中国科学: 化学 2023 年 第 53 卷 第 2 期: 174 ~ 195
SCIENTIA SINICA Chimica chemcn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS

评 述

基于深度学习的3D分子生成模型研究进展

姚少伦[1], 宋杰[2*], 冯尊磊[2], 贾凌翔[3], 钟子鹏[3], 宋明黎[3]

1. 浙江大学人工智能省部共建协同创新中心, 杭州 310007
2. 浙江大学软件学院, 宁波 315020
3. 浙江大学计算机科学与技术学院, 杭州 310027
*通讯作者, E-mail: sjie@zju.edu.cn

收稿日期: 2022-10-27; 接受日期: 2022-12-07; 网络版发表日期: 2023-01-09
浙江省基础公益项目(编号: LGG22F020007), 浙江大学上海高等研究院繁星基金(编号: SN-ZJU-SIAS-001)和中央高校基本业务费(编号: 2021FZZX001-23)

摘要 分子生成作为药物设计领域的一个基本问题, 旨在以低成本和高效率的方式设计出具有理想生物活性和药代动力学属性的新颖分子. 近年来深度生成模型在药物设计中得到了广泛应用, 大量的模型结构和优化策略得到探索, 其中大多是生成一维或二维的分子结构. 随着深度学习在处理几何图形数据上的快速发展, 面向3D分子的生成模型被提出, 因其在直接生成3D分子构象和基于结构的药物设计上的优势和潜力而越来越受到关注. 本文对近年来国内外学者在3D分子生成上取得的成果进行了系统的总结和分析, 从3D分子生成算法输入的角度将其分为基于隐变量的生成、基于2D分子图的生成和基于3D分子构象的生成; 接着从3D分子生成算法输出的角度将其分类为定向生成和非定向生成; 随后总结了相关生成模型在主要的公开数据集上的性能, 以探究各种生成模型的优缺点; 最后对未来可能的研究方向进行了展望.

关键词 3D分子生成, 等变图神经网络, 深度生成模型, 定向生成, 非定向生成

1 引言

新药研发被公认是一个时间长、耗费大、风险大的漫长过程. 一款新药从药物发现、临床前试验、临床实验到审批上市, 平均需要花费10年时间和26亿美元研发费用, 投资回报率不到2%[1]. 药物发现的主要目标是针对人类疾病研发出具有所需功能且对人体安全的药物, 包括靶标发现和验证、虚拟筛选、先导化合物设计和优化、临床前候选化合物优化等一系列步骤. 先导化合物设计和优化的常用策略是从具有潜力的已知活性分子入手来进行改造, 其目标是通过优化已有的苗头化合物来寻找药物活性更高、药代动力学性质更好且副作用更小的新型分子. 传统的计算机辅助药物设计技术如虚拟筛选, 通常利用专家制定的规则, 需要大量的数据和特征工程, 且很难在巨大和离散的类药化学空间(10^{23}~10^{60}量级)[1]中搜索到新颖的化合物.

随着深度学习技术的发展和定量构效关系(Quantitative structure–activity relationship, QSAR)数据的剧增, 多种深度生成模型联合强化学习或贝叶斯优化算

引用格式: Yao S, Song J, Feng Z, Jia L, Zhong Z, Song M. Advances in deep learning-based 3D molecular generative models. *Sci Sin Chim*, 2023, 53: 174–195, doi: 10.1360/SSC-2022-0204

© 2023 《中国科学》杂志社

法在生成具有理想属性、新颖且多样的分子上显示出较大潜力，这为学术界和制药行业大幅缩减了新药研发的时间和成本. 近年来基于深度学习的分子生成方法主要包括基于SMILES和基于分子图表示的方法. 尽管这些分子表示方法在各种应用中都很有效，但更本质和更富含信息的分子表示是3D几何，其中原子由其笛卡尔坐标表示. 在化学信息学和计算化学等领域处理3D分子构象至关重要，这是因为它决定了分子的生物、化学和物理性质，如电荷分布、势能、分子对接、药效团搜索等. 比如在药物设计中，理解分子如何与特定的靶蛋白结合至关重要，而这一过程很大程度上取决于这两者的3D构象，包括几何(形状匹配)和化学(疏水、亲水)相互作用[2]. 预测3D分子构象的传统方法包括实验测定和基于力场或量子力学的第一性原理预测, 这些方法十分耗时且成本高昂. 随着几何深度学习在处理几何图形数据上的快速发展[3]，近年来面向3D分子构象的深度生成模型逐渐成为研究热点.

分子生成(molecular generation)利用深度学习技术来学习分子数据集的概率分布，并生成具有理想生物活性和药代动力学属性的新颖分子. 分子生成通过在类药化学空间中搜索和采样多目标属性约束的新分子, 有效减小了搜索空间, 进而加速上述药物发现的进程. 现有的绝大多数基于字符串或基于图的分子生成方法由于缺乏空间信息而无法识别出空间异构体、原子间非键相互作用、以及分子和靶标蛋白在三维空间中的相互作用，因而在实际药物设计场景中受到限制. 3D分子生成(3D molecular generation)有助于解决上述挑战，旨在设计出具有理想生物化学属性和几何属性的新颖分子. 比如在基于受体结构的药物设计中，不同于2D分子图的简化结构表示, 3D分子生成算法生成的3D构象可以在生成过程中考虑分子内或分子间相互作用, 以及来自蛋白口袋的局部约束, 因而能更好地保证预测的结构具有生物活性或药物属性. 根据生成算法的输入将现有的3D分子生成方法总结为三大类, 第1类基于隐变量的生成是指在没有给出参考分子的情况下生成3D分子构象，即从零开始生成3D分子构象；第2类基于2D分子图的生成是指给定2D分子图来生成或优化3D分子构象；第3类基于3D分子构象的生成是指给定3D分子构象来生成或优化3D分子构象. 根据生成算法的输出将3D分子生成方法划分为定向生成和非定向生成. 非定向生成是指利用生成模型学习训练数据集的分布并从中采样出具有新颖、多样、化学有效以及几何有效(稳定态)等通用属性的3D分子构象. 定向生成是指在符合这些通用属性的基础上还需满足多种特定的生物活性和药物属性，如对特定靶蛋白具有高结合亲和力、靶标选择性、化学可合成性、安全性等. 具体分类以及代表模型如图1和表1所示, 各方法的创新点和优劣性将在第3~5节着重讨论.

3D分子生成是药物发现领域一个极具挑战的任务，其存在的问题和难点主要包括以下几个方面：(1)搜索空间巨大：由键长、键角和扭转角组成的三维结构空间巨大，虽然分子图对可能的三维构象施加了特定的约束(例如键长范围取决于键类型, 四面体中心决定了局部原子的空间排列), 然而可能的构象空间随着图尺寸和可旋转键数量的增加呈指数增长，这阻碍了穷尽探索的可能性. (2)对称性归纳偏置：3D分子生成需要处理几何图形式的数据. 不同于一般的图数据, 几何图不仅为每个节点分配一个特征，还为每个节点分

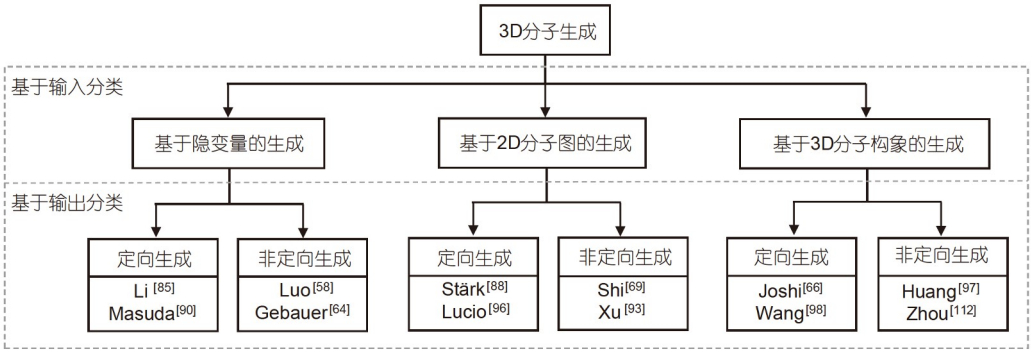


图 1　3D分子生成算法分类 (网络版彩图)

Figure 1　Classification of 3D molecular generation algorithm (color online).

表 1 3D分子生成算法分类

Table 1 Classification of 3D molecular generation algorithm

算法分类		算法名称	输入	生成方法	输出[a)]
基于隐变量的生成	非定向生成	G-SchNet[64]	Attribute	AR	原子间距离
		EDMNet[54]	White noise	GAN	原子间距离
		3DMolNet[89]	Gaussian distribution	VAE	原子间距离
		MolGym[76]	None	RL	原子间距离、键角、扭转角
		COVARIANT[77]	None	RL	原子坐标
		E-NFs[57]	Gaussian distribution	Flow	原子坐标
		G-SphereNet[58]	Gaussian distribution	Flow+AR	原子间距离、键角、扭转角
		EDM[72]	Gaussian distribution	Diffusion	原子坐标
	定向生成	AutoGrow4[83]	None	GA	原子坐标
		liGAN[90]	None	VAE+GAN	原子坐标
		DeepLigBuilder[85]	None	MCTS+RL	原子坐标
		DSBDD[92]	None	AR	原子坐标
基于2D分子图的生成	非定向生成	CVGAE[49]	2D Graph	VAE	原子坐标
		GraphDG[50]	2D Graph	VAE	原子间距离
		ConfVAE[51]	2D Graph+3D Geometry	VAE	原子间距离
		ConfGF[69]	2D Graph	Diffusion	距离梯度场
		DGSM[70]	2D Graph	Diffusion	距离梯度场
		ClofNet[110]	2D Graph	Diffusion	距离梯度场
		GeoDiff[71]	2D Graph	Diffusion	原子坐标
		Torsional Diffusion[73]	2D Graph	Diffusion	扭转角
		GeoMol[87]	2D Graph	OT	原子坐标
		CGCF[93]	2D Graph	Flow	原子间距离
		BOKEI[80]	2D Graph	BO	扭转角
	定向生成	DeepDock[96]	2D Graph	EBM	原子坐标
		EquiBind[88]	2D Graph	OT	原子坐标
基于3D分子构象的生成	非定向生成	3DLinker[97]	3D Geometry	VAE	原子坐标
		Uni-Mol[112]	3D Geometry	Transformer	原子坐标
	定向生成	BOA[79]	3D Geometry	BO	原子坐标
		3D-Scaffold[66]	3D Geometry	AR	原子坐标
		cG-SchNet[67]	3D Geometry	AR	扭转角
		Coarse-GrainingVAE[98]	3D Geometry	VAE	原子坐标
		TorsionNet[99]	3D Geometry	RL	扭转角
		RGA[84]	3D Geometry	GA	原子坐标

a) 输出列中原子坐标表示生成模型直接输出原子坐标, 其余表示生成模型首先生成中间几何元素(包括原子间距离或其梯度场、键角、扭转角), 然后根据中间几何元素来生成原子坐标.

配一个几何向量. 分子可以用几何图来表示, 其中原子的3D坐标是几何向量. 分子的3D构象在空间中具有旋转、平移的几何对称性, 这是因为分子在3D空间中发生旋转、平移后其3D坐标会发生变化, 但控制原子动力学的物理定律是不变的. 在处理分子的3D几何图数据时, 将对称性这种归纳偏置纳入模型设计是至关重

要的，这也推动了几何等变图神经网络的发展. (3)构象分布多模态: 一个分子体系的总能量随其结构的变化而改变的方式可以通过势能面来描述，势能面上存在多个能量高点和低点，其中最低处称为全局能量最低点，其他能量较高的极小点则称为局部能量最低点. 真实的分子倾向于以能量较低的稳定态构象存在，对应于势能面上的能量低点，这个数量在一到几千之间变化，如何有效搜索这些势能面上的局部能量最低点是一大挑战. 最近，扩散模型凭借其强大的生成能力在建模复杂多模态构象分布上显示出优势，能更好地生成高质量和多样性的结构. 我们将在下文介绍扩散模型在3D分子生成建模上的出色表现. 现有的机器学习模型正试图从不同角度解决上述挑战，目前已有一些综述对此领域进行了报道[4~7]. 为了促进更多研究者的参与，我们对3D分子生成模型从算法到应用的最新进展进行系统回顾. 本文主要内容安排如下: 首先，概述了3D分子生成的基本概念; 其次，梳理了基于隐变量的生成、基于2D分子图的生成和基于3D分子构象的生成这三种方法的研究进展，每一种方法又根据3D分子生成算法的输出分为定向生成和非定向生成进行讨论; 随后, 总结了相关生成模型在主要的公开数据集上的性能, 来说明各种生成模型的优缺点; 最后, 对3D分子生成领域进行总结和展望.

2　分子生成的基本概念

2.1　分子表示

神经网络能将复杂的输入表示转换成解决特定任务所需的潜在空间的表示，因此输入表示的选择对于模型如何学习分子信息至关重要. 在本节中，我们将介绍分子表示(molecular representation)的三类主流方法: 1D字符串, 2D分子图和3D分子几何. 图2给出了一个特定分子(青霉烷)的三种表示方法示例.

2.1.1　1D字符串表示

最常见的1D字符串表示类型为简化分子线性输入规范(Simplified Molecular Input Line Entry System, SMILES)，由David Weininger和Arthur Weininger于1986年开发，并由日光化学信息系统(Daylight Chemical Information Systems)改进和扩展[8]. SMILES是一种简单的基于ASCII字符的表示，根据预定义的原子排序规则将一个分子转换成一个字符序列，能够明确地描述原子、键、环、芳香性和立体化学等信息. 同一个分子由于起点原子差异可以表示成多个不同的SMILES, 实际使用时选择统一的规范化SMILES (Canonical SMILES)表示法，而通用SMILES可用于数据增强. 由于SMILES基于序列表示的特性，自然语言处理算法可以自然地被移植到分子生成领域. 先前的研究工作[9~12]利用循环神经网络(Recurrent Neural Network, RNN)、长短期记忆网络(Long Short-Term Memory, LSTM)等自回归生成模型生成分子的SMILES字符串. 这些方法忽略了SMILES语法的复杂性，容易生成无效的SMILES，即生成的SMILES不符合化合价规则导致不能转换成分子结构. 为了解决这个问题，Krenn等人[13]设计了一种改进的字符串表示方法，称为自引用嵌入式字符串(Self-Referencing Embedded Strings, SELFIES)，基于SELFIES的RNN模型可以生成100%有效的分子. 另外，在化学反应预测任务中，上述针对单一分子设计的方法忽略了反应物和生成物SMILES之间的关系，我们课题组[14]提出了一种用于化学反应预测的紧凑的分子字符串表示R-SMILES，通过将输入和输出SMILES的根原子进行对齐的方式，

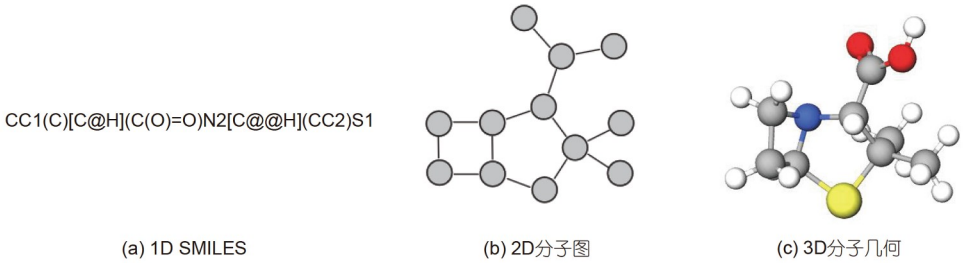


图 2　一个示例分子(青霉烷)的三种分子表示 (网络版彩图)

Figure 2　Three molecular representations for an exemplary molecule (penicillin) (color online).

避免模型学习复杂的SMILES语法, 从而专注于学习化学反应相关的化学知识. 此外, 1D字符串表示还包括国际化学物标识(International Chemical Identifier, InChI)[15]和SMARTS(SMiles ARbitrary Target Specification)[16].

2.1.2 2D分子图表示

表示分子更直观的方法是根据分子的路易斯结构将其转化成一个分子图(molecular graph), 其中原子表示成节点, 化学键表示成边, 并根据原子类型和化学键类型为节点和边赋值对应的标签. 与1D字符串表示相比, 2D分子图表示具有如下优势: 一方面, 图表示能直接捕获原子之间的连接, 而该信息在1D字符串表示中需要模型学习其复杂的语法规则; 另一方面, 如上文所述, 基于SMILES的分子生成方法无法确保生成的分子具有100%化学有效性, 而基于2D分子图的生成方法很容易确保生成的分子是100%化学有效的. 近年来已有许多研究工作致力于开发能生成真实分子图的神经网络架构. 2D分子生成根据生成策略可分为三类: (1)一次输出图的节点类型和邻接矩阵来生成分子图, 有许多深度生成模型可以一次性生成整个分子, 如基于变分自编码器(Variational Autoencoder, VAE)的VGAE[17]、RVAE[18]、GraphVAE[19], 基于生成对抗网络(Generative Adversarial Network, GAN)的MolGAN[20], 基于流(Flow)模型的GraphNVP[21]等. (2)基于子结构生成分子图, 即将子结构或官能团作为构建块来生成高质量的分子. 代表性模型有基于VAE的JTVAE[22]、HierVAE[23], 基于强化学习(Reinforcement Learning, RL)的RationaleRL[24]、MolEvol[25], 基于采样的MARS[26]、MIMOSA[27]等. (3)通过顺序添加节点和边生成分子图, 代表性模型有基于VAE的CGVAE[28], 基于RL的GCPN[29]、DeepGraphMolGen[30], 基于Flow的GraphAF[31]、GraphDF[32]等.

2.1.3 3D分子几何表示

现有的基于1D字符串或2D分子图的生成模型没有考虑化合物分子的空间构象, 例如同分异构体具有一样的拓扑结构, 但其3D空间结构是不同的. 捕获3D空间结构特征对于准确预测分子的物理化学性质如量子性质、分子构象生成至关重要. 因此, 更本质和更有信息的分子表示是3D几何(3D geometry), 又称3D分子构象(3D molecular conformation), 它包含了3D几何信息(如位置)和重要的物理规律(如对称性). 3D几何将原子坐标加入到它们的表示中, 并能够描述三维空间中原子之间的相对位置. 这里我们考虑处于平衡态的构象, 对应于势能面上的局部最低点. 与1D字符串和2D分子图表示相比, 3D几何表示显著扩展了有待探索的分子空间, 同时也增加了难度, 主要挑战在于模型需要在3D空间中对旋转和平移操作保持等变性. 最近, 3D分子几何表示的研究工作在探索等变性上主要通过以下两种策略, 一种是等变图神经网络, 代表性工作有SE(3)-Transformers[33], EGNN[34]等; 另一种是借助中间几何元素如原子间距离和角度信息来使用E(3)不变的三维信息, 代表性工作有SchNet[2], PhysNet[35], DimNet[36]等. 下文将详细介绍该领域现有的研究进展并对其进行分类整理. 此外, 分子的3D表示还包括3D密度图(3D density map)和3D表面网格(3D surface mesh). 分子的3D密度图由组成分子的所有原子的体积组成[37], 常用3D网格(3D grid)表示, 可用3D卷积神经网络(3D Convolutional Neural Network, 3DCNN)对其进行编码[38]. 3D表面网格由定义分子表面形状的节点、边和面组成, 也常用3D网格表示, 可用测地线卷积神经网络(Geodesic Convolutional Neural Network, GCNN)[39]或3D图神经网络(3D Graph Neural Network, 3DGNN)[2]对其进行编码. 这些分子表示方法在分子生成任务中应用较少, 在这里仅作简单介绍以便读者了解.

2.2 等变图神经网络

2.2.1 等变性

物体(或系统)的对称性(symmetry)是指对其变换操作后, 物体的某一特性保持等变(equivariant)或不变(invariant)的性质[40]. 对称性在物理系统中无处不在, 例如动量守恒中的空间平移不变性, 分子构象、蛋白质或点云的旋转对称性. 在建模物理系统时, 赋予深度学习模型这种归纳偏置对于模型的训练和泛化能力都至关重要. 下面给出等变性的定义和相关符号说明.

$E(n)$等变($E(n)$-equivariance)是指模型在n维欧式空间中对平移、旋转、反射具有等变性[41]. 设有一个函数$\phi: X \to Y$, X是输入, Y是输出, 定义某变换群G, 对于任意群元素$g \in G$, 若有

$$\phi\left(T_g(x)\right) = S_g(\phi(x)) \tag{1}$$

则称函数ϕ对变换群G是等变的, 其中T_g指变换g在输入空间中的群表示, S_g是指同一个变换g在输出空间中的群表示, 上式说明了对输入的变换操作等同于对输出施加相同的变换. 当$S_g = I$, 即输出的群表示是单位矩阵时, 式(1)表示不变性(invariance), 即对输入做群中的任何变换, 输出保持不变. 神经网络$\phi: y = \phi(x)$的$E(n)$等变性包括以下三种对称性变换, 图3展示了其中的旋转等变性.

(1)平移等变性(translation equivariance): 对输入施加一段平移$g \in \mathbb{R}^n$, 则输出也发生了同样的平移, 有$y + g = \phi(x + g)$.

(2)旋转等变性(rotation equivariance): 对输入进行旋转导致输出也发生同样的旋转, 用正交矩阵$Q \in \mathbb{R}^{n\times n}$表示旋转变换, 则有$Qy = \phi(Qx)$.

(3)反射等变性(reflection equivariance): 对输入关于平面(或直线)作反射(翻转)后得到输出, 即输入和输出互为镜像, 用正交矩阵$P \in \mathbb{R}^{n\times n}$表示反射变换, 则有$Py = \phi(Px)$.

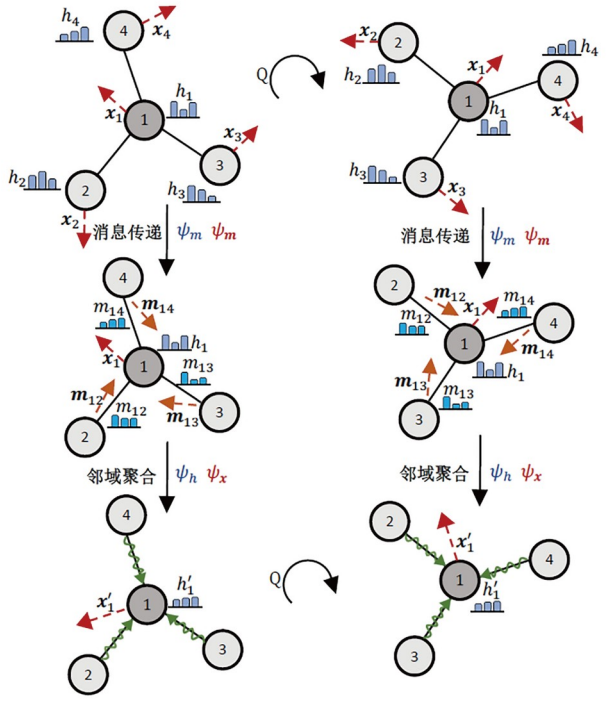


图 3　几何等变消息传递(旋转情况)示意图[41] (网络版彩图)
Figure 3　An illustration of the geometrically equivariant message passing in the case of rotation[41] (color online).

2.2.2　图神经网络

图神经网络(Graph Neural Network, GNN)是近年来出现的一种利用神经网络来学习图结构数据, 提取发掘图结构数据中特征和模式的框架, 其优异的性能引起了学者的高度关注和深入探索[42]. 通过在图中的节点和边上制定一定的策略, GNN将图结构数据转化为规范而标准的表示, 并输入到多种不同的神经网络中进行训练, 在节点分类、边信息传播和图分类等任务上取得优良的效果. 当应用于分子时, GNN通过逐步聚合原子及分子环境的信息来提取特征. 下面给出GNN 中相关符号的说明和定义.

图(Graph)是由节点(Vertex)和连接节点的边(Edge)构成, 记为$G = (V, E)$. 其中$V = \{v_1, v_2, ..., v_n\}$表示节点集合, $E = \{e_1, e_2, ..., e_m\}$表示边集合. 每个节点i的节点特征表示为$h_i$, 连接节点$i$和$j$的边特征表示为$e_{ij}$. 近年来多种GNN框架被提出[43], 如图卷积神经网络(Graph Convolutional Network, GCN)、图注意力网络(Graph Attention Network, GAT)和图同构网络(Graph Isomorphism Network, GIN)等, 其中最流行的是Gilmer等人[44]提出的消息传递神经网络(Message Passing Neural Network, MPNN). MPNN将GNN统一到了一个通用的框架中, 它迭代地对每个节点或边进行消息传递和邻域聚合. MPNN的消息传播机制表示如下:

$$m_{ij} = \phi_m\left(h_i, h_j, e_{ij}\right)$$
$$h_i = \phi_h\left(\sum_{j\in\mathcal{N}(i)} m_{ij}, h_i\right) \tag{2}$$

其中m_{ij}是节点i经消息传递后生成的信息, ϕ_m和ϕ_h分别是消息传递函数和邻域聚合函数, $\mathcal{N}(i)$是节点i的邻居. 邻域聚合操作ϕ_h满足置换等变性, 即改变输入顺序, 输出顺序也发生相应改变. 有几项研究工作利用该置换等变性在降低节点排列组合复杂度上的优势, 将GNN应用于具有几何输入的复杂系统, 例如粒子的三维坐标. 如Li等人[45]提出的动态粒子交互网络DPI-Net和Mrowca等人[46]提出的层次化关系网络HRN通过构建动态交互图来模拟粒子动力学. Kipf等人[47]提出的神经关系推断模型NRI通过自动推断潜在交互图, 在多体粒子系统动力学模拟和人体运动捕获方面取得了优异性能. 此外, 在量子化学任务中, GNN通常借助中间几何元素如原子间距离和角度信息来使用$E(3)$不变的三维信息, 进而提高平衡态和非平衡态分子构象的力

和能量的预测精度, 代表性方法有Schütt等人提出的通用量子化学深度学习架构SchNet[2]和极化原子交互网络PaiNN[3].

当节点特征除了标量还包含实体坐标, 如分子图的节点特征包括原子类型及其3D空间坐标信息时, 还需要考虑原子的3D空间坐标会随着分子在空间中变换而发生变换. 也就是说除了上述讨论的GNN标准置换等变性, 还需要模型满足平移、旋转和反射等变性. 关于$E(3)$等变性在分子3D空间中如何发挥作用, 我们将在下文介绍等变图神经网络及其在3D分子构象生成任务中的应用.

2.2.3 等变图神经网络

基于图的方法中, 最近发展的一个领域是$E(n)$等变图神经网络(Equivariant Graph Neural Network, EGNN), 它们处理n维图的绝对坐标系. 因此, 这些网络可能适合应用于分子的3D表示. 根据2.2.2对图的定义, 将图表示为$G=(V,E)$, 具有节点$v_i \in V$和边$e_{ij} \in E$. 在分子3D几何图中, 对于每个节点的特征, 除了之前的节点特征$h_i \in R^H$, 还包括节点的三维坐标$x_i \in \mathbb{R}^3$. Victor等人[34]提出的EGNN不仅能在节点坐标x_i上保持平移、旋转和反射等变性, 还能在节点上保持前文所述的GNN标准置换等变性. 具体来说, EGNN采用了基于标量化建模等变性质的方法, 通过将几何特征转化为不变的标量, 利用多层感知机(Multilayer Perceptron, MLP)得到标量变化, 然后将此标量变化加回到原有的几何特征上从而得到等变性. EGNN的通用框架表示如下:

$$\begin{aligned} m_{ij} &= \phi_m\left(h_i, h_j, \left\|x_j - x_i\right\|^2, e_{ij}\right) \\ x_i &= x_i + C\sum_{j\neq i}(x_i - x_j)\phi_x(m_{ij}) \\ h_i &= \phi_h\left(h_i, \sum_{j\in\mathcal{N}(i)} m_{ij}\right) \end{aligned} \tag{3}$$

其中$\left\|x_j - x_i\right\|$是几何特征的标量化, ϕ函数为不同的MLP, 式(3)的邻域聚合步骤将几何信息和非几何信息进行了关联, 保证了几何特征和非几何特征在消息传播过程中的等变性.

2.3 模型框架

分子生成的模型框架主要有两类. 第一类是深度生成模型, 它旨在利用深度学习模型对数据分布进行建模. 第二类是利用启发式策略来实现目标的组合优化方法. 两者在利用结构化数据上的主要区别在于, 深度生成模型利用的是连续隐空间的表示, 而组合优化方法直接在结构化数据空间中进行搜索. 在这里我们简要描述这些模型框架的原理, 以及它们在3D分子生成上的应用.

2.3.1 深度生成模型

深度生成模型通过学习可观测数据的概率密度来生成新的合成数据, 包括图像、视频和文本等. 这里主要关注药物发现中3D分子生成的深度生成模型. 深度生成模型的目标是最小化生成数据分布与真实数据分布之间的距离, 可以用极大似然法进行求解. 从处理极大似然函数方法的角度, 可将生成模型分为三类: 第一类方法通过变分或抽样的方法求似然函数的近似分布, 主要包括直接优化似然函数变分下界的变分自编码器和采用抽样方法近似计算似然函数的玻尔兹曼机; 第二类方法是避开求极大似然过程的隐式方法, 其代表模型是通过生成器和判别器之间的对抗行为来优化模型参数从而巧妙避开求解似然函数的生成对抗网络; 第三类方法是对似然函数进行适当变形以简化计算, 包括流模型和自回归模型, 流模型利用可逆函数构造似然函数后直接优化模型参数, 自回归模型将目标函数分解为条件概率乘积的形式. 下面将分别介绍这三类模型的代表性模型框架: 变分自编码器、生成对抗网络、流模型、自回归模型、扩散模型, 如图4所示.

(1) 变分自编码器

变分自编码器(Variational Autoencoder, VAE)[48]的基本结构包括一个编码器和一个解码器, 其中编码器学习将分子映射到连续低维空间的隐变量, 然后解码器将隐变量重建为新分子. 训练VAE时最小化的损失函数包括重构项和正则化项两部分, 重构项迫使解码器从隐变量中重建出正确的分子, 正则化项则使用Kullback–Leibler散度(KL散度)来衡量重建的样本分布与数据分布之间的距离. 在3D分子生成中, Mansimov等人[49]提出的条件变分图自编码器(CVGAE)首次以2D分子图为条件, 使用VAE学习原子表征并直接生成原子坐标, 但不能保证构象的旋转平移等变性, 需将生成后的构象与真实构象进行比对, 性能较差. 为了解

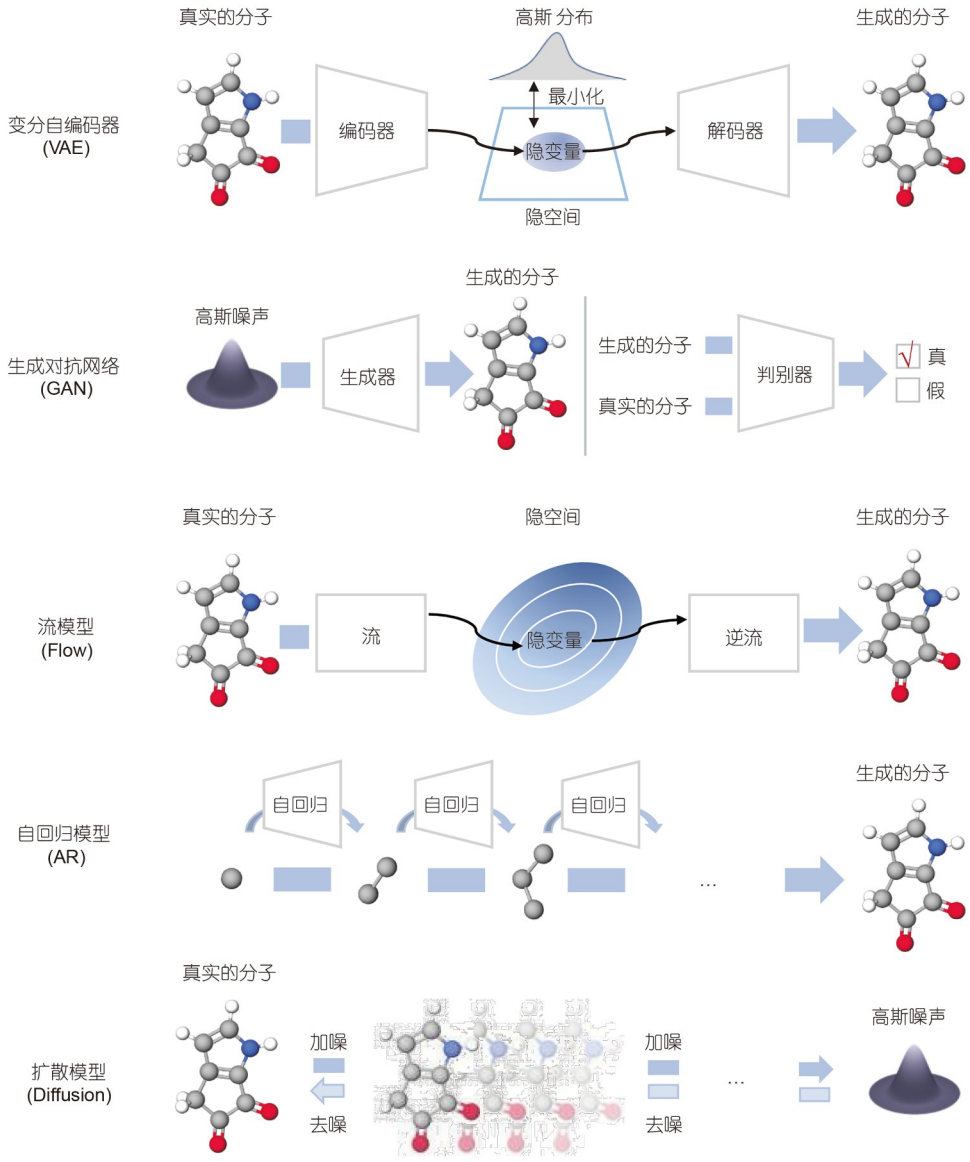


图 4　深度生成模型比较: 变分自编码器、生成对抗网络、流模型、自回归模型、扩散模型 (网络版彩图)
Figure 4　Comparison of the deep generative models: VAE, GAN, Flow, AR, and Diffusion models (color online).

决这个问题, Simm等人[50]提出了图距离几何模型(GraphDG)利用中间几何元素(键长)来生成构象. 具体来说, GraphDG使用VAE预测键长, 再通过距离几何算法利用先前计算得到的键长来求解构象. Xu等人[51]提出的ConfVAE模型使用条件VAE和距离几何算法将GraphDG中分离的两个步骤结合起来, 改进成端到端的双层优化框架.

(2) 生成对抗网络

生成对抗网络(Generative Adversarial Network, GAN)[52]是当前深度学习领域最热门的生成模型, 广泛应用于图像生成领域. GAN本质上是将难以求解的似然函数转化成神经网络, 让模型自己训练出合适的参数来拟合似然函数, 这个神经网络就是GAN的判别器. GAN由一个生成器和一个判别器组成, 生成器的目标是从高斯噪声中生成逼真的分子让判别器无法判别出真伪, 判别器的目标是正确区分分子是真实的(来自训练数据集)还是虚假的(由生成器生成). 这两个网络在训练过程中互相博弈, 不断优化自身的生成能力

和判别能力, 最终达到两者之间的纳什均衡, 即当判别器的判别能力达到一定程度却无法判断数据来源时, 就获得了一个学习到真实数据分布的生成器. 先前已有一些研究工作使用GAN来生成分子的SMILES字符串[53]或分子图[18,20]的邻接矩阵. 最近, 研究人员采用这种方法来生成3D分子构象, 如Hoffmann等人[54]提出的欧几里得距离矩阵网络(EDMNet)使用GAN来生成原子间距离矩阵, 进而生成分子3D结构. 然而, 使用GAN生成序列、分子图或3D构象仍然具有挑战性, 主要体现在以下几点: (1) 模型难以训练, 构建序列、图或3D坐标需要离散选择反向传播梯度, 容易出现梯度消失; (2) 出现模式崩溃, 表现为生成的分子多样性低; (3) 目标函数的形式导致模型在训练过程没有可以指示训练进度的指标.

(3) 流模型

主流深度生成模型中, VAE推导出了似然函数的变分下界来代替真实的数据分布, 这种近似模型并不能生成最精确的分子; GAN使用对抗和交替训练的方法避免了优化似然函数, 保留了模型的精确性, 但在训练过程中会出现训练困难、模式崩塌等问题, 因此研究一种既能保证模型精度又容易训练的深度生成模型具有重要意义. 标准化流(Normalizing Flow)[55]通过学习先验分布(如高斯分布)和真实世界的高维数据(如分子)之间一系列可逆变换来生成分子, 同时该可逆变换函数可计算得到. 流模型相对于VAE的主要优势在于其可逆映射能够计算精确的数据似然, 有望生成质量足够好的分子. 得益于该优势, 近年来已有一些研究将流模型应用于分子图的生成[21,31,56]. 如Shi等人[31]提出了基于流的自回归模型GraphAF来生成分子图, 其性能优于VAE. 最近, 研究人员采用流模型来生成3D分子构象, 如Satorras等人[57]提出的$E(n)$等变标准化流模型(E-NFs)将EGNN集成到标准化流模型, 能一次性同时生成所有原子的特征和3D坐标. Luo等人[58]提出了生成3D分子几何结构的自回归流模型G-SphereNet, 将分子生成建模成一个序列决策过程, 根据条件信息逐步生成原子的类型和相对位置信息(距离、角度和扭转角), 保证了模型的不变性和等变性.

(4) 自回归模型

自回归(Autoregressive, AR)模型可以对序列数据进行密度估计, 用同一变量在之前各个时刻的观测值预测该变量在当前时刻的观测值. AR将目标函数分解为条件概率乘积, 即首先用条件概率表示可见层数据相邻元素之间的关系, 然后以条件概率乘积来表示联合概率分布. 分子图数据可以转化成一系列节点和边的序列数据, 这些节点和边作为分子图的子组件存在潜在的依赖关系. AR将分子图数据的联合分布分解为一系列子组件的概率乘积, 可通过前面的子组件预测下一个子组件. AR最大的优势是可以对序列数据进行很好的密度估计, 在图像、文本和语音等领域的生成任务中得到成功应用, 包括循环神经网络(RNN)[59], PixelCNN[60], WaveNet[61], Transformer[62]和BERT[63]等. 近年来, AR也被广泛应用于3D分子生成领域, 如Gebauer等人[64]提出的G-SchNet模型将欧氏空间约束和原子分布的旋转不变性作为先验知识, 使用基于SchNet[65]的自回归神经网络模型依次生成新原子, 并将其放置在焦点原子的局部网格点. Joshi等人[66]提出的3D-Scaffold模型建立在G-SchNet基础上, 从现有的3D骨架中自回归地生成新原子, 生成的新候选分子保证具有所需的骨架. 这种方法需要分子骨架和目标性质之间关系的先验知识, 并且可能会因为对特殊分子的采样而影响模型性能. 该论文的作者在此基础上又提出了条件G-SchNet (cG-SchNet)[67], 能够以多种结构和化学特性的任意组合为条件, 有针对性地生成3D分子结构. 然而, 使用AR生成序列、图或3D构象仍然具有挑战性, 由于AR的每一步输出都需要按顺序等待前一步的输出, 这种按顺序生成的方式会阻碍AR利用并行运算, 导致训练和生成样本所需要的计算量远大于VAE和GAN等其他通用模型. 实际使用时可采用数据并行来利用并行计算资源从而加速生成.

(5) 扩散模型

扩散模型(Diffusion models)[68]的提出受到物理学中非平衡热力学的启发, 它定义了一个扩散步骤的马尔科夫链, 可分为正向扩散过程和逆向扩散过程. 扩散模型在正向扩散过程中逐步对数据添加随机噪声, 直至最后一步达到各向同性高斯分布; 逆向扩散过程则反转这一过程, 从噪声分布中构建所需的数据样本. 得益于生成真实样本的强大能力, 扩散模型在建模复杂的多模态分子构象分布上显示出优势, 能更好地生成高质量和多样性的结构. 如Shi等人[69]提出的ConfGF是基于扩散模型的分子构象生成早期工作, 在保证旋转和平移等变性的同时, 利用朗之万动力学方法对分子构象进行采样. 然而, ConfGF只建模了一

阶、二阶和三阶邻居之间的局部距离，忽略了分子内部的长程非键相互作用力. 由此, Luo等人[70]提出的动态图评分匹配(DGSM)模型通过动态地建模三维空间中原子对之间的相互作用力，包括化学键和非键相互作用，将ConfGF扩展到了长程作用力的建模. 然而, DGSM在扩散学习过程中输入的扰动距离矩阵可能违反三个原子之间距离的三角不等式. 为了解决这一挑战, Xu等人[71]提出的几何扩散模型(GeoDiff)引入了旋转–平移不变马尔可夫过程来对坐标分布施加约束. Hoogeboom等人[72]提出的$E(3)$等变扩散模型(EDM)进一步扩展了上述方法，通过一个等变网络学习去噪扩散过程，能同时对连续特征(原子坐标)和分类特征(原子类型)进行操作. 由于构象生成的难点主要在于扭转角的自由度，Jing等人[73]提出的扭转扩散模型(Torsional Diffusion)将超环面的扩散过程和一个从外向内的评分模型作用在扭转角空间来生成分子构象，显著提升了预测性能和速度.

2.3.2 组合优化方法

深度生成模型需要大量的标注数据进行训练，但药物发现场景难以满足这种数据要求，因为标注数据需要资源和时间密集型的生化实验. 在这种情况下，组合优化方法可以通过将分子生成问题转化成离散优化问题来帮助缓解这种数据稀疏的现状，但难点在于需要在巨大的化学空间中找到探索和利用的平衡. 在这里我们主要讨论5种组合优化方法: (1) 强化学习; (2) 贝叶斯优化; (3) 遗传算法; (4) 蒙特卡洛树搜索; (5) 最优传输.

(1) 强化学习

强化学习(Reinforcement Learning, RL)通常被描述为马尔可夫决策过程(Markov decision process, MDP), 用来建模智能体与环境的交互过程. MDP的公式MDP = (S,A,P,R)描述了RL中的几个基本概念: 状态集合S、动作集合A、状态转移概率P和奖励函数R. 状态集合S表示所求解问题中所有可能出现的状态所构成的有限或无限的集合; 动作集合A表示所求解问题中智能体能够采取的所有动作所构成的有限或无限的集合; 状态转移概率P$(S_{t+1}\mid S_t,A_t)$表示在状态S_t下采取了动作A后进入下一个状态S_{t+1}的概率，满足马尔可夫性; 奖励函数R(S_t,A_t,S_{t+1})表示在状态S_t执行了动作A后到达状态S_{t+1}时，智能体能够得到的奖励. RL的目标在于学习一种能使奖励最大化的最优策略. 在2D分子生成任务中，状态通常被定义为部分生成的分子图，动作可以是在某个位置添加一个原子、化学键或子结构，奖励可以是部分生成的分子的某种属性. 如You等人[29]提出的图卷积策略网络(GCPN)和Zhou等人[74]提出的分子深度Q网络(MolDQN)使用强化学习在分子图上迭代地添加或删除原子和化学键以最大化类药性(Quantitative Estimate of Drug-likeness, QED)分数，从而生成具有高QED的类药化合物. Olivecrona等人[75]在SMILES字符串上操作，使用RL来微调先前预训练的生成模型，以最大化生成分子与多巴胺受体的结合亲和力. 近期，利用强化学习进行3D分子结构从头设计也取得了很大进展. Simm等人[76]提出的MolGym模型使用强化学习在笛卡尔坐标下从头生成分子. MolGym利用用量子力学性质能量作为奖励函数，使用基于平移和旋转不变状态–动作表示的行动者–批评神经网络来保证模型的平移旋转等变性质. 该模型在MolGym RL环境的一系列分子设计任务中能生成有效的分子. 在此基础上，作者进一步提出了一种行动者–批评神经网络COVARIANT[77]，通过基于球面谐波级数展开的旋转协变状态–行动者表示保证了分子设计过程的对称性，结果表明利用分子设计过程的对称性显著提升了生成分子的质量和模型泛化能力.

(2) 贝叶斯优化

贝叶斯优化(Bayesian Optimization, BO)是一种调整超参的技术，常用于优化计算成本高或不可导的目标函数. BO通过代理优化的方式，根据新的信息来更新前信念以产生一个更新的后信念. BO有两个关键步骤: 1) BO为目标函数建立代理函数, 常用高斯过程表示的概率分布来量化代理函数的不确定性; 2) 采集函数在由代理函数确定的可能最小值附近做更多采样，并随之迭代更新代理函数，最终达到全局最小值. BO[8,22,78]通过学习分子的连续嵌入，将离散优化问题转化为连续优化问题. 这些方法包括首先训练一个VAE将离散的分子映射到连续的嵌入空间，然后训练另一个神经网络从其连续嵌入向量中预测原分子的化学属性. 然后，他们在连续嵌入空间中使用BO来找到一个关联属性得分最好的嵌入，该嵌入进而被解码器解码成一个离散的分子. 在3D分子构象采样中，一个核心挑战是使用较少的能量评价指标来找到低能稳定

构象，Chan等人[79]提出了一种叫做贝叶斯优化算法(BOA)的随机搜索方法来寻找分子的低能稳定构象. 与穷举搜索或随机搜索方法相比，BOA由于较好平衡了利用和探索之间的关系，能使用更少的能量评价指标来找到更低的能量构象. 该论文作者在此基础上又提出了基于先验知识的贝叶斯优化模型BOKEI[80]，使用相关扭转角知识作为约束条件，将低能构象搜索过程定义为一个约束BO问题, 与基准相比该方法能在减少搜索空间的同时找到更低能量构象.

(3) 遗传算法

遗传算法(Genetic Algorithm, GA)是借鉴生物的自然选择和遗传进化机制而开发出的一种全局优化自适应概率搜索算法. GA使用群体搜索技术，通过对当前群体施加选择、交叉、变异等一系列遗传操作，从而产生出新一代的群体，并逐步使群体进化到包含或接近最优解的状态. 由于其具有思想简单、易于实现、应用效果明显等优点而被众多应用领域所接受, 并在自适应控制、组合优化、模式识别、机器学习、人工生命、管理决策等领域得到广泛应用. 在分子生成任务中, GA通过分子突变来寻找具有理想属性的化合物分子. GA的核心由两部分组成: 一组变异规则和一个适应度函数. 适应度函数包括一个加权插值的属性预测分数和一个用于惩罚长期存活分子的惩罚分数, 其中惩罚分数鼓励模型探索多样的分子. 对现有分子应用突变规则得到新分子, 在每次迭代中, 适应度分数较低的旧分子会被移除. 如Nigam等人[81]通过神经网络参数化适应度函数，使用遗传算法设计出高油水分配系数(log*P*)的分子. 不同于Nigam等人使用一组固定的突变规则，Ahn等人[82]提出的遗传专家指导学习(GEGL)框架使用策略函数执行分子突变来生成高回报的分子, 进一步拓展了GA在分子生成上的应用. GA也被广泛应用于3D分子生成领域，如Spiegel等人[83]提出的半自动化计算机辅助药物设计开源程序AutoGrow4利用遗传算法和反应库来对种群中的分子进行突变，迭代应用虚拟反应来最大化期望适应度函数来生成或优化药物分子. Fu等人[84]提出的增强遗传算法(RGA)将遗传算法中的进化过程建模成MDP，用神经网络代替随机游走方法来指导遗传算法中的交叉和变异操作, 以更智能地探索化学空间. 此外, 该网络使用多种配体–靶标复合物进行预训练来学习不同靶标共享的结合亲和力物理特性，然后在优化过程中用强化学习进行微调.

(4) 蒙特卡洛树搜索

蒙特卡洛树搜索(Monte Carlo Tree Search, MCTS)方法在分子优化方面显示出优越性. 不同于GA, MCTS可以从头生成分子. MCTS包括四个步骤: 选择、扩展、模拟和反向传播. 具体来说, MCTS通过迭代以上四个步骤来构建搜索树，树中每个节点代表分子生成过程的一个中间状态. 每次迭代中，模型首先从搜索树中选择一个有希望的状态，枚举该状态的可能操作，并模拟生成原子、化学键或子结构来得到下一个中间状态，通过收集奖励信息并反向传播. 如Li等人[85]提出的DeepLigBuilder模型能直接在靶蛋白的3D结合口袋内构建和优化配体的3D分子结构. 该模型使用配体神经网络(L-Net)生成有效的3D类药分子，用MCTS搜索高结合亲和力的分子，可以直接输出类药物分子的3D结构和拓扑结构，而不需要额外的原子位置或键序推断.

(5) 最优传输

最优传输(Optimal Transport, OT)是一种比较概率分布的几何工具，即利用几何方法来建模概率分布和衡量概率分布之间的距离，是连接几何和概率的桥梁. OT的核心是研究如何用最经济的方式将一个概率分布变换到另外一个概率分布. 由于深度神经网络本质上是在训练概率分布间的传输映射，常需比较两个概率分布间的距离，OT为此提供了解决思路. 近年来, OT被广泛应用于深度学习领域，如Arjovsky等人[86]提出的WGAN利用生成器计算OT映射, 将隐空间的均匀分布映射成数据分布，判别器计算数据分布和生成分布间的Wasserstein距离. 最近OT被应用于3D分子生成，如Ganea等人[87]提出的分子3D构象扭转几何生成模型GeoMol是一种端到端、非自回归和*SE*(3)不变的机器学习方法，用于生成低能构象异构体分布. GeoMol利用MPNN捕获局部和全局图信息，预测局部原子3D结构和扭转角来组装相邻的局部3D结构对, 避免由于几何自由度导致不必要的过度参数化. 模型在局部预测中使用基于非对抗性OT的损失函数来促进多样化构象异构体的生成，最后测试时将所有预测的相邻局部3D结构对组合成完整构象. Stärk等人[88]提出的*SE*(3)等变几何深度学习模型EquiBind利用*E*(3)协变网络生成配体和受体上的结合关键点，而后在三维空间中使用OT损失对这两组点进行匹配从而计算其差

异. 利用上述得到的关键点, 模型将输出的小分子构象叠合至蛋白口袋中来确定结合模式. 同时, 该网络还能够对配体的三维结构进行变换实现小分子柔性对接.

3 基于隐变量的生成

基于隐变量的生成是指在没有给出参考分子的情况下在3D空间中生成分子, 即从零开始生成3D分子构象, 生成的构象需同时满足化学有效与几何有效, 以及多种理想属性, 包括生物活性、靶标选择性、结构新颖性、化学可合成性、成药性和安全性等. 根据3D分子生成算法的输出, 可将基于隐变量的生成任务分为定向生成和非定向生成. 表1从各方面对这些方法进行了比较.

3.1 非定向生成

该方法的目标是在没有给出参考分子的情况下从给定数据集中学习3D分子构象分布, 并从中采样出新颖、多样、化学有效以及几何有效的3D分子构象.

Gebauer等人[64]提出的G-SchNet模型将欧氏空间约束和原子分布的旋转不变性作为先验知识, 使用基于SchNet[65]的自回归神经网络模型依次生成新原子, 并将其放置在焦点原子的局部网格点. 作者将生成模型的分布偏向于具有小HOMO–LUMO间隙的分子, 表明模型能捕捉到3D几何结构与电子特性之间的复杂关系. Hoffmann等人[54]提出的欧几里得距离矩阵网络(EDMNet)和Nesterov等人[89]提出的3D分子结构生成模型3DMolNet分别使用GAN和VAE来生成原子间距离, 进而生成分子3D结构. Satorras等人[57]提出的$E(n)$等变标准化流模型(E-NFs)将EGNN集成到Flow模型来保证生成模型的对称性. 具体来说, 通过将EGNN集成到连续时间可微的神经常微分方程获得Flow的可逆等变函数. 该模型能一次性同时生成所有原子的特征和3D坐标, 生成的分子结构在准确性和稳定性上大大优于现有的Flow模型. Luo等人[58]提出的生成3D分子几何结构的自回归流模型G-SphereNet, 将分子生成建模成一个序列决策过程, 根据条件信息逐步生成原子的类型和相对位置信息, 保证了模型的不变性和等变性. 此外, 在序列生成过程中, 使用球形消息传递网络SphereNet和注意机制来提取条件信息. G-SphereNet在随机分子几何结构生成和以量子性质为目标的分子发现任务方面优于E-NFs和G-SchNet. 近期, 利用RL进行3D分子结构从头设计也取得了很大进展. Simm等人[76]提出的MolGym模型使用RL在笛卡尔坐标下从头生成分子. MolGym将量子力学性质能量作为奖励函数, 使用基于平移和旋转不变状态–动作表示的行动者–批评神经网络来保证模型的平移旋转等变性质. 该模型在MolGym RL环境的一系列分子设计任务中能生成有效的分子. 在此基础上, 作者进一步提出了一种行动者–批评神经网络COVARIANT[77], 通过基于球面谐波级数展开的旋转协变状态–行动者表示保证了分子设计过程的对称性, 结果表明利用分子设计过程的对称性显著提升了生成分子的质量和模型泛化能力.

3.2 定向生成

该方法的目标是在没有给出参考分子的情况下从头生成出同时满足化学有效和几何有效, 以及具有多种理想药物属性或能与特定蛋白口袋结合的3D分子构象.

Masuda等人[90]提出的一种基于条件变分自编码器(CVAE)与GAN的分子生成模型LiGAN, 能够在蛋白口袋中生成分子, 主要的策略是先生成分子的原子密度图, 再将其转换成分子结构. 该实验室的研究者们在此基础上又提出了一步生成基于受体结合位点的3D分子结构[91]. 具体来说, 首先使用CVAE来学习蛋白质–配体复合物的原子密度网格表示, 然后用原子拟合和键推断的方法从生成的原子密度构建出有效的分子构象. Li等人[85]提出的DeepLigBuilder模型使用深度生成模型直接在靶蛋白的3D结合口袋内构建和优化配体的3D分子结构. 该模型提出了一个配体神经网络(L-Net)来生成有效的3D类药物分子, 用MCTS搜索强结合分子, 可以直接输出类药分子的3D和拓扑结构, 而不需要额外的原子位置或键序推断. Luo等人[92]提出的DSBDD模型将给定的蛋白结合位点作为3D上下文, 估计原子在3D空间中出现的概率密度, 即更有可能出现原子的位置将被赋予更高的概率, 然后使用自回归采样方案从学习的分布中顺序地采样原子, 直到没有空间容纳新的原子. 结合该采样方案, 该模型可以生成有效且多样的分子. 实验结果表明, 即使模型没有明确地被优化, 从该模型中采样的分子也表现出与特定靶点的高亲和力和良好类药性.

4 基于2D分子图的生成

不同于上述基于隐变量的方法, 基于2D分子图的生成是指从给定的2D分子图生成或优化出同时满足化学有效和几何有效, 以及具有多种理想的生物活性和药物属性的3D分子构象. 类似于基于隐变量的生成方法的子分类, 根据3D分子生成算法的输出, 上述每一大类生成方法可二次划分为定向生成和非定向生成. 表1按照不同的算法和生成方法进行了组织和分类.

4.1 非定向生成

该方法的目标是从给定的2D分子图生成新颖、多样、化学有效以及几何有效的3D分子构象. 从2D分子图生成3D分子构象对许多下游任务如受体–配体结合亲和力预测、基于受体结构的药物设计等具有重要意义.

Mansimov等人[49]提出的条件变分图自编码器(CVGAE)首次以2D分子图为条件, 使用VAE模型学习原子表征并直接生成3D原子坐标, 但不能保持构象的旋转平移等变性, 需将生成后的构象与真实构象进行比对, 性能较差. 为了解决这个问题, 大多数后续的模型都是基于中间几何元素(如原子间距离和扭转角)来生成构象. 这些元素在分子平移、旋转时能保持不变, 这种平移旋转不变性已被证明是分子几何建模中的一个重要的归纳偏置. 如Simm等人[50]提出的图距离几何模型(GraphDG)和Xu等人[93]提出的条件图连续流模型(CGCF)分别使用VAE和Flow模型预测原子间距离, 再通过距离几何算法利用原子间距离来求解构象. Xu等人[51]提出的条件变分自编码器框架ConfVAE使用条件VAE和距离几何算法将GraphDG、CGCF中分离的两个步骤结合起来, 改进成端到端的双层优化框架. 但这些方法都存在累积误差问题, 即预测原子间距离过程的噪声会传导到后续的坐标搜索过程, 影响生成的3D结构的准确性. 为了克服上述缺陷, Shi等人[69]提出的ConfGF模型受到经典计算化学中力场的启发, 使用去噪得分匹配算法来学习原子间距离的对数密度梯度场, 并通过链式法则计算出原子坐标, 然后利用朗之万动力学方法对分子构象进行采样. 然而, ConfGF忽略了分子内部的长程非键相互作用力, 比如分子图中相距较远的两个原子可能在三维结构中形成氢键. Luo等人[70]提出的动态图评分匹配(DGSM)模型通过动态地建模三维空间中原子对之间的相互作用力, 包括化学键和非键相互作用, 将ConfGF扩展到了长程作用力的建模. 这种基于中间几何元素方法的优势是能保证旋转和平移的不变性. 然而, 预测的原子间距离可能会相互冲突, 违反三个原子之间距离的三角不等式, 例如有研究发现GraphDG在GEOM–Drugs测试集上对8.65%的分子产生了违反三角不等式的距离矩阵[94]. 此外, 最新最前沿的方法是最近提出的GeoMol[87], GeoDiff[71]和Torsional Diffusion[73]. GeoMol首先学习预测几何元素的最小集合(即键长和键角)得到局部3D结构, 然后通过预测扭转角来组装相邻的局部3D结构对, 最后在测试时将所有预测的相邻局部3D结构对组合成完整构象. 为了生成不同的构象, 随机高斯噪声被附加到每个初始节点, 并使用基于OT的损失函数对模型进行训练. GeoDiff模拟扩散过程将分子坐标扰乱成噪声分布, 再用去噪扩散模型模拟逆向扩散过程来恢复真实构象. 作者将去噪扩散模型与几何表示相结合, 将逆向扩散过程建模成马尔可夫链, 并将平移旋转等变性引入马尔可夫核, 以保证模型的平移旋转等变性质. 由于构象生成的难点主要在于扭转角的自由度, Jing等人[73]提出的扭转扩散模型(Torsional Diffusion)将扩散过程只作用于扭转角, 而键长和键角则通过标准的化学信息学方法进行计算. Torsional Diffusion成为首个在GEOM–Drugs数据集上效果超越商业软件OMEGA[95]的深度生成模型, 并且在训练速度上比GeoDiff快了两个数量级.

4.2 定向生成

该方法的目标是从给定的2D分子图生成或优化出同时满足化学有效和几何有效, 以及具有多种理想药物属性或能与特定蛋白口袋结合的3D分子构象.

Méndez-Lucio等人[96]提出的DeepDock模型利用几何深度学习方法来预测配体和靶蛋白结合构象. 该模型通过学习每个配体–靶体对基于距离似然的统计力学势, 使用能量模型(EBM)作为全局优化算法对其进行优化, 从而构建出配体的实验结合构象. 同已有的用于对接和筛选任务中的评分函数相比, 此处基于距离似然的统计势有类似或更好的效果, 这种方法代表了人工智能用于改进基于结构的药物设计的一个案例. Stärk等人[88]提出的$SE(3)$等变几何深度学习模型EquiBind利用$E(3)$协变网络生成配体和受体上的结合

关键点，而后在三维空间中对这些点进行匹配来确定结合模式. 同时，该网络还能够对配体的三维结构进行变换实现小分子柔性对接.

5 基于3D分子构象的生成

基于3D分子构象的生成是指从给定的3D分子生成或优化出同时满足化学有效和几何有效，以及具有多种理想的生物活性和药物属性的3D分子构象. 根据生成算法的输出，二次分类为定向生成和非定向生成. 表1按照不同的算法和生成方法进行了分类比较.

5.1 非定向生成

该方法的目标是从给定的3D分子生成或优化出新颖、多样、化学有效以及几何有效的3D分子构象.

针对两个分子片段的连接问题，Huang等人[97]提出了一种基于条件变分自编码器的生成模型3DLinker, 利用EGNN来结合分子的不变特征与等变特征, 将分子图的生成与分子三维信息的表征进行整合，从而有效地预测连接器的锚点与边的具体连接状态. 该模型在分子片段连接器的恢复上表现出良好的性能优势, 并能够准确预测分子的3D坐标信息.

5.2 定向生成

该方法的目标是从给定的3D分子生成或优化出具有理想生物活性或药物属性的3D分子构象，如低能分子构象优化、基于骨架的分子构象优化、由粗到细的分子构象优化等.

Chan等人[79]提出了一种叫做贝叶斯优化算法(BOA)的随机搜索方法来搜索分子的低能稳定构象. Wang等人[98]提出的粗粒化变分自编码器(Coarse-Graining VAE)使用VAE模型学习从粗粒度分子构象生成细粒度分子构象. Joshi等人[66]提出的3D-Scaffold模型建立在G-SchNet基础上，从现有的3D骨架中自回归地生成新原子，生成的新候选分子保证具有所需的骨架. 这种方法需要分子骨架和目标性质之间关系的先验知识，并且可能会因为对特殊分子的采样而影响模型性能. 该论文的作者在此基础上又提出了条件G-SchNet (cG-SchNet)[67]，能够以多种结构和化学特性的任意组合为条件，有针对性地生成3D分子结构. 与以前的方法相比，该模型不需要在特定目标属性的数据集上微调, 显式条件使cG-SchNet能够从所有可用的参考数据集中有效地学习. Fu等人[84]提出的增强遗传算法(RGA)将遗传算法中的进化过程建模成MDP，用神经网络代替随机游走方法来指导遗传算法中的交叉和变异操作, 以更智能地探索化学空间. 此外，该网络使用多种配体–靶标复合物进行预训练来学习不同靶标共享的结合亲和力物理特性，然后在优化过程中用强化学习进行微调. Gogineni等人[99]提出了一种基于RL的顺序构象搜索技术TorsionNet，将3D构象生成问题建模成MDP，使用基于热力学的吉布斯分数作为奖励, 使用近端策略优化(PPO)算法和课程学习策略(curriculum learning strategy)来训练模型. TorsionNet是第一个用于大分子构象搜索的深度学习模型，其在大支链烷烃和木质素聚合物数据集上的性能明显优于经典的化学信息学和分子动力学采样方法.

6 3D分子生成算法评价

6.1 常用数据集

近几十年来，3D分子数据库得到不断积累. 这里我们主要关注基于深度学习的3D分子设计任务相关的公开数据集, 并将其分为以下两类: 第一类化学空间数据集是指以类似于列举化学空间的方式包含特定原子组成的化合物；第二类任务数据集用于药物设计中的特定任务，如分子动力学模拟，能量、力等量子属性预测和亲和力预测. 第3、4、5节中讨论的大量工作已经证明了这些数据集的有效性，在准确预测或生成目标的同时大大降低了传统方法密度函数理论(DFT)所需的计算量. 下面列出这些数据集, 并在表2展示其统计信息.

6.1.1 化学空间数据集

(1) GEOM[100]: 一个用于3D分子生成的基准数据集, 包含了超过430000个分子的数百万个构象, 由第一性原理模拟得到. GEOM被拆分成了GEOM-QM9和GEOM-Drugs这两个子集. GEOM-QM9是QM9数据集的扩展，它仅限于含有9个重原子(总共29个原子)的小分子，分子质量小，可旋转键少. GEOM-Drugs数据集由中等大小的药物分子组成，最多181个原子(91个重原子)，平均包含44.2个原子，每个分子平均有6.5个可旋转键.

表 2 3D分子常用数据集及其统计信息

Table 2 Available datasets and statistics for 3D molecular generation

数据集		任务	分子数[a)]	对称性	链接
化学空间数据集	GEOM-QM9[100]	3D分子生成	133232	等变性	https://doi.org/10.7910/DVN/JNGTDF
	GEOM-Drugs[100]	3D分子生成	304340	等变性	https://doi.org/10.7910/DVN/JNGTDF
	Molecule3D[101]	3D分子生成	3899647	等变性	https://github.com/divelab/MoleculeX
	QM9[102]	量子属性预测	133885	不变性	https://moleculenet.org/datasets-1
任务数据集	ISO17[2]	分子动力学模拟	645000	不变性	http://www.quantum-machine.org
	MD17[103]	分子动力学模拟	3611115	不变性	http://www.quantum-machine.org
	Atom3D[104]	属性预测	–	不变性	https://www.atom3d.ai
	PDBbind[105]	亲和力预测	15193 (3890)	不变性	http://www.pdbbind.org.cn
	CrossDock2020[38]	亲和力预测	22584102 (2922)	不变性	https://github.com/gnina/models
	scPDB[106]	亲和力预测	16034 (4782)	不变性	http://bioinfo-pharma.u-strasbg.fr/scPDB
	DUD-E[107]	亲和力预测	1114300 (102)	不变性	http://dude.docking.org

a) 括号里的数字表示靶蛋白数量. Atom3D是一组包含小分子、蛋白质和RNA的基准数据集, 这里不统计其分子数.

(2) Molecule3D[101]: 一个用于3D分子生成的基准数据集, 包含大约400万个分子的精确基态几何构象, 由DFT计算得到. 该基准数据集提供了一套软件工具用于数据处理、拆分、训练和评估等操作.

(3) QM9[102]: 一个用于量子属性预测的基准数据集, 来自GDB17数据集, 包含133885个有机小分子, 由不超过9个重原子(C, O, N和F)组成. 每个分子都具有几何、能量、电子和热力学等12种量子属性, 由DFT计算得到.

6.1.2 任务数据集

(1) ISO17[2]: 一个分子动力学轨迹数据集, 包含了从QM9中提取的原子组成为$C_7O_2H_{10}$的129个同分异构体的分子动力学轨迹, 每条轨迹包含5000个构象, 数据集由64.5万个带标记的实例组成. 其中的分子反映了平衡态周围的构象分布, 可用于评估模型分布与平衡态周围数据分布之间的相似性.

(2) MD17[103]: 一个分子动力学轨迹数据集, 包含8个小分子的分子动力学轨迹及其对应状态的能量和相互作用力信息. 每个数据都由单个分子的轨迹组成, 覆盖了各种构象.

(3) Atom3D[104]: 一个涵盖小分子、蛋白质和核酸三维结构的综合数据集, 包含了8个预测任务涉及分子结构、分子之间相互作用、分子功能和设计等.

(4) PDBbind[105]: 是蛋白质数据库(PDB)的一个子集. PDBBind v2020包含19443个蛋白–配体复合物结构及其各自的势能(半峰抑制浓度、解离常数等), 其中包含3890个独特受体和15193个独特配体.

(5) CrossDock2020[38]: 一个利用Pocketome数据库和smina对接程序对PDBbind进行扩增得到的数据集, 包含总计22500000个蛋白-配体的结合构象.

(6) scPDB[106]: 一个蛋白–配体结合位点数据库, 共16034个条目, 包含由3678个独特蛋白质和5608个独特配体组成的9283个结合位点. 该数据库提供了蛋白质、配体、结合位点和结合模式的全原子描述.

(7) DUD-E[107]: 一个广泛用于分子对接和虚拟筛选的数据集, 包含22886种活性化合物及其对102个靶点的结合亲和力分数.

6.2 3D分子生成评价指标

3D分子生成的评价指标包括均方根偏差, 覆盖分数和匹配分数, 下面分别介绍.

(1) 均方根偏差(Root Mean Squared Deviations, RMSD): 衡量两个3D分子构象的相似性. 这里使用了RMSD的一种类型Kabsch-RMSD, 利用Kabsch算法[108]旋转和平移两个构象使其距离最小, 即将两个3D坐标矩阵叠加对齐后来计算RMSD. 公式为:

$$\mathrm{RMSD}(R,\hat{R})=\left(\frac{1}{n}\sum_{i=1}^{n}\left\|R_i-\hat{R}_i\right\|^2\right)^{\frac{1}{2}} \tag{4}$$

RMSD越小, 两个构象差异越小; RMSD为零时两个构象相同.

(2)覆盖分数(Coverage score, COV)和匹配分数(Matching score, MAT)[93]: COV用来衡量生成构象与稳定态构象的相似性, 表示生成构象的准确性. MAT用来衡量模型生成多样性构象的能力, 这对于发现新的构象至关重要. 上述Kabsch-RMSD在计算中考虑了对称性, 因此这两个基于构象之间Kabsch-RMSD的指标COV和MAT也满足对称性. 公式为:

$$\mathrm{COV}\left(S_g, S_r\right)=\frac{1}{|S_r|}\left|\left\{R \in S_r \mid \mathrm{RMSD}(R, \hat{R})<\delta, \hat{R} \in S_g\right\}\right| \tag{5}$$

$$\mathrm{MAT}\left(S_g, S_r\right)=\frac{1}{|S_r|} \sum_{R \in S_r} \min_{\hat{R} \in S_g} \mathrm{RMSD}(R, \hat{R}) \tag{6}$$

其中S_g表示生成构象, S_r表示参考的稳定态构象, δ是给定的RMSD阈值. 一个好的生成方法具有高COV和低MAT.

6.3 3D分子生成效果对比

由于第3节讨论的生成模型除了3.2.1以2D分子图为条件的非定向分子生成方法采用了GEOM数据集并使用了相同的实验设置, 其他模型大多没有使用统一的基准难以进行比较, 因此这里仅对3.2.1中以2D分子图为条件的非定向分子生成任务的实验结果进行总结. 实验设计沿用了Shi等人[69]的设置, 使用GEOM-QM9和GEOM-Drug基准数据集来进行3D分子构象生成, 两个数据集的训练集、验证集和测试集分别为200000、2500和22408 (GEOM-QM9)/14324 (GEOM-Drug)分子图–构象对. 采用COV和MAT对生成的构象进行质量评估, GEOM-QM9和GEOM-Drug的RMSD阈值δ分别被设定为0.5和1.25.

实验结果均来自相关论文, 如表3所示. RDKit[109]是一款广泛应用的传统开源软件, 基于距离几何来生成构象. 其余基于深度学习的方法可以分为三类, 第一类是基于原子间距离的方法, 首先生成原子间距离或关于距离的梯度场, 然后根据距离或梯度生成原子坐标. 其中基于距离的方法有: GraphDG[50]和CGCF[93]分别通过VAE和Flow预测原子间距离矩阵, 然后通过距离几何技术预测原子坐标; ConfVAE[51]通过二层优化将上述流程改进成一个端到端的框架; EGNN[34]是一个等变模型, 将非线性变换作用在节点距离上, 并利用邻居节点坐标为标架来表示向量信息. 基于梯度场的方法有: ConfGF[69]通过学习原子间距离矩阵的梯度场来生成构象并保证等变性, DGSM[70]在ConfGF的基础上提出了一种动态图构建策略来建模长程相互作用, ClofNet[110]采用了与ConfGF相同的基于分数的框架, 通过一种具有完备局部标架的$SE(3)$等变图网络来估计梯度场. 第二类是直接生成分子中所有原子的坐标: CVGAE[49]是生成原子坐标的早期尝试, 不能保持构象的旋转平移等变性; GeoMol[87]首先预测分子中原子的局部3D结构, 然后在一个复杂的过程中组装全局结构; GeoDiff[71]模拟扩散过程将分子坐标扰乱成噪声分布, 再用去噪扩散模型模拟逆向扩散过程来恢复真实构象; DMCG[94]通过设计专门的损失函数来保证生成构象的旋转平移和置换不变性, 其模型包括多个堆叠块, 每个块细化其前一个块生成的构象. 第三类是预训练模型: Zhu等人[111]提出了一种新的联合2D和3D表示的预训练方法, 设计了3个预训练任务: 掩码原子和坐标的重建、基于二维图生成三维构象、基于三维构象生成二维图. Uni-Mol[112]是一个大规模3D分子表征学习框架, 由具有$SE(3)$等变transformer框架的分子预训练模型和口袋预训练模型组成, 可用于多种下游任务, 如分子性质预测、构象生成等.

从表3可以看出, 第三类的两种预训练模型在生成构象的多样性指标COV和准确性指标MAT上基本全面超越了现有的基准模型, 表明了预训练模型在构象生成任务上的有效性. 第一类基于原子间距离的方法中, 总体上基于梯度场的模型在生成效果上优于基于距离的模型, 在GEOM-Drug数据集上优势更大, 说明基于梯度场的模型能更好地建模多模态分布并生成准确而多样的构象, 在GEOM-Drug数据集上优势说明其能较好地模拟包含更多原子和官能团的复杂分布的梯度场. 第二类直接生成原子坐标的方法中, GeoDiff和DMCG在构象多样性和准确性上的表现接近或超越了预训练模型, 展现出直接生成方法的巨大潜力. GeoDiff和DMCG也超越了第一类基于原子间距离的方法中表现最优的DGSM和ClofNet, 可能是因为第一类方法中预测的原子间距离矩阵存在违反三角不等式约束的风险, 而直接预测原子坐标能避免各种距离违反情况因而在结构预测上表现出色.

表 3 3D分子生成代表算法效果表

Table 3 Performance table of representative 3D molecular generation algorithm

数据集		GEOM-QM9				GEOM-Drugs			
算法		COV (%)↑		MAT (Å)↓		COV(%)↑		MAT (Å)↓	
		Mean	Median	Mean	Median	Mean	Median	Mean	Median
开源软件	RDKit[109]	83.26	90.78	0.3447	0.2935	60.91	65.70	1.2026	1.1252
基于原子间距离	GraphDG[50]	73.33	84.21	0.4245	0.3973	8.27	0.00	1.9722	1.9845
	CGCF[93]	78.05	82.48	0.4219	0.3900	53.96	57.06	1.2487	1.2247
	ConfVAE[51]	80.42	85.31	0.4066	0.3891	53.14	53.98	1.2392	1.2447
	EGNN[34]	80.93	86.27	0.3832	0.3898	40.71	33.01	1.3574	1.3346
	ConfGF[69]	88.49	94.13	0.2673	0.2685	62.15	70.93	1.1629	1.1596
	DGSM[70]	91.49	95.92	0.2139	0.2137	78.73	94.39	1.0154	0.9980
	ClofNet[110]	90.21	93.14	0.2430	0.2457	88.64	97.56	0.9040	0.9023
直接生成原子坐标	CVGAE[49]	0.09	0.00	1.6713	1.6088	0.00	0.00	3.0702	2.9937
	GeoMol[87]	71.26	72.00	0.3731	0.3731	67.16	71.71	1.0875	1.0586
	GeoDiff[71]	91.68	95.82	0.2099	0.2026	89.13	97.88	0.8629	0.8529
	DMCG[94]	96.34	99.53	0.2065	0.2003	96.69	**100.00**	0.7223	0.7236
预训练模型	Zhu等人[111]	96.93	**100.00**	0.1958	0.1849	**97.05**	**100.00**	0.7056	0.6973
	Uni-Mol[112]	**98.68**	**100.00**	**0.1806**	**0.1510**	92.69	**100.00**	**0.6596**	**0.6215**

7 总结与展望

基于深度学习的3D分子生成旨在利用深度学习技术设计出安全有效且新颖的3D分子构象, 在新药研发的实际应用中发挥着越来越重要的作用. 本文对3D分子结构设计进行了系统的梳理分析: 首先, 根据3D分子生成算法的输入将3D分子结构设计方法分为基于隐变量的生成、基于2D分子图的生成和基于3D分子构象的生成三大类, 接着根据算法输出的不同将其分类为定向生成和非定向生成; 总结了相关生成模型在主要公开数据集上的算法性能. 希望通过以上介绍, 读者能够对3D分子生成有一个较为全面的了解, 并在相关基于深度模型的实际任务中加以利用.

虽然基于深度学习的3D分子结构设计发展迅速并取得了很大的成功, 但由于分子结构的复杂性, 仍存在一些挑战有待解决. 我们列出了几个可能的未来发展方向, 以供参考.

(1) 更好地建模3D分子表征

在过去几年里, 生成模型取得了许多进展, 能从1D 字符串、2D分子图和3D分子几何表示中学习. 生成模型从3D表示中能学习到更丰富的信息, 其中包含更多关于分子潜在的物理信息, 但这些框架仍存在技术上的挑战. 目前建模3D分子表征主要通过EGNN或者将几何信息编码进MPNN这两种方法. 然而, 不同的几何信息对分子表征学习的影响还缺乏严格论证, 尚未建立起标准的空间信息学习方法, 这将是3D分子表示学习领域中一个很有潜力的研究方向. 此外, 大多数3D分子生成模型以自回归的方式生成3D构象, 而分子结构实际上是无序的, 这种在原子之间强加顺序的规则会引入额外偏差, 进而影响模型性能. 最后, 化合物在结构上呈现出层次结构, 如小分子由多个官能团组成, 蛋白质由氨基酸组成. 利用这些层次结构, EGNN可以尝试在多个粒度上对化合物进行建模, 以提高模型的效率和通用性.

(2) 药物设计大模型

近年来, 利用GNN进行分子表示学习取得了显著进展. 为了充分利用未标记的分子数据, 研究人员首先在大规模的分子数据库上预训练GNN以获得一个可以处理各种分子相关任务的通用编码器, 然后在下游任务中微调这些预训练GNN. 隐式编码在模型参数中的知识可以赋能多种下游任务, 有助于缓解药物设计中标记数据不足的问题. 在2D分子图上, GROVER[113]

和MPG[114]等具有巨量参数的预训练GNN在学习通用分子图表示上表现出强大的表达能力. 目前, 已有一些研究将预训练模型用于3D分子表征学习, 如Zhu等人[115]提出的基于多视角对比学习的分子预训练框架MEMO, 能够自适应地集成来自2D拓扑图、3D几何、SMILES字符串和分子指纹这四种不同分子特征空间的信息, 在下游分子性质预测基准任务上性能优于SOTA, 并对学习到的模型权重提供了符合先验化学知识的解释. 近期深势科技提出的大规模3D分子表征学习框架Uni-Mol[112]包含两个相同架构的模型, 一个是由2.09亿个3D分子构象数据训练的分子预训练模型, 另一个是由320万个蛋白口袋数据训练的口袋预训练模型. Uni-Mol在14/15分子性质预测任务中表现优于SOTA, 在蛋白质–配体结合位点预测、分子构象生成等3D相关下游任务中表现尤其出色. 此外, 开发生物多模态预训练模型用于生物医药领域中的靶点发现、药物开发等环节, 将成为AI制药领域的基础设施.

(3) 干湿结合

干湿结合是指由AI计算模拟的干实验, 和实际中进行生化操作的湿实验相结合的研发方式. 首先, 由计算机在化学空间中对化合物进行模拟生成、预测和筛选. 随后, 由干实验筛选得到的化合物将提交到实验室中, 由研究人员进行合成、测试、检验等实际操作来进行优化筛选. 同时, 湿实验数据将被反馈给AI, 用于优化迭代干实验模型. 基于这种循环, 只需少量针对性的湿实验数据, AI就能够不断地生成、迭代, 直至得到具有理想活性的先导化合物. 如瑞典的阿斯利康AstraZeneca iLab对药物发现的设计–制造–测试–分析(DMTA)循环自动化进行了探索和尝试. DMTA包括设计新分子、化学合成、实验测试和分析改进四个环节, 然后再开始下一轮设计, 其中AI主要参与了DMTA循环中的设计和分析环节, 湿实验则参与了制造和测试环节. 他们提出的条件递归神经网络使化学家能够与计算机交互工作, 以加快化学空间的探索和潜在新药分子的设计[9]. 除了提升效率和节约成本, 干湿结合也是新药研发领域解决数据不足问题的一种方式.

(4) 更多真实世界的应用和数据集

虽然各种深度生成模型在理论和计算上发展迅速, 但在实际药物研发中使用AI生成模型的例子鲜有耳闻. 这是因为实际场景中仍存在一系列困难有待攻克, 比如实际分子设计中的多目标优化问题, 设计出的分子必须是可合成的, 用于筛选或估计分子属性的预测模型本身容易出错等. 另外, 目前使用的公开数据集主要关注低能的真空、水相模拟构象, 缺少实际中和靶标蛋白结合的药效构象数据集, 针对该问题除了上文提到的预训练大模型和干湿结合方法可以显著减少样本量的需求, 在算法方面还可以使用迁移学习、多任务学习、小样本学习、数据增强等方法来解决该领域数据不足的问题. 此外, 最近涌现了许多关于蛋白质折叠结构的工作, 如AlphaFold2[116]和RoseTTAFold[117]在蛋白质结构预测上取得了突破性进展, 推进了几何等变性在蛋白质等大分子结构领域的研究. 蛋白质主要是线性骨架结构, 小分子结构高度分支且常包含多种环, 这使得蛋白质折叠结构预测方法与本文讨论的小分子结构预测方法设置上存在差异. 本文讨论的方法虽然只在有限规模和复杂性的小分子数据集上得到了评估, 也可能有利于其他具有更复杂数据结构、更多样化约束和更复杂交互的应用.

参考文献

1 Gaudelet T, Day B, Jamasb AR, Soman J, Regep C, Liu G, Hayter JBR, Vickers R, Roberts C, Tang J, Roblin D, Blundell TL, Bronstein MM, Taylor-King JP. *Briefings BioInf*, 2021, 22: 1–22

2 Gainza P, Sverrisson F, Monti F, Rodolà E, Boscaini D, Bronstein MM, Correia BE. *Nat Methods*, 2020, 17: 184–192

3 Bronstein MM, Bruna J, LeCun Y, Szlam A, Vandergheynst P. *IEEE Signal Process Mag*, 2017, 34: 18–42

4 Xie W, Wang F, Li Y, Lai L, Pei J. *J Chem Inf Model*, 2022, 62: 2269–2279

5 Hadfield TE, Deane CM. *Curr Opin Struct Biol*, 2022, 73: 102326

6 Wang M, Li D, Hou T, Kang Y. *Sci Sin-Chim*, 2023, DOI: 10.1360/SSC-2022-0135 (in Chinese) [王明阳, 李丹, 侯廷军, 康玉. 中国科学:化学, 2023, DOI: 10.1360/SSC-2022-0135]

7 Shen L, Jia L, Tang D, Zhang ZY, Peng D. *Sci Sin-Chim*, 2022, 52: 858–868 (in Chinese) [申林, 贾璐阳, 汤典东, 张战运, 彭鼎. 中国科学:化学, 2022, 52: 858–868]

8 Weininger D. *J Chem Inf Model*, 1988, 28: 31–36
9 Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A. *ACS Cent Sci*, 2018, 4: 268–276
10 Kotsias PC, Arús-Pous J, Chen H, Engkvist O, Tyrchan C, Bjerrum EJ. *Nat Mach Intell*, 2020, 2: 254–265
11 Segler MHS, Kogej T, Tyrchan C, Waller MP. *ACS Cent Sci*, 2018, 4: 120–131
12 Arús-Pous J, Patronov A, Bjerrum EJ, Tyrchan C, Reymond JL, Chen H, Engkvist O. *J Cheminform*, 2020, 12: 38
13 Krenn M, Häse F, Nigam AK, Friederich P, Aspuru-Guzik A. *Mach Learn-Sci Technol*, 2020, 1: 045024
14 Zhong Z, Song J, Feng Z, Liu T, Jia L, Yao S, Wu M, Hou T, Song M. *Chem Sci*, 2022, 13: 9023–9034
15 Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D. *J Cheminform*, 2015, 7: 23
16 https://www.daylight.com
17 Kipf TN, Welling M. 2016, preprint at https://doi.org/10.48550/arXiv.1611.07308.
18 Ma T, Chen J, Xiao C. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, Canada, 2018, 7113–7124
19 Simonovsky M, Komodakis N. In Proceedings of the 27th International Conference on Artificial Neural Networks, Rhodes, Greece, 2018, 412–422
20 De Cao N, Kipf T. 2018, preprint at https://doi.org/10.48550/arXiv.1805.11973.
21 Madhawa K, Ishiguro K, Nakago K, Abe M. 2019, preprint at https://doi.org/10.48550/arXiv.1905.11600.
22 Jin W, Barzilay DR, Jaakkola T. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 2018, 2323–2332
23 Jin W, Barzilay DR, Jaakkola T. In Proceedings of the 37th International Conference on Machine Learning, Virtual Event, 2020, 4839–4848
24 Jin W, Barzilay DR, Jaakkola T. In Proceedings of the 37th International Conference on Machine Learning, Virtual Event, 2020, 4849–4859
25 Chen B, Wang T, Li C, Dai H, Song L. In Proceedings of the 9th International Conference on Learning Representations, Virtual Event, Austria, 2021
26 Xie Y, Shi C, Zhou H, Yang Y, Zhang W, Yu Y, Li L. 2021, preprint at https://doi.org/10.48550/arXiv.2103.10432.
27 Fu T, Xiao C, Li X, Glass LM, Sun J. In Proceedings of the 35th AAAI Conference on Artificial Intelligence, Virtual Event, 2021, 125–133
28 Liu Q, Allamanis M, Brockschmidt M, Gaunt A. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, Canada, 2018
29 You J, Liu B, Ying Z, Pande V, Leskovec J. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, Canada, 2018
30 Khemchandani Y, O'Hagan S, Samanta S, Swainston N, Roberts TJ, Bollegala D, Kell DB. *J Cheminform*, 2020, 12: 53
31 Shi C, Xu M, Zhu Z, Zhang W, Zhang M, Tang J. 2020, preprint at https://doi.org/10.48550/arXiv.2001.09382.
32 Luo Y, Yan K, Ji S. In Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 2021, 7192–7203
33 Fuchs FB, Worrall DE, Fischer V, Welling M. 2020, preprint at https://doi.org/10.48550/arXiv.2006.10503.
34 Satorras VG, Hoogeboom E, Welling M. In Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 2021, 9323–9332
35 Unke OT, Meuwly M. *J Chem Theor Comput*, 2019, 15: 3678–3693
36 Gasteiger J, Groß J, Günnemann S. In Proceedings of the 10th International Conference on Learning Representations, Virtual Event, 2022
37 Orlando G, Raimondi D, Duran-Romaña R, Moreau Y, Schymkowitz J, Rousseau F. *Nat Commun*, 2022, 13: 961
38 Francoeur PG, Masuda T, Sunseri J, Jia A, Iovanisci RB, Snyder I, Koes DR. *J Chem Inf Model*, 2020, 60: 4200–4215
39 Masci J, Boscaini D, Bronstein M, Vandergheynst P. In Proceedings of the 15th IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 2015, 37–45
40 Marsden J, Weinstein A. *Rep Math Phys*, 1974, 5: 121–130
41 Han J, Rong Y, Xu T, Huang W. 2022, preprint at https://doi.org/10.48550/arXiv.2202.07230.
42 Sperduti A, Starita A. *IEEE Trans Neural Netw*, 1997, 8: 714–735
43 Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, Wang L, Li C, Sun M. *AI Open*, 2020, 1: 57–81
44 Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. In Proceedings of the 34th International Conference on Machine Learning, Sydney,

Australia, 2017, 1263–1272
45 Li Y, Wu J, Tedrake R, Tenenbaum JB, Torralba A. 2018, preprint at https://doi.org/10.48550/arXiv.1810.01566.
46 Mrowca D, Zhuang C, Wang E, Haber N, Fei-Fei LF, Tenenbaum J, Yamins DL. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, Canada, 2018, 8799–8810
47 Kipf T, Fetaya E, Wang KC, Welling M, Zemel R. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 2018, 2688–2697
48 Kingma DP, Welling M. 2019, preprint at https://doi.org/10.48550/arXiv.1810.01566.
49 Mansimov E, Mahmood O, Kang S, Cho K. *Sci Rep*, 2019, 9: 20381
50 Simm G, Hernandez-Lobato JM. In Proceedings of the 37th International Conference on Machine Learning, Virtual Event, 2020, 8949–8958
51 Xu M, Wang W, Luo S, Shi C, Bengio Y, Gomez-Bombarelli R, Tang J. In Proceedings of the 38th International Conference on Machine Learning, Virtual Event. 2021, 11537–11547
52 Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. 2014, preprint at https://doi.org/10.48550/arXiv.1406.2661.
53 Guimaraes GL, Sanchez-Lengeling B, Outeiral C, Farias PLC, Aspuru-Guzik A. 2017, preprint at https://doi.org/10.48550/arXiv.1705.10843.
54 Hoffmann M, Noé F. 2019, preprint at https://doi.org/10.48550/arXiv.1910.03131.
55 Kobyzev I, Prince SJD, Brubaker MA. 2019, preprint at https://doi.org/10.48550/arXiv.1908.09257.
56 Zang C, Wang F. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, New York, USA. 2020, 617–626
57 Satorras VG, Hoogeboom E, Fuchs FB, Posner I, Welling M. In Proceedings of the 35th International Conference on Neural Information Processing Systems, Online Event, 2021
58 Luo Y, Ji S. In Proceedings of the 10th International Conference on Learning Representations, Virtual Event, 2022
59 Hochreiter S, Schmidhuber J. In Proceedings of the 10th International Conference on Neural Information Processing Systems, Denver, USA, 1996
60 Oord A van den, Kalchbrenner N, Kavukcuoglu K. In Proceedings of the 33rd International Conference on Machine Learning, New York, USA, 2016, 1747–1756
61 Oord A van den, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K. 2016, preprint at https://doi.org/10.48550/arXiv.1609.03499.
62 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017
63 Devlin J, Chang MW, Lee K, Toutanova K. 2018, preprint at https://doi.org/10.48550/arXiv.1810.04805.
64 Gebauer N, Gastegger M, Schütt K. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, Canada, 2019
65 Schütt K, Kindermans PJ, Sauceda Felix HE, Chmiela S, Tkatchenko A, Müller KR. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017
66 Joshi RP, Gebauer NWA, Bontha M, Khazaieli M, James RM, Brown JB, Kumar N. *J Phys Chem B*, 2021, 125: 12166–12176
67 Gebauer NWA, Gastegger M, Hessmann SSP, Müller KR, Schütt KT. *Nat Commun*, 2022, 13: 973
68 Ho J, Jain A, Abbeel P. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Online Event, 2020, 6840–6851
69 Shi C, Luo S, Xu M, Tang J. In Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 2021
70 Luo S, Shi C, Xu M, Tang J. In Proceedings of the 35th International Conference on Neural Information Processing Systems, Online Event, 2021, 19784–19795
71 Xu M, Yu L, Song Y, Shi C, Ermon S, Tang J. In Proceedings of the 10th International Conference on Learning Representations, Virtual Event, 2022
72 Hoogeboom E, Satorras VG, Vignac C, Welling M. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, USA, 2022
73 Jing B, Corso G, Chang J, Barzilay R, Jaakkola T. 2022, preprint at https://doi.org/10.48550/arXiv.2206.01729.

74 Zhou Z, Kearnes S, Li L, Zare RN, Riley P. *Sci Rep*, 2019, 9: 10752

75 Olivecrona M, Blaschke T, Engkvist O, Chen H. *J Cheminform*, 2017, 9: 48

76 Simm G, Pinsler R, Hernandez-Lobato JM. In Proceedings of the 37th International Conference on Machine Learning, Virtual Event, 2020, 8959–8969

77 Simm GNC, Pinsler R, Csányi G, Hernández-Lobato JM. In Proceedings of the 9th International Conference on Learning Representations, Virtual Event, Austria, 2021

78 Griffiths RR, Hernández-Lobato JM. *Chem Sci*, 2020, 11: 577–586

79 Chan L, Hutchison GR, Morris GM. *J Cheminform*, 2019, 11: 32

80 Chan L, Hutchison GR, Morris GM. *Phys Chem Chem Phys*, 2020, 22: 5211–5219

81 Nigam A, Friederich P, Krenn M, Aspuru-Guzik A. 2020, preprint at https://doi.org/10.48550/arXiv.1909.11655.

82 Ahn S, Kim J, Lee H, Shin J. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Online Event, 2020, 12008–12021

83 Spiegel JO, Durrant JD. *J Cheminform*, 2020, 12: 25

84 Fu T, Gao W, Coley CW, Sun J. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, USA, 2022

85 Li Y, Pei J, Lai L. *Chem Sci*, 2021, 12: 13664–13675

86 Arjovsky M, Chintala S, Bottou L. 2017, preprint at https://doi.org/10.48550/arXiv.1701.07875.

87 Ganea OE, Pattanaik L, Coley CW, Barzilay R, Jensen KF. In Proceedings of the 35th International Conference on Neural Information Processing Systems, Online Event. 2021

88 Stärk H, Ganea OE, Pattanaik L, Barzilay R, Jaakkola T. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, USA, 2022

89 Nesterov V, Wieser M, Roth V. 2020, preprint at https://doi.org/10.48550/arXiv.2010.06477.

90 Masuda T, Ragoza M, Koes DR. In Proceedings of the 34th International Conference on Neural Information Processing Systems workshop, Online Event. 2020

91 Ragoza M, Masuda T, Koes DR. *Chem Sci*, 2022, 13: 2701–2713

92 Luo S, Guan J, Ma J, Peng J. In Proceedings of the 35th International Conference on Neural Information Processing Systems, Online Event, 2021

93 Xu M, Luo S, Bengio Y, Peng J, Tang J. In Proceedings of the 9th International Conference on Learning Representations, Virtual Event, Austria. 2021

94 Zhu J, Xia Y, Liu C, Wu L, Xie S, Wang T, Wang Y, Zhou W, Qin T, Li H, Liu TY. 2022, preprint at https://doi.org/10.48550/arXiv.2202.01356.

95 Hawkins PCD. *J Chem Inf Model*, 2017, 57: 1747–1756

96 Méndez-Lucio O, Ahmad M, del Rio-Chanona EA, Wegner JK. *Nat Mach Intell*, 2021, 3: 1033–1039

97 Huang Y, Peng X, Ma J, Zhang M. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, USA, 2022

98 Wang W, Xu M, Cai C, Miller BK, Smidt T, Wang Y, Tang J, Gómez-Bombarelli R. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, USA, 2022

99 Gogineni T, Xu Z, Punzalan E, Jiang R, Kammeraad J, Tewari A, Zimmerman P. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Online Event, 2020, 20142–20153

100 Axelrod S, Gómez-Bombarelli R. *Sci Data*, 2022, 9: 185

101 Xu Z, Luo Y, Zhang X, Xu X, Xie Y, Liu M, Dickerson K, Deng C, Nakata M, Ji S. 2021, preprint at https://doi.org/10.48550/arXiv.2110.01717.

102 Ramakrishnan R, Dral PO, Rupp M, von Lilienfeld OA. *Sci Data*, 2014, 1: 140022

103 Chmiela S, Tkatchenko A, Sauceda HE, Poltavsky I, Schütt KT, Müller KR. *Sci Adv*, 2017, 3: e1603015

104 Townshend RJL, Vögele M, Suriana P, Derry A, Powers A, Laloudakis Y, Balachandar S, Jing B, Anderson B, Eismann S, Kondor R, Altman RB, Dror RO. In Proceedings of the 35th International Conference on Neural Information Processing Systems, Sydney, Australia, 2021

105 Liu Z, Su M, Han L, Liu J, Yang Q, Li Y, Wang R. *Acc Chem Res*, 2017, 50: 302–309

106 Desaphy J, Bret G, Rognan D, Kellenberger E. *Nucleic Acids Res*, 2015, 43: D399–D404

107 Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. *J Med Chem*, 2012, 55: 6582–6594

108 Kabsch W. *Acta Cryst A*, 1976, 32: 922–923

109 Riniker S, Landrum GA. *J Chem Inf Model*, 2015, 55: 2562–2574

110 Du W, Zhang H, Du Y, Meng Q, Chen W, Shao B, Liu TY. In Proceedings of the 39th International Conference on Machine Learning, Baltimore,

USA, 2022, 5583–5608
111 Zhu J, Xia Y, Wu L, Xie S, Qin T, Zhou W, Li H, Liu TY. 2022, preprint at https://doi.org/10.48550/arXiv.2207.08806
112 Zhou G, Gao Z, Ding Q, Zheng H, Xu H, Wei Z, Zhang L, Ke G. 2022, preprint at https://doi.org/10.26434/chemrxiv-2022-jjm0j-v3
113 Rong Y, Bian Y, Xu T, Xie W, WEI Y, Huang W, Huang J. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Online Event, 2020, 12559–12571
114 Li P, Wang J, Qiao Y, Chen H, Yu Y, Yao X, Gao P, Xie G, Song S. *Briefings BioInf*, 2021, 22: bbab109
115 Zhu Y, Chen D, Du Y, Wang Y, Liu Q, Wu S. In Proceedings of the 39th International Conference on Machine Learning workshop, Baltimore, USA, 2022
116 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. *Nature*, 2021, 596: 583–589
117 Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, Millán C, Park H, Adams C, Glassman CR, DeGiovanni A, Pereira JH, Rodrigues AV, van Dijk AA, Ebrecht AC, Opperman DJ, Sagmeister T, Buhlheller C, Pavkov-Keller T, Rathinaswamy MK, Dalwadi U, Yip CK, Burke JE, Garcia KC, Grishin NV, Adams PD, Read RJ, Baker D. *Science*, 2021, 373: 871–876

Advances in deep learning-based 3D molecular generative models

Shaolun Yao[1], Jie Song[2*], Zunlei Feng[2], Lingxiang Jia[3], Zipeng Zhong[3], Mingli Song[3]

[1] *Collaborative Innovation Center of Artificial Intelligence by MOE and Zhejiang Provincial Government, Zhejiang University, Hangzhou 310007, China*
[2] *College of Software Technology, Zhejiang University, Ningbo 315020, China*
[3] *College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*
**Corresponding author (email: sjie@zju.edu.cn)*

Abstract: Molecular generation, aiming at designing novel molecules with desired properties such as biological activity and drug metabolism and pharmacokinetic (DMPK) properties in a low-cost and high-efficient manner, is a fundamental problem in drug discovery. Recently, deep generative models have been widely used in drug discovery, along with numerous model architectures and optimization strategies being explored, most of which are developed to generate one-dimensional or two-dimensional molecular structures. Inspired by the rapid development of deep learning in processing geometric graph data, 3D generative models for molecular generation have also been proposed, gaining attention for their advantages and potential on direct 3D molecular conformation generation and structure-based drug discovery. This survey has offered a systematic summarization of existing research achievements of the domestic and foreign researchers in recent years in the aspects of 3D molecular generation. According to the input of 3D molecular generation algorithms, it is divided into latent variable-based generation, 2D graph-based generation and 3D geometry-based generation. Then, according to the output of 3D molecular generation algorithms, it is divided into goal-directed generation and non-goal-directed generation. Furthermore, the performance of different generative models on main public datasets is summarized, which proves the advantages and disadvantages of the various models. Finally, some promising research directions are proposed.

Keywords: 3D molecular generation, equivariant graph neural networks, deep generative models, goal-directed generation, non-goal-directed generation

doi: 10.1360/SSC-2022-0204