



基于深度学习的全新药物设计研究进展

王明阳, 李丹, 侯廷军*, 康玉*

浙江大学药学院, 杭州 310058

*通讯作者, E-mail: tingjunhou@zju.edu.cn; yukang@zju.edu.cn

收稿日期: 2022-06-16; 接受日期: 2022-07-18; 网络版发表日期: 2022-10-26

浙江省自然科学基金重大项目(编号: LD22H300001)和国家自然科学基金(编号: 21575128, 81773632)资助项目

摘要 先导化合物的设计和发现是新药研发中最具挑战性和创造性的阶段, 其过程需考虑候选分子的结构新颖性、生物活性、靶标选择性、可合成性、成药性和安全性等多种属性的优化。虽然计算机辅助药物设计方法的发展和应用大大节省了先导化合物发现阶段的时间和经济成本, 但仍未能扭转新药研发成功率低的现状。近年来, 随着深度学习技术的不断发展, 基于深度学习的全新药物设计方法为先导化合物的发现带来新的契机, 前景巨大。这些全新药物设计模型使用的深度学习框架包括编码-解码器、循环神经网络、生成对抗网络、强化学习等。本文综述了这些深度学习框架的基本原理、模型输入分子表征以及效果评测指标, 并对其在全新药物设计领域的应用前景进行了展望。

关键词 全新药物设计, 计算机辅助药物设计, 深度学习, 机器学习, 分子生成

1 引言

新药研发是一项周期长、投入高、风险大的复杂系统工程。据美国药品研究与制造商协会报告显示^[1], 自21世纪以来, 临床试验复杂性的提高以及药品市场监管模式的变化, 新药研发的平均成本急剧增加, 2010~2015年间达到了平均26亿美元, 也对药物的多属性优化及其对作用靶点机制的明确性提出了更高的要求。

先导化合物的发现与优化在新药研发过程中至关重要, 高质量的先导化合物能够大大缩短药物探索的时间, 提高成药的可能性。在先导化合物的设计过程中, 要充分考虑候选分子的结构新颖性、生物活性、靶标选择性、化学可合成性、成药性和安全性等, 这

些性质直接影响药物开发的成败, 因此先导化合物的发现一直是创新药物研发的主要瓶颈。传统的先导化合物发现的方法主要基于对天然产物的结构修饰; 近20年来, 随着计算机硬件、软件和算法的飞速发展, 高通量虚拟筛选和全新药物设计等计算机辅助药物设计技术开始取代传统方法, 并大大减少了先导化合物发现的时间和经济成本^[2-4]。虚拟筛选技术常用于从海量化合物数据库中搜索潜在的候选分子; 因其主要针对已注册过的分子(目前大部分化合物数据库为 $10^7 \sim 10^9$ 量级^[5,6]), 搜索的化学空间远小于现推测的类药化学空间(10^{60} 量级^[7]), 因此虚拟筛选技术并不能完全发掘全新的化合物结构^[8-10]。

全新药物设计(*de novo drug design*, 也被称为药物从头设计)则使用分子生成方法对化合物结构不断优

引用格式: Wang M, Li D, Hou T, Kang Y. Deep learning approaches for *de novo* drug design: an overview. *Sci Sin Chim*, 2023, 53: 95–106, doi: [10.1360/SSC-2022-0135](https://doi.org/10.1360/SSC-2022-0135)

化,使新生成化合物的预设属性获得最优解.它不依赖已有的化学数据库,可实现对类药空间更加全面的探索和发掘.传统的全新药物设计方法通常与片段生长法或遗传算法相结合^[11~15],但生成的分子往往难以同时满足新颖性与多种理想属性(如生物活性、化学可合成性、成药性等)的要求^[16,17].深度学习(deep learning, DL)的引入为全新药物设计注入了新的活力.作为近期发展最快的人工智能技术,DL能够更高效地处理数据,对化合物属性深度特征的提取能力更强^[18],已在不同的药物研发任务中取得了显著的成果^[19~22].目前已有越来越多的基于DL框架的全新药物设计模型被开发出来,旨在产生符合多种理想性质的新颖骨架分子.

基于DL的全新药物设计方法按照算法基本框架主要分为四类,包括编码-解码器(encoder-decoder, Enc-Dec)^[23,24]、循环神经网络(recurrent neural network, RNN)^[25~27]、生成对抗网络(generative adversarial network, GAN)^[28]、强化学习(reinforcement learning, RL)^[29,30].尽管近五年内已经开发出近百种基于DL的全新药物设计模型,但是这些模型均是在上述四种DL框架下进行两个方面的改造:分子表征以及多目标优化策略.基于DL框架的全新药物设计模型可以使用不同的分子表征作为输入,如SMILES (simplified molecule input line entry specification)^[23,31]、分子指纹(molecular fingerprint, MFP)^[25]、二维(two dimension, 2D)^[32,33]或三维(three dimension, 3D)^[34,35]的分子图(molecular graph, MG)等.多目标优化则是使用不同策略来同时优化生成分子的多种属性.例如,使用不同的DL框架约束控制生成的步骤,使得分子对某一靶标保持活性的同时也具有很高的可合成性^[30,36].目前已经有研究者成功使用DL框架设计并合成出有活性的苗头化合物^[31,37,38],且耗时远远小于传统的全新药物设计方法,潜力巨大.本文将针对不同的DL算法在全新药物设计方法开发中的应用进行回顾和总结,以期促进更多研究者的参与.

2 分子表征

2.1 化合物数据库

基于DL框架的全新药物设计模型的构建需要包含大量分子数据的训练集.目前,DL模型使用的训练

集分子可以从公开的化学数据库中搜集,常用的化学数据库总结于表1中.

2.2 SMILES字符串

SMILES^[43]字符串是目前已报道的DL框架分子生成模型最常用的分子表征.这种化学字符串由于具有易读性和高逻辑性等特点,是一种计算化学领域的分子通用表征形式.SMILES字符串可被看作文本,需要转换为独热(one-hot)编码之后才能被DL框架处理,转换的方式如图1a示.生成SMILES的过程实际上模拟了文本的生成,也就是将含有不同字母的SMILES字符串作为文本的方式进行处理;模型可以通过最大化似然或最小化交叉熵的方法学习离散文本数据输入的形式^[23,26],最终生成语法上有效的SMILES字符串.但是,SMILES具有单向唯一性:虽然每个SMILES字符串只能转换成唯一的分子,但是每个分子可以用多个不同的SMILES字符串表示.生成模型不能理解这种单向唯一性,这导致一个分子的SMILES分散在化学空间的不同位置^[44],使得训练之后模型对应的化学空间内存在高度重复的分子,造成生成分子基本骨架的高度重复性.为解决上述问题,可采用分子SMILES的枚举法^[45,46],也就是在训练之前枚举一个分子的多重不同的SMILES表征作为训练数据,这样就能避免生成的分子骨架和片段的重复,使得结构更加多样.

2.3 分子指纹图谱MFP

与SMILES不同,MFP是将分子转换为固定长度的比特数据,其可直接用于DL框架的输入,如MACCS

表 1 基于DL框架的全新药物设计模型中常用的代表性化学数据库

Table 1 The representative chemical libraries used in DL-based *de novo* drug design

名称	描述	数量
GDB-13 ^[5]	生成库	$>9.7 \times 10^8$
ZINC15 ^[6]	商用库	$>7.5 \times 10^8$
GDB-17 ^[39]	生成库	$\sim 5.0 \times 10^7$
ChEMBL ^[40]	生物活性分子数据库	2×10^6
QM9 ^[41]	GDB-17数据库的子集	1.3×10^5
DrugBank ^[42]	药物数据库	1×10^4

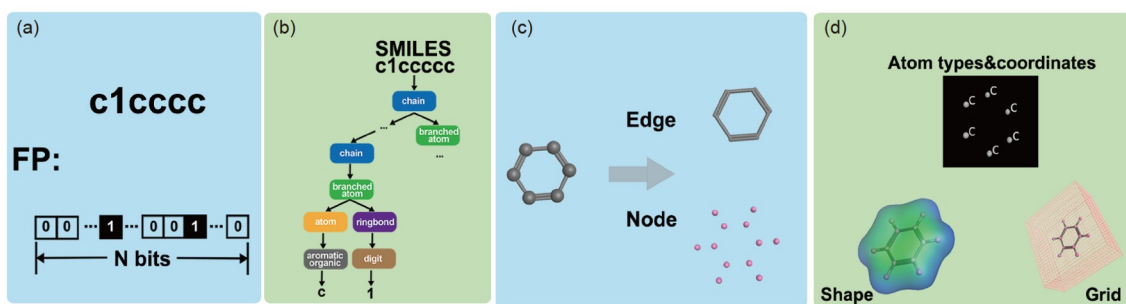


图 1 基于DL框架的全新药设计模型中常用的分子表征(使用苯环作为例子). (a) SMILES与FP; (b) CFG; (c) 2D-MG; (d) 3D-MG (网络版彩图)

Figure 1 Representative molecular representations in DL-based *de novo* drug design (benzene as an example). (a) SMILES and FP; (b) CFG; (c) 2D-MG; (d) 3D-MG (color online).

(molecular ACCess system)和ECFP (extended-connectivity fingerprint)指纹已被Enc-Dec框架成功用于生成模型的构建^[28,47]. 使用这两种MFP作为分子表征时需要注意一些问题: 首先, 要将模型生成的MACCS指纹转化为有效分子, 即需要在数据库(通常采用PubChem数据库)中搜索之后再转换, 当数据库内不存在相似的分子时, 则寻找最为相似的分子^[48], 这就限制了生成分子只能存在于已有的数据库中, 从而偏离了全新药设计的初衷; 其次, 由于ECFP是将分子结构特征与化学属性结合起来, 所以生成模型的隐藏层需要与活性预测器联用时才能进行全新药设计的任务^[28,49–51].

2.4 SMILES树

SMILES树也称为上下文无关语法(context-free grammar, CFG)^[52,53], 是对SMILES字符串的一种扩展, 使用CFG作为输入的生成模型首先需要定义分子生成的语法规则. 图1b显示了苯环“c1ccccc1”的构建规则: 如“chain”表示连接化学键, “branched”表示支链原子等. 这些分子构建的规则将被转换为解析树的形式, 随后在解析树分支按照从左到右进行预序遍历, 分解为一系列的构建规则并编码成one-hot形式作为深度学习模型的输入^[52,53]. 如2.2小节所述, 使用SMILES作为生成模型的分子表征只能够保证生成字符串的语法有效, 并不能保证字符串的内在含义. 有研究报道, 基于SMILES的生成模型经常会生成一些字符串语法有效但性质不稳定或不易合成的分子, 如烯醇化合物、重氮盐或四环等. 而基于CFG构建过程的语法实质上是在生成分子的过程中加入了约束掩码, 这样就能够

控制模型生成较优的分子^[52,54].

2.5 2D-MG

2D-MG是将分子的结构信息转换为平面图片的形式作为DL框架的输入, 所以需要使用不同图元素来表示一个分子^[32,33]. 经典的2D-MG表征方式如图1c所示, 即将分子的所有原子和化学键分别转换为节点(node)和边(edge)的分子图形式. 在这个分子图中, 所有的节点和边都带有原子类型和化学键类型的标签. 在训练模型之前需要选择一种方式对分子图进行处理: (1) 制定分子图的顺序生成策略: 添加/删除节点概率策略、添加/删除边概率策略、连接/断开节点概率策略^[32]; 在训练过程中, 这些概率决策模块将通过已知分子图进行参数化, 并考虑了节点排列的固定排序和均匀随机排序; (2) 将节点和边转换为全连接矩阵后, 使用one-hot编码作为DL框架的输入^[55]. 使用2D-MG表征的分子相比于其他分子表征更加模块化和概念化. 为了能够生成有效的分子, 分子的生成需要以多原子或片段为基本单位, 相比于其他以单原子为基本单位的生成模型, 其多样性较低^[32,33], 更加适合于先导化合物的骨架跃迁, 结构优化与改造.

2.6 3D-MG

使用上述分子表征的生成模型几乎都是基于配体的方法, 即没有在生成过程中考虑到配体分子与蛋白的相互作用. 由于这些生成模型都以SMILES或者2D-MG作为训练集, 而这些二维的数据难以与受体蛋白的三维构象相结合, 因此使得生成的分子仅仅与训练集中配体的属性较为相似. 3D-MG则能够在模型构建

分子的过程中关注分子的立体几何构象, 也能够生成的过程中融入受体的信息, 使得分子生成的过程考虑到与受体的结合构象, 进而能够执行靶标特异性的药物设计任务。目前在生成模型中使用的3D-MG主要包括: 原子类型与坐标标注、形状曲面(shape-mesh)、三维网格(3D-grid)等(图1d)。

原子类型与坐标标注是最常使用的3D-MG, 以最直观的方式将一个分子表示成原子类型和坐标的数字集合; 生成模型需要根据已生成的前一个原子, 在原子库内选出概率分布最高的下一个原子的类型并确定坐标^[34,56-58]。这种表征其实是上述2D-MG的一种特殊形式, 相当于在2D-MG中加入了原子坐标。Shape-mesh是按照形状将分子转换为三维图片, 然后以图片形式来处理分子^[24,59], 由于图片可以被设置成多通道, 因此可使用不同的通道来代表不同的分子性质, 如疏水基团、氢键供体、氢键受体等^[24]。3D-grid则是将分子放入到一个三维的网格中作为立体图片, 网格带有固定的体素分辨率; 与shape-mesh类似, 在一个体素内的原子使用不同的属性作为通道^[60]。由于3D-MG涉及分子的立体构象, 使用时需注意分子的旋转不变性^[57]。

3 全新药物设计模型中使用的DL框架

我们将目前基于DL框架的模型总结在表2中, 按照基本框架将其分成四类: Enc-Dec、RNN、GAN以及

及RL。在本节中, 我们将对这些模型框架的原理进行简要介绍, 并选取几个代表性的模型作为应用实例的补充。

3.1 Enc-Dec框架

Enc-Dec是由编码器和解码器组成的无监督DL框架, 编码器将输入的分子表征映射到具有隐藏层中, 解码器则将隐藏层的向量映射回原来的输入表征。

在生成模型中使用最多的Enc-Dec结构是变分自编码器(variational autoencoders, VAE)。如图2a所示, 在分子生成任务中, 对VAE模型的训练主要是为了减小生成的分子表征与原始分子表征之间的误差^[23], 如式(1)所示。

$$\max LB = \int q(z) \log p(x|z) dz - KL(q(z|x) \| p(z)) \quad (1)$$

训练的过程主要根据最大化下界(lower bound, LB)结合编码器 $q(z|x)$ 和解码器 $p(x|z)$ 的公式(KL表示KL散度^[61], 在这里 $p(z)$ 为高斯分布)。ChemVAE是第一个报道的用于全新药物设计的VAE框架, 在经过训练后的多层感知器对分子的属性进行预测后, 隐藏层 z 中的分子分布将会出现基于属性值(类药性以及可合成性)的梯度分布, 能够控制采样之后的分子属性。

随后, 许多研究者对VAE模型进行了改进, 如GVAE^[52]、JTVAE^[53]、SDVAE^[54]和CVAE^[62,63], 这些模型通常是使用了不同的分子表征, 并且添加了一些

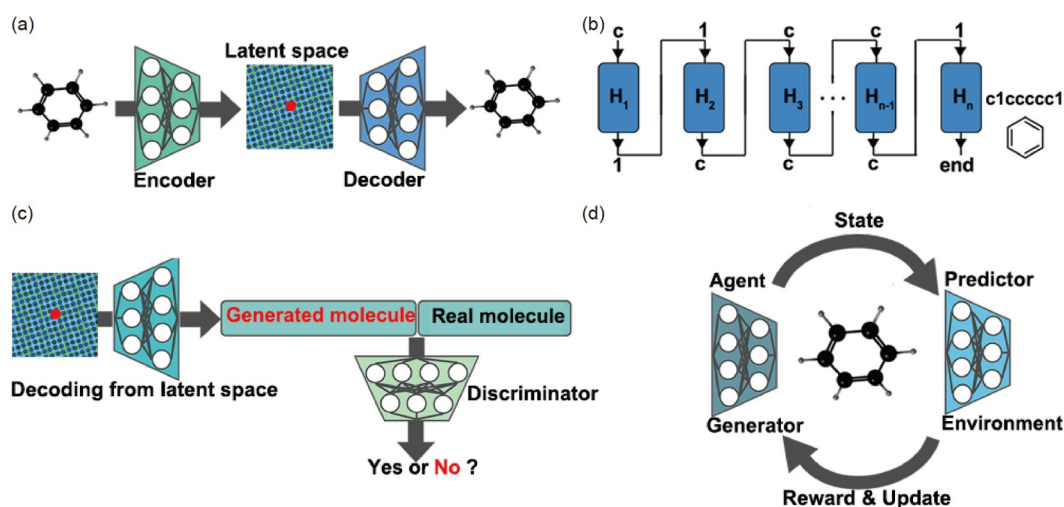


图2 全新药物设计模型中使用的DL框架。(a) Enc-Dec; (b) RNN; (c) GAN; (d) RL (网络版彩图)

Figure 2 The DL frameworks in *de novo* drug design. (a) Enc-Dec; (b) RNN; (c) GAN; (d) RL (color online).

表 2 基于DL框架的代表性全新药物设计模型^{a)}Table 2 Representative DL-based *de novo* drug design models^{a)}

模型(框架)	分子表征+特点	标注	模型(框架)	分子表征+特点	标注
ChemVAE ^[23] (Enc-Dec, CNN-GRU)	SMILES, 在隐藏层应用多层感知机	首次将深度学习框架应用于全新药物设计	DeepLigBuilder ^[58] (MPNN+RL)	3D-MG, 使用打分函数作为RL的奖励函数	用于设计SARS-CoV-2抑制剂
CVAE ^[63] (Enc-Dec, RNN-RNN)	附带理化性质的SMILES, Conditional VAE	用户可以定制生成分子	cG-SchNET ^[34] (CNN)	3D-MG, 使用CNN作为框架预测生成的原子类型和坐标	—
GrammarVAE ^[52] (Enc-Dec, CNN-GRU)	CFG	应用CFG	ChemTS ^[72] (RNN+RL)	SMILES, RNN+MCTS	低计算资源消耗
CGVAE ^[64] (Enc-Dec, GNN-GRU)	2D-MG, 在VAE框架中使用了GNN框架	2D-MG	REINVENT ^[69] (RNN+RL)	SMILES, REINFORCE	—
JTVAE ^[77] (Enc-Dec)	同时使用了2D-MG+SMILES树作为输入	—	REINVENT2.0 ^[29] (RNN+RL)	SMILES, 在RL的奖励系统中引入了多目标优化打分	对REINVENT的改进
JAEGER ^[77] (Enc-Dec)	对JT-VAE进行了改良	成功设计并合成出有活性的PIK4抑制剂	DrugEX ^[73] (RNN+RL)	SMILES, 在RNN中引入了探索策略	用于设计A2AR抑制剂
NeVAE ^[78] (Enc-Dec)	2D-MG, 在解码器中加入了基于梯度的生成算法	设计了实用与2D-MG的VAE框架	ReLeaSE ^[70] (RNN+RL)	SMILES, REINFORCE+stack-RNN	用于设计JAK2抑制剂
Shape-VAE ^[24] (Enc-Dec, CNN-LSTM)	3D-MG	以3D形状的形式表示分子; 用于设计A2AR和THRB KIT抑制剂	DeepFMPO ^[71] (RNN+RL)	SMILES, 使用了TD策略	以片段为基本单位构建分子
Heteroencoder ^[44] (Enc-Dec, LSTM-LSTM)	SMILES, Hetero-encoder	在一个生成模型中使用了多种分子表征	ECAAE ^[38] (AAE)	SMILES, 在隐藏层中使用了缠绕-分解策略	成功设计并合成出有活性的JAK家族选择性抑制剂
GTM-RNN ^[79] (Enc-Dec, LSTM-LSTM)	SMILES, 使用GTM作为采样工具	生成的分子在药效团匹配度和分子对接打评测结果较好	DruGAN ^[28] (AAE)	MFP	产生具有预先定义的抗癌属性的分子
Gentri ^[31] (Enc-Dec, GRU+CNN)+RL	2D-MG和SMILES加入SOM作为RL框架的打分	成功设计并合成出有活性的DDR1抑制剂	LatentGAN ^[76] (AAE)	SMILES, Hetero-encoder	用于设计EGFR、HTR1A以及S1PR1抑制剂
GraphINVENT ^[80] (GNN)	2D-MG, GNN+全局读取模块	使用GNN迭代生成2D-MG	LiGANN ^[24] (CNN+GAN)	3D-MG, CNN+GAN	根据蛋白口袋形状生成分子; 用于设计环化酶、A2aR以及IRAK-4抑制剂
MolRNN ^[32] (RNN)	2D-MG	与几种多目标优化的生成模型进行了对比	ORGAN ^[74] (GAN+RL)	SMILES, GAN+RL	生成的分子具有多种理想的理化性质
GEN ^[81] (RNN)	SMILES, 双向LSTM	生成的分子具有更高的有效性和新颖性	ORGANIC ^[75] (GAN+RL)	SMILES, GAN+RL	对ORGAN的改进
cRNN ^[25] (RNN)	附带理化性质的MFP, cRNN框架	目标导向的分子生成	ATNC ^[37] (GAN+RL)	SMILES, GAN+RL	ORGAN的衍生框架; 成功设计并合成了多个靶点的活性抑制剂
CLM ^[27] (RNN)	MFP	低数据训练			

a) A2AR: A2a腺苷酸受体 (A2a adenosine receptor); DDR1: 盘状结构域受体1 (discoidin domain receptor 1); EGFR: 表皮生长因子受体 (epidermal growth factor receptor); GNN, 图神经网络(graph neural network); HTR1A: 5-羟色胺受体1A (5-hydroxytryptamine (serotonin) receptor 1A); IRAK-4: 白细胞介素-1受体相关激酶4 (interleukin-1 receptor associated kinase-4); JAK: Janus激酶(Janus kinases); KIT: 酪氨酸激酶(tyrosine kinase); MPNN, 消息传递神经网络(message passing neural networks); PI4K: 磷脂酰肌醇4-激酶(phosphatidylinositol 4-kinases); S1PR1: 鞘氨醇1磷酸酯受体1 (sphingosine-1-phosphate receptor 1); SOM, 自组织映射网络 (self-organizing maps); THRB: 甲状腺激素受体beta (thyroid hormone receptor beta).

实用的功能, 如SDVAE^[54]和GVAE^[52]使用了CFG作为分子输入, 能够使隐藏层内邻近分子的结构相似性更加接近, 从而随机采样能够产生更多合理的分子. 在同样的数据集训练下, SDVAE、GVAE和JTVAE生成的化合物明显比ChemVAE生成的化合物在结构上更加合理且新颖. CGVAE^[64]、ShapeVAE^[59]、JTVAE^[53]以及GraphVAE^[65]则以MG的方式表征分子. 这些表征方式的概率分布相比于SMILES更适用于VAE模型的连续潜伏层, 与CFG一样也会使得隐藏层内分子结构的分布更为平滑, 从而提高采样的合理性和新颖性. CVAE^[63]和SSVAE^[62]则在VAE框架中添加了药物设计任务中的实用功能, 用户可通过预设特定的骨架或特定属性值来定制生成的分子. 这些模型虽然在一定程度上提高了生成分子的有效性和新颖性, 但其基本框架和原理并没有较大的改变.

除VAE外, 数据转化器(transformer)这种Enc-Dec框架也被用于全新药物设计的模型中^[30,66]. Transformer生成模型的框架与经典的基于CNN或者RNN的Enc-Dec生成模型完全不同: 首先, 在嵌入输入分子表征的词汇时, 使用的是注意力(attention)机制, 嵌入的方式与经典的Enc-Dec相同; 其次, transformer通过训练之后以顺序以及自回归的方式输出SMILES字符^[67], 根据之前生成的部分序列确定下一个要生成的SMILES字符, 并采用掩码多头自注意层(masked multi-head self-attention layer)来防止序列未解码部分的分子表征的词汇信息泄露. 多头自注意层是transformer的核心, 该模型由多个可缩放点积注意力函数(scaled dot-product attention functions)组成, 便于模型按顺序获取关键信息, 注意机制的计算过程如式(2)所示:

$$\text{attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (2)$$

其中 $\mathbf{Q}$ 、 $\mathbf{K}$ 和 $\mathbf{V}$ 分别表示查询矩阵、键矩阵和值矩阵, d_k 表示 K 的维数.

3.2 RNN框架

RNN模型以类似于文本的方式来生成分子, 将分子的原子或片段等基本单元作为基本词汇, 按照时间顺序来生成分子. RNN模型生成下一词汇的概率取决于前面几层的结果, 其训练过程如图2b以及式(3)所示:

$$\max P(x) = P(s_1) \prod_{t=2}^T P(s_t | s_{t-1}, \dots, s_1) P \quad (3)$$

也就是说, RNN分子生成模型试图最大化分子字符串($y_1, y_2, \dots, y_t$)的有效组合的概率, 以最高概率选择生成下一词汇. 为了解决梯度消失问题, 在RNN层中会使用长短期记忆神经网络(long short-term memory, LSTM)或其简化形式门控神经网络(gate recurrent unit, GRU). 在RNN模型中通常会使用迁移学习(transfer learning, TL), 其传统意义是指通过从相关任务中学习到的知识进行迁移来提高新任务的学习, 而在基于RNN的模型中, TL指使用活性化合物在特定药物设计任务中对生成模型进行微调再训练. 许多研究^[26,68]已经成功地使用TL微调RNN模型, 生成对靶点有活性的抑制剂分子. RNN相对于其他DL框架的优势在于生成的分子长度不受限制, 模型的搭建和训练也比较简单.

3.3 RL框架

RL是用来处理动态决策问题的人工智能框架, 在全新药物设计模型中, RL模型利用一个奖励函数(通常是基于机器学习的属性预测器)对生成器的一个分子构建的动作 s_T 策略进行评分, 评分后的结果 r 返回到生成器, 以指导下一轮的策略行动. RL的训练过程如图2d以及式(4)所示:

$$\max R = \sum_{s_T \in S} P(s_T) r(s_T) \quad (4)$$

其训练过程主要是利用预测器对生成器的每个状态 s_T 进行奖励, 使终端状态 s 的总得分 R 最大化.

REINFORCE^[69~71]以及蒙特卡洛树搜索(Monte Carlo tree search, MCTS)^[72,73]是在RL模型中经常使用的策略函数. 例如, REINVENT^[69]生成模型中就引入了改进的REINFORCE策略函数, 既能保证生成分子的有效性, 又考虑了生成分子的理化性质与训练集的一致性. 为了考察REINVENT的性能, 作者首先使用REINVENT设计了塞来昔布衍生物, 证明该模型具有使用单一活性化合物进行结构改造及骨架跃迁的能力. 随后他们又设计了多巴胺D2受体(D2 dopamine receptor, DRD2)拮抗剂, 证明该模型同时具有设计靶标特异性化合物的能力. DeepFMPO^[71]也使用了REINFORCE策略函数, 在数据处理阶段, 他们将DRD2和

DRD4的活性分子数据转换为片段库, 基于片段而非原子生成分子. 他们最终生成了理化性质与DRD2和DRD4的活性分子相近的分子集.

MCTS策略函数能够同时评估在分子构建过程中当前的部分生成片段以及后续的完整生成片段. ChemTs^[72]是结合了RNN和MCTS策略函数的生成模型. 通过比较ChemTs与RNN、ChemVAE和GrammarVAE, 发现ChemTs可以用较少的计算资源生成理化性质较优的分子. DrugEX^[73]则将MCTS与一个双向RNN结合, 将一个未经训练的RNN和一个经过训练的RNN集成到一个生成器中, 以保证生成分子的多样性和新颖性.

3.4 GAN框架

GAN模型由生成器 G 和鉴别器 D (也可以是属性预测器)组成, 如图2c和式(5)所示:

$$\min_G \max_D V(D, G) \\ = E_{x \sim P_{\text{data}}(x)} [\log D(x)] + E_{z \sim P_z(z)} [\log(1 - D(G(z)))] \quad (5)$$

其原理是生成器和鉴别器两方之间的零和博弈. 鉴别器试图将生成器生成的分子鉴别为假($G(D(x)) \rightarrow 0$, $D(G(z)) \rightarrow 0$), 而生成器则试图从噪声中生成分子“欺骗”鉴别器($D(G(z)) \rightarrow 1$). 在全新药物设计的模型中, GAN通常与其他DL框架联合使用. 例如, 在训练过程中结合RL, ORGAN^[74]和ORGANIC^[75]通过GAN鉴别器和RL奖励系统的结合对生成的分子进行评分; 或者与VAE联合成对抗自编码器(adversarial autoencoder, AAE)这种融合算法^[28,76]. AAE通过引入GAN鉴别器作用于VAE的潜在层, 使得针对特定目标采样的生成化合物成为可能. 与VAE相比, AAE更适合使用MFP这种分子表征作为框架的输入^[28].

4 基于DL框架的全新药物设计模型中使用的评测指标

为了对上述基于DL框架的全新药物设计模型的性能进行全面了解, 在完成生成任务之后需对计算结果进行评测. 由于这些生成模型一次生成的有效分子数量通常比传统的方法多1~2个数量级, 如果直接对生成分子进行模拟或者活性测试, 将耗费巨大的时间

和资源. 目前通常是选用一些计算资源消耗相对较小、计算复杂度相对简单的评价指标对生成分子的质量或者模型性能进行初步评测. 我们将常用的评测指标总结在表3中.

有效性(validity)、新颖性(novelty)和唯一性(uniqueness)最常用的统计学指标, 这些指标来源于机器学习领域的文本和图片生成方向, 最初用于评价图片和文本的生成质量. 目前几乎所有的生成模型都在使用这些指标来初步评测生成模型的性能. 弗雷歇化学空间距离(Frechet chemNet distance, FCD)也是一个常用的指标^[82], 其使用了一个基于神经网络的生物活性预测器, 分别将训练集以及生成的分子集放入到活性预测器中, 然后计算两个集合在生物活性预测器倒数第二层的分布距离, 使用距离值来表示生成分子与训练集的理化性质差异.

也有一些综合性评测平台如MOSES和Guacamol^[83,84]结合各种评测指标对生成的分子集进行综合评价. MOSES平台则是对上述一些指标进行了整理, 以增加用户的便利性. Guacamol平台则更多关注评测生成算法执行给定任务的能力, 如模型生成某一特定分子的能力、模型生成同分异构体的能力、模型生成预设的SMILES相似分子的能力等.

5 总结与展望

在文本、图像、音乐以及视频的分类和生成领域, DL算法相比于传统的机器学习算法已展现出明显的优势. 目前在全新药物设计领域使用的DL框架也多是分子的构建过程类比于文本或者图片的生成过程. 由于全新药物设计的过程与这些任务大体相似, 这些DL模型的框架不需要进行过多的修改而直接用到药物设计领域, 但效果整体上仍不尽如人意. 目前已经报道的DL框架用于全新药物设计模型已经接近100例, 这些模型均能够在各种评测指标上取得让人满意的结果, 但是成功设计、合成并通过体外实验证明化合物有活性的案例仅仅只有4例^[31,37,38,77]. 这说明研究者们虽然已经开发出种类多样的生成模型, 但目前DL框架用于全新药物设计仍然处于起步阶段.

DL能否成功用于全新药物设计需要注意到分子构建与文本、图像等生成任务之间存在的一些关键性差异, 从而对DL框架进行药物设计任务的“适配化”.

表 3 用于评测生成模型性能的评测指标(其中生成数据集为 G , 训练集为 E)^{a)}Table 3 The evaluation metrics for generative models (G represents the generated molecular set and E represents the training set)

类别	评测指标	注释
分子集	Validity	G 中有效化合物的比例
	Novelty	G 中只存在于 G 中而不存在于 E 中的化合物的比例
	Uniqueness	G 中不重复的化合物的比例
	Frag	G 和 E 中化合物的片段相似性
	Scaff	G 和 E 中化合物的骨架相似性
	SNN	G 和 E 中Tanimoto相似性系数最高化合物的平均值
	Div	G 中化合物的骨架相似性
	FCD ^[82]	E 和 G 中化合物的理化性质的相似性
	Completeness ^[45]	生成 E 中所有的分子
	Uniformity ^[45]	生成分子的等概率
分子	Closedness ^[45]	只生成 E 中存在的分子
	理化性质	G 中化合物的 M_w 、 $\log P$ 、TPSA属性
	SA	G 中化合物的合成难易度(值为0~1, 易到难)
	NP	G 中化合物的类别(值为0~5时, 越大则更类似于天然产物; 值为-5~0时, 越大则更类似于合成化合物)
	QED	G 中化合物的类药性
综合评测平台	JS ^[72]	对于 G 中的化合物 m , 通过下式计算: $JS(m) = \log P(m) - SA(m) - \text{RingPenalty}(m)$ 其中RingPenalty指的是不真实的大环类型化合物的惩罚系数
	Guacamol ^[84]	模型再现训练集的属性分布的能力; 生成特定分子结构的能力; 探索和开发化学空间的能力; 单目标和多目标优化任务的多样性能力
	MOSES ^[83]	包含了几种经典的分子生成模型; 一套衡量所生成分子的多样性和质量的指标

a) Div: 内部多样性(internal diversity); Frag: 片段相似性(fragment similarity); JS: 联合打分(jointly score); $\log P$: 脂水分布系数(the lipid-aqueous partition coefficient); M_w : 重均分子量(molecular weight); NP: 天然产物相似性打分(natural product-likeness score); QED: 类药性(quantitative estimation of drug-likeness); SA: 可合成系数(synthetics accessibility score); Scaff: 骨架相似性(scaffold similarity); SNN: 最近邻相似(nearest neighbor similarity); TPSA: 拓扑分子极性表面积(topological molecular polar surface area).

首先是分子表征与图片、文本的差异, 分子表征最主要的特点是多样性高以及容错性差. (1) 多样性高: 图片和文本是电子化的信息, 其可以被计算机程序完整记录; 而分子是真实存在的物质, 由于其具备多种理化性质, 所以不存在一种分子表征能够完美地描述一个分子, 需要根据不同的任务选择对分子结构、理化性质等某些方面进行描述, 所以存在种类繁多的分子表征方式. (2) 容错性差: 在生成的文本或者图像只有单个像素或单词错误的情况下, 模型不会完全否定生成的文本或图像; 但是在全新药物设计模型中, 在分子表征中只要出现任何微小的错误, 模型就会将其判断为无效的分子, 这使模型会避免生成很多相对复杂的分子来减少错误的发生, 但最终会导致对化学空间探索的不够充分, 而使得生成的分子多样性

降低.

其次是活性预测器的假阳性问题. 如在RL和GAN的框架中, 基于机器学习的QSAR (quantitative structure-activity relationship)模型通常作为奖励函数来指导分子的生成, 但QSAR在预测分子的生物活性和理化性质时存在假阳性率高的问题. 现在已经有模型使用假阳性率低的QSAR模型用于生成模型的奖励系统, 如上面提到的Gentrl模型实际上就是运用SOM来降低假阳性率^[31].

最后是评测指标, 虽然已经有很多指标用来评价生成的分子质量以及生成模型的性能, 但目前仍然缺乏一个“金标准”. 目前使用的所有评测指标都有一定的缺点. 例如, 有效性、新颖性和单一性是评估模型利弊最常用的指标. 然而, 生成模型可能会通过生成

不具有类药性质的化合物(如短/长链烷烃和大环烷烃)来过度优化这些指标. 而FCD实质上是一种基于神经网络的生物活性预测器, 也无法避免假阳性的问题. Guacamol平台指标关注的一些任务在药物设计中其实并不必要. 最近已经有研究者发现, 许多应用广泛的评测指标实际上无法有效评价生成模型的性能^[85].

训练集的选择对DL模型的表现也至关重要, 使用同类型的训练集只能得到同类型的结果. 如果训练过程中使用的数据集不能充分覆盖类药化学空间, 分子生成模型只能学到不完整的语法, 那生成的分子也会

分布在有限的化学空间中. 但目前还没有一个评测指标能够在模型训练之前对训练集的质量进行评估.

基于DL的全新药物设计方法展现出广阔的发展前景, 但并不代表传统的方法已经过时, 传统的基于配体的遗传算法也可以获得与基于DL的生成算法相似的结果^[86,87], 也有评测结果表明基于DL的生成算法对化学空间的探索能力要低于传统的方法^[88]. 总而言之, 我们应该清楚地认识到, 要将DL用于药物设计实际应用场景仍有许多问题亟待解决. 但DL在其他领域取得的优势给了我们信心, 同时也需要注意进行药物设计应用场景的“适配化”.

参考文献

- 1 Takebe T, Imai R, Ono S. *Clin Transl Sci*, 2018, 11: 597–606
- 2 Cheng T, Li Q, Zhou Z, Wang Y, Bryant SH. *AAPS J*, 2012, 14: 133–141
- 3 Sliwoski G, Kothiwale S, Meiler J, Lowe EW. *Pharmacol Rev*, 2014, 66: 334–395
- 4 Tian S, Wang J, Li Y, Li D, Xu L, Hou T. *Adv Drug Deliver Rev*, 2015, 86: 2–10
- 5 Blum LC, Raymond JL. *J Am Chem Soc*, 2009, 131: 8732–8733
- 6 Sterling T, Irwin JJ. *J Chem Inf Model*, 2015, 55: 2324–2337
- 7 Polishchuk PG, Madzhidov TI, Varnek A. *J Comput Aided Mol Des*, 2013, 27: 675–679
- 8 Davies M, Nowotka M, Papadatos G, Dedman N, Gaulton A, Atkinson F, Bellis L, Overington JP. *Nucleic Acids Res*, 2015, 43: W612–W620
- 9 Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J. *Nucleic Acids Res*, 2016, 44: D1045–D1053
- 10 Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE. *Nucleic Acids Res*, 2019, 47: D1102–D1109
- 11 Böhm HJ. *J Comput-Aided Mol Des*, 1992, 6: 593–606
- 12 Clark DE, Frenkel D, Levy SA, Li J, Murray CW, Robson B, Waszkowycz B, Westhead DR. *J Comput-Aided Mol Des*, 1995, 9: 13–32
- 13 Gillet VJ, Newell W, Mata P, Myatt G, Sike S, Zsoldos Z, Johnson AP. *J Chem Inf Model*, 1994, 34: 207–217
- 14 Pierce AC, Rao G, Bemis GW. *J Med Chem*, 2004, 47: 2768–2775
- 15 Schneider G, Lee ML, Stahl M, Schneider P. *J Comput-Aided Mol Des*, 2000, 14: 487–494
- 16 Fischer T, Gazzola S, Riedl R. *Expert Opin Drug Dis*, 2019, 14: 791–803
- 17 Hassan Baig M, Ahmad K, Roy S, Mohammad Ashraf J, Adil M, Haris Siddiqui M, Khan S, Amjad Kamal M, Provaznik I, Choi I. *Curr Pharm Des*, 2016, 22: 572–581
- 18 Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, Lanctot M, Sifre L, Kumaran D, Graepel T, Lillicrap T, Simonyan K, Hassabis D. *Science*, 2018, 362: 1140–1144
- 19 Liu Z, Du J, Fang J, Yin Y, Xu G, Xie L. *Database(Oxford)*, 2019, 2019: baz104
- 20 Pérez-Sianes J, Pérez-Sánchez H, Díaz F. *Curr Comput-Aid Drug*, 2019, 15: 6–28
- 21 Winkler DA, Le TC. *Mol Inf*, 2017, 36: 1600118
- 22 Uesawa Y. *Bioorg Med Chem Lett*, 2018, 28: 3400–3403
- 23 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
- 24 Skalic M, Sabbadin D, Sattarov B, Sciabola S, de Fabritiis G. *Mol Pharm*, 2019, 16: 4282–4291
- 25 Kotsias PC, Arús-Pous J, Chen H, Engkvist O, Tyrchan C, Bjerrum EJ. *Nat Mach Intell*, 2020, 2: 254–265
- 26 Segler MHS, Kogej T, Tyrchan C, Waller MP. *ACS Cent Sci*, 2018, 4: 120–131

- 27 Moret M, Friedrich L, Grisoni F, Merk D, Schneider G. *Nat Mach Intell*, 2020, 2: 171–180
- 28 Kadurin A, Nikolenko S, Khrabrov K, Aliper A, Zhavoronkov A. *Mol Pharm*, 2017, 14: 3098–3104
- 29 Blaschke T, Arús-Pous J, Chen H, Margreitter C, Tyrchan C, Engkvist O, Papadopoulos K, Patronov A. *J Chem Inf Model*, 2020, 60: 5918–5922
- 30 Wang J, Hsieh CY, Wang M, Wang X, Wu Z, Jiang D, Liao B, Zhang X, Yang B, He Q, Cao D, Chen X, Hou T. *Nat Mach Intell*, 2021, 3: 914–922
- 31 Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, Terentiev VA, Polykovskiy DA, Kuznetsov MD, Asadulaev A, Volkov Y, Zhohus A, Shayakhmetov RR, Zhebrak A, Minaeva LI, Zagribelnyy BA, Lee LH, Soll R, Madge D, Xing L, Guo T, Aspuru-Guzik A. *Nat Biotechnol*, 2019, 37: 1038–1040
- 32 Li Y, Zhang L, Liu Z. *J Cheminform*, 2018, 10: 33
- 33 Pogány P, Arad N, Genway S, Pickett SD. *J Chem Inf Model*, 2019, 59: 1136–1146
- 34 Gebauer NWA, Gastegger M, Hessmann SSP, Müller K-R, Schütt KT. *Nat Commun*, 2022, 13: 973
- 35 Ragoza M, Masuda T, Koes DR. *Chem Sci*, 2022, 13: 2701–2713
- 36 Imrie F, Hadfield TE, Bradley AR, Deane CM. *Chem Sci*, 2021, 12: 14577–14589
- 37 Putin E, Asadulaev A, Vanhaelen Q, Ivanenkov Y, Aladinskaya AV, Aliper A, Zhavoronkov A. *Mol Pharm*, 2018, 15: 4386–4397
- 38 Polykovskiy D, Zhebrak A, Vetrov D, Ivanenkov Y, Aladinskiy V, Mamoshina P, Bozdaganyan M, Aliper A, Zhavoronkov A, Kadurin A. *Mol Pharm*, 2018, 15: 4398–4405
- 39 Ruddigkeit L, van Deursen R, Blum LC, Reymond JL. *J Chem Inf Model*, 2012, 52: 2864–2875
- 40 Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, Magariños MP, Mosquera JF, Mutowo P, Nowotka M, Gordillo-Marañón M, Hunter F, Junco L, Mugumbate G, Rodriguez-Lopez M, Atkinson F, Bosc N, Radoux CJ, Segura-Cabrera A, Hersey A, Leach AR. *Nucleic Acids Res*, 2019, 47: D930–D940
- 41 Ramakrishnan R, Dral PO, Rupp M, von Lilienfeld OA. *Sci Data*, 2014, 1: 1–7
- 42 Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. *Nucleic Acids Res*, 2018, 46: D1074–D1082
- 43 Weininger D. *J Chem Inf Comput Sci*, 1988, 28: 31–36
- 44 Bjerrum EJ, Sattarov B. *Biomolecules*, 2018, 8: 131
- 45 Arús-Pous J, Johansson SV, Prykhodko O, Bjerrum EJ, Tyrchan C, Reymond JL, Chen H, Engkvist O. *J Cheminform*, 2019, 11: 71
- 46 Hemmerich J, Asilar E, Ecker GF. *J Cheminform*, 2020, 12: 18
- 47 Kim K, Kang S, Yoo J, Kwon Y, Nam Y, Lee D, Kim I, Choi YS, Jung Y, Kim S, Son WJ, Son J, Lee HS, Kim S, Shin J, Hwang S. *npj Comput Mater*, 2018, 4: 67
- 48 Kadurin A, Aliper A, Kazennov A, Mamoshina P, Vanhaelen Q, Khrabrov K, Zhavoronkov A. *Oncotarget*, 2017, 8: 10883–10890
- 49 Wild DJ, Blankley CJ. *J Chem Inf Comput Sci*, 2000, 40: 155–162
- 50 Rogers D, Hahn M. *J Chem Inf Model*, 2010, 50: 742–754
- 51 Awale M, Reymond JL. *Nucleic Acids Res*, 2014, 42: W234–W239
- 52 Kusner MJ, Paige B, Hernández-Lobato JM. *Grammar Variational Autoencoder*. New York: PMLR, 2017. 1945–1954
- 53 Wengong J, Barzilay R, Jaakkola T. 2018, arXiv: 1802.04364
- 54 Dai H, Tian Y, Dai B, Skiena S, Song L. 2018, arXiv: 1802.08786
- 55 De Cao N, Kipf T. 2018, arXiv: 1805.11973
- 56 Joshi RP, Gebauer NWA, Bontha M, Khazaieli M, James RM, Brown JB, Kumar N. *J Phys Chem B*, 2021, 125: 12166–12176
- 57 Atz K, Grisoni F, Schneider G. *Nat Mach Intell*, 2021, 3: 1023–1032
- 58 Li Y, Pei J, Lai L. *Chem Sci*, 2021, 12: 13664–13675
- 59 Skalic M, Jiménez J, Sabbadin D, De Fabritiis G. *J Chem Inf Model*, 2019, 59: 1205–1214
- 60 Ragoza M, Masuda T, Koes DR. 2020, arXiv: 2010.08687
- 61 Hershey JR, Olsen PA. Approximating the Kullback Leibler divergence between Gaussian mixture models. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing. Honolulu, 2007. IV317–IV320
- 62 Kang S, Cho K. *J Chem Inf Model*, 2019, 59: 43–52
- 63 Lim J, Ryu S, Kim JW, Kim WY. *J Cheminform*, 2018, 10: 31

- 64 Liu Q, Allamanis M, Brockschmidt M, Gaunt A. *Constrained Graph Variational Autoencoders for Molecule Design*. New York: PMLR, 2018
- 65 Simonovsky M, Komodakis N. 2018, arXiv: 1802.03480
- 66 He J, You H, Sandström E, Nittinger E, Bjerrum EJ, Tyrchan C, Czechtizky W, Engkvist O. *J Cheminform*, 2021, 13: 26
- 67 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. 2017, arXiv: 1706.03762
- 68 Gupta A, Müller AT, Huisman BJH, Fuchs JA, Schneider P, Schneider G. *Mol Inf*, 2018, 37: 1700111
- 69 Olivecrona M, Blaschke T, Engkvist O, Chen H. *J Cheminform*, 2017, 9: 48
- 70 Popova M, Isayev O, Tropsha A. *Sci Adv*, 2018, 4: 7885
- 71 Ståhl N, Falkman G, Karlsson A, Mathiason G, Boström J. *J Chem Inf Model*, 2019, 59: 3166–3176
- 72 Yang X, Zhang J, Yoshizoe K, Terayama K, Tsuda K. *Sci Tech Adv Mater*, 2017, 18: 972–976
- 73 Liu X, Ye K, van Vlijmen HWT, IJzerman AP, van Westen GJP. *J Cheminform*, 2019, 11: 35
- 74 Guimaraes GL, Sanchez-Lengeling B, Outeiral C, Farias PLC, Aspuru-Guzik A. 2018, arXiv: 1705.10843
- 75 Sanchez-Lengeling B, Outeiral C, Guimaraes GL, Aspuru-Guzik A. *ChemRxiv*, 2017, arXiv: 5309668.v5309663
- 76 Prykhodko O, Johansson SV, Kotsias PC, Arús-Pous J, Bjerrum EJ, Engkvist O, Chen H. *J Cheminform*, 2019, 11: 1–3
- 77 Godinez WJ, Ma EJ, Chao AT, Pei L, Skewes-Cox P, Canham SM, Jenkins JL, Young JM, Martin EJ, Guiguemde WA. *Nat Mach Intell*, 2022, 4: 180–186
- 78 Samanta B, De A, Jana G, Gómez V, Chattaraj PK, Ganguly N, Gomez-Rodriguez M. *J Mach Learn Res*, 2020, 21: 1–33
- 79 Sattarov B, Baskin II, Horvath D, Marcou G, Bjerrum EJ, Varnek A. *J Chem Inf Model*, 2019, 59: 1182–1196
- 80 Mercado R, Rastemo T, Lindelöf E, Klambauer G, Engkvist O, Chen H, Bjerrum EJ. *Appl AI Lett*, 2020, 1: a12.18
- 81 van Deursen R, Ertl P, Tetko IV, Godin G. *J Cheminform*, 2020, 12: 22
- 82 Preuer K, Renz P, Unterthiner T, Hochreiter S, Klambauer G. *J Chem Inf Model*, 2018, 58: 1736–1741
- 83 Polykovskiy D, Zhebrak A, Sanchez-Lengeling B, Golovanov S, Zhavoronkov A. 2019, arXiv: 1811.12823
- 84 Brown N, Fiscato M, Segler MHS, Vaucher AC. *J Chem Inf Model*, 2019, 59: 1096–1108
- 85 Skinnider MA, Stacey RG, Wishart DS, Foster LJ. *Nat Mach Intell*, 2021, 3: 759–770
- 86 Jensen JH. *Chem Sci*, 2019, 10: 3567–3572
- 87 Bai Q, Tan S, Xu T, Liu H, Huang J, Yao X. *Briefings Bioinf*, 2020, 22: 161
- 88 Wang M, Sun H, Wang J, Pang J, Chai X, Xu L, Li H, Cao D, Hou T. *Briefings Bioinf*, 2022, 23: bbab544

Deep learning approaches for *de novo* drug design: an overview

Mingyang Wang, Dan Li, Tingjun Hou^{*}, Yu Kang^{*}

School of Pharmacy, Zhejiang University, Hangzhou 310058, China

^{*}Corresponding authors (email: tingjunhou@zju.edu.cn; yukang@zju.edu.cn)

Abstract: The design and discovery of lead compounds is the most challenging and creative stage in drug development, and multiple attributes of candidate molecules need to be optimized in this process, such as structural novelty, biological activity, target selectivity, synthesability, druggability, and safety. Although the development and application of computer-aided drug design methods have substantially reduced the time and cost in the lead compound discovery stage, it has not yet been able to reverse the current situation of low success rate of drug development. In recent years, with the continuous development of deep learning (DL) technology, DL-based *de novo* drug design methods have brought new opportunities for the discovery of lead compounds. The DL frameworks used by these new drug design models include encoder-decoder, recurrent neural network, generative adversarial networks, and reinforcement learning. In this paper, the basic theory, the input molecular representation and the evaluation metrics of their methods are reviewed, and the application prospect of DL-based *de novo* drug design methods is also discussed.

Keywords: *de novo* drug design, computer-aided drug design, deep learning, machine learning, molecular generation

doi: [10.1360/SSC-2022-0135](https://doi.org/10.1360/SSC-2022-0135)