\mathbf{m}_t(i, j, C, C, :))$ 为 3, 则它们由一条三键连接。表示键的布尔矩阵记为 $\text{ISBOND}_t \in \mathbb{R}^{N \times N \times nf \times nf}$ 。

一旦我们获得了原子对的类型概率和键的存在情况, 就可以估计原子的可能价:

$$\hat{V}_{\text{pred}}(t) = \text{sum}(\mathbf{p}_t(\hat{X}_{\text{pair}}) \odot \text{ISBOND}_t) \in \mathbb{R}^N. \quad (1-25)$$

由于输入数据受到不同水平噪声的影响, GFLoss 被定义为预测价 V_{pred} 与真实价 V 之间的均方误差:

$$\mathcal{L}_{GF}(t) = ||\alpha_t(\hat{V}_{\text{pred}}(t) - V_t)||^2, \quad (1-26)$$

其中 α_t 是扩散过程中噪声数据中真实数据的水平。

算法 1.2: GFLoss 损失函数项伪代码

Input 预测的原子序数 \hat{X}_t , 化合价 \hat{V}_t 和几何坐标 \hat{P}_t ;

根据预测的原子序数 \hat{X}_t 计算任一原子类型 $\mathbf{p}_t(\hat{X}_{\text{atom}})$;

根据任一原子类型计算任一原子对的原子类型 $\mathbf{p}_t(\hat{X}_{\text{pair}})$;

计算任一原子对距离 \mathbf{d}_t ;

计算原子对距离与 3 种化学键的典型键长之间的差距 \mathbf{m}_t ;

根据差距 \mathbf{m}_t 判断原子对是否形成某种化学键, 并记为布偶型矩阵 ISBOND_t ;

根据矩阵 ISBOND_t 和原子对的原子类型 $\mathbf{p}_t(\hat{X}_{\text{pair}})$ 计算由几何构象和化学规则

预测的原子化合价 $\hat{V}_{\text{pred}}(t)$;

Return $\mathcal{L}_{GF}(t) = ||\alpha_t(\hat{V}_{\text{pred}}(t) - V_t)||^2$.

1.3 扩散及去噪过程

扩散过程对于 $t = 0, \dots, T$ 需要定义 α_t, σ_t 。由于 $\alpha_t = \sqrt{1 - \sigma_t^2}$, 只需定义 α_t 即可。这些值应该单调递减, 从 $\alpha_0 \approx 1$ 开始, 最终到达 $\alpha_T \approx 0$ 。在本文中, 我们设定

$$\alpha_t = (1 - 2s) \cdot f(t) + s \quad \text{其中} \quad f(t) = (1 - (t/T)^2)$$

精度取 10^{-5} , 以避免数值不稳定情况。这种计划在符号中与中引入的余弦噪声计划非常相似, 但我们的符号表述要简单一些。为了避免采样过程中的数值不稳定性, 我们遵循 nichol2021improved 的剪切过程, 并计算 $\alpha_{t|t-1} = \alpha_t / \alpha_{t-1}$, 其中我们定义 $\alpha_{-1} = 1$ 。然后, $\alpha_{t|t-1}^2$ 的值被剪切到下限 0.001。这样做是为了在采样过程中避免数值不稳定性, 因为 $1/\alpha_{t|t-1}$ 现在在采样过程中保持有界。然后, α_t 的值可以使用累积乘积 $\alpha_t = \prod_{\tau=0}^t \alpha_{\tau|\tau-1}$ 来重新计算。

回想一下, $\text{SNR}(t) = \alpha_t^2 / \sigma_t^2$ 。就像中一样, 我们计算负对数 SNR 曲线定义为 $\gamma(t) = -(\log \alpha_t^2 - \log \sigma_t^2)$, 其中 $\sigma_t^2 = 1 - \alpha_t^2$ 。 $\gamma(t)$ 是一个单调递增的函数, 可以用高数值精度计算出所有所需的分量。例如, $\alpha_t^2 = \text{sigmoid}(-\gamma(t))$, $\sigma_t^2 = \text{sigmoid}(\gamma(t))$, 以及 $\text{SNR}(t) = \exp(-\gamma(t))$ 。

在提出的 GFMDiff 的前向过程中, 位置、原子序数和原子价逐渐被噪声先验分布所扰动。通过具有预定义方差序列 $\beta_t \in (0, 1), t = 0, 1, \dots, T$ 的马尔可夫链将真实构象 $G_0 = (P_0, X_0, A_0, V_0)$ 进行转换。后验分布定义为:

$$q(G_{1:T}|G_0) = \prod_{t=1}^T q(G_t|G_{t-1}), \quad (1-27)$$

$$q(G_t|G_{t-1}) = \mathcal{N}_p(p_t; \sqrt{\alpha_t} p_{t-1}, \beta_t I) \cdot \mathcal{N}_h(h_t; \sqrt{\alpha_t} h_{t-1}, \beta_t I), \quad (1-28)$$

为方便起见, 其中 $h_t = \text{concat}(x_t, a_t, v_t)$ 表示原子特征。

为了在去噪过程中的满足几何坐标的等变性要求, 在将原子坐标引入扩散之前, 它们必须先转换到质心为零的线性子空间内^[1, 3-4]。因此, 噪声分布和上述后验分布都受限于相同的线性子空间。此外, 原子特征对于对神经网络是等变的, 故它们不需要进一步处理即可在整个前向和后向过程中与几何坐标一起进行运算。

考虑欧几里得变量 $\mathbf{p} \in \mathbb{R}^{M \times n}$ 在线性子空间上满足 $\sum_i \mathbf{x}_i = \mathbf{0}$ 。故可以认为 \mathbf{p} 是 n 维空间内的一个质心为零的点云。对于在此子空间上的一个正态分布 \mathcal{N}_x , 欧式变量服从

的分布可以表示为：

$$\mathcal{N}_x(\mathbf{x}|\mu, \sigma^2 I) = (\sqrt{2\pi}\sigma)^{-(M-1)\cdot n} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{x} - \mu\|^2\right), \quad (1-29)$$

其中 μ 位于与 \mathbf{x} 相同的子空间中。从这个正态分布中采样有多种方法，例如可以从维度为 $(M-1) \cdot n$ 的正态分布中采样，然后将采样映射到 $M \cdot n$ 维的环境空间中，使得它的重心为零。但是还有一种更简单的方法：可以直接在 $M \cdot n$ 维的环境空间中进行采样，并减去 $\sum_i \mathbf{x}_i$ 。由于正态分布是各向同性的（即无论选择哪个方向，其方差都是 σ^2 ），故本文采用这样的方法对原子几何坐标进行采样。

在去噪过程中，我们使用上述的去噪核函数来逼近每个时间步的分子构象，

$$p_\theta(G_{t-1}|G_t) = \mathcal{N}(G_{t-1}; \mu_\theta(G_t, t), \sigma_t^2 I), \quad (1-30)$$

$$\hat{G}_t = G_t / \alpha_t - \hat{\epsilon}_t \times \sigma_t / \alpha_t, \quad (1-31)$$

其中 $\hat{\epsilon}_t$ 是参数化神经网络的输出。

1.4 目标函数与评价指标

对于 DDPMs，典型的目标函数是数据对数似然的变分下界。在先前相关研究方法的基础上，我们将目标函数与 GFLoss 相结合：

$$\mathcal{L}_t = E_{\epsilon_t \sim \mathcal{N}(0, I)} \left[\frac{1}{2} \omega(t) \left(\|\epsilon_t - \hat{\epsilon}_t\|^2 + \lambda \|\alpha_t (\hat{V}_{pred}(t) - V_t)\|^2 \right) \right], \quad (1-32)$$

其中 $\omega(t) = (1 - \text{SNR}(t)/\text{SNR}(t-1))$ 。

此外，在评价模型参数的好坏时，需要比较模型学习的去噪分布于前向噪声分布的差异。该指标具体表现为参数的负对数似然函数，由于负对数似然函数无法直接计算得到，本文通过算法 1.3 估计该负对数似然函数的取值。

算法 1.3: GFMDiff 参数的负对数似然函数估计

Input 原子坐标 p , 原子特征 $h = [x, a, v]$, 去噪神经网络 ϕ ;

采样时间步 $t \sim \mathcal{U}(1, T)$ 和噪声 $\varepsilon_t = [\varepsilon_t^{(p)}, \varepsilon_t^{(h)}] \sim \mathcal{N}(\mathbf{0}, I)$;

混合输入样本与噪声 $z_t = \alpha_t[p, h] + \sigma_t \varepsilon_t$;

计算神经网络损失函数 $\mathcal{L}_t = \frac{1}{2}(1 - \text{SNR}(t-1)/\text{SNR}(t))\|\varepsilon_t - \phi(z_t, t)\|^2$;

采样 0 时刻样本噪声 $\varepsilon_0 = [\varepsilon_0^{(p)}, \varepsilon_0^{(h)}] \sim \mathcal{N}(\mathbf{0}, I)$;

混合 0 时刻输入样本与噪声 $z_0 = \alpha_0[p, h] + \sigma_0 \varepsilon_0$;

计算 0 时刻损失函数 $\mathcal{L}_0 = \mathcal{L}_0^{(x)} + \mathcal{L}_0^{(h)} = -\frac{1}{2}\|\varepsilon_0 - \phi(z_0, 0)\|^2 - \log Z + \log p(h|z_0^{(h)})$;

$\mathcal{L}_{\text{base}} = -\text{KL}(q(z_T|x, h)|p(z_T)) = -\text{KL}(\mathcal{N}_{xh}(\alpha_T[x, h], \sigma_T^2 I) | \mathcal{N}_{xh}(\mathbf{0}, I))$;

Return $\hat{\mathcal{L}} = T \cdot \mathcal{L}_t + \mathcal{L}_0 + \mathcal{L}_{\text{base}}$.

其中对于前向和反向过程的样本 KL 散度的计算, 无法直接得到。对于两个各向同性正态分布 $q = \mathcal{N}(\mu_1, \sigma_1^2 I)$ 和 $p = \mathcal{N}(\mu_2, \sigma_2^2 I)$, 标准的 KL 散度可表示为:

$$\text{KL}(q||p) = d \cdot \log \frac{\sigma_2}{\sigma_1} + \frac{1}{2} \left[\frac{d \cdot \sigma_1^2 + \|\mu_1 - \mu_2\|^2}{\sigma_2^2} - d \right], \quad (1-33)$$

其中 d 是分布的维度。在扩散模型中, 扩散过程和去噪过程具有相同的方差 σ^2 。如果 $\sigma_1 = \sigma_2 = \sigma$, 则这两个分布的 KL 散度可简化为:

$$\text{KL}(q||p) = \frac{1}{2} \left[\frac{\|\mu_1 - \mu_2\|^2}{\sigma^2} \right]. \quad (1-34)$$

现假设 $\mathcal{N}_1(\tilde{\mu}_1, \sigma^2 I)$ 和 $\mathcal{N}_2(\tilde{\mu}_2, \sigma^2 I)$ 在一个线性子空间上, 其中均值 $\tilde{\mu}$ 是相对于子空间中的任意坐标系定义的, 则这两个分布之间的 KL 散度包含一个包含欧几里得距离项 $\|\tilde{\mu}_1 - \tilde{\mu}_2\|^2$ 。

类似于 E-NF^[3] 和 GeoDiff^[4] 中的论证, 可以构造一个正交变换 Q , 将满足 $\sum_i \mu_i = \mathbf{0}$ 条件的环境空间映射到子空间上, 以满足 $\begin{bmatrix} \tilde{\mu} \\ \mathbf{0} \end{bmatrix} = Q\mu$ 。由于 $\|\tilde{\mu}\| = \left\| \begin{bmatrix} \tilde{\mu} \\ \mathbf{0} \end{bmatrix} \right\| = \|\mu\|$, 故 $\|\tilde{\mu}_1 - \tilde{\mu}_2\|^2 = \|\mu_1 - \mu_2\|^2$ 。这表明式 1-34 在环境空间内可以通过计算得到。

对于诸如 \mathcal{N} 这样的分布, 只要两个分布之间的方差相同, KL 散度仍然可以在环境空间中计算。现在考虑分布 $q = \mathcal{N}_{ph}(\mu_1, \sigma^2 I)$ 和 $p = \mathcal{N}_{ph}(\mu_2, \sigma^2 I)$ 的组合 KL 散度。注意到这里的均值由两部分组成 $\mu = [\mu^{(p)}, \mu^{(h)}]$, 其中 p 部分位于一个子空间中, 而 h 作为原子特征分布没有特定要求。因此该分布可以分解为 $\mathcal{N}_{ph}(\mu, \sigma^2 I) = \mathcal{N}_p(\mu^{(p)}, \sigma^2 I) \cdot$

$\mathcal{N}(\mu^{(h)}, \sigma^2 I)$ 。因此扩散和去噪过程样本分布的 KL 散度可以被简化为：

$$\begin{aligned} \text{KL}(q||p) &= \text{KL}\left(\mathcal{N}_p(\mu_1^{(p)}, \sigma^2 I) || \mathcal{N}_p(\mu_2^{(p)}, \sigma^2 I)\right) + \text{KL}\left(\mathcal{N}(\mu_1^{(h)}, \sigma^2 I) || \mathcal{N}(\mu_2^{(h)}, \sigma^2 I)\right) \\ &= \frac{1}{2} \left[\frac{\|\mu_1^{(p)} - \mu_2^{(p)}\|^2}{\sigma^2} \right] + \frac{1}{2} \left[\frac{\|\mu_1^{(h)} - \mu_2^{(h)}\|^2}{\sigma^2} \right] = \frac{1}{2} \left[\frac{\|\mu_1 - \mu_2\|^2}{\sigma^2} \right]. \end{aligned} \quad (1-35)$$

2 实验结果及分析

在此部分，本文将首先全方位检验提出的用于创新 3D 药物分子设计的几何促进的分子扩散框架 GFMDiff 在生成 3D 分子任务中的表现，用于性能比较的三个任务和两个公开数据集都具备极强的代表性。同时为了进一步探究本文提出的 DTN 去噪内核作为分子学习模型，在分子性质预测任务上的表现。本文将 DTN 与时下性能最优的 3D 分子学习模型，在 2 个公开数据集上的表现进行对比。

2.1 药物分子设计

在本节中，我们报告了 GFMDiff 在两个主流数据集 GEOM-QM9^[5] 和 GEOM-Drugs^[6] 上的三个生成任务中的表现。本文采用了时下最前沿的相关研究模型作为基线模型，在采取一致的实验条件的前提下，直接引用他们生成的实验结果。结果表明，我们的方法在多个方面显著优于该领域最优秀的模型，展现出在 3D 分子生成任务中时下最优的性能。

2.1.1 实验设置

数据集：为了进行全面且公平的比较，我们在两个基准数据集（GEOM-QM9 和 GEOM-Drugs）上进行了三组实验：在 GEOM-QM9 上的药物分子设计，在 GEOM-QM9 上有条件的药物分子设计，和在 GEOM-Drugs 上的药物分子设计。具体的数据集处理工作较为琐碎，本文将在对应的生成任务实验分析中分别介绍。

GEOM-QM9 数据集（简称：QM9）是一个广泛运用在基于深度学习的分子学习领域的数据集，包含超过 13 万个分子及其对应的构象，以及每个分子对应的 19 种性质。数据集中，在包含氢原子条件下，平均每个分子有 19 个原子，其中最大的分子有 29 个原子。同时，整个数据集的分子仅包括氢，碳，氮，氧，氟这五种原子。

GEOM-Drugs（简称：Drugs）是一个规模相对更大的数据集，其包括的分子数量和每个分子所含平均原子数都较多。该数据集记录了超过 45 万个分子和它们对应的 3700 万个不同构象。在包含氢原子的条件下，平均每个分子由 44 个原子构成，最大的由 181 个原子构成。同时，整个数据集的分子包含 16 类原子，较 QM9 比原子种类更丰富。

评价指标：为公平的评价 GFMDiff 的性能并与时下最优秀的模型对比，本文采用目前该领域通用的评价指标，具体介绍如下。

- 负对数似然函数（Negative log-likelihood / NLL）：是一个被广泛应用于评价参数估计的指标，其在深度学习中也可以被用作损失函数。

- 稳定性（Stability）：是 3D 分子设计领域最重要的评价指标，其衡量了生成的分子构型和化学键在几何空间内是否稳定。对于一个化学键而言，根据其两端连接的原子的类型和化学键类型的不同，在计算化学上存在不同的理想稳定键长^{①②}。例如对两个碳原子，其可能形成单键，双键和三键。对于碳碳单键，双键和三键，其理论键能分别为 346, 602, 835 (千焦/摩尔)，对应键长分别为 154, 134, 120 (皮米)。对于模型生成的分子几何构象，若一对碳原子间距离小于 120 皮米，则认为这两个碳原子由三键连接，若距离大于 120 皮米，且小于 134 皮米，则认为它们由双键连接。若原子间距离超过 154 皮米，则认为两者之间不存在键相连。根据计算化学相关知识，若一个原子经过预测，连接的所有键总计化合价预期理论值一致，则认为该原子稳定。以碳原子为例，若其经过预测后与 4 个其他原子形成单键，或形成 2 个单键 1 个双键，则认为该碳原子是稳定的。如果一个分子中所有的原子都具备正确的化合价，则该分子也被认为是稳定的。该指标有效的同时衡量生成的三维分子构象的几何和拓扑性质。在具体评价中，稳定性可分为原子稳定性（Atom Stability）和分子稳定性（Molecule Stability），他们各自代表稳定原子/分子在所有原子/分子中的数量占比。

- 有效性（Validity）：是在 2D 和 3D 分子生成领域通用的评价指标，其衡量了有效分子在所有分子中的数量占比。一个分子的有效与否，决定于 RDKit 中对分子中化合价和度的判断。

- 唯一性（Uniqueness）：同样对 2D 和 3D 分子设计通用，该指标衡量了生成结果中非重复的分子的数量占比。

- 平均绝对误差（Mean absolute error / MAE）：是一个常用于回归任务的评价指标。在本文中，该指标仅用于评价有条件的分子生成的样本属性。

基线模型：本文采用了本领域最具代表性且最前沿的模型，包含基于自回归模型的，流型模型的和扩散模型的方法。

- E-NF^[3] 是基于流型模型的 3D 分子设计的模型。其中用于图学习的网络为等变的

① http://www.wiredchemist.com/chemistry/data/bond_energies_lengths.html

② https://www.mrbigler.com/documents/Chemistry_Reference_Tables.pdf

EGNNs^[7]。该网络将原子间距离视作键的特征，通过信息传递更新相应节点特征。

- G-SchNet^[8] 在 SchNet^[9] 的基础上设计了一个自回归模型，通过逐步生成原子和键的方式生成分子结构。SchNet 已经被证明其对角度信息具备良好提取能力。

- EDM^[1] 是最早将扩散模型引入到药物分子设计的模型。其去噪内核同样采用 EGNNs^[7]。

- Bridge 和 Bridge+Force^[10] 提出能量方式用于更有效的引入几何信息，通过分子间作用力，引导分子生成更有效更稳定的构型。

- GCDM^[11] 是距今该领域表现最好且拥有完整开源代码的方法。其参照 ColfNet 的思路^[12] 引入了完整的空间几何信息。

为了公平客观的评价 GFMDiff 的性能，本文选取的基线模型都是具备完整可复现开源代码的模型，并在采用一致基本参数设置的条件进行实验。本文引用了 EDM^③实验部分中 E-NF, G-SchNet 和 EDM 的表现，引用了 Bridge 和 Bridge+Force 的实验结果，以及 GCDM^④的实验结果。

实验环境：本文在生成实验上基于的硬件平台如下：

GEOM-QM9 实验: Intel(R) Xeon(R) Platinum 8358 和 2 * NVIDIA A100 SXM4 40GB;

GEOM-Drugs 实验: AMD EPYC 7742 和 4 * NVIDIA A100 PCIe 80GB。

此外，相关实验依赖的一些重要的包有：CUDA 11.4, Python 3.9.13, PyTorch 1.10.0, PyG 2.0.4, RDKit 2022.03.5。

2.1.2 在 GEOM-QM9 上的药物分子设计

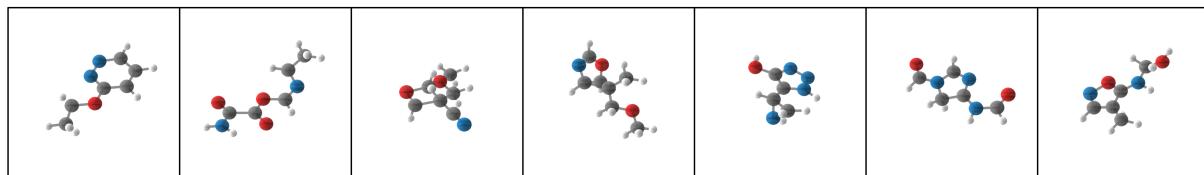


图 2.1 GFMDiff 在 GEOM-QM9 上药物分子设计样本示意

在该任务的实验中，本文依照先前研究，将 QM9 数据集分别划分为样本大小为 10 万，1.8 万和 1.3 万的训练集，验证集和测试集。表 2.1 为 GFMDiff 和基线模型在分子

③ https://github.com/ehoogeboom/e3_diffusion_for_molecules

④ <https://github.com/BioinfoMachineLearning/bio-diffusion>

表 2.1 GEOM-QM9 上药物分子设计结果对比

Type	Method	NLL↓	Atom Stability (%)↑	Mol Stability (%)↑	Validity (%)↑	Uniqueness × Validity(%)↑
NF	E-NF	-59.7	85.0	4.9	40.2	39.4
AR	G-SchNet	N/A	95.7	68.1	85.5	80.3
DDPM	EDM	-110.7±1.5	98.7±0.1	82.0±0.4	91.9±0.5	90.7±0.6
	Bridge	N/A	98.7±0.1	81.8±0.2	N/A	N/A
	Bridge+Force	N/A	<u>98.8±0.1</u>	84.6±0.3	N/A	N/A
	GCDM	-171.0±0.2	98.7±0.0	85.7±0.4	94.8±0.2	93.3±0.0
DDPM (Ours)	GFMDiff w/o tri	-123.1±0.4	98.7±0.1	85.9±0.2	94.9±0.2	94.2±0.2
	GFMDiff w/o GFLoss	-127.5±0.4	98.7±0.0	<u>87.5±0.1</u>	<u>96.0±0.0</u>	<u>95.2±0.0</u>
	GFMDiff	<u>-132.5±0.2</u>	99.1±0.0	91.3±0.2	97.0±0.3	96.1±0.2
	Data		99.0	95.2	97.7	97.7

在 QM9 数据集上的表现对比，指标旁的箭头代表更优取值的方向。同时图 2.1 展示了部分生成分子的三维构象。NLL 的取值为模型在测试集上的得分，而剩余指标的计算基于模型随机生成的 10000 个样本。此外，部分基线模型的特定指标可能在先前研究没有说明，在此用 N/A 替代。在表中，最好和次好的结果分别以粗体和下划线的形式予以强调。

除了在测试集上的 NLL 得分，本文在其他的所有指标上都显著优于基线模型。稳定性作为创新 3D 药物分子设计任务中的核心指标，GFMDiff 在这两个评价指标中较先前研究有明显优势，分子稳定性较表现最好的基线模型有 6.5% 的提升，这证明了 GFMDiff 对几何信息的有效利用极大的提升了模型性能。在有效性层面，GFMDiff 较表现最好的基线模型的提升分别为 2.3% 和 3%。这说明了 GFMDiff 能够生成有效且非重复的分子。值得一提的是，GFMDiff 生成结果已经十分接近，甚至在个别指标上超过了数据本身，实验结果充分表明了本文提出的 GFMDiff 在生成中型分子时具备的优异性能。

为衡量本文提出的 DTN 去噪内核和 GFLoss 损失函数的有效性，本文进行了两组消融实验。在此，将 GFLoss 损失函数的权重 λ 设置为 0，则可以得到 GFMDiff w/o GFLoss 模型。在此基础上，本文将 DTN 中原子对-三元组间的注意力机制模块取代为一个原子对自注意力机制模块，得到 GFMDiff w/o tri 模型。GFMDiff w/o GFLoss 消融实验的结果表明本文引入的 GFLoss 损失函数对生成稳定、有效的药物分子三维构象有积极促进作用。其在评价指标上全面落后于 GFMDiff。GFMDiff w/o tri 模型的结果较另外两个 GFMDiff 模型有更大的降幅。这样的结果表明 DTN 中对分子三元组信息的提取，并将其融入原子对距离学习是有效的。作为表现最差的 GFMDiff 模型变体，该模型仍优于其他基线模型，这也足以说明双轨 Transformer 网络中的双轨设计和基于 Transformer 的分子图学习的有效性。

2.1.3 在 GEOM-QM9 上有条件的药物分子设计

除了最基本的稳定性和有效性，生成具备良好属性的药物分子是药物分子生成模型应当具备的一项能力。为评价模型在有条件的药物分子生成任务上的表现，本文根据通行做法开展实验。在 QM9 数据集包含的所有性质中，选取各向同性极化率 α ，最高占据分子轨道 $\varepsilon_{\text{HOMO}}$ ，最低未占分子轨道 $\varepsilon_{\text{LUMO}}$ ，两者间能量隙 $\Delta\varepsilon$ ，偶极矩 μ 和 298.15K 时的热容 C_v 。

表 2.2 GEOM-QM9 上有条件的药物分子设计结果对比

Task	α	$\Delta\varepsilon$	$\varepsilon_{\text{HOMO}}$	$\varepsilon_{\text{LUMO}}$	μ	C_v
Units	Bohr ³	meV	meV	meV	D	$\frac{\text{cal}}{\text{mol K}}$
Naive (Upper-bound)	9.01	1470	645	1457	1.616	6.857
# Atom	3.86	866	426	813	1.053	1.971
EDM	2.76	655	356	584	1.111	1.101
GCDM	<u>1.97</u>	<u>602</u>	<u>344</u>	<u>479</u>	<u>0.844</u>	<u>0.689</u>
GFMDiff	1.74	558	321	430	0.728	0.593
QM9 (Lower-bound)	0.10	64	39	36	0.043	0.040

为检验生成的药物分子是否具备良好属性，在实验时，需要将 QM9 数据集的训练集部分平均分为两部分。在训练阶段，第一部分被用作一个 EGNN^[7] 神经网络的训练。

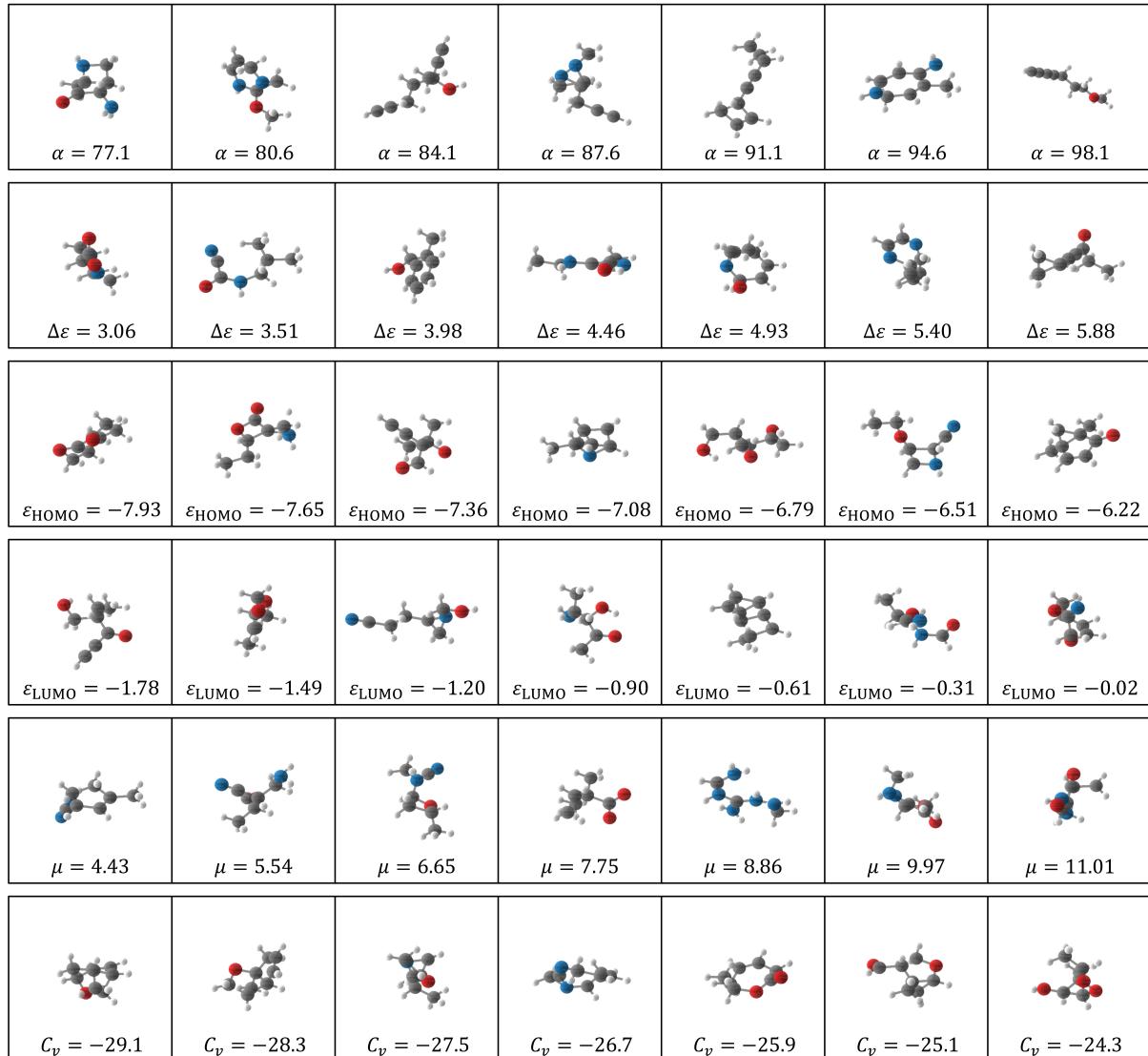


图 2.2 GFMDiff 在 GEOM-QM9 上有条件的药物分子设计样本示意

该网络通过学习正确的分子及对应的属性，故将具备根据输入分子预测相应性质的能力。第二部分数据集被用作生成模型的训练，即在普通的生成模型基础上，在训练时增加对应性质信息。在两部分都完成训练后，则预先根据训练集中性质的均值和方差，采样得到需要生成样本具备的性质。根据这部分性质，生成模型会生成 10000 个采样结果。这部分采样结果将通过之前训练好的 EGNN 神经网络，进行对应的性质预测，最终的衡量标准为 EGNN 预测的分子性质和分子生成时依照的性质间的平均绝对误差。

在基线模型上，本文引入了三个参照模型：Naive, #Atom 和 QM9。Naive 模型在测试阶段时，EGNN 模型根据随机打乱原有的采样样本对应性质并预测相应结果，这代表了在此任务下预测结果 MAE 的取值上限。#Atom 模型在测试阶段，EGNN 模型仅依据

分子所含原子数对性质进行预测。如果生成模型的测试 MAE 较 #Atom 模型低，则说明该模型能够根据特定性质生成相应的分子。参照模型 QM9 则代表 EGNN 在 QM9 数据集上的预测性能，该表现代表分子性质预测结果的 MAE 下界。

根据表 2.2 所示，本文提出的 GFMDiff 在测试任务中预测结果的 MAE 均低于目前最优的 GCDM 模型，在所有六个指标中的领先幅度分别为：11.7%，7.3%，6.7%，10.2% 和 13.7%。该表现足以证明，GFMDiff 具备根据指定条件生成分子的能力。图 2.2 中全面的展示了 GFMDiff 依据不同的性质取值，生成对应的药物分子样本。以各向同性极化率 α 为例，随着 α 取值增加，模型倾向于生成更长的碳链。

2.1.4 在 GEOM-Drugs 上的药物分子设计

在 GEOM-Drugs 数据集基础上进行创新药物分子设计是一个具有挑战性的任务，因为该数据集不仅包含分子多，也因其中分子主要为大分子，这对分子图中键的形成提出了较高的要求。在对 GEOM-Drugs 的实验中，我们将 GFMDiff 与 E-NF、EDM、Bridge+Force 和 GCDM 进行了比较。此外，由于当前方法面对如此庞大且自身数据不够精良的生成任务时，较 QM9 数据集上的表现有较大差距，常常无法生成新的分子，故在此讨论生成分子的有效性是缺乏意义的。在评价指标方面，本文采用了生成 10000 个药物分子的稳定性用作性能比较。

表 2.3 GEOM-Drugs 上药物分子设计结果对比

Type	Method	Atom Stability (%) ↑	Mol Stability (%) ↑
DDPM	E-NF	75.0	0
	EDM	81.3	0.0
	Bridge	81.0±0.7	0.0
	Bridge+Force	82.4±0.8	0.0
	GCDM	86.4±0.2	3.7±0.3
Ours	GFMDiff	86.5±0.2	3.9±0.2
Data		86.5	2.8

由于 GEOM-Drugs 中分子庞大，且该数据集不如 QM9 精确，其自身的稳定性就比 QM9 中要低得多。由表 2.3 与 QM9 上实验展现出的优秀表现一致，本文提出的 GFMDiff

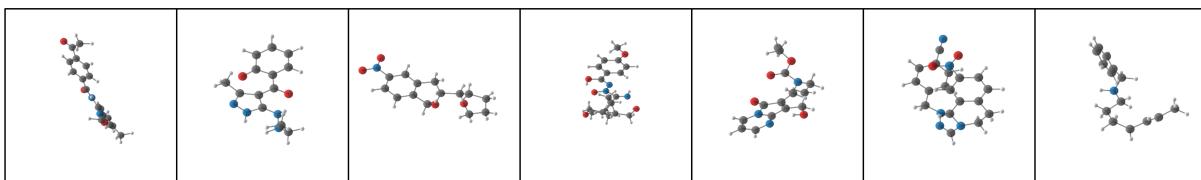


图 2.3 GFMDiff 在 GEOM-Drugs 上药物分子设计样本示意

在原子稳定性方面略优于 GCDM，同时显著优于其他基线模型。在分子的整体稳定性方面，GFMDiff 的表现超过第二名结果 5.4%。图 2.3 展示了部分生成分子的三维构象。在 Drugs 上的结果也足以证明了本文提出的 GFMDiff 在捕捉三维原子相互作用并生成的分子键时的强大能力。

2.2 分子性质预测

在 AI 辅助药物发现领域，分子性质预测也是十分基础的研究问题。为验证本文提出的双轨 Transformer 网络（DTN）对三维分子几何信息的学习能力，本文在两个大型公开分子数据集上进行了全面的实验。

2.2.1 实验设置

数据集:在分子性质预测实验中，本文使用了两个广泛使用的基准公开数据集：OC20 和 GEOM-QM9。在此，本节只介绍 OC20 数据集的详细信息。

- Open Catalyst 2020 (OC20)^[13] 数据集是一个较新的公开大规模数据集，用于建模和发现催化剂。具体而言，其目标是对结构弛豫进行高效的密度泛函理论 (Density functional theory / DFT) 近似计算。在催化剂研究中，结构弛豫是一项基础计算任务，其被用于确定结构的活性和选择性。数据集中的所有结构都包含一个表面和一个吸附物，表面由一个周期性的晶胞定义。该数据集中包括三个任务，分别是结构到能量和力 (S2EF)，初始结构到弛豫结构 (IS2RS) 和初始结构到弛豫能量 (IS2RE)。本文中，我们聚焦于初始结构到弛豫能量 (IS2RE) 任务，这是催化剂研究中最常见的任务，因为弛豫能量通常与催化剂的活性和选择性相关。与相关研究采取的设置类似，本文将 IS2RE 的数据集分为训练集、验证集和测试集。训练集包含 460,328 个分子结构，验证集分为域内 (ID)、域外吸附物 (OOD Ads)、域外催化剂 (OOD Cat) 和域外吸附物和催化剂 (OOD Both) 四个子集，分别包含 24,733、24,961、24,738、24,971 个结构。

评价指标：在回归任务中，本文主要采取预测结果的 MAE 作为衡量评价模型预测准确程度的指标。此外在 OC20 的任务上，本文还引入了真实能量阈值内占比（Energies within a Threshold）作为另一衡量指标。

基线模型：在基线模型的选择上，本文选取了时下最优的有可复现代码的模型用于性能对比。包括 CGCNN^[14], SchNet^[9], PhysNet^[15], MGNC^[16], DimeNet++^[17], GemNet^[18], PaiNN^[19], SphereNet^[20] 和 ComENet^[21]。

2.2.2 在 GEOM-QM9 数据集上的分子性质预测

为检测 DTN 在分子性质预测上的性能和在量子化学系统中的预测能力，我们将 DTN 应用于 QM9 数据集上的分子性质预测任务。表 2.4 展示了本文 DTN 和基线模型在回归任务上结果的 MAE 对比，任务包括 12 种性质及所有任务上的总体均方根误差 (std. MAE)。其中每项任务表现最佳和次佳的结果分别以粗体和下划线的形式予以强调。

表 2.4 GEOM-QM9 上药物分子性质预测结果对比

Property	Unit	SchNet	PhysNet	MGNC	DimeNet++	PaiNN	SphereNet	ComENet	DTN
μ	D	0.033	0.0529	0.0560	0.0297	0.012	0.0245	0.0245	<u>0.0162</u>
α	a_0^3	0.235	0.0615	<u>0.0300</u>	0.0435	0.045	0.0449	0.0452	0.0279
$\varepsilon_{\text{HOMO}}$	meV	41	32.9	42.1	24.6	27.6	<u>22.8</u>	23.1	20.7
$\varepsilon_{\text{LUMO}}$	meV	34	24.7	57.4	19.5	20.4	<u>18.9</u>	19.8	16.6
$\Delta\epsilon$	meV	63	42.5	64.2	32.6	45.7	<u>31.1</u>	32.4	28.8
$\langle R^2 \rangle$	a_0^2	0.073	0.765	<u>0.110</u>	0.331	0.066	0.268	0.259	0.145
ZPVE	meV	1.7	1.39	<u>1.12</u>	1.21	1.28	<u>1.12</u>	1.20	1.08
U_0	meV	14	8.15	12.9	6.32	<u>5.85</u>	6.26	6.59	5.34
U	meV	19	8.34	14.4	6.28	<u>5.83</u>	6.36	6.82	5.46
H	meV	14	8.42	14.6	6.53	<u>5.98</u>	6.33	6.86	5.60
G	meV	14	9.4	16.2	7.56	<u>7.35</u>	7.78	7.98	6.69
c_v	$\frac{\text{cal}}{\text{mol K}}$	0.033	0.028	0.038	0.023	0.024	<u>0.022</u>	0.024	0.021
std. MAE	%	1.76	1.37	1.86	0.98	1.01	<u>0.91</u>	0.93	0.089

DTN 在 11 个性质预测任务上取得了最佳性能，并在 1 个性质预测任务上取得了次

佳性能。同时 DTN 将 QM9 数据集的总体预测结果的均方根误差从 0.91 降低到 0.89，实现更稳定的预测结果。实验结果说明了本文的 DTN 网络在分子图学习领域同样具备良好的性能。

2.2.3 在 OC20 数据集上的分子性质预测

Open Catalyst 2020 (OC20) 数据集^[13] 是一个新发布的大规模数据集，其被用于催化剂的发现和优化。该数据集包含了数百万个 DFT 驰豫计算结果，涵盖了巨大的化学结构空间，以便可以完全训练机器学习模型。

本研究专注于 IS2RE 任务。Chanussot 等^[13] 的研究中的研究中提供了 CGCNN、SchNet 和 DimeNet++ 的结果。原始的 GemNet 论文中没有其在 OC20 数据集的结果，故本文使用 OC 项目网站上公开可用的代码^⑤ 来生成 GemNet-T 的结果。本文对 DTN 的实验设置直接采用与基线模型一致的设定，以实现公平的性能对比。此外，本文使用的评估指标是能量的平均绝对误差 (MAE) 和能量阈值内的占比 (EwT)。虑的是基于边缘的 2 跳信息，时间复杂度极高，在大型催化剂分子上可能配置不当。

表 2.5 OC20 上分子性质预测结果对比

Model	Energy MAE [eV] ↓					EwT ↑				
	ID	OOD Ads	OOD Cat	OOD Both	Average	ID	OOD Ads	OOD Cat	OOD Both	Average
CGCNN	0.6203	0.7426	0.6001	0.6708	0.6585	3.36%	2.11%	3.53%	2.29%	2.82%
SchNet	0.6465	0.7074	0.6475	0.6626	0.6660	2.96%	2.22%	3.03%	2.38%	2.65%
DimeNet++	0.5636	0.7127	0.5612	0.6492	0.6217	4.25%	2.48%	4.40%	2.56%	3.42%
GemNet-T	0.5561	0.7342	0.5659	0.6964	0.6382	<u>4.51%</u>	2.24%	4.37%	2.38%	3.38%
SphereNet	0.5632	0.6682	0.5590	0.6190	0.6024	4.56%	2.70%	<u>4.59%</u>	2.70%	<u>3.64%</u>
ComENet	0.5558	0.6602	0.5491	0.5901	0.5888	4.17%	<u>2.71%</u>	4.53%	<u>2.83%</u>	3.56%
DTN	<u>0.5560</u>	0.6591	0.5443	0.5872	0.5867	4.35%	2.82%	4.76%	2.98%	3.73%

表 2.5 显示了 DTN 在能量 MAE 方面在 4 个子任务中有 3 个最佳表现，并在平均值上表现最好。在 EwT 方面，DTN 在 3 个子任务都表现最佳。具体而言，它将预测的平均能量 MAE 降低了 0.021。此外，DTN 也将平均 EwT 从 3.64% 提高到 3.73%，考虑到本身较低的 EwT 值，这是一个很大的改进。

^⑤ <https://github.com/Open-Catalyst-Project/ocp>

值得注意的是，近年涌现了新量子系统学习模型 ForceNet^[22] 和 GemNet^[18]。ForceNet 的一个显着优势是其在大分子上的高效性和可扩展性，其专注于 S2EF 任务，故没有 IS2RE 任务的原始结果。经过考察，DimeNet++ 和 SphereNet 在性能上略优于 ForceNet，而我们的 DTN 在性能上显著优于这些基线模型。

GemNet 有两个变体，GemNet-T 和 GemNet-Q。GemNet-T 将距离和角度信息作为输入，并包含有效的架构和新颖的网络组件，如双线性层和缩放因子。从表 2.5 可知，GemNet-T 在性能上与 DimeNet++ 相似。GemNet-Q 声称能够捕捉到分子的通用表示，然而其考虑的是基于边缘的 2 跳信息，时间复杂度极高，在大型催化剂分子上可能配置不当。

总而言之，在两个分子学习数据集上的表现充分证明了 DTN 在大多数指标上达到了时下最优的水平，引入的完全的几何信息提取机制能够有效的提升在分子性质预测和分子生成任务上的表现。

参考文献

- [1] HOOGEBOOM E, SATORRAS V G, VIGNAC C, et al. Equivariant Diffusion for Molecule Generation in 3D[C]//Proceedings of Machine Learning Research: Proceedings of the 39th International Conference on Machine Learning: vol. 162. [S.I.]: PMLR, 2022: 8867-8887.
- [2] HUANG L, ZHANG H, XU T, et al. Mdm: Molecular diffusion model for 3d molecule generation[C/OL]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 37: 4. [S.I. : s.n.], 2023: 5105-5112. <https://ojs.aaai.org/index.php/AAAI/article/view/25639>. DOI: 10.1609/aaai.v37i4.25639.
- [3] GARCIA SATORRAS V, HOOGEBOOM E, FUCHS F, et al. E(n) Equivariant Normalizing Flows[C] //Advances in Neural Information Processing Systems: vol. 34. [S.I.]: Curran Associates, Inc., 2021: 4181-4192.
- [4] XU M, YU L, SONG Y, et al. GeoDiff: A Geometric Diffusion Model for Molecular Conformation Generation[C]//International Conference on Learning Representations. [S.I. : s.n.], 2022.
- [5] RAMAKRISHNAN R, DRAL P O, RUPP M, et al. Quantum chemistry structures and properties of 134 kilo molecules[J]. Scientific data, 2014, 1(1): 1-7.
- [6] AXELROD S, GOMEZ-BOMBARELLI R. GEOM, energy-annotated molecular conformations for property prediction and molecular generation[J]. Scientific Data, 2022, 9(1): 185.
- [7] SATORRAS V G, HOOGEBOOM E, WELLING M. E(n) Equivariant Graph Neural Networks[C] //Proceedings of Machine Learning Research: Proceedings of the 38th International Conference on Machine Learning: vol. 139. [S.I.]: PMLR, 2021: 9323-9332.
- [8] GEBAUER N, GASTEGGER M, SCHÜTT K. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules[C/OL]//Advances in Neural Information Processing Systems: vol. 32. [S.I.]: Curran Associates, Inc., 2019. <https://proceedings.neurips.cc/paper/2019/file/a4d8e2a7e0d0c102339f97716d2fdfb6-Paper.pdf>.
- [9] SCHÜTT K, KINDERMANS P J, SAUCEDA FELIX H E, et al. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions[C]//Advances in Neural Information Processing Systems: vol. 30. [S.I.]: Curran Associates, Inc., 2017.
- [10] GONG C, WUL, LIUX, et al. Diffusion-based Molecule Generation with Informative Prior Bridges[C/OL] //NeurIPS 2022 AI for Science: Progress and Promises. [S.I. : s.n.], 2022. <https://openreview.net/forum?id=QagNEt9k8Vi>.
- [11] MOREHEAD A, CHENG J. Geometry-Complete Diffusion for 3D Molecule Generation[C]//ICLR 2023 - Machine Learning for Drug Discovery workshop. [S.I. : s.n.], 2023.

- [12] DU W, ZHANG H, DU Y, et al. SE(3) Equivariant Graph Neural Networks with Complete Local Frames[C]//Proceedings of Machine Learning Research: Proceedings of the 39th International Conference on Machine Learning: vol. 162. [S.l.]: PMLR, 2022: 5583-5608.
- [13] CHANUSSOT L, DAS A, GOYAL S, et al. Open Catalyst 2020 (OC20) Dataset and Community Challenges[J/OL]. ACS Catalysis, 2021, 11(10): 6059-6072. eprint: <https://doi.org/10.1021/acscatal.0c04525>. <https://doi.org/10.1021/acscatal.0c04525>. DOI: 10.1021/acscatal.0c04525.
- [14] XIE T, GROSSMAN J C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties[J]. Phys. Rev. Lett., 2018, 120: 145301. DOI: 10.1103/PhysRevLett.120.145301.
- [15] UNKE O T, MEUWLY M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges[J/OL]. Journal of Chemical Theory and Computation, 2019, 15(6): 3678-3693. eprint: <https://doi.org/10.1021/acs.jctc.9b00181>. <https://doi.org/10.1021/acs.jctc.9b00181>. DOI: 10.1021/acs.jctc.9b00181.
- [16] LU C, LIU Q, WANG C, et al. Molecular Property Prediction: A Multilevel Quantum Interactions Modeling Perspective[C/OL]//AAAI'19/IAAI'19/EAAI'19: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. Honolulu, Hawaii, USA: AAAI Press, 2019. <https://doi.org/10.1609/aaai.v33i01.33011052>. DOI: 10.1609/aaai.v33i01.33011052.
- [17] GASTEIGER J, GIRI S, MARGRAF J T, et al. Fast and uncertainty-aware directional message passing for non-equilibrium molecules[J]. ArXiv preprint arXiv:2011.14115, 2020.
- [18] GASTEIGER J, BECKER F, GÜNNEMANN S. GemNet: Universal Directional Graph Neural Networks for Molecules[C/OL]//RANZATO M, BEYGELZIMER A, DAUPHIN Y, et al. Advances in Neural Information Processing Systems: vol. 34. [S.l.]: Curran Associates, Inc., 2021: 6790-6802. https://proceedings.neurips.cc/paper_files/paper/2021/file/35cf8659cfcb13224cbd47863a34fc58-Paper.pdf.
- [19] SCHÜTT K, UNKE O, GASTEGGER M. Equivariant message passing for the prediction of tensorial properties and molecular spectra[C/OL]//MEILA M, ZHANG T. Proceedings of Machine Learning Research: Proceedings of the 38th International Conference on Machine Learning: vol. 139. [S.l.]: PMLR, 2021: 9377-9388. <https://proceedings.mlr.press/v139/schutt21a.html>.
- [20] LIU Y, WANG L, LIU M, et al. Spherical Message Passing for 3D Molecular Graphs[C]//International Conference on Learning Representations. [S.l. : s.n.], 2022.

- [21] WANG L, LIU Y, LIN Y, et al. ComENet: Towards Complete and Efficient Message Passing for 3D Molecular Graphs[C]//Advances in Neural Information Processing Systems. [S.l. : s.n.], 2022.
- [22] HU W, SHUAIBI M, DAS A, et al. ForceNet: A Graph Neural Network for Large-Scale Quantum Calculations[J]., 2021. arXiv: 2103.01436 [cs.LG].

A 化学键长

本文在生成药物分子中的化学键时，依据计算化学领域常见的原子键长作为参考。表 A.1，表 A.2 和表 A.3 列出了出现在 GEOM-QM9 和 GEOM-Drugs 中原子类型间所有可能的化学键的典型键长。表中数值代表对应化学键的典型长度（皮米），而横线代表该两种原子间不可能存在稳定的对应化学键连接，原子间距离短于典型键长则可以被认为被对应的键连接。在实际计算中，对单键，双键和三键典型键长分别有 10, 5 和 3 皮米的冗余。例如，对于两个距离 136 皮米的碳原子，他们之间距离虽然大于 134 皮米，小于 154 皮米，但本文认定 136 皮米小于 134+5 皮米，则认为这两个碳原子由双键连接。

表 A.1 典型单键键长

	H	C	O	N	P	S	F	Si	Cl	Br	I	B	As
H	74	109	96	101	144	134	92	148	127	141	161	119	152
C	109	154	143	147	184	182	135	185	177	194	214	-	-
O	96	143	148	140	163	151	142	163	164	172	194	-	-
N	101	147	140	145	177	168	136	-	175	214	222	-	-
P	144	184	163	177	221	210	156	-	203	222	-	-	-
S	134	182	151	168	210	204	158	200	207	225	234	-	-
F	92	135	142	136	156	158	142	160	166	178	187	-	-
Si	148	185	163	-	-	200	160	233	202	215	243	-	-
Cl	127	177	164	175	203	207	166	202	199	214	-	175	-
Br	141	194	172	214	222	225	178	215	214	228	-	-	-
I	161	214	194	222	-	234	187	243	-	-	266	-	-
B	119	-	-	-	-	-	-	-	175	-	-	-	-
As	152	-	-	-	-	-	-	-	-	-	-	-	-

表 A.2 典型双键键长

	C	O	N	P	S
C	134	120	129	-	160
O	120	121	121	150	-
N	129	121	125	-	-
P	-	150	-	-	186
S	-	-	-	186	-

表 A.3 典型三键键长

	C	O	N
C	120	113	116
O	113	-	-
N	116	-	110

B 超参数设置

由于完整跑完一轮训练所需时间较长，本文未对超参数设置进行完整的实验。但经过初步实验，本文确定了实验的相关超参数。在 QM9 和 Drugs 数据集上的模型训练时，学习率设定为 0.001，DTN 层数设定为 5 层，特征维度设定为 256，注意力头数设定为 8，神经元遗忘率为 0.1，独热编码原子类型标准化系数为 0.25，原子序数标准化系数为 0.1，化合价标准化系数为 0.1。在 QM9 数据集上的模型训练中，扩散步数为 500，每批样本数量设定为 32，而在 Drugs 上的训练中，扩散步数为 1000，每批样本数量为 1，这一设定受制于 GPU 显存容量限制。为实现等效训练的效果，本文通过梯度累积的方式，实现了等效批样本数量 64 的训练效果。

攻读硕士学位期间取得的研究成果

一、学术论文

1. XU C, ZHANG Y, WANG W, DONG L. Pursuit and evasion strategy of a differential game based on deep reinforcement learning[J]. *Frontiers in Bioengineering and Biotechnology*, 2022, 10: 827408.
2. ZHANG Y, XU C, WU X, ZHANG Y, DONG L, WANG W. LFGCF: Light folksonomy graph collaborative filtering for tag-aware recommendation[J]. *Expert Systems with Applications*, 2022, Under Review.
3. XU C, ZHANG Y, CHEN H, DONG L, WANG W. A fairness-aware graph contrastive learning recommender framework for social tagging systems[J]. *Information Sciences*, 2023: 119064.
4. CHEN H, XU C, ZHENG L, ZHANG Q, LIN X. Diffusion-based graph generative methods[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2023, Under Review.
5. XU C, WANG H, ZHENG P, WANG W, CHEN H. Geometric-facilitated Denoising Diffusion Model for 3D Molecule Generation[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Under Review.

二、科研项目

1. 国家重点研发计划之先进计算与新兴软件：面向分布式异构计算系统内存池化关键技术

三、发明专利

1. 信号干扰下的超带宽精确定位方法，专利申请号：202210119309.7

四、科研竞赛

1. “华为杯”第十八届中国研究生数学建模竞赛，三等奖，排名：1/3。
2. 第五届全国应用统计专业学位研究生案例大赛，三等奖，排名：1/3。
3. OGB-LSC @NeurIPS 2022 (PCQM4Mv2 Track), NO.11, 排名：1/5。
4. 第十一届“泰迪杯”数据挖掘挑战赛，三等奖，排名：1/3。

致谢

志之所趋，无远弗届，穷山距海，不能限也。

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含本人为获得浙江工商大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名:

导师签名:

签字日期: 年 月 日

签字日期: 年 月 日

关于论文使用授权的说明

本学位论文作者完全了解浙江工商大学有关保留、使用学位论文的规定：浙江工商大学有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅，可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文，并且本人电子文档的内容和纸质论文的内容相一致。

本论文提交 即日起/ 半年/ 一年以后，同意发布。

“内部”学位论文在解密后也遵守此规定。

学位论文作者签名:

导师签名:

签字日期: 年 月 日

签字日期: 年 月 日