基于人工智能的小分子生成模型在药物发现中的研究进展

唐 谦[1], 陈柔菜[2], 沈哲远[2], 池幸龙[3], 车金鑫[2*], 董晓武[2**]

([1] 浙江省药品化妆品审评中心, 杭州 310012; [2] 浙江大学药学院, 杭州 310058; [3] 杭州医学院, 杭州 310013)

摘 要 随着人工智能技术的快速发展,小分子生成模型已成为药物发现领域的重要研究方向。该类模型,包括生成对抗网络(GANs)、变分自编码器(VAEs)和扩散模型等,已被证明在优化药物属性和生成复杂分子结构方面具有显著能力。本文综合分析了上述先进技术在药物发现过程中的应用,展示了其如何补充和改进传统药物设计方法。同时,提出了当前方法在数据质量、模型复杂性、计算成本及泛化能力等方面的挑战,并对未来的研究方向进行了展望。

关键词 小分子生成模型;药物发现;人工智能技术

中图分类号 TP18;R914.2 **文献标志码** A **文章编号** 1000−5048(2024)03−0295−11

doi: 10.11665/j.issn.1000−5048.2024031501

引用本文 唐谦,陈柔菜,沈哲远,等. 基于人工智能的小分子生成模型在药物发现中的研究进展 [J]. 中国药科大学学报,2024,55(3):295 − 305.

Cite this article as: TANG qian, CHEN Roufen, SHEN Zheyuan, *et al*. Research progress of artificial intelligence-based small molecule generation models in drug discovery[J]. *J China Pharm Univ*, 2024, 55(3): 295 − 305.

Research progress of artificial intelligence-based small molecule generation models in drug discovery

TANG qian[1], CHEN Roufen[2], SHEN Zheyuan[2], CHI Xinglong[3], CHE Jinxin[2*], DONG Xiaowu[2**]

[1]Zhejiang Center for Drug & Cosmetic Evaluation, Hangzhou 310012; [2]College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058; [3]School of Pharmacy, Hangzhou Medical College, Hangzhou 310013, China

Abstract With the rapid development of artificial intelligence technology, small molecule generation models have emerged as a significant research direction in the field of drug discovery. These models, including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models, have proven to possess remarkable capabilities in optimizing drug properties and generating complex molecular structures. This article comprehensively analyzes the application of the aforementioned advanced technologies in the drug discovery process, demonstrating how they supplement and enhance traditional drug design methods. At the same time, it addresses the challenges facing current methods in terms of data quality, model complexity, computational cost, and generalization ability, with a prospect of future research directions.

Key words small molecule generation model; drug discovery; artificial intelligence technology

This study was supported by Zhejiang Provincial Soft Science Research Program (No. 2024C35015)

目前计算机辅助药物设计(computer-aided drug design, CADD)与传统药物设计方法结合[1],已逐渐成为药物设计的主流方法,弥补了高通量筛选(high-throughput screening, HTS)[2]的缺陷。然而,传统的计算机辅助药物设计方法无法有效穿透庞大的药物化学空间(约 10^{60} 量级)。因此,生成新颖、独特且有效的分子仍然是一项挑战性的任务[3]。近年来,随着人工智能技术的蓬勃发展,将分子生

收稿日期 2024-03-15 **通信作者** *Tel: 15068846367 E-mail: chejx@zju.edu.cn
**Tel: 13588478539 E-mail: dongxw@zju.edu.cn

基金项目 浙江省软科学研究计划项目(No. 2024C35015)

成模型引入药物发现过程已成为分子设计、合成的研究热点，为药物设计和开发打开了新的视角和途径。当前常用的生成模型包括生成对抗网络(GAN)、变分自编码器(VAE)和扩散模型(diffusion model)等[4]。本文对近 5 年药物设计工作中的小分子生成模型进行介绍，并根据不同药物设计场景对其进行归纳，并展望小分子生成模型在药物设计场景的发展方向及挑战。

1 分子生成模型概念

1.1 分子生成模型结构

现有的分子生成方法分为基于配体的分子生成(ligand-based molecular generation, LBMG)和基于结构的分子生成(structure-based molecular generation, SBMG)(图 1-A 和 1-B)[5]。其中，基于配体的分子生成方法主要根据现有活性分子骨架及其属性来生成分子。然而，由于 LBMG 方法是在缺少靶点口袋信息的情况下进行训练和生成，其无法考虑分子与靶蛋白的相互作用，且可能会因训练集中的活性分子骨架多样性不足而产生偏差。因此，该方法生成的分子将有可能缺乏多样性，不利于突破专利壁垒。而基于结构的分子生成方法是近年来药物设计领域的一个重要发展方向，该方法利用靶蛋白的三维结构信息，如靶蛋白的活性位点或口袋等关键区域的详细结构信息，来引导新分子的生成和优化。与基于配体的方法相比，基于结构的方法能够更直接地考虑分子与靶蛋白的相互作用，从而更精准地设计出类药分子。

1.2 数据来源和特征

1.2.1 数据来源 高质量的数据在分子生成领域发挥着至关重要的作用。在这一过程中，数据库(database)和数据集(dataset)是关键的组成部分。

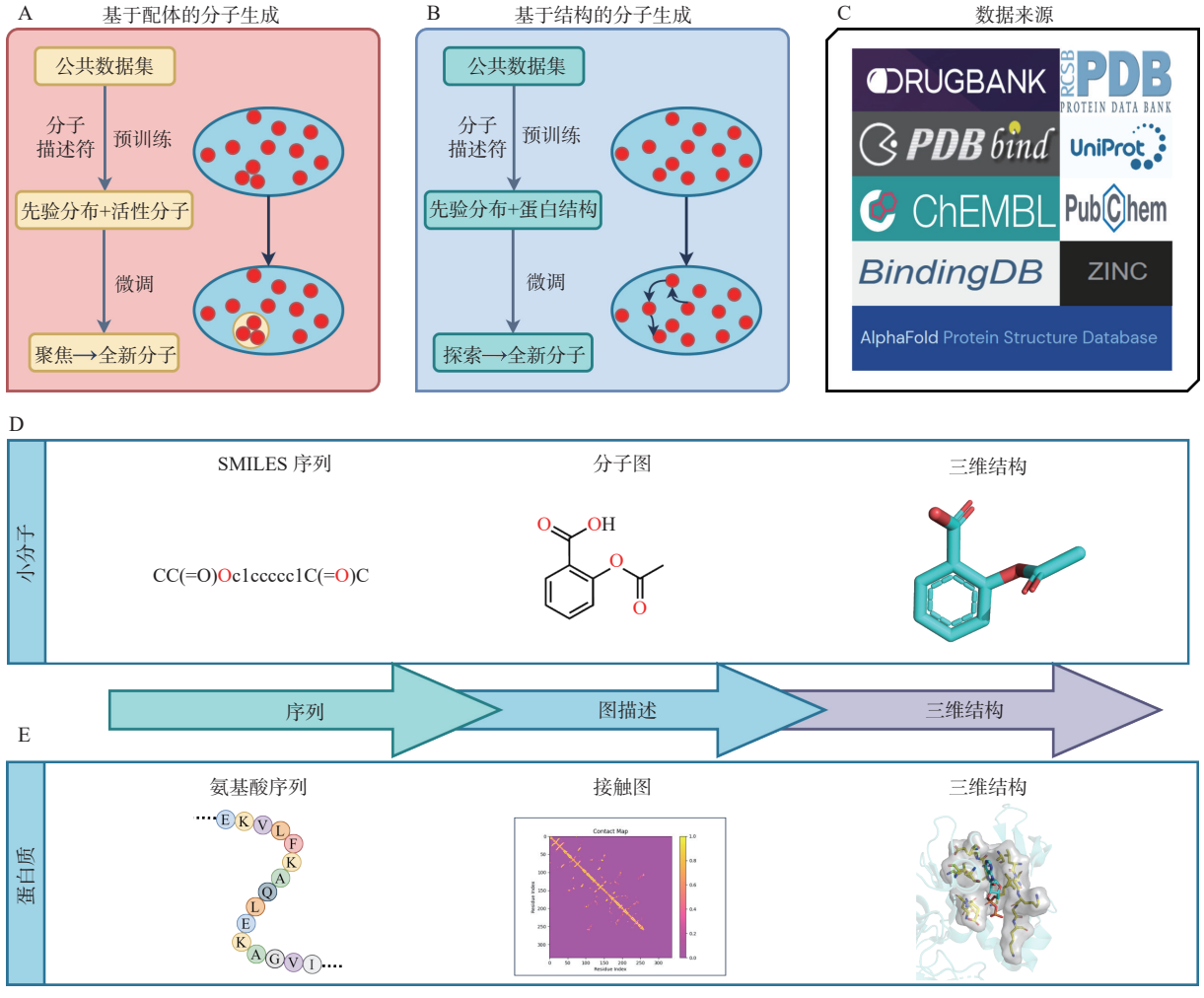

图 1 分子生成模型分类、数据来源和特征

A:基于配体的分子生成示意图; B:基于结构的分子生成示意图; C:数据来源; D:配体的数据表示方法; E:靶蛋白的数据表示方法

如图 1-C 所示, 在分子生成研究中, 常用的小分子数据库包括 ZINC 数据库、ChEMBL 数据库以及基准数据集如 MOSES 数据集和 QM9 数据集等。这些数据集分别提供了不同类型的分子数据, 以支持化学规则、药物化学过滤和 3D 构象生成模型的研究。此外, 一些带有靶标信息的数据库和数据集主要为基于靶蛋白口袋的分子生成提供帮助, 包括提供序列信息的 Uniprot 数据库, 提供蛋白质-配体复合物数据的 CrossDock 数据集和 PDBbind 数据库。RCSB PDB 和 AlphaFold 蛋白质结构数据库则为分子生成提供了大量的实验解析和人工智能预测的靶标结构数据。

1.2.2　*数据特征*　在分子和蛋白质的表示法方面, 一维、二维和三维方法各展其优, 并针对不同的应用场景具有独特的应用价值和优势[6]。对于一维表示法, 其广泛应用于小分子和蛋白质的表示中。小分子通过 SMILES 格式进行表示, 这种方法利用 ASCII 字符串以直观和简洁的方式描述化学结构(图 1-D)[6]。蛋白质的一维表示通常依赖于氨基酸序列, 以反映其组成和序列顺序(图 1-E)[7]。在二维表示法中, 分子结构通过图表示法进行分析, 其中分子被视为图, 原子和化学键分别作为节点和边(图 1-D)[8]。尽管这种方法在形式上是二维的, 它却能够编码三维信息, 例如通过将原子坐标和键角编入节点和边的属性中(图 1-E)[9]。三维表示法是从物理模型到计算技术的重大变革。3D 分子图是一种常见的三维表示方式, 直接将原子的三维坐标作为三维特征, 而 3D 分子网格则通过体素(体积元素, volume pixel)的方式来表示分子构象中的不同元素或属性(图 1-D)。而通常, 三维表征的坐标会利用等变图神经网络, 如 E(n) 或 SE(n) 模型进行处理, 这保持了输入数据的几何不变性和对称性, 使得它们在蛋白质结构预测、功能分析和药物设计等方面展现出了高效率和准确性(图 1-E)[10]。

2　主流分子生成模型算法

2.1　变分自编码器

变分自编码器(variational autoencoders, VAE)包含两个核心组件: 编码器与解码器(图 2-A)。在 VAE 的训练过程中, 关键是最小化包含重构损失与正则化项的损失函数。重构损失确保解码器根据隐变量准确重建原始分子, 而正则化项则利用 Kullback-Leibler(KL)散度衡量重构分子的分布与原始数据分布之间的偏差。这种方法有效地确保了模型能够生成既高质量又在统计上与真实分子匹配的分子[11]。

2.2　生成对抗网络

生成对抗网络(generative adversarial network, GAN)构建了一个由两个相互竞争的网络组成的系统: 一个是负责生成新的分子结构的生成器(generator), 另一个是专注于区分合成分子结构与真实分子结构的判别器(discriminator)(图 2-B)[12]。这一竞争机制不仅促进了模型生成逼真分子的能力, 也提高了模型在学习和理解复杂分子特征方面的效率[13]。

2.3　扩散模型

扩散模型(diffusion models)的设计灵感来源于物理学中的非平衡热力学原理, 它们通过定义一个马尔科夫链的扩散步骤, 涵盖了正向扩散过程和逆向扩散过程两个阶段(图 2-C)[14]。在正向扩散过程中, 这些模型会逐步向数据注入随机噪声, 直到数据逐渐演变成各向同性的高斯分布状态; 而在逆向扩散过程中, 则执行相反的操作, 从噪声状态重构出原始的数据样本。这种方法使扩散模型在生成真实样本方面展现出显著的能力[15]。

2.4　标准化流模型

标准化流(normalizing flow)通过实施一连串的可逆变换, 实现了从简单的先验分布(比如高斯分布)到复杂高维数据(例如分子结构)的高效桥接。与变分自编码器(VAE)相比, 其能够进行精确的数据似然计算, 显著提升了新分子结构的生成能力, 并在化学及药物设计领域中对生产高品质分子展现出巨大潜力(图 2-D)[16]。

2.5　递归神经网络模型

递归神经网络(RNN)是一类用于处理序列数据的神经网络, 通过其独特的反馈连接结构, 能够在序列的每个步骤中传递先前的信息, 从而有效捕捉时间序列或序列数据中的动态特性(图 2-E)。RNN 在处理序列化的分子表示(如 SMILES 字符串)方面显示出独特的优势, 使其在生成具有特定化学性质的分子方面具有更高的灵活性和准确性[17]。

2.6　Transformer 模型

Transformer 模型(自注意力模型)自 2017 年提出以来, 已经彻底改变了自然语言处理(NLP)等序

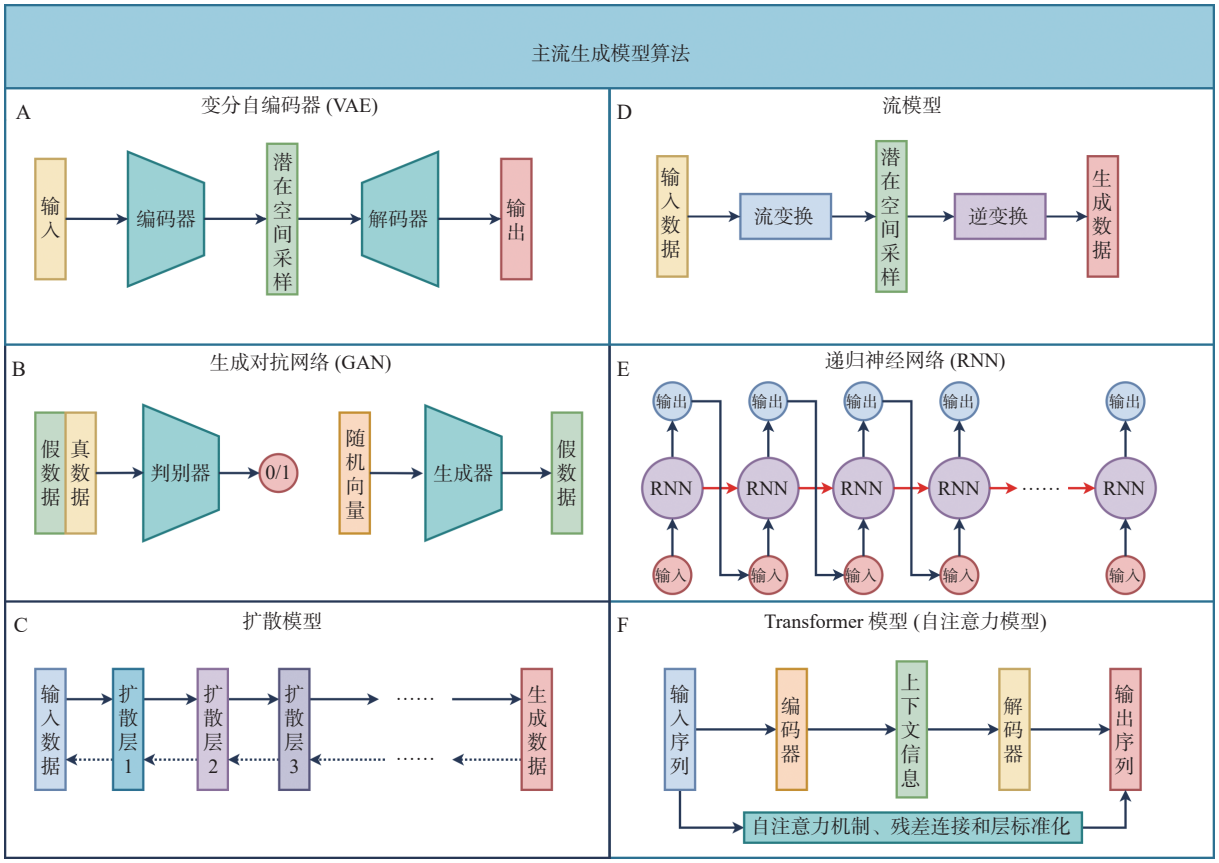

图 2 主流生成模型算法示意图

A:变分自编码器; B:生成对抗网络; C:扩散模型; D:标准化流模型; E:递归神经网络;F:Transformer 模型

列处理任务的面貌(图 2-F)。近年来, Transformer 的应用领域不断扩展, 分子生成领域也开始尝试利用其强大的序列建模能力。此外, Transformer 的并行处理能力显著提高了模型的训练和推理速度, 这对于处理大规模分子数据集尤为重要[18]。

3 分子生成模型的评价基准

在分子生成模型的研究领域, 评价指标对于衡量模型性能和指导模型改进至关重要。这些评价指标主要分为两大类: 药物样性和多元化评价指标。药物样性指标评估分子成药潜质, 如物理化学属性(如相对分子质量、脂水分配系数)等[19]。Bickerton 等[20] 提出的药物样性的定量估算(QED)是一种基于批准药物分子属性的评估方法, 涵盖了分子质量、疏水性和极性等。Weng 等[21] 开发的 RediscMol 基准用于评估生成模型在生物属性方面的表现, 为药物设计提供指导。Handa 等[22] 探讨了分子生成模型实际验证的挑战, 通过案例研究发现, 公共数据上模型的表现优于专有数据, 指出评估新化合物设计方法在药物发现中的难度。Ertl 等[23] 提出的 SAscore 基于片段贡献和复杂性惩罚, 与药物化学家估计的合成可能性显示出良好一致性。Thakkar 等[24] 开发的 RAscore, 一个基于机器学习的快速合成可行性估计方法, 至少比传统方法快 4 500 倍, 适用于早期药物发现阶段。Wang 等[25] 的 DeepSA 使用深度学习预测化合物合成可达性, 性能优于现有方法, 特别是在区分难合成分子方面。除了传统和合成可行性评价指标, 新兴的评价方法如空间得分(SPS)[26−28] 等。

4 小分子生成模型在药物发现中的应用与进展

4.1 分子从头生成

分子从头生成(*de novo* molecular generation)是一种旨在自动提出新化学结构的方法, 常用于药物发现以获取具有理想的生物效应和药代动力学特性的分子。近期研究进展包括 Guo 等[29] 通过结合课程学习(CL)与基于 REINVENT 的方法, 提升了从头分子设计的质量, 显示出 CL 在加速学习和

提高生成质量方面的优势。Zhang 等[30] 提出的 ResGen 模型, 利用并行多尺度建模原则, 通过分层的自回归过程直接在蛋白质口袋内生成 3D 分子, 有效捕捉分子与口袋间的相互作用, 提高了计算效率和生成的物理合理性。此外, 该团队还以“锁钥”原理的方式开发了包含两个等变神经网络(Geodesic-GNN 和 Geoattn-GNN)的 SurfGen 模型, 其通过结合拓扑和几何结构学习, 分别捕获口袋表面的拓扑相互作用和配体原子与表面节点之间的空间相互作用[31]。Mokaya 等[32] 探索了基于 SMILES 字符串的分子生成极限, 通过课程学习和深度强化学习, 提出了循环迭代优化程序(rIOP)来优化分子生成, 允许更好地控制生成分子集的组成。Moret 等[33] 利用化学语言模型(CLM)结合结构和活性信息, 生成了具有中到低纳摩尔级别活性的 PI3Kγ 抑制剂, 验证了其在脑癌细胞模型中对 PI3K/Akt 通路的有效抑制作用。Qian 等[34] 开发的 KGDiff 模型, 将化学知识引导融入扩散模型中, 通过知识指导的去噪过程, 有效提高了结合亲和力。Xu 等[35] 开发了 Tree-Invent 模型, 通过强化学习和拓扑树约束, 灵活地探索目标化学空间, 除了从头生成外, 还可以应用于骨架跃迁等。

4.2　基于片段修饰的分子生成

4.2.1　基于骨架　在现代药物发现领域, 骨架修饰和骨架跃迁是关键技术, 用于创造具有新颖生物活性的分子。Lim 等[36] 的研究展示了如何使用 VAE 作为图生成模型来进行基于骨架的分子设计, 该方法能够顺序地添加原子和键, 生成衍生分子, 在保留主骨架结构的同时控制分子属性。Hu 等[37] 开发了基于 Transformer 的分子生成模型 SyntaLinker-Hybrid, 通过片段杂交化和迁移学习步骤以及对 BRAF 激酶的对接实验验证了其有效性。Zheng 等[38] 提出的 DeepHop 模型, 基于结合分子 3D 构象和蛋白质序列信息的多模式 Transformer 架构, 专注于骨架跃迁, 生成具有高生物活性的新型骨架分子, 展示了优异的性能。Fialková等[39] 开发的基于强化学习的生成模型 LibINVENT, 通过指定化学反应快速创建共享相同核心的化合物库。Loeffler 等[40] 报道的 REINVENT 4, 基于自注意力机制的变换器架构, 设计用于生成小分子的开源框架。Liao 等[41] 开发的 Sc2Mol 模型采用 VAE 生成骨架, 以及使用 Transformer 进行结构修饰的两步流程, 展示了其在学习化学规则和优化类药物分子设计方面的能力。Liu 等[42] 的 DrugEx v3 模型, 通过多目标强化学习, 侧重于生成类似药物的配体, 引入了新位置编码方案以在一个骨架中生长多个片段。Xu 等[43] 报道的 3D-SMG 模型, 结合了交叉聚合连续滤波卷积(ca-cfconv)实现高效的空间特征提取, 并引入了用于 ADMET 特性预测的数据自适应多模型方法, 取得了优异的性能。Hu 等[44] 提出的 ScaffoldGVAE, 通过变分自动编码器原理和多视图神经网络, 实现药物分子的骨架生成和跃迁, 侧重于保留分子侧链同时修饰骨架。Xie 等[45] 基于扩散模型开发的 DiffDec, 通过骨架修饰优化分子, 结合三维口袋约束, 提高结构感知 R-Group 的生成。

4.2.2　基于片段　基于片段的药物设计(fragment-based drug design, FBDD), 通过 3 种主要策略: 片段生长、片段连接、片段合并, 将片段分子优化为先导化合物。Bilsland 等[46] 提出了一个基于长短期记忆(LSTM)网络的“双重”自动编码器, 能同时重构 SMILES 和分子指纹。这种模型通过对分子指纹解码器层应用迁移学习, 生成了一个分类器模型, 识别预测为频繁命中的新片段。Hadfield 等[47] 开发 STRIFE 模型, 从靶蛋白结构提取片段热点图(FHMs), 引导片段生长, 且允许用户指定设计的片段结构并生成拓展物。Du 等[48] 的 3D-MCTS 模型利用蒙特卡洛树搜索(MCTS)算法, 通过多线程并行模拟和实时能量约束的剪枝策略, 高效识别具有优化目标结合亲和力的分子。Powers 等[49] 的 FRAME 模型, 基于 SE(3)-等变神经网络, 用于片段生长, 展现出在预测配体亲和力和选择性方面的显著优势。Sauer 等[50] 提出了一种基于演员-评论家模型(actor-critic model)的片段强化学习变体, 引入冷冻片段(freezing fragments)和使用试剂作为片段源, 结合基于化学反应的拆分方案, 提高了可合成性, 调整网络输出概率以平衡新颖性和多样性。Wang 等[51] 的 Frag-G/M 框架, 基于条件 Transformer、递归神经网络(RNN)和强化学习, 减少了模型训练中标签数据的使用, 生成的分子展示了良好的骨架多样性。

此外, 在生成片段化合物库方面, 目前也取得良好的进展, 如 Eguida 等[52] 报道的 POEM 方法, 利用所有公开的蛋白质-配体复合物结构信息, 生成针对特定目标的化合物库, 该模型已成功应用于生成

150 万个潜在的 CDK8 抑制剂库。Buehler 等[53] 通过分析公共数据库中的分子，探索了生物活性片段空间的扩展，通过生成数据库 GDB-13s，发现了许多新的、结构简单且具有合成可及性的片段。Diao 等[54] 的 MacFrag 模型，通过改进的 BRICS 规则和高效的子图提取算法，快速枚举分子片段空间，为大规模数据库的片段化提供了一种高效方法。

4.2.3 基于连接子 基于片段的药物发现已经成为早期药物开发的一个有效范式。其中，在靶向蛋白降解嵌合体（PROTAC）或抗体偶联药物（ADC）的设计中，连接子（linker）的选择和设计至关重要。Imrie 等[55] 提出了 DeLinker，一个基于图的深度生成模型，结合三维结构信息应用于片段连接、骨架跃迁和 PROTAC 设计等场景。Yang 等[56] 开发的 SyntaLinker，一个基于约束的 Transformer 架构，通过编解码层处理分子片段序列，自动连接片段并生成完整分子，验证了其在片段连接、骨架跃迁等方面的应用。Tan 等[57] 提出了 DRlinker，基于强化学习，控制药物设计中片段的连接，生成具有特定属性（如 Linker 长度和 log P）的化合物，展示了其在骨架跃迁场景中的潜力。Li 等[58] 开发的 PROTAC-INVENT，一个三维生成模型，专为 PROTAC Linker 设计，通过强化学习生成二维结构及与靶蛋白和 E3 连接酶的三维结合构象（PTS），且包含对 PROTAC 的对接协议。Kao 等[59] 提出的 AIMLinker，基于变分自编码器，专门为 PROTAC Linker 设计，通过整合输入片段的结构信息来优化先导化合物，内置过滤机制确保生成具有优良化学属性的分子。Zhang 等[60] 的 GRELinker，结合基于图的神经网络（GGNN）、强化学习和课程学习（CL）算法，提高了生成满足多重属性约束条件分子的比例，CL 策略使其能生成更复杂的 Linker 结构。

4.3 基于药效团

药效团，作为药物分子中与靶蛋白相互作用的关键部分，包括氢键给体或受体、疏水区域和电荷区域等，能够精确捕获配体与靶蛋白之间的相互作用。Imrie 等[61] 提出了一种基于图的生成模型 DEVELOP (DEep Vision-Enhanced Lead OPtimisation)，该模型结合了图神经网络（GNN）和卷积神经网络（CNNs），将 3D 药效团信息约束纳入分子生成过程。Zhu 等[62] 则提出了以药效团引导的分子生成模型（pharmacophore-guided molecule generation, PGMG）。该模型利用药效团作为药物-靶标相互作用所必需的空间分布化学特征，通过不需要特定活性数据进行训练来克服数据稀缺的限制。

4.4 基于组学

组学（Omics）是一种全面表征和定量分析生物体内分子群体的方法，它深入探究了生物分子层面的结构、功能和动态变化[63]。Born 等[64] 提出了基于强化学习和靶细胞或癌症部位的转录组特征的分子生成框架 PaccMann^{RL}。该方法利用变分自编码器来定制针对特定转录组轮廓的分子，使用抗癌药物敏感性预测模型作为奖励函数，从而优化分子生成过程。Pravalphruekul 等[65] 提出了一个基于变分自编码器的生成模型 BiCEV，旨在从差异基因表达数据中设计具有多重作用潜力的新分子，并通过提供基因敲除图谱来验证 BiCEV，并评估生成的分子与已确认的沉默基因抑制剂的相似程度。此外，Das 等[66] 基于变分自编码器提出了 Gex2SGen，其以非细胞特异性方式接受所需的基因表达谱作为输入，并设计出能激发所需转录组谱的类药分子。

4.5 基于化学反应

在药物发现过程中，化学合成扮演着核心角色。Dolfus 等[67] 提出了 Synthesia，一种基于逆合成分析引导的结构修饰方法。它通过交换合成路线中的前体分子并前向合成重建，使得可生成所需的特定分子属性（如 ADMET 性质）且不影响可合成性的类似物。此外，Qiang 等[68] 提出一种通过统一模型 Uni-RXN 弥合化学反应预处理和条件分子生成之间差距的新方法。Uni-RXN 在反应分类准确性和多种可合成分子的生成方面取得了显著改进，展示了其在简化药物发现过程方面的潜力。

4.6 基于多目标的分子生成与基于多靶点的分子生成

在药物发现过程中，不仅需提升分子的活性，还需保证分子的 ADMET 等成药性质。近年来，多目标分子生成方法在药物发现领域也取得了显著进展。Khemchandani 等[69] 提出了采用图卷积和强化学习的多目标分子生成方法 DeepGraphMolGen，用于生成具有期望属性的分子。该模型使用图卷积网络（GCNs）学习从蛋白、配体结合数据（binding data）中的分子与目标的相互作用模型，并通过强化学习优化包括药物相似性和可合成性在内的多个

目标。Lamanna 等[70] 提出的 GENERA 算法结合了深度学习算法 DeLA-Drug 和遗传算法, 针对特定靶标快速生成有前景的候选分子, 并通过多目标优化验证其能力。GENERA 生成的聚焦库显示了与已知 ACE-2 结合剂相比更好的得分, 证明了其在目标导向的 *de novo* 设计中的创新潜力。

4.7　分子生成模型发现的化合物

分子生成模型已在药物发现领域成为关键的研究工具, 特别是在加速新药候选化合物的发现与优化过程中展现了其独特价值[71]。例如, 由英国的 Exscientia 和住友大日本制药(Sumitomo Dainippon Pharma)共同开发的 DSP-1181, 作为首个由 AI 设计且于 2020 年 1 月在日本进入临床Ⅰ期研究的分子, 是一种长效的 5-HT_{1A} 受体激动剂。遗憾的是, 该分子尚未披露结构且在临床Ⅰ期试验中未达到预期效果, 因此已停止其开发[72]。此外, 英矽智能公司运用其多模态生成式强化学习平台 Chemistry42, 成功鉴定 TRAF 相互作用激酶(TNIK)为抗纤维化靶标, 并开发了小分子抑制剂 INS018_055(图 3-A)。该药物可以通过口服、吸入或局部应用, 已在多个体内研究中证明具有显著的抗纤维化和抗炎效果, 并在覆盖 78 名健康志愿者的随机、双盲、安慰剂对照的临床Ⅰ期试验中验证了其安全性和药代动力学。从靶点发现到药物候选仅用时 18 个月, 体现了生成式 AI 药物发现流程的高效性, 目前该分子正在临床Ⅱ期试验中[73]。英矽智能公司也成功优化了一系列基于氮杂环的 CDK8 抑制剂, 经过多轮优化, 这些抑制剂在体外的微粒体稳定性、激酶选择性等特性均得到显著提升。其中, 先导化合物 **1** 在 CDK8 LanthaScreen 结合测定中显示出的卓越抑制活性, IC_{50} 为 0.7 nmol/L, 并在 MV4-11 细胞活力测试中也表现出 IC_{50} 为 11.8 nmol/L 的优秀抑制活性(图 3-B)。此外, 化合物 **1** 在多个体内肿瘤模型中也展示了剂量依赖性的抑制效果, 目前该分子仍在临床前研究中[74]。2024 年, 麦克马斯特大学与斯坦福大学的研究团队共同开发了 SyntheMol 模型, 成功设计针对超级细菌鲍曼不动杆菌(*Acinetobacter baumannii*)且易于合成的分子, 并成功合成了 58 种分子。其中 6 种结构新颖的分子(化合物 **2**~**7**)对包括肺炎克雷伯菌在内的多种细菌病原体表现出显著的抗菌活性, 并具有良好的安全性(图 3-C)。然而, SyntheMol 模型目前仍有局限

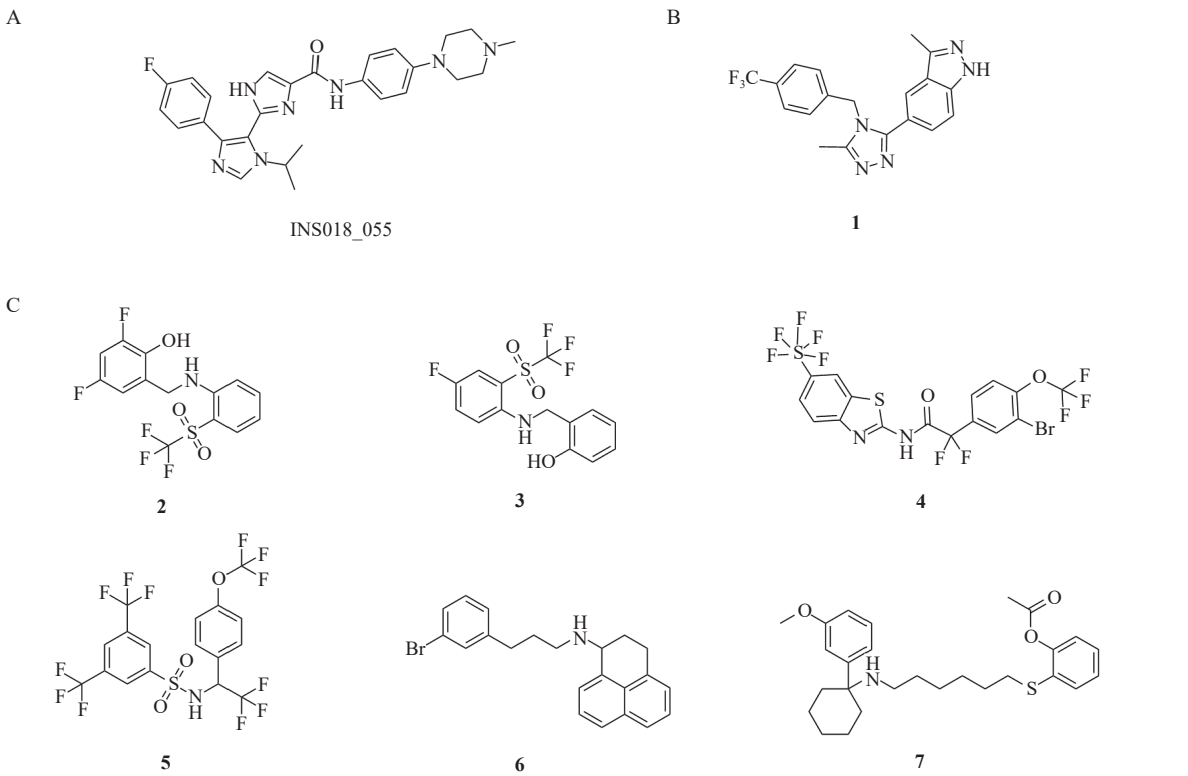

图 3　部分由 AI 设计的化合物结构

A: TNIK 抑制剂 INS018_055; B: CDK8 抑制剂 Compound **23**; C: SyntheMol 模型设计的 6 种结构新颖的抗生素

性，如有 4 个分子因水溶性不足未能进入动物实验阶段[75]。尽管 AI 设计的分子在转化为临床应用方面仍面临挑战，但生成式 AI 的应用仍在不断推动着药物研发的边界。例如，英矽智能公司通过其 Chemistry42 平台在抗纤维化治疗领域取得的突破等。这些进展不仅彰显了 AI 在药物发现过程中的应用潜力，也突显了持续创新与技术改进的必要性，以克服现有挑战并最终实现新药的临床成功。

5 结论和展望

深度生成模型已在分子生成等领域显示出巨大的潜力，尤其在计算化学、药物设计及人工智能等多学科交叉的研究领域中，这些模型在生成复杂分子结构与优化药物属性方面展示了其显著的能力。特别是变分自编码器、生成对抗网络、流模型以及扩散模型等，其在模拟和生成具有药物活性的新分子结构方面取得了重要进展。然而，这些先进模型在转化为实际应用时，仍然面临诸多挑战。这些挑战包括但不限于数据的质与量、模型的复杂性与计算成本之间的权衡、泛化能力的强弱，以及模型的可解释性、透明度和隐私保护的重要性等。

在上述背景下，模型决策过程的解释性和透明度显得尤其关键。因此，未来的研究应不仅聚焦于提升模型性能，而且应深入探讨如何优化其泛化能力、透明度等方面，以确保这些技术在实际应用中的有效性与可靠性。此外，对于药物化学空间的探索，深度生成模型有潜力打破传统的界限，提供超出现有知识库的新分子。这不仅可以加速新药的发现过程，而且有助于探索未被传统方法所涵盖的化学空间，从而揭示未知的药效关系和分子机制。有效克服当前面临的挑战，并充分利用深度生成模型的这一潜力，将极大推动药物发现和分子设计的创新，为未来的药物研发带来革命性的变化。

References

[1] Sabe VT, Ntombela T, Jhamba LA, *et al*. Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: a review[J]. *Eur J Med Chem*, 2021, **224**: 113705.

[2] Macarron R, Banks MN, Bojanic D, *et al*. Impact of high-throughput screening in biomedical research[J]. *Nat Rev Drug Discov*, 2011, **10**(3): 188-195.

[3] Zeng XX, Wang F, Luo Y, *et al*. Deep generative molecular design reshapes drug discovery[J]. *Cell Rep Med*, 2022, **3**(12): 100794.

[4] Bilodeau C, Jin WG, Jaakkola T, *et al*. Generative models for molecular discovery: recent advances and challenges[J]. *WIREs Comput Mol Sci*, 2022, **12**(5): e1608.

[5] Thomas M, Smith RT, O'Boyle NM, *et al*. Comparison of structure- and ligand-based scoring functions for deep generative models: a GPCR case study[J]. *J Cheminform*, 2021, **13**(1): 39.

[6] Wigh DS, Goodman JM, Lapkin AA. A review of molecular representation in the age of machine learning[J]. *WIREs Comput Mol Sci*, 2022, **12**(5): e1603.

[7] Carracedo-Reboredo P, Liñares-Blanco J, Rodríguez-Fernández N, *et al*. A review on machine learning approaches and trends in drug discovery[J]. *Comput Struct Biotechnol J*, 2021, **19**: 4538-4558.

[8] Cereto-Massagué A, Ojeda MJ, Valls C, *et al*. Molecular fingerprint similarity search in virtual screening[J]. *Methods*, 2015, **71**: 58-63.

[9] David L, Thakkar A, Mercado R, *et al*. Molecular representations in AI-driven drug discovery: a review and practical guide[J]. *J Cheminform*, 2020, **12**(1): 56.

[10] Coley CW, Barzilay R, Green WH, *et al*. Convolutional embedding of attributed molecular graphs for physical property prediction[J]. *J Chem Inf Model*, 2017, **57**(8): 1757-1772.

[11] Igashov I, Pavlichenko N, Grudinin S. Spherical convolutions on molecular graphs for protein model quality assessment[J]. *Mach Learn: Sci Technol*, 2021, **2**(4): 045005.

[12] Zhang Y, Huang W, Wei Z, *et al*. EquiPocket: an E(3)-Equivariant Geometric Graph Neural Network for Ligand Binding Site Prediction[EB/OL]. *arXiv*, 2023. http://arXiv.org/abs/2302.12177.

[13] Chen C, Chen X, Morehead A, *et al*. 3D-equivariant graph neural networks for protein model quality assessment[J]. *Bioinformatics*, 2023, **39**(1): btad030.

[14] MohammadiS, O'Dowd B, Paulitz-Erdmann C, *et al*. Penalized variational autoencoder for molecular design[EB/OL]. *ChemRxiv*, 2019. https://ChemRxiv.org/engage/ChemRxiv/article-details/60c74169f96a0012ee286438.

[15] Prokhorov V, Shareghi E, Li YZ, *et al*. On the importance of the kullback-leibler divergence term in variational autoencoders for text generation[C]//Proceedings of the 3rd Workshop on Neural Generation and Translation. Hong Kong, China. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 118–127.

[16] Choi J, Seo S, Choi S, *et al*. ReBADD-SE: multi-objective molecular optimisation using SELFIES fragment and off-policy self-critical sequence training[J]. *Comput Biol Med*, 2023, **157**:

106721.

[17] He DK, Liu Q, Mi Y, *et al*. De novo generation and identification of novel compounds with drug efficacy based on machine learning[J]. *Adv Sci*, 2024, **11**(11): e2307245.

[18] Kutsal M, Ucar F, Kati ND. Computational drug discovery on human immunodeficiency virus with a customized long short-term memory variational autoencoder deep-learning architecture[J]. *CPT Pharmacometrics Syst Pharmacol*, 2024, **13**(2): 308-316.

[19] Bian YM, Wang JM, Jun JJ, *et al*. Deep convolutional generative adversarial network (dcGAN) models for screening and design of small molecules targeting cannabinoid receptors[J]. *Mol Pharm*, 2019, **16**(11): 4451-4460.

[20] Bickerton GR, Paolini GV, Besnard J, *et al*. Quantifying the chemical beauty of drugs[J]. *Nat Chem*, 2012, **4**(2): 90-98.

[21] Weng GQ, Zhao HF, Nie D, *et al*. RediscMol: benchmarking molecular generation models in biological properties[J]. *J Med Chem*, 2024, **67**(2): 1533-1543.

[22] Handa K, Thomas MC, Kageyama M, *et al*. On the difficulty of validating molecular generative models realistically: a case study on public and proprietary data[J]. *J Cheminform*, 2023, **15**(1): 112.

[23] Ertl P, Schuffenhauer A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions[J]. *J Cheminform*, 2009, **1**(1): 8.

[24] Thakkar A, Chadimová V, Bjerrum EJ, *et al*. Retrosynthetic accessibility score (RAscore) - rapid machine learned synthesizability classification from AI driven retrosynthetic planning[J]. *Chem Sci*, 2021, **12**(9): 3339-3349.

[25] Wang SH, Wang L, Li FL, *et al*. DeepSA: a deep-learning driven predictor of compound synthesis accessibility[J]. *J Cheminform*, 2023, **15**(1): 103.

[26] Krzyzanowski A, Pahl A, Grigalunas M, *et al*. Spacial Score─A comprehensive topological indicator for small-molecule complexity[J]. *J Med Chem*, 2023, **66**(18): 12739-12750.

[27] Preuer K, Renz P, Unterthiner T, *et al*. Fréchet ChemNet distance: a metric for generative models for molecules in drug discovery[J]. *J Chem Inf Model*, 2018, **58**(9): 1736-1741.

[28] Moret M, Grisoni F, Katzberger P, *et al*. Perplexity-based molecule ranking and bias estimation of chemical language models[J]. *J Chem Inf Model*, 2022, **62**(5): 1199-1206.

[29] Guo J, Fialková V, Arango JD, *et al*. Improving *de novo* molecular design with curriculum learning[J]. *Nat Mach Intell*, 2022, **4**: 555-563.

[30] Zhang O, Zhang JT, Jin JY, *et al*. ResGen is a pocket-aware 3D molecular generation model based on parallel multiscale modelling[J]. *Nat Mach Intell*, 2023, **5**: 1020-1030.

[31] Zhang O, Wang TY, Weng GQ, *et al*. Learning on topological surface and geometric structure for 3D molecular generation[J]. *Nat Comput Sci*, 2023, **3**(10): 849-859.

[32] Mokaya M, Imrie F, van Hoorn WP, *et al*. Testing the limits of SMILES-based *de novo* molecular generation with curriculum and deep reinforcement learning[J]. *Nat Mach Intell*, 2023, **5**: 386-394.

[33] Moret M, Pachon Angona I, Cotos L, *et al*. Leveraging molecular structure and bioactivity with chemical language models for *de novo* drug design[J]. *Nat Commun*, 2023, **14**(1): 114.

[34] Qian H, Huang WJ, Tu SK, *et al*. KGDiff: towards explainable target-aware molecule generation with knowledge guidance[J]. *Brief Bioinform*, 2023, **25**(1): bbad435.

[35] Xu MY, Chen HM. Tree-invent: a novel multipurpose molecular generative model constrained with a topological tree[J]. *J Chem Inf Model*, 2023, **63**(22): 7067-7082.

[36] Lim J, Hwang SY, Moon S, *et al*. Scaffold-based molecular design with a graph generative model[J]. *Chem Sci*, 2019, **11**(4): 1153-1164.

[37] Hu LZ, Yang YY, Zheng SJ, *et al*. Kinase inhibitor scaffold hopping with deep learning approaches[J]. *J Chem Inf Model*, 2021, **61**(10): 4900-4912.

[38] Zheng SJ, Lei ZR, Ai HT, *et al*. Deep scaffold hopping with multimodal transformer neural networks[J]. *J Cheminform*, 2021, **13**(1): 87.

[39] Fialková V, Zhao JX, Papadopoulos K, *et al*. LibINVENT: reaction-based generative scaffold decoration for *in silico* library design[J]. *J Chem Inf Model*, 2022, **62**(9): 2046-2063.

[40] Loeffler HH, He JZ, Tibo A, *et al*. Reinvent 4: modern AI-driven generative molecule design[J]. *J Cheminform*, 2024, **16**(1): 20.

[41] Liao ZR, Xie L, Mamitsuka H, *et al*. Sc_2Mol: a scaffold-based two-step molecule generator with variational autoencoder and transformer[J]. *Bioinformatics*, 2023, **39**(1): btac814.

[42] Liu XH, Ye K, van Vlijmen HWT, *et al*. DrugEx v3: scaffold-constrained drug design with graph transformer-based reinforcement learning[J]. *J Cheminform*, 2023, **15**(1): 24.

[43] Xu C, Liu RD, Huang SH, *et al*. 3D-SMGE: a pipeline for scaffold-based molecular generation and evaluation[J]. *Brief Bioinform*, 2023, **24**(6): bbad327.

[44] Hu C, Li S, Yang CX, *et al*. ScaffoldGVAE: scaffold generation and hopping of drug molecules via a variational autoencoder based on multi-view graph neural networks[J]. *J Cheminform*, 2023, **15**(1): 91.

[45] Xie JJ, Chen S, Lei JP, *et al*. DiffDec: structure-aware scaffold decoration with an end-to-end diffusion model[J]. *J Chem Inf Model*, 2024, **64**(7): 2554-2564.

[46] Bilsland AE, McAulay K, West R, *et al*. Automated generation of novel fragments using screening data, a dual SMILES au-

toencoder, transfer learning and syntax correction[J]. *J Chem Inf Model*, 2021, **61**(6): 2547-2559.

[47] Hadfield TE, Imrie F, Merritt A, *et al*. Incorporating target-specific pharmacophoric information into deep generative models for fragment elaboration[J]. *J Chem Inf Model*, 2022, **62**(10): 2280-2292.

[48] Du HY, Jiang DJ, Zhang O, *et al*. A flexible data-free framework for structure-based *de novo* drug design with reinforcement learning[J]. *Chem Sci*, 2023, **14**(43): 12166-12181.

[49] Powers AS, Yu HH, Suriana P, *et al*. Geometric deep learning for structure-based ligand design[J]. *ACS Cent Sci*, 2023, **9**(12): 2257-2267.

[50] Sauer S, Matter H, Hessler G, *et al*. Integrating reaction schemes, reagent databases, and virtual libraries into fragment-based design by reinforcement learning[J]. *J Chem Inf Model*, 2023, **63**(18): 5709-5726.

[51] Wang JK, Zeng YD, Sun HY, *et al*. Molecular generation with reduced labeling through constraint architecture[J]. *J Chem Inf Model*, 2023, **63**(11): 3319-3327.

[52] Eguida M, Schmitt-Valencia C, Hibert M, *et al*. Target-focused library design by pocket-applied computer vision and fragment deep generative linking[J]. *J Med Chem*, 2022, **65**(20): 13771-13783.

[53] Buehler Y, Reymond JL. Expanding bioactive fragment space with the generated database GDB-13s[J]. *J Chem Inf Model*, 2023, **63**(20): 6239-6248.

[54] Diao YY, Hu F, Shen ZH, *et al*. MacFrag: segmenting large-scale molecules to obtain diverse fragments with high qualities[J]. *Bioinformatics*, 2023, **39**(1): btad012.

[55] Imrie F, Bradley AR, van der Schaar M, *et al*. Deep generative models for 3D linker design[J]. *J Chem Inf Model*, 2020, **60**(4): 1983-1995.

[56] Yang YY, Zheng SJ, Su SM, *et al*. SyntaLinker: automatic fragment linking with deep conditional transformer neural networks[J]. *Chem Sci*, 2020, **11**(31): 8312-8322.

[57] Tan YH, Dai LX, Huang WF, *et al*. DRlinker: deep reinforcement learning for optimization in fragment linking design[J]. *J Chem Inf Model*, 2022, **62**(23): 5907-5917.

[58] Li BQ, Ran T, Chen HM. 3D based generative PROTAC linker design with reinforcement learning[J]. *Brief Bioinform*, 2023, **24**(5): bbad323.

[59] Kao CT, Lin CT, Chou CL, *et al*. Fragment linker prediction using the deep encoder-decoder network for PROTACs drug design[J]. *J Chem Inf Model*, 2023, **63**(10): 2918-2927.

[60] Zhang H, Huang JC, Xie JJ, *et al*. GRELinker: a graph-based generative model for molecular linker design with reinforcement and curriculum learning[J]. *J Chem Inf Model*, 2024, **64**(3): 666-676.

[61] Imrie F, Hadfield TE, Bradley AR, *et al*. Deep generative design with 3D pharmacophoric constraints[J]. *Chem Sci*, 2021, **12**(43): 14577-14589.

[62] Zhu HM, Zhou RY, Cao DS, *et al*. A pharmacophore-guided deep learning approach for bioactive molecular generation[J]. *Nat Commun*, 2023, **14**(1): 6234.

[63] Dahal S, Yurkovich JT, Xu H, *et al*. Synthesizing systems biology knowledge from omics using genome-scale models[J]. *Proteomics*, 2020, **20**(17/18): e1900282.

[64] Born J, Manica M, Oskooei A, *et al*. PaccMann^{RL}: *De novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning*[J]. *iScience*, 2021, **24**(4): 102269.

[65] Pravalphruekul N, Piriyajitakonkij M, Phunchongharn P, *et al*. De novo design of molecules with multiaction potential from differential gene expression using variational autoencoder[J]. *J Chem Inf Model*, 2023, **63**(13): 3999-4011.

[66] Das D, Chakrabarty B, Srinivasan R, *et al*. Gex2SGen: designing drug-like molecules from desired gene expression signatures[J]. *J Chem Inf Model*, 2023, **63**(7): 1882-1893.

[67] Dolfus U, Briem H, Rarey M. Synthesis-aware generation of structural analogues[J]. *J Chem Inf Model*, 2022, **62**(15): 3565-3576.

[68] Qiang B, Zhou YR, Ding YH, *et al*. Bridging the gap between chemical reaction pretraining and conditional molecule generation with a unified model[J]. *Nat Mach Intell*, 2023, **5**: 1476-1485.

[69] Khemchandani Y, O'Hagan S, Samanta S, *et al*. DeepGraphMolGen, a multi-objective, computational strategy for generating molecules with desirable properties: a graph convolution and reinforcement learning approach[J]. *J Cheminform*, 2020, **12**(1): 53.

[70] Lamanna G, Delre P, Marcou G, *et al*. GENERA: a combined genetic/deep-learning algorithm for multiobjective target-oriented *de novo* design[J]. *J Chem Inf Model*, 2023, **63**(16): 5107-5119.

[71] Jayatunga MKP, Xie W, Ruder L, *et al*. AI in small-molecule drug discovery: a coming wave[J]? *Nat Rev Drug Discov*, 2022, **21**(3): 175-176.

[72] Lv Q, Zhou FL, Liu XH, *et al*. Artificial intelligence in small molecule drug discovery from 2018 to 2023: does it really work[J]? *Bioorg Chem*, 2023, **141**: 106894.

[73] Ren F, Aliper A, Chen J, *et al*. A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models[J]. *Nat Biotechnol*, 2024: 1-13

[74] Li YG, Liu YT, Wu JP, *et al*. Discovery of potent, selective, and orally bioavailable small-molecule inhibitors of CDK8 for the treatment of cancer[J]. *J Med Chem*, 2023, **66**(8): 5439-5452.

[75] Swanson K, Liu G, Catacutan DB, *et al*. Generative AI for designing and validating easily synthesizable and structurally novel antibiotics[J]. *Nat Mach Intell*, 2024, **6**: 338-353.

［专家介绍］车金鑫, 博士, 特聘研究员, 主要从事合理药物设计和新药发现研究工作, 聚焦药物重定位方向, 围绕混合药物设计和多技术交叉赋能, 成功发现了针对不同靶标和适应证的具有进一步开发价值的候选分子。相关成果以第一/通信作者发表于 *Adv Sci*、*J Med Chem*、*Brief Bioinform* 等国际学术期刊, 授权专利 2 项, 主持国家自然科学基金、浙江省自然科学基金等项目。此外, 担任浙江省药学会药物化学与抗生素专委会青年委员(兼秘书)、浙江省抗癌协会抗癌药物专委会青年委员、《中国药科大学学报》和《药学进展》青年编委以及多个学术期刊客座编辑等学术兼职。获浙江省药学会科学技术一等奖 1 项。

［专家介绍］董晓武, 博士, 教授, 博士生导师, 浙江大学药学系副主任、创新药物研究中心副主任, 浙江省杰出青年基金获得者。主要从事药物化学和化学生物学的研究, 致力于合理药物设计和新药发现, 聚焦靶向蛋白降解、人工智能药物设计、药物重定位等技术及其在新型先导分子发现中的应用, 主导研发的多个候选药物获得化药 1 类新药临床试验批件, 并进入Ⅰ、Ⅱ期临床研究, 部分品种已实现成果转化。迄今, 在 *Nucleic Acids Res*、*Adv Sci*、*J Med Chem*、*Brief Bioinformatics* 等国际著名刊物上发表论文 100 余篇。授权国家发明专利 20 项, 其中国际授权专利 4 项。主持了国家自然科学基金面上项目 3 项、“十三五”国家新药创制重大专项、浙江省“领雁”研发攻关计划等项目, 先后获浙江省科技进步奖二等奖 2 项。

· 本刊讯 ·

《中国药科大学学报》获评“RCCSE中国核心学术期刊”

2024 年 4 月, RCCSE《中国学术期刊评价研究报告(第七版)》正式发布,《中国药科大学学报》被评为“RCCSE 中国核心学术期刊(A-)”。

RCCSE 中国学术期刊评价体系是目前国内公认的七大学术期刊评价体系之一, 由中国科教评价研究院、武汉大学中国科学评价研究中心、武汉大学图书馆、中国科教评价网联合研制, 每隔 2 年推出新的一版, 2023 年为第七版。经研究人员对相关文献的检索、统计和分析, 以及学科专家评审,《中国药科大学学报》入编 RCCSE《中国学术期刊评价研究报告(第七版)》药学类核心学术期刊, 位列 61 种药学类期刊第 14 名, 排名较 2020 年版上升 8 位。

编辑部全体成员对各位编委、审稿专家、作者和读者一直以来对《中国药科大学学报》的支持和帮助表示衷心感谢。编辑部将持续提升期刊的学术质量和影响力, 努力搭建广大药学科研工作者的高质量学术交流平台。

(本刊编辑部)