# 人工智能技术在先导化合物发现与优化中的应用进展

李紫玥, 丛开源, 吴诗琪, 朱启华, 徐云根*, 邹 毅*

(中国药科大学药学院, 江苏 南京 211198)

**摘要**: 近年来, 人工智能 (artificial intelligence, AI) 技术发展突飞猛进, 被广泛应用于医学、药学等多个领域, 加速了药物研发的进程。本文聚焦AI在先导化合物发现与优化环节的应用, 详细介绍了AI辅助虚拟筛选以及分子生成方法发现先导化合物, 特别是AI驱动药物进入临床试验的应用案例, 同时简略阐述AI基本算法模型在定量构效关系 (quantitative structure-activity relationship, QSAR) 和药物重定位中的应用, 为基于AI的药物发现提供参考。

**关键词**: 人工智能; 药物发现; 先导化合物; 虚拟筛选; 分子生成

**中图分类号**: R914 **文献标识码**: A **文章编号**: 0513-4870(2024)09-2443-11

# Advances of artificial intelligence technology in the discovery and optimization of lead compounds

LI Zi-yue, CONG Kai-yuan, WU Shi-qi, ZHU Qi-hua, XU Yun-gen*, ZOU Yi*

(*School of Pharmacy, China Pharmaceutical University, Nanjing 211198, China*)

**Abstract**: In recent years, artificial intelligence (AI) technology has advanced rapidly and has been widely applied in various fields such as medicine and pharmacy, accelerating the drug development process. Focusing on the application of AI in the discovery and optimization of lead compounds, this review provides a detailed introduction to AI-assisted virtual screening and molecular generation methods for discovering lead compounds, while particularly highlighting the cases of AI-drived drugs into clinical trials. Additionally, we briefly outline the application of AI basic algorithm models in quantitative structure-activity relationship (QSAR) and drug repurposing, offering insights for AI-based drug discovery.

**Key words**: artificial intelligence; drug discovery; lead compound; virtual screening; molecular generation

人工智能 (artificial intelligence, AI) 是一门利用计算机技术、理论、方法和软件等来研究和开发能够用来扩展人的智能、模拟人类行为和思维, 从而对数据进行分析的科学, 主要包括机器学习 (machine learning, ML) 和深度学习 (deep learning, DL) 两个子领域[1]。随着计算机技术的快速发展和大数据时代的到来, AI在多个领域取得了一定的突破, 比如阿尔法围棋 (AlphaGo) 利用DL技术在人机围棋对决中战胜职业围棋选手, 以及近期大热的聊天机器人程序ChatGPT (chat generative pre-trained transformer) 和文本转视频生成模型Sora的诞生, 在各个领域都产生了一定的影响。

伴随着AI技术的快速发展与推广, AI+已经渗透到医疗健康领域的各个方面。起初, AI大规模应用于医疗影像, 后逐渐渗透到药物研发领域, 使药物开发模式产生一系列的变化。近年来, 随着生物医药数据的不断积累, AI技术在药学领域中的应用越来越广泛。新药研发过程中产生的数据涉及从靶点发现、先导化合物的发现与优化、临床前研究到临床试验的各个阶段。AI可以从海量的药物研发原始数据中快速挖掘信息密度高的数据, 并通过整合分析这些数据为药物研发提供更多新的见解。如今, AI技术正在逐步参与

收稿日期: 2024-02-21; 修回日期: 2024-05-19.
基金项目: 中国药科大学教学改革研究课题 (2023XJQN16).
*通讯作者 E-mail: zouyi@cpu.edu.cn; xyg@cpu.edu.cn
DOI: 10.16438/j.0513-4870.2024-0221

药物发现过程的各个阶段[2-7], 学术界和产业界都在尝试研究使用AI来辅助药物研发, 为新药的发现与开发寻求助力 (图1)。

药物发现是一个漫长而又复杂的过程, 其中先导化合物的发现与优化是新药研发中的核心环节, 需要通过“设计、合成、活性测试”的反复循环来提高化合物的活性、选择性和成药性。目前, 先导化合物的发现和优化过程高度依赖药物化学家的经验和大量资源的投入, 速度较慢且工作量大, 而在人工智能驱动下先导化合物的发现与优化则有望同时兼顾到速度和精度, 为提升效率和降低成本提供机会。本文将聚焦AI在先导化合物的发现与优化过程中的应用, 在简要介绍机器学习及深度学习的基本概念和算法后, 着重介绍ML/DL算法在先导化合物发现和优化中的具体案例, 以阐明AI在药物设计过程中的重要作用。

## 1 机器学习和深度学习概念及算法介绍

机器学习是实现人工智能的一种方式, 是人工智能的子领域。机器学习基于已有的数据、知识或者经验, 自动识别和解析数据, 总结有意义的模式, 并以此在相似的环境里做出预测或决策, 主要分为监督学习和无监督学习两大类[8]。其中, 监督学习是指可以从带标签的训练集中学习或建立一个模型, 并根据该模型对新的实例进行预测[9]。而无监督学习方法中的模型则使用未标记的数据集进行训练, 并允许在没有任何监督的情况下对该数据进行操作[10]。

### 1.1 常用的机器学习算法介绍

**1.1.1 朴素贝叶斯 (naive bayes, NB)** NB是一种常见的分类方法, 其核心思想是通过考虑特征概率来预测分类, 即对于给出的待分类样本, 求解在此样本出现的条件下各个类别出现的概率[11]。

**1.1.2 K最近邻 (k-nearest neighbours, KNN)** KNN算法由Cover和Hart[12]在1967年提出, 是一种分类算法。其总体思路为, 样本的邻域都是带标签样本, 对于一个待分类样本X, 距离它最近的邻域中何种类别样本数量最多, 那么该样本将被划分为同一类 (图2A)。

**1.1.3 支持向量机 (support vector machines, SVM)** SVM由Cortes等[13]在1995年提出, 多用于分类问题。SVM的思想是, 在其原始低维输入空间中的训练数据可以在通过映射构建的高维潜在空间中分离 (图2B)。

**1.1.4 决策树 (decision tree, DT) 与随机森林 (random forest, RF)** DT是在已知特征取值的基础上, 通过构建树型决策结构来进行分析的一种常用的分类算法[14]。决策树模型具有可读性、分类速度快的特点, 在各种实际业务建模过程中广泛使用 (图2C)。

RF是通过构建多个DT对样本进行训练并预测的一种分类器, DT的分类结果中何种类别最多, 那么RF就会把这种类别当作最终的结果 (图2D)。

**1.1.5 极限梯度提升 (extreme gradient boosting, XGBoost)** XGBoost算法是一种基于决策树的集成学习算法, 可以通过迭代地训练弱学习器, 通常是决策树, 并根据前一轮迭代的错误来调整模型, 从而逐步提升整体模型的性能。XGBoost算法是一种强大、灵活、高效的机器学习算法, 可以应用于各种领域和问题[15]。

### 1.2 深度学习概念及模型介绍

深度学习是机器学习研究中的一个新的领域, 利用深度人工神经网络可以从海量的数据中进行自动学习、分类和预测, 找寻数据的特征。

**1.2.1 深度神经网络 (deep neural networks, DNN)** 人工神经网络是一种受人脑的生物神经网络启发而设计的计算模型, 擅长从输入数据中学习映射关系, 从而完成预测或者分类问题, 它类似于生物神经网络, 由人工神经元构成[8]。每个神经元由简单的数学模型来模拟生物神经细胞的信号传递与激活, 一般认为超过三层的神经网络就可以叫做深度神经网络。前馈是指神

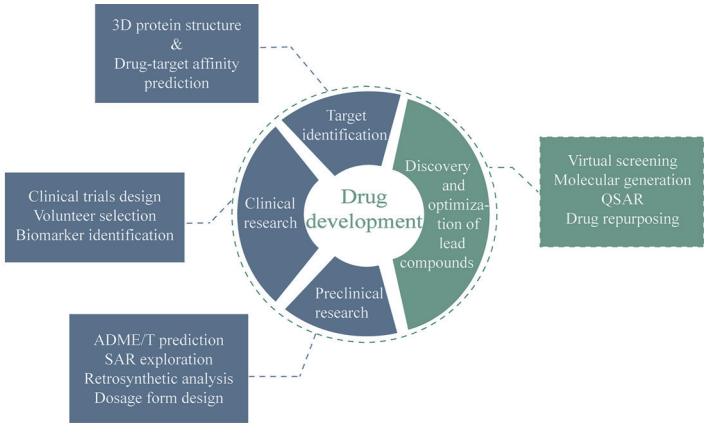


**Figure 1** Application of artificial intelligence technology in drug development process. QSAR: Quantitative structure-activity relationship; ADME/T: Absorption, distribution, metabolism, excretion/toxicity; SAR: Structure-activity relationship

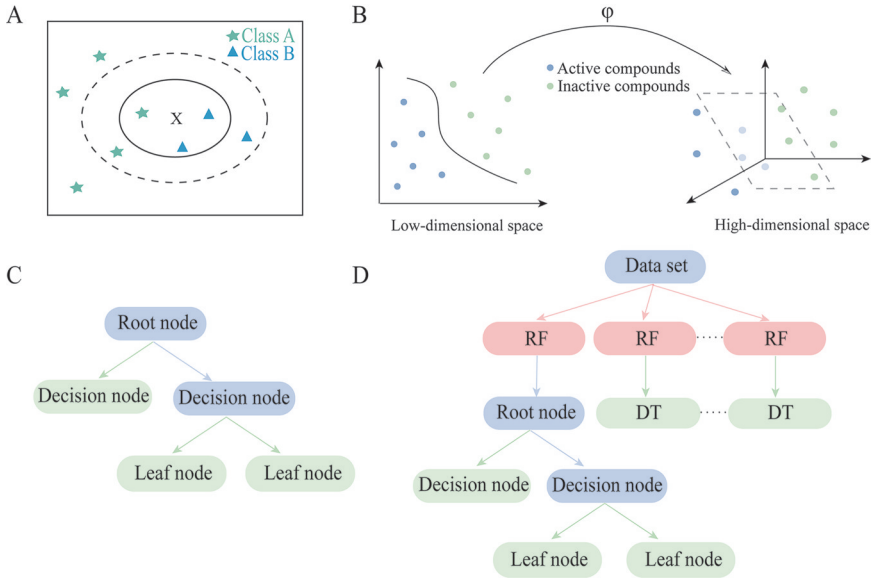


**Figure 2** Introduction to machine learning algorithms. A: The basic principle of KNN algorithm; B: SVM maps low-dimensional data to high-dimensional space for separation; C: The diagram of DT algorithm model; D: The diagram of RF algorithm model. KNN: K-nearest neighbours; SVM: Support vector machines; DT: Decision tree; RF: Random forest

经网络的传播方向是单向的, 信号仅在一个方向上传播, 虽然前馈网络结构简单, 但不能很好地处理时间序列数据, 于是循环神经网络应运而生。

**1.2.2 循环神经网络 (recurrent neural networks, RNN)** RNN 由 Hopfield[16]在 1982 年提出, 是指在全连接神经网络的基础上增加了前后时序上的关系, 可以更好地处理比如机器翻译等与时序相关的问题。

传统的神经网络模型是从输入层到隐藏层再到输出层, 层与层之间是全连接的, 每层之间的节点是无连接的, 这种普通的神经网络对时序数据束手无策。相对于前馈神经网络, RNN 可以接收上一个时间点的隐藏状态, RNN 的实质是上一个时刻的网络状态将会作用并影响到下一个时刻的网络状态, 表明 RNN 和序列数据密切相关 (图 3A)。

长短时记忆神经网络 (long short-term memory networks, LSTM) 是特殊的 RNN, 尤其适合顺序序列数据的处理, LSTM 由 Hochreiter 和 Schmidhuber[17]在 1997 年提出。LSTM 明确旨在避免 RNN 长期依赖性问题, 在语音识别、图片描述、自然语言处理等许多领域中成功应用。

2020 年, Yang 等[18]使用基于 LSTM 的神经网络模型从 ChEMBL 数据库中训练二十万种化合物, 然后使用包含 135 种已发表的 E1A 结合蛋白 (E1A-binding protein, p300) 抑制剂和 576 个大环分子的数据集对模型进行微调, 以生成新的 p300/CREB 结合蛋白 (CREB-binding protein, CBP) 抑制剂。该模型生成了

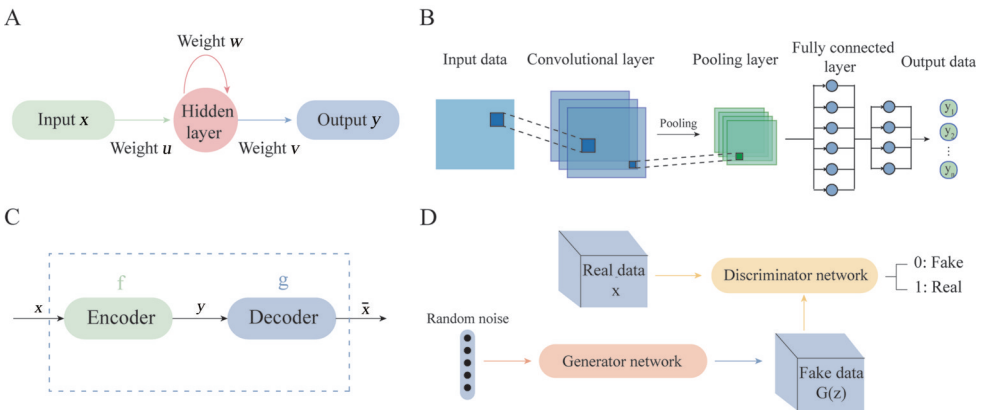


**Figure 3** Introduction to deep learning algorithms. A: The structure of RNN; B: The structure of CNN; C: The components of AE; D: The mechanism of GAN model. RNN: Recurrent neural networks; CNN: Convolutional neural networks; AE: Autoencoder; GAN: Generative adversarial networks

一个包含672个化学结构的集中库, 从中选择一些化合物进行合成。经过进一步的系统优化, 获得了一套高效的抑制剂。其中, 潜在的候选药物B026在人类癌症动物模型中对p300/CBP组蛋白乙酰转移酶表现出较高的抑制活性和显著的肿瘤生长抑制作用, 已被确定用于进一步的临床前开发。

2022年, Li等[19]提出了一种生成式深度学习 (generative deep learning, GDL) 模型, GDL模型是基于LSTM算法的分布学习循环神经网络架构, 可以生成与训练集分子遵循相同化学分布的新分子, 它的作用是为给定的生物靶标生成量身定制的虚拟化合物库。随后, 将GDL模型应用于受体相互作用蛋白激酶1 (receptor-interacting protein kinase 1, RIPK1) 激酶抑制剂的筛选, 并成功发现了一种全新骨架的RIPK1激酶抑制剂, 证明深度神经网络在早期药物发现中的重要作用。

**1.2.3 卷积神经网络 (convolutional neural networks, CNN)** CNN是一类包含卷积计算且具有深度结构的前馈神经网络[20], 由输入层−卷积层−激活函数−池化层−全连接层组成。CNN的核心是卷积层, 它能够对特定的预测问题进行建模, 针对训练数据集学习大量的卷积核; 卷积计算完成后, 往往会加入一个修正线性单元 (rectified linear unit, ReLU) 函数, 将数据非线性化; 池化层可以有效地缩小矩阵的尺寸, 既加快了计算速度, 也能防止过拟合问题; 最后使用全连接层进行最终的分类 (图3B)。

2022年, Noguchi等[21]提出了像素卷积神经网络 (pixel convolutional neural network, PixelCNN) 模型, 将简化分子线性输入规范 (simplified molecular input line entry system, SMILES) 字符串转换为二维矩阵数据, 应用掩蔽神经网络层建立模型。他们对PixelCNN的性能进行了多方面的分析, 并将其与RNN在生成期望性质的分子方面和基于片段生长优化的化学空间探索方面进行了详尽的比较。尽管RNN在直接预测与目标分子性质相对应的分子结构方面优于PixelCNN, 但基于PixelCNN的框架在分子结构的片段生长优化方面明显优于RNN方法, 可以很好地应用于基于片段的药物发现任务。

**1.2.4 自编码器 (autoencoder, AE)** AE作为深度神经网络的一类方法, 主要用于数据降维、压缩以及获取低维度表征等。自编码器与传统机器学习中的主成分分析等降维方法的作用相同, 但与之相比更为灵活, 效果往往更好。

自编码器可以学习输入数据的隐含特征, 称为编码, 同时用学习到的新特征重构出原始输入数据, 称为解码。其主要目的是将输入数据$x$转换成中间变量$y$, 然后再将$y$转换成$\bar{x}$, 对比输入$x$和输出$\bar{x}$, 使它们无限接近 (图3C)。

2020年, Chenthamarakshan等[22]提出了一种称为分子受控生成 (controlled generation of molecules, CogMol) 的生成模型, 通过在变分自编码器模型中引入多属性受控采样方案, 设计一组所需属性的新型病毒蛋白的分子。他们使用CogMol为严重急性呼吸综合征冠状病毒2 (severe acute respiratory syndrome coronavirus 2, SARS-CoV-2) 的三种靶蛋白生成新分子, 并限制其靶点亲和力和选择性、药物相似性、合成可行性和毒性。结果表明, 生成的分子能够与靶点相关的药物口袋有利结合, 并表现出较低的预测代谢物毒性和较高的合成可行性。

**1.2.5 生成式对抗网络 (generative adversarial networks, GAN)** GAN是生成模型的一种, 通过获取样本来训练模型, 然后按照定义的目标数据分布去生成数据 (图3D)。GAN包括一个生成器和一个判别器, 生成器捕捉真实数据样本的潜在分布, 并生成新的数据样本; 判别器是一个二分类器, 判别输入是真实数据还是生成的样本。生成器和判别器需要不断优化, 提高各自的生成能力和判别能力, 这个学习优化过程就是寻找二者之间的一个纳什均衡[23]。

2018年, Polykovskiy等[24]提出了纠缠条件对抗自动编码器 (entangled conditional adversarial autoencoder, ECAAE) 模型, 它基于各种属性生成分子结构, 例如针对特定蛋白质的活性、溶解度或合成性能。研究人员用ECAAE生成一种新型的Janus激酶3 (Janus kinase 3, JAK3) 抑制剂, 发现的分子经体外测试, 显示出良好的活性和选择性。

## 2 AI技术在先导化合物发现与优化中的应用

### 2.1 虚拟筛选

虚拟筛选是指在进行生物活性筛选之前, 根据预先设定的条件, 在计算机上对化合物分子进行预筛选, 以识别出最可能与靶标结合的小分子, 从而大大降低实际筛选化合物的数目, 同时提高先导化合物的发现效率。一直以来, 虚拟筛选已成为苗头化合物和先导化合物发现的有力手段之一。然而, 虚拟筛选的不足之处在于随着虚拟化合物库规模的扩大, 其筛选的速度和效率相对低下; 对于基于分子对接的虚拟筛选策略而言, 其打分函数的准确性有待提高。近年来, 将AI的各大算法应用于虚拟筛选中, 一方面可以扩大虚拟化合物库中分子的数量及多样性; 另一方面, 改进筛选算法, 开发基于AI的打分函数将使筛选结果更加准确。

**2.1.1 大规模虚拟筛选** 虚拟筛选的商业化合物库一般只有不到一千万个可用化合物, 与潜在的$10^{60}$个类药化合物空间相比只是一小部分, 这种局限性降低了虚拟筛选的药物发现效率。后来, 亿级级别的虚拟化合物库被开发出来, 如Enamine公司通过分子砌块与化学反应组合构建虚拟库, 其化合物数量可到达百亿级别[25,26]。但随着虚拟库的规模增加到亿级, 使用传统的虚拟筛选算法进行筛选的效率有限, 因此需要开发高效、快速的方法来进行大规模化合物库的虚拟筛选。

2021年, Kalliokoski开发了软件macHine leArning booSTEd dockiNg, 即HASTEN, HASTEN是首个在亿级基准测试上进行评估的工具, 通过迭代训练机器学习模型, 预测分子对接评分以加速基于结构的虚拟筛选[27]。随后他们在抗菌分子和抗病毒激酶的案例中使用了HASTEN工具对Enamine REAL类先导化合物虚拟库中的15.6亿分子进行筛选研究: 首先随机选择0.1%的虚拟化合物作为起始数据进行常规分子对接, 以此训练机器学习模型对化合物库中的所有分子进行对接打分预测并排序, 继续选择排名前0.1%的化合物进行分子对接, 重复该过程九次, 最终的训练数据集总量达到整个虚拟库的1%。结果显示, 当仅对接库中1%的化合物时, HASTEN能够以90%的高召回率找出真正一千个高分虚拟命中。由此可见, HASTEN是一种有效的、适用于日常药物发现中亿级库筛选的策略[28]。

同年, Sadybekov等[25]在《Nature》上发表了一种虚拟合成分层枚举筛选 (virtual synthon hierarchical enumeration screening, V-SYNTHES) 方法, 可对超过110亿个化合物的虚拟库进行基于层次结构的筛选, 大大减少了搜索潜在苗头化合物 (hits) 时需要评估的分子数量, 使用的计算资源是标准方法的1/100。该方法可轻松扩展以适应组合库的快速增长, 并且可能适用于任何对接算法。随后在Rho相关卷曲螺旋激酶1 (Rho-associated protein kinase 1, ROCK1) 中测试了V-SYNTHES的筛选能力, 在对21个命中化合物进行合成和体外测试发现, 有6个化合物可以与ROCK1激酶结合, 且抑制常数$K_i$值小于10 μmol·L$^{-1}$。

2022年, Beroza等[29]提出了一种虚拟筛选方法, 将潜在的产物存储为构建块 (分子片段) 和动态生成化合物的连接规则, 再利用特征树算法搜索构建块并进行比较, 将最好的构建块与互补的构建块根据化学规则连接起来, 并考虑合成可行性来生成新的分子。该方法将基于结构的片段评估方法扩展到更为广大的化学空间, 随后应用这种方法从近10亿种虚拟化合物中筛选ROCK1抑制剂, 挑选了69个分子进行购买和验证, 其中27个分子的抑制常数$K_i$值小于10 μmol·L$^{-1}$, 命中率高达39%。而且这种对接方法比传统的对接方法快好几个数量级。

2023年, Gryniukova等[30]应用基于AI的新方法PharmAI DiscoveryEngine对含有260万个分子的化合物库进行虚拟筛选, 挑选出了434个潜在靶向抑制沉寂信息调节因子1 (silent information regulator 1, Sirtuin-1) 蛋白的小分子, 仅占全库的0.02%。经多阶段体外验证, 成功发现了9种化学结构新颖的Sirtuin-1抑制剂。最后使用液相色谱仪/质谱仪 (liquid chromatograph/mass spectrometer, LC/MS) 无标记法测试化合物的抑制活性, 测试结果显示有两个化合物抑制活性最好 (图4)。由此可见, AI工具可将大规模的筛选数据集缩减到只有几百个针对特定靶标的小分子化合物, 从而使先导化合物的发现过程更快速、更经济。

**2.1.2 虚拟筛选打分函数** 为了高效地获得高活性分子, 除了需要快速的筛选计算方法外, 对于基于分子对接的虚拟筛选而言, 开发更可靠的打分函数显得尤为重要。

2020年, 浙江大学侯廷军教授课题组[31]通过机器学习中的XGBoost方法结合分子对接打分中的能量辅助项, 提出了一种名为EAT-Score (energy auxiliary terms score) 的打分函数, 包括五种不同的经典对接打分的能量评分和蛋白质–配体相互作用指纹项 (图5)。经DUD-E (database of useful decoys-enhanced) 数据集验证了EAT-Score在虚拟筛选中的表现远优于经典打分函数, 其受试者工作曲线ROC的曲线下面积 (AUC) 值比传统方法提高了约0.3。

2022年, Morris等[32]创建了机器学习一致性对接工具 (machine learning consensus docking tool, MILCDock), MILCDock采用多层感知器方法, 将五种传统分子对接工具的结合亲和力和结合模式预测相融合, 赋予一个综合分数, 较大的打分值预示着更高的结合亲和力。经DUD-E和LIT-PCBA (一种用于机器学习和虚拟筛选的无偏数据集) 两个数据集测试, 显示MILCDock在DUD-E数据集上的性能超过了传统分子对接工具和其他一致性对接方法。由此可见, MILCDock不仅集成了传统工具的优点, 还通过机器学习模型实现了更高的预测精度。

2023年, Zhang等[33]提出了一种新的分子对接打分函数, 称为基于理论的相互作用能量组合打分 (theory-based interaction energy component score, TB-IECS), 该函数结合了来自Smina和NNScore2 (neural-network scoring function version 2) 的能量项, 并使用

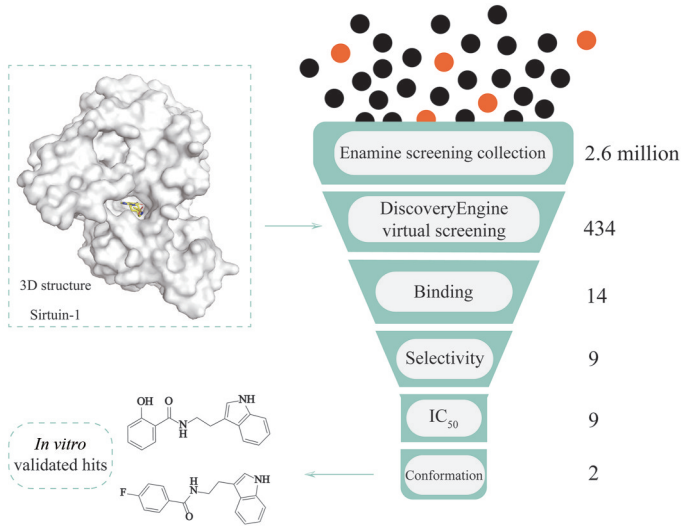


**Figure 4** Artificial intelligence (AI)-assisted virtual screening process for novel Sirtuin-1 small molecule inhibitors. Sirtuin-1: Silent information regulator 1

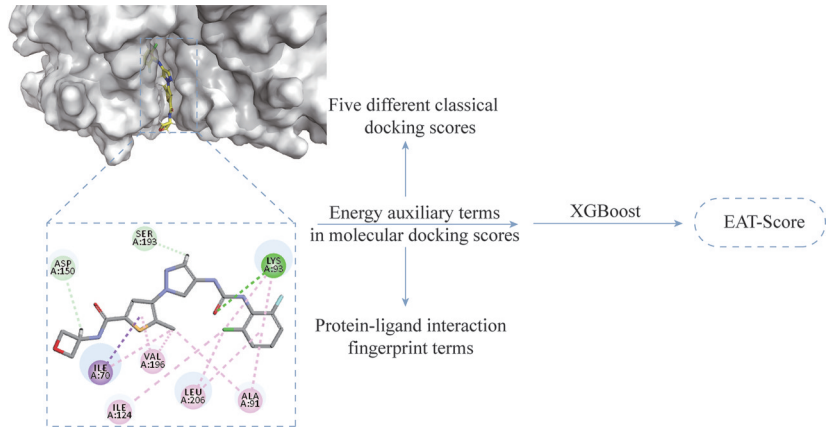


**Figure 5** The design process of EAT-Score. XGBoost: Extreme gradient boosting; EAT-Score: Energy auxiliary terms score

XGBoost构建模型。TB-IECS在DUD-E数据集、LIT-PCBA数据集和实际情景的虚拟筛选中的表现优于Glide SP (standard precision Glide docking)和Dock等传统打分函数, 有效平衡了效率和准确性。TB-IECS在机器学习打分函数中表现出优越性, 有望成为一种准确的虚拟筛选方法。

**2.1.3 AI辅助虚拟筛选发现临床候选化合物** Pharos iBio是一家专门从事罕见和难治性疾病治疗的新药开发公司, 利用其基于大数据和AI技术的新药开发平台Chemiverse开发新药 (图6A)。Chemiverse平台支持使用多种算法、10个不同模块和超过2.3亿个大数据条目, 涉及从靶标识别、先导化合物发现和优化以及临床阶段的多个方面, 从而加快药物发现并缩短研发周期。Pharos iBio利用该平台开发了多款新药, 其中一款是下一代FMS样酪氨酸激酶3 (FMS-like tyrosine kinase 3, FLT3)抑制剂PHI-101。在临床前研究中, 相比于上市的FLT3抑制剂米哚妥林 (midostaurin)和吉瑞替尼 (gilteritinib), PHI-101具有更高的抗白血病活性和较好的选择性, 同时显示出更低的毒性。目前PHI-101正在开展针对急性髓细胞白血病和高级别浆液性卵巢癌的两项I期临床试验 (NCT04842370/NCT04678102)。

Relay Therapeutics是一家靶向肿瘤小分子药物发现的精准治疗公司, 该公司整合业内前沿的实验和计算方法, 建立了Dynamo平台。Dynamo平台主要分为3个模块, 分别是药物靶点识别、先导化合物发现和先导化合物优化 (图6B)。Dynamo平台首先将蛋白质的结构生物学信息输入到Anton 2计算平台中, 生成长时间运动的全长蛋白质的虚拟模拟, 有助于更好地了解蛋白质的动态变化并寻找药物靶点。之后使用云计算从虚拟化合物库中对数十亿个分子进行虚拟筛选, 以获得大量结构新颖、具有潜在活性的小分子。Dynamo

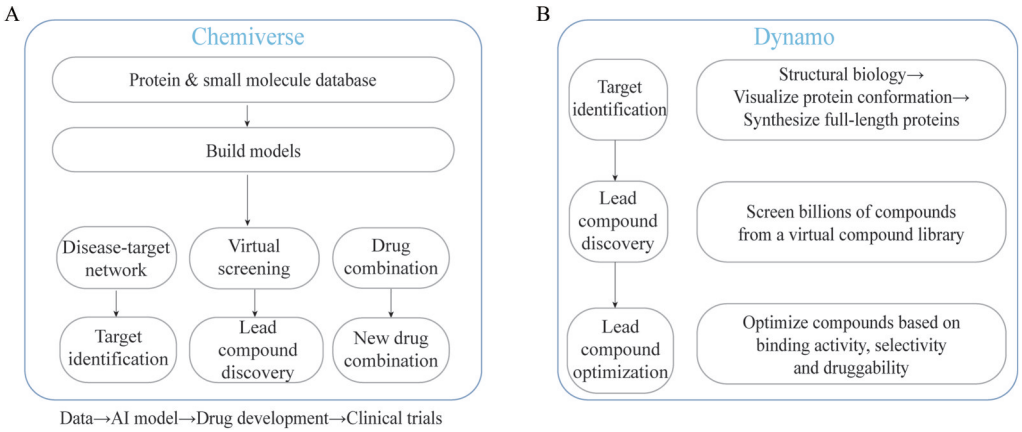


**Figure 6** The introduction of AI-assisted virtual screening research and development platform. A: The drug development process of Chemiverse platform; B: The drug development process of Dynamo platform

平台还可结合药效、选择性、生物利用度和成药性等指标, 对先导化合物进行优化, 并在实验室进行化合物的合成及活性测试; 由此获得的湿实验结果再与计算预测值进行比较来训练机器学习模型, 从而改进计算预测模型。目前Relay Therapeutics通过该平台开发出了4个临床候选药物, 分别是成纤维细胞生长因子受体2 (fibroblast growth factor receptor 2, FGFR2) 抑制剂RLY-4008 (临床I期, NCT04526106)、含Src同源2结构域蛋白酪氨酸磷酸酶 (Src homology-2-containing protein tyrosine phosphatase 2, SHP2) 抑制剂RLY-1971 (临床I期, NCT04252339)、磷脂酰肌醇3-激酶$\alpha$ (phosphoinositide 3-kinase, PI3K$\alpha$) 抑制剂RLY-2608 (临床I期, NCT05216432) 和RLY-5836 (临床I期, NCT05759949)。

### 2.2 分子生成

近年来, 深度学习被广泛用于分子生成中, 通过分子生成进行从头药物设计, 可以获得全新骨架的药物分子, 提高药物分子的新颖性。

2021年, 北京大学来鲁华教授课题组[34]开发了一种的新型药物设计方法DeepLigBuilder, 使用深度生成模型直接在靶点结合口袋内构建和优化配体。首先训练一个可以生成具有有效三维结构的类药分子, 再将基于靶标的信息引入模型中, 从而得到有良好预测结合力的分子。他们使用DeepLigBuilder设计了SARS-CoV-2的主蛋白酶 (main protease, Mpro) 的潜在抑制剂, 实验结果表明DeepLigBuilder发现了与靶蛋白有新相互作用的化合物, 但这些化合物能否有效抑制Mpro还需要进一步的实验验证。

2023年, Yu等[35]通过应用生成模型和基于结构的药物设计发现了一种高效和高选择性的大环细胞周期蛋白依赖性激酶2 (cyclin-dependent kinase 2, CDK2) 抑制剂QR-6401。首先针对10种已发表的CDK2抑制剂开发了一个基于片段的变分自编码器生成模型 (fragment-based variational autoencoder generative model, FBVAE), 生成了3 220个分子文库, 并通过Glide Docking筛选得到10个化合物, 通过结构改造得到了具有一定活性和选择性的化合物**13**, 复合物晶体结构显示CDK2与化合物**13**呈U型结合。随后将化合物**13**利用MacroTransformer方法进行大环化衍生得到了活性和选择性最优的化合物QR-6401 (图7), 该化合物在OVCAR3 (人卵巢癌细胞) 卵巢癌异种移植模型中表现出较好的抗肿瘤疗效。值得注意的是, 近期中国科学院上海药物研究所[36]也报道了一种可以快速识别和生成环状化合物的深度生成方法D3Rings, 为药物发现和开发提供重要的支持。

目前国内外许多大型AI制药公司都开发了自己的分子生成平台, 加速先导化合物的发现与优化进程。2017年英矽智能团队使用美国癌症研究所的公开数据, 训练了一个对抗式自动编码器, 该深度神经网络能够根据目标分子特征生成具有潜在抗癌特性的候选分子。在这项研究中, 英矽智能利用该算法预测了69种化合物, 并合成和测试了其中的4种, 结果表明预测化合物具有良好的活性[37]。2019年英矽智能团队提出了生成张量增强学习 (generative tensorial reinforcement learning, GENTRL) 模型, 并在ZINC分子数据集上训练GENTRL模型后, 利用公开可用的激酶抑制剂数据集, 针对盘状蛋白结构域受体1 (discoidin domain receptor 1, DDR1) 靶点生成了40个候选分子, 并从中筛选出6个化合物用于合成和验证, 体外实验显示其中4个化合物对DDR1具有较好的酶抑制活性, $IC_{50}$值可达到10 nmol·L$^{-1}$。这一耗时仅35天的研究验证了GENTRL模型良好的预测性能[38]。2020年, 研究团队

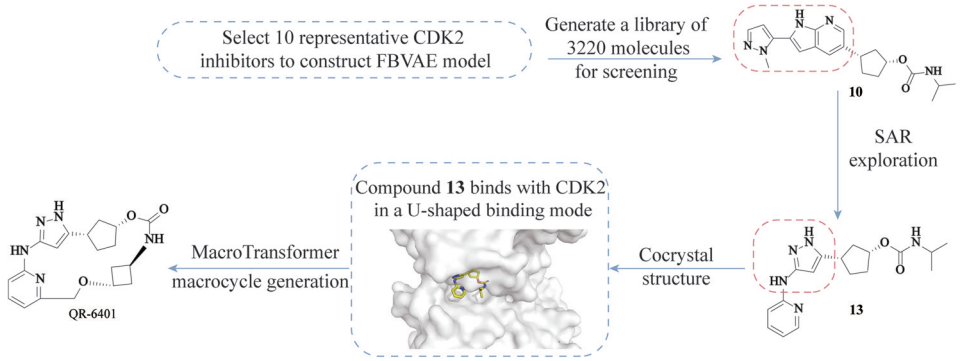


**Figure 7** The design and discovery process of QR-6401. CDK2: Cyclin-dependent kinase 2; FBVAE: Fragment-based variational autoencoder generative model

以GENTRL模型为核心, 发布了集成多种前沿算法模型的AI分子生成工具Chemistry42, 包括生成自编码器、生成式对抗网络、基于流的生成模型、进化算法、语言模型等[39]。该平台的主要优势是个性化奖励机制, Chemistry42会持续采用奖励机制和三维物理结构模块对生成的分子结构进行评估, 并在生成算法辅助下进行多维度评分和优化, 涵盖药效、合成难度、药代动力学特性等。

2021～2023年间, 英矽智能有多个临床阶段的内部自研项目, 涉及纤维化、肿瘤、抗病毒多个领域, 进一步验证其AI分子生成平台的能力。其中最引人注目的是治疗特发性肺纤维化的候选药物ISM001-055, 在不到30个月的时间里, 英矽智能仅用了传统药物开发成本的一小部分, 就将由AI发现和设计的新型药物带入了I期临床试验 (NCT05154240)。2023年2月, ISM001-055获得FDA孤儿药认定, 目前正处于II期临床 (NCT05975983)。值得注意的是, 这是世界上首例完全通过AI设计, 且进入II期临床试验的药物。此外, 英矽智能还利用其生成式AI平台与AlphaFold相结合来识别新型先导化合物, 在细胞周期蛋白依赖性激酶20 (cyclin-dependent kinase 20, CDK20) 抑制剂[40]和盐诱导激酶2 (salt-inducible kinase 2, SIK2) 抑制剂[41]领域取得了重大突破。

### 2.3 AI技术的其他应用

**2.3.1 定量构效关系 (quantitative structure-activity relationship, QSAR)** 1962年, Hansch做出了关于QSAR的开创性工作, QSAR是药物发现的重要组成部分, 可以高效、低成本地预测分子活性和性质。经典的QSAR方法主要通过数学模型建立各种描述符与生物活性之间的定性/定量关系。其中, 描述符包括分子指纹、图或其他数学表示等; 生物活性包括酶活、亲和力、结合自由能、药代动力学性质/毒性 (absorption, distribution, metabolism, excretion/toxicity, ADME/T) 等。由于广泛使用数学模型, QSAR很早就结合了机器学习算法, 应用于QSAR的两个最成功的机器学习算法是RF和DNN[42]。

2019年, Shi等[43]建立了一个基于二维分子图像的CNN模型, 可以有效预测药物分子的药代动力学和毒性等特性, 包括人细胞色素P450酶1A2 (human cytochrome P450 enzyme 1A2, CYP1A2) 和P-糖蛋白抑制活性、血脑屏障穿透性和Ames致突变性。结果表明所建立的CNN模型的预测能力与基于手动结构描述和特征选择的现有机器学习模型相当。说明CNN可以有效地提取与分子ADME/T特性相关的关键图像特征, 并为虚拟筛选和药物设计研究提供有用的工具。

2020年, Arian等[44]使用KNN模型识别活性蛋白激酶抑制剂, 并使用遗传算法提取最佳描述符。为了评估所提出模型的性能, 利用SVM和NB模型进行检验, 结果表明KNN模型的输出结果明显优于其他QSAR模型。

2021年, Zhou等[45]收集了1 785种潜在的人类免疫缺陷病毒1 (human immunodeficiency virus 1, HIV-1) 抑制剂, 通过随机抽样将数据库分为训练集和测试集。基于训练集, 使用NB模型建立HIV-1抑制剂的分类器。通过测试集验证, NB模型预测了88.3%的抑制剂和87.2%的非抑制剂, 相关系数为85.2%, 并且还获得了关键分子片段。

2021年, Kumar等[46]开发了一种预测血脑屏障穿透肽 (blood-brain barrier penetrating peptides, B3PPs) 的方法B3Pred。在从B3Pdb (database of blood brain barrier crossing peptides) 数据库中获得的血脑屏障肽上训练、测试和评估了开发的模型, 其中, 基于RF的模型在前80个选定特征方面表现最佳, 最高准确率为85.08%, 受试者工作曲线下面积 (area under the receiver operating characteristic, AUROC) 值为0.93。B3pred目前可在网页端使用, 包括预测、设计B3PP和扫描蛋白

质序列中的B3PP三个主要模块。

2022年, Wu等[47]建立基于机器学习的分层支持向量回归 (hierarchical support vector regression, HSVR) 模型来预测皮肤渗透系数并揭示内在渗透机制。建立的HSVR模型在训练集、测试集和异常值集这三个数据集中表现出优异的性能, 各种统计评估和验证评估证实了HSVR模型的准确性和预测性。

在过去的几十年中, QSAR已成为各大制药公司药物研发过程中不可或缺的工具之一。QSAR和机器学习方法的融合将药物发现的概念从基于规则驱动转变为数据驱动, 促进了新化合物的发现。但机器学习方法依赖于实验数据, 由于实验条件没有标准化, 实验数据通常是不平衡的、有噪声的, 未来应建立规模大但一致的实验数据集和新的机器学习算法。

**2.3.2 药物重定位** 药物重定位也称老药新用, 是指已经上市的药物, 以及曾经或正处于临床前或临床研究中的候选药物应用于新疾病的过程。与传统的药物研发方法相比, 药物重定位可以节省将药物推向市场所需的早期成本和时间, 进而加快了从基础研究工作到临床治疗的过程。目前进展非常快的AI辅助研发的药物几乎都是老药新用 (表1)。

其中BioXcel Therapeutics是一家利用人工智能方法识别和开发神经科学和免疫肿瘤学领域变革性药物的生物制药公司。基于高选择性$\alpha$2肾上腺素受体激动剂右美托咪定 (dexmedetomidine), BioXcel利用其人工智能平台开发了BXCL501, 直接针对因果激动机制, BXCL501有望快速缓解与精神分裂症及双相障碍相关的急性激越。值得注意的是, BXCL501从获批临床试验到药物上市仅用了4年。

为了快速发现对SARS-CoV-2感染患者具有治疗作用的小分子药物, 药物重定位策略是一种快速而有效的方法。利用人工智能算法和真实世界患者数据验证, 可加速候选药物的确定, 以便顺利推进临床试验[48]。2020年, BenevolentAI公司[49]借助生物医学知识图谱确定了一种药物“巴瑞替尼 (baricitinib)”, 可以减轻新型冠状病毒感染 (corona virus disease 2019, COVID-19) 并减少炎症损伤。巴瑞替尼是Eli Lilly和Incyte联合开发用于治疗类风湿关节炎的上市药物, 目前该药已获批用于单药或与现有抗病毒药物联合治疗COVID-19感染[50]。

2023年, Yasir等[51]使用经过训练的深度学习模型, 对FDA批准的药物库进行新型环氧合酶2 (cyclo-oxygenase-2, COX-2) 抑制剂的筛选。首先, 从DUD-E数据库中获取了由活性和非活性化合物组成的COX-2参考数据集, 对化合物进行特征提取后, 运用来自DeepChem的图卷积网络 (graph convolutional network model, GraphConvMol) 模型构建预测模型。在模型训练完成后, 进一步预测了FDA批准药物的COX-2抑制活性, 结果显示抗癌药物维莫德吉 (vismodegib) 具有COX-2抑制活性。另外预测结果中有一些化合物为已知的COX-2抑制剂, 进一步证明了模型的可靠性。

## 3 总结与展望

本综述首先回顾了基本的ML和DL算法模型, 介绍DL模型在药物研发中的应用, 之后详细介绍了AI技术在虚拟筛选以及分子生成中的应用, 这是目前AI制药行业常用的发现和优化先导化合物的方法, 还对其他的方法如QSAR、药物重定位进行应用介绍, 展现AI在药物研发领域的重要作用。

虽然AI技术在药物研发领域取得了一定的成就, 其仍存在一些问题, 首先数据问题是限制当前AI+新药行业发展的最大壁垒, 如何获取高质量的大样本数据是亟需解决的问题。另外, 小样本数据的有效学习是未来AI技术重要的发展方向。其次, 利用AI技术进行药物设计的内部模型的可解释性也是人们关注的问题,

**Table 1** Examples of AI-assisted drug repurposing. SARS-CoV-2: Severe acute respiratory syndrome coronavirus 2

| Company | Candidate | Original drug | Indication | Clinical stage |
|---|---|---|---|---|
| Healx | HLX-0201 | Sulindac | Fragile X syndrome | Phase II |
| BioXcel Therapeutics | BXCL501 | Dexmedetomidine | Acute treatment of agitation associated with schizophrenia or bipolar I or II disorder in adults | Launched |
| | | | Acute treatment of agitation associated with Alzheimer's dementia | Phase III |
| | | | Opioid use disorder | Phase II |
| | | | Post traumatic stress disorder | Phase I |
| AI Therapeutics | LAM-001 | Rapamycin | Pulmonary hypertension/bronchiolitis obliterans syndrome | Phase II |
| | | | Pulmonary sarcoidosis | Phase I |
| Pharnext | PXT864 | Baclofen & acamprosate | Amyotrophic lateral sclerosis | Phase II |
| | PXT3003 | Baclofen & naltrexone & sorbitol | Charcot-marie-tooth disease type 1A | Phase III |
| Eli Lilly | Baricitinib | Baricitinib | SARS-CoV-2 infection | Launched |
| Evergreen Therapeutics | EG-007 | Unspecified | Carcinoma of endometrium | Phase III |

针对此, 近年来相继报道了 LIME (local interpretable model-agnostic explanations)、SHAP (shapley additive explanations) 等用于从深度神经网络模型中获取解释性信息的方法框架[52]。

总的来看, 人工智能在药物研发的各个领域, 特别是先导化合物的发现和优化环节发挥着重要的作用, AI 辅助药物设计将降低药物发现成本并加快药物研发速度。传统的新药从研发初期到上市可能需要十余年, 消耗数十亿美金, 随着时代发展, 药物研发的成本也在不断提高, 研发过程耗资大、风险高、周期长。随着众多研究团队以及各大知名药企纷纷布局基于 AI 的新药开发, AI+制药行业的发展突飞猛进, 越来越多的 AI 辅助设计药物进入临床管线, 体现 AI 技术在药物设计领域的强大驱动能力。

**作者贡献**: 李紫玥完成了文献收集及论文初稿; 邹毅和徐云根是文章的构思者并进行修改和定稿; 丛开源和吴诗琪参与了文章修改; 朱启华提供了指导性意见。

**利益冲突**: 所有作者均声明没有利益冲突。

## References

[1] Mak KK, Pichika MR. Artificial intelligence in drug development: present status and future prospects [J]. Drug Discov Today, 2019, 24: 773-780.

[2] Bao L, Wang Z, Wu Z, et al. Kinome-wide polypharmacology profiling of small molecules by multi-task graph isomorphism network approach [J]. Acta Pharm Sin B, 2023, 13: 54-67.

[3] Xu Z, Wang X, Zeng S, et al. Applying artificial intelligence for cancer immunotherapy [J]. Acta Pharm Sin B, 2021, 11: 3393-3405.

[4] Yang Y, Zhou D, Zhang X, et al. D3AI-CoV: a deep learning platform for predicting drug targets and for virtual screening against COVID-19 [J]. Brief Bioinform, 2022, 23: 1-16.

[5] Shi Y, Zhang X, Yang Y, et al. D3CARP: a comprehensive platform with multiple-conformation based docking, ligand similarity search and deep learning approaches for target prediction and virtual screening [J]. Comput Biol Med, 2023, 164: 107283-107290.

[6] Vora LK, Gholap AD, Jetha K, et al. Artificial intelligence in pharmaceutical technology and drug delivery design [J]. Pharmaceutics, 2023, 15: 1916-1961.

[7] Zhang B, Zhang L, Chen Q, et al. Harnessing artificial intelligence to improve clinical trial design [J]. Commun Med (Lond), 2023, 3: 191-193.

[8] Yang X, Wang Y, Byrne R, et al. Concepts of artificial intelligence for computer-assisted drug discovery [J]. Chem Rev, 2019, 119: 10520-10594.

[9] Van Engelen JE, Hoos HH. A survey on semi-supervised learning [J]. Mach Learn, 2019, 109: 373-440.

[10] Dike HU, Zhou Y, Deveerasetty KK, et al. Unsupervised learning based on artificial neural network: a review [C]// 2018 IEEE International Conference on Cyborg and Bionic Systems (CBS). New York: IEEE, 2018: 322-327.

[11] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers [J]. Mach Learn, 1997, 29: 131-163.

[12] Cover T, Hart P. Nearest neighbor pattern classification [J]. IEEE Trans Inf Theory, 1967, 13: 21-27.

[13] Cortes C, Vapnik V. Support-vector networks [J]. Mach Learn, 1995, 20: 273-297.

[14] Berk RA. Statistical Learning from A Regression Perspective [M]. Cham, Switzerland: Springer, 2020: 157-337.

[15] Chen T, Guestrin C. XGBoost: a scalable tree boosting system [C]// KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery, 2016: 785-794.

[16] Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities [J]. Proc Natl Acad Sci U S A, 1982, 79: 2554-2558.

[17] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Comput, 1997, 9: 1735-1780.

[18] Yang Y, Zhang R, Li Z, et al. Discovery of highly potent, selective, and orally efficacious p300/CBP histone acetyltransferases inhibitors [J]. J Med Chem, 2020, 63: 1337-1360.

[19] Li Y, Zhang L, Wang Y, et al. Generative deep learning enables the discovery of a potent and selective RIPK1 inhibitor [J]. Nat Commun, 2022, 13: 6891-6908.

[20] LeCun Y, Bengio Y. Convolutional networks for images, speech, and time series [M]// Arbib MA. The Handbook of Brain Theory and Neural Networks. Cambridge: MIT Press, 1998: 255-258.

[21] Noguchi S, Inoue J. Exploration of chemical space guided by pixelCNN for fragment-based *de novo* drug discovery [J]. J Chem Inf Model, 2022, 62: 5988-6001.

[22] Chenthamarakshan V, Payel Das SCH, Strobelt H, et al. CogMol: target-specific and selective drug design for COVID-19 using deep generative models [C]// NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2020: 4320-4332.

[23] Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [C]// NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2672-2680.

[24] Polykovskiy D, Zhebrak A, Vetrov D, et al. Entangled conditional adversarial autoencoder for *de novo* drug discovery [J]. Mol Pharm, 2018, 15: 4398-4405.

[25] Sadybekov AA, Sadybekov AV, Liu Y, et al. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds [J]. Nature, 2022, 601: 452-459.

[26] Lyu J, Wang S, Balius TE, et al. Ultra-large library docking for

discovering new chemotypes [J]. Nature, 2019, 566: 224-229.

[27] Kalliokoski T. Machine learning boosted docking (HASTEN): an open-source tool to accelerate structure-based virtual screening campaigns [J]. Mol Inform, 2021, 40: e2100089.

[28] Sivula T, Yetukuri L, Kalliokoski T, et al. Machine learning-boosted docking enables the efficient structure-based virtual screening of giga-scale enumerated chemical libraries [J]. J Chem Inf Model, 2023, 63: 5773-5783.

[29] Beroza P, Crawford JJ, Ganichkin O, et al. Chemical space docking enables large-scale structure-based virtual screening to discover ROCK1 kinase inhibitors [J]. Nat Commun, 2022, 13: 6447-6456.

[30] Gryniukova A, Kaiser F, Myziuk I, et al. AI-powered virtual screening of large compound libraries leads to the discovery of novel inhibitors of Sirtuin-1 [J]. J Med Chem, 2023, 66: 10241-10251.

[31] Ye WL, Shen C, Xiong GL, et al. Improving docking-based virtual screening ability by integrating multiple energy auxiliary terms from molecular docking scoring [J]. J Chem Inf Model, 2020, 60: 4216-4230.

[32] Morris CJ, Stern JA, Stark B, et al. MILCDock: machine learning enhanced consensus docking for virtual screening in drug discovery [J]. J Chem Inf Model, 2022, 62: 5342-5350.

[33] Zhang X, Shen C, Jiang D, et al. TB-IECS: an accurate machine learning-based scoring function for virtual screening [J]. J Cheminform, 2023, 15: 63-79.

[34] Li Y, Pei J, Lai L. Structure-based de novo drug design using 3D deep generative models [J]. Chem Sci, 2021, 12: 13664-13675.

[35] Yu Y, Huang J, He H, et al. Accelerated discovery of macrocyclic CDK2 inhibitor QR-6401 by generative models and structure-based drug design [J]. ACS Med Chem Lett, 2023, 14: 297-304.

[36] Ma M, Zhang X, Zhou L, et al. D3Rings: a fast and accurate method for ring system identification and deep generation of drug-like cyclic compounds [J]. J Chem Inf Model, 2024, 64: 724-736.

[37] Kadurin A, Aliper A, Kazennov A, et al. The cornucopia of meaningful leads: applying deep adversarial autoencoders for new molecule development in oncology [J]. Oncotarget, 2017, 8: 10883-10890.

[38] Zhavoronkov A, Ivanenkov YA, Aliper A, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors [J]. Nat Biotechnol, 2019, 37: 1038-1040.

[39] Ivanenkov YA, Polykovskiy D, Bezrukov D, et al. Chemistry42: an AI-driven platform for molecular design and optimization [J]. J Chem Inf Model, 2023, 63: 695-701.

[40] Ren F, Ding X, Zheng M, et al. AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor [J]. Chem Sci, 2023, 14: 1443-1452.

[41] Zhu W, Liu X, Li Q, et al. Discovery of novel and selective SIK2 inhibitors by the application of AlphaFold structures and generative models [J]. Bioorg Med Chem, 2023, 91: 117414-117426.

[42] Soares TA, Nunes-Alves A, Mazzolari A, et al. The (re)-evolution of quantitative structure-activity relationship (QSAR) studies propelled by the surge of machine learning methods [J]. J Chem Inf Model, 2022, 62: 5317-5320.

[43] Shi T, Yang Y, Huang S, et al. Molecular image-based convolutional neural network for the prediction of ADMET properties [J]. Chemometr Intell Lab Syst, 2019, 194: 103853-103861.

[44] Arian R, Hariri A, Mehridehnavi A, et al. Protein kinase inhibitors' classification using k-nearest neighbor algorithm [J]. Comput Biol Chem, 2020, 86: 107269-107275.

[45] Zhou J, Hao J, Peng L, et al. Classification and design of HIV-1 integrase inhibitors based on machine learning [J]. Comput Math Methods Med, 2021, 2021: 5559338.

[46] Kumar V, Patiyal S, Dhall A, et al. B3Pred: a random-forest-based method for predicting and designing blood-brain barrier penetrating peptides [J]. Pharmaceutics, 2021, 13: 1237-1249.

[47] Wu YW, Ta GH, Lung YC, et al. In silico prediction of skin permeability using a two-QSAR approach [J]. Pharmaceutics, 2022, 14: 961-984.

[48] Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI [J]. Nature, 2018, 555: 604-610.

[49] Richardson P, Griffin I, Tucker C, et al. Baricitinib as potential treatment for 2019-nCoV acute respiratory disease [J]. Lancet, 2020, 395: e30-e31.

[50] Marconi VC, Ramanan AV, De Bono S, et al. Efficacy and safety of baricitinib for the treatment of hospitalised adults with COVID-19 (COV-BARRIER): a randomised, double-blind, parallel-group, placebo-controlled phase 3 trial [J]. Lancet Respir Med, 2021, 9: 1407-1418.

[51] Yasir M, Park J, Han ET, et al. Vismodegib identified as a novel COX-2 inhibitor *via* deep-learning-based drug repositioning and molecular docking analysis [J]. ACS Omega, 2023, 8: 34160-34170.

[52] Harren T, Matter H, Hessler G, et al. Interpretation of structure-activity relationships in real-world drug design data sets using explainable artificial intelligence [J]. J Chem Inf Model, 2022, 62: 447-462.