基于人工智能模型筛选与生成先导化合物的研究进展

顾志浩,郭文浩,姚和权,李宣仪*,林克江**

(中国药科大学药学院药物化学系,南京 211198)

摘 要 良好的先导化合物对于药物研发具有深远影响,可以提高药物上市的成功率。利用传统方法发现先导化合物存在成本高且耗时的问题,而人工智能(artificial intelligence,AI)可以高效发现良好的先导化合物。本文系统地总结了通过人工智能的筛选模型与生成模型获得先导化合物的研究进展,按照输入信息的类型归纳整理不同的模型,重点介绍了利用筛选模型实现药物重定位和利用生成模型实现多目标药物设计,探讨了人工智能在先导化合物研究领域的发展前景,为人工智能在先导化合物方面的应用提供新的研究思路。

关键词 人工智能;先导化合物;筛选;生成

中图分类号 TP18;R914.2 **文献标志码** A **文章编号** 1000-5048(2023)03-0294-11

doi:10.11665/j.issn.1000-5048.2023042201

引用本文 顾志浩,郭文浩,姚和权,等.基于人工智能模型筛选与生成先导化合物的研究进展[J].中国药科大学学报,2023,54(3):294-304.

Cite this article as: GU Zhihao, GUO Wenhao, YAO Hequan, *et al*. Research progress of the screening and generation of lead compounds based on artificial intelligence model[J]. *J China Pharm Univ*, 2023, 54(3): 294-304.

Research progress of the screening and generation of lead compounds based on artificial intelligence model

GU Zhihao, GUO Wenhao, YAO Hequan, LI Xuanyi*, LIN Kejiang**

Department of Medicinal Chemistry, School of Pharmacy, China Pharmaceutical University, Nanjing 211198, China

Abstract Excellent lead compounds have a profound influence on drug development and can improve the success rate of product launch. It is expensive and time-consuming to discover lead compounds by traditional methods, yet artificial intelligence (AI) can discover good lead compounds efficiently. This article systematically summarizes the research progress of obtaining lead compounds through the screening and generation models of AI, classifies different models according to the type of information input, focuses on drug repurposing by screening model and multi-objective drug design by generation model, and discusses the development prospect of AI in the research field of lead compounds, aiming to provide new research ideas for the application of AI in lead compounds.

Key words artificial intelligence; lead compound; screening; generation

This study was supported by the National Natural Science Foundation of China (No. 81903439)

药物研发是一个复杂且曲折的过程,在很多阶段都存在失败的可能性。药物科学家往往需要花费十余年的时间,耗费数十亿美元,才可能成功上市一个药物[1-2]。随着新药研发难度的加大,新药上市所需的研发成本持续增加。在药物的研发周期中,先导化合物的发现研究至关重要。先导化合物的药理活性和理化性质会直接影响化合物进入临床研究所需时间,从而影响药物的研发进

收稿日期 2023-04-22 **通信作者** *Tel:15261483658 E-mail:xyl_cpu2021@163.com
**Tel:15996210593 E-mail:link@cpu.edu.cn

基金项目 国家自然科学基金资助项目(No. 81903439)

程。高通量筛选（high-throughput screening，HTS）和组合化学是发现先导化合物的常用研究方法，但是这些方法存在昂贵且耗时的缺点[3]。因此，需要其他技术手段助力先导化合物的发现研究以解决现有方法的缺点，从而加快药物研发的进程，降低药物研发的成本。

近年来人工智能（AI）应用于药物研发的各个环节（见图 1），有效地改善了药物研发成本高、成功率低等情况[4-6]。同时，随着药物相关信息的急速膨胀和相关数据库的规范化，人工智能在先导化合物的发现研究中发挥着越来越重要的作用。例如在靶点确证环节，AlphaFold2[7]和 RoseTTA-Fold[8]模型可以预测仅知一维氨基酸序列的蛋白质三维结构；DrugnomeAI[9]可以预测药物靶点及其成药性；DeepPhos[10]可以预测蛋白质的磷酸化位点；DeepCoSI[11]可以预测蛋白质中与配体化合物形成共价键的反应位点，利用这些模型能为先导化合物的设计提供相关的靶点信息。

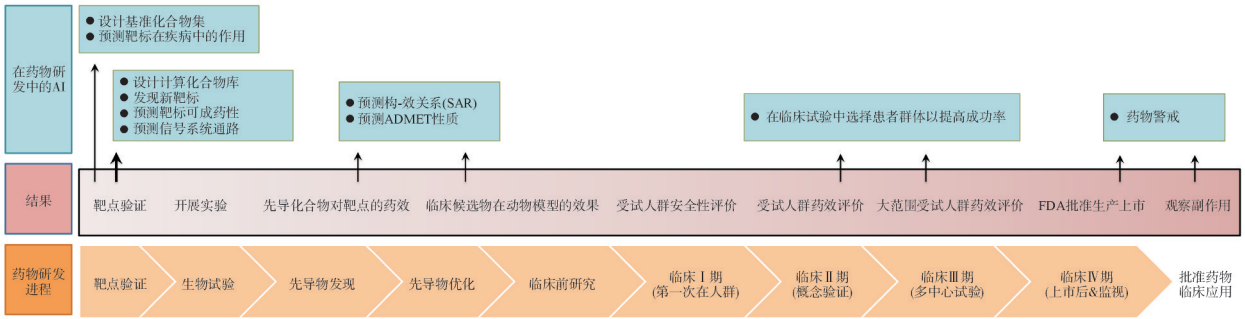

图 1　人工智能(AI)在药物研发过程中的应用

在先导化合物的发现研究中，人工智能模型主要分为两大类：筛选模型和生成模型。筛选模型能快速且低成本地从海量化合物中筛选出良好的先导化合物，生成模型能生成结构新颖且符合要求的先导化合物，并且模型的命中率较传统模型有明显的提高。因此，借助人工智能模型发现先导化合物成为研究的新方向。本文汇总了近几年用于筛选与生成先导化合物的人工智能模型，按照输入信息的类型归纳不同的模型，总结目前人工智能模型发现先导化合物的发展方向，并进行了展望，期望能为该领域的研究人员提供研究思路上的启发。

1　筛选模型

筛选模型能针对特定靶点从海量化合物中筛选出理想的先导化合物。根据输出结果，筛选模型分为分类模型和回归模型。分类模型和回归模型的构建流程见图 2。模型所需要的数据主要来源于 PDB、PDBbind、DrugBank、BindingDB 和 ChEMBL 等化合物或者蛋白质数据库。获取数据后选择满足需求的数据和数据特征，数据特征一般为化合物的一维、二维、三维特征和蛋白质的相对应特征。利用提取的特征训练模型并用相关的指标评价模型的性能。分类模型的性能评价指标与模型混淆矩阵的真正性、假正性、真负性和假负性有关，有准确率（accuracy）、灵敏度（sensitivity，SE）、特效度（specificity，SP）和 AUC 等。准确率表示模型预测正确的结果占总样本的比例；灵敏度表示模型预测的真正性样本占所有正性样本的比例；特效度表示模型预测的真负性样本占所有负性样本的比例；AUC 是受试者工作特征曲线下的面积，AUC 越大，表示分类模型性能越好。回归模型的性能评价指标有平均绝对误差（mean absolute error，MAE）、均方误差（mean squared error，MSE）、均方根误差（root mean square error，RMSE）和决定系数（R^2）等。MAE 反映预测误差的实际情况，值越小说明模型预测结果更接近真实值；MSE 反映数据的变化程度；RMSE 反映预测结果的精密度；R^2 介于 0 至 1 之间，反映回归模型的拟合效果，越接近于 1 说明模型回归拟合的效果越好。

1.1　分类模型

分类模型能定性预测化合物与靶点是否有良好的相互作用。分类模型主要是基于机器学习的模型，常见的机器学习模型有支持向量机（support

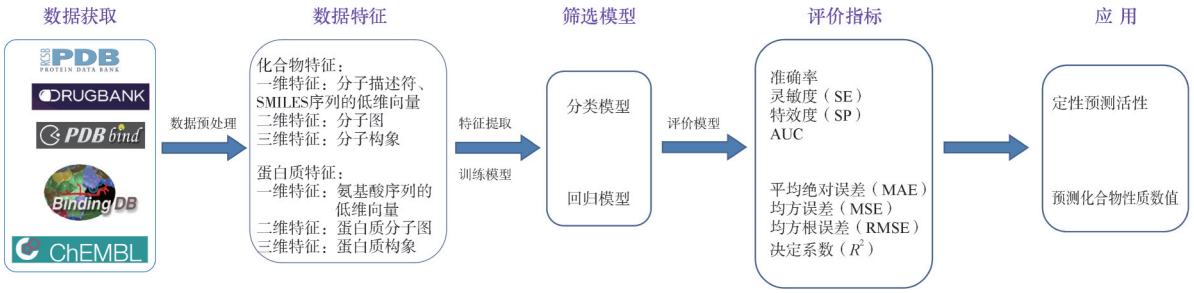

图2 分类模型和回归模型的构建流程图

vector machine，SVM）、随机森林（random forest，RF）、贝叶斯、XGBoost和深度学习等[12]。下面将按照输入信息的类型介绍不同的分类模型。

1.1.1 化合物的定量构-效关系和分子指纹 化合物的定量构-效关系（QSAR）信息和分子指纹是常见的分类模型输入方式。Xie等[13]利用SVM建立筛选间质表皮转化因子（c-Met）抑制剂的模型。模型筛选出75种化合物，其中8种化合物对c-Met有良好的抑制活性。Chen等[14-15]也利用SVM，通过输入QSAR建立模型筛选组织蛋白酶L和止血因子Ⅻa的抑制剂，成功筛选出在体外具有生物活性的先导化合物。Singh等[16]建立根据QSAR信息筛选表皮生长因子受体活性抑制剂的RF模型，得到有较高准确率的模型。Prathipati等[17]输入最大直径为12的扩展类指纹和最大直径为4的功能类指纹训练贝叶斯模型，筛选出抑制结核杆菌的化合物。Tuerkova等[18]建立了基于XGBoost的先导化合物筛选模型，通过向模型中输入QSAR信息训练模型，用于筛选靶向肝脏有机阴离子转运多肽的化合物，得到抑制活性较好的先导化合物。

1.1.2 化合物分子的像素图片 深度学习模型是机器学习模型的一种，其中的卷积神经网络（convolutional neural network，CNN）是擅长处理图像的经典分类模型，利用卷积操作提取特征信息的原理见图3。林克江课题组将化合物分子的像素图片作为输入训练卷积神经网络模型，利用训练好的模型筛选出两个对周期蛋白依赖性激酶4（CDK4）有抑制作用的化合物——吲哚菁绿（IC_{50} = 2 μmol/L）和坎地沙坦酯（IC_{50} = 5 μmol/L）[19]。

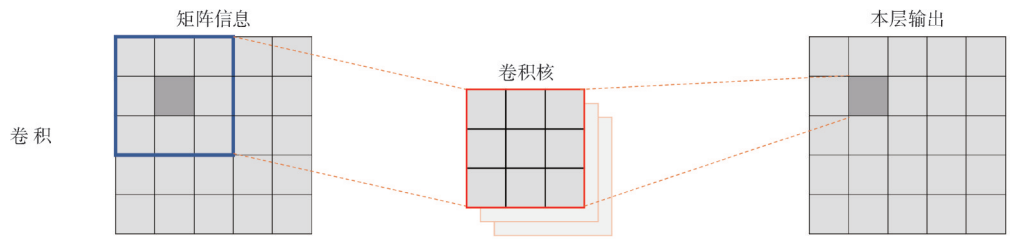

图3 CNN中卷积操作提取特征信息的原理图

1.1.3 化合物-靶点复合物信息 化合物分子的像素图片作为模型的输入方式存在一定的不足：一是像素图片中化学键的长短和粗细影响模型的训练结果；二是像素图片的清晰度也会影响模型的分类性能。而图卷积网络（graph convolutional network，GCN）既避免像素图片表示化合物的不足，也利用卷积操作强大的特征提取能力。图卷积网络是把化合物分子中的原子和化学键分别编码表示成图片中的“节点”和“连接节点的边”，在节点级别上提取特征信息并使用邻接矩阵对相邻节点上提取的信息进行整合[20]。

将图卷积表示的化合物相关信息和靶点的相关信息进行整合，可以输入到分类模型，定性评价药物与靶点之间的相互作用。Peng等[21]提出一种“端到端”的图卷积网络EEG，输入药物和靶点的低维特征向量训练网络模型，用于药物-靶点相互作用（DTI）预测。评估结果表明，EEG在DTI预测方面优于其他模型。Shao等[22]将DTI预测视为链路预测问题，提出一种包含注意力机制的异构图端到端的GCN模型。模型利用图卷积操作获得药

物和靶点的嵌入表示，并在特征信息聚合时引入了图注意机制，使得模型的分类性能好于其他模型。Wu 等[23]从全局和局部两个方面学习药物与蛋白质在分子层面上的特征信息，然后设计了一种虚拟节点，构建了 BridgeDPI 模型用于药物与蛋白质相互作用预测。BridgeDPI 弥补了药物和蛋白质之间在特征表示方面的差距，并使用图卷积操作捕获特征，结果表明模型有较好的分类性能。Tsubaki 等[24]对化合物和相对应的靶点蛋白质用不同的方式表征：用图卷积提取化合物二维分子图的特征向量；用一维卷积提取蛋白质氨基酸序列的特征向量。选择两个常用于比较分类模型性能的数据集——Human 数据集和 C. elegans 数据集作为数据来源，训练模型。训练结果表明网络模型的分类性能好于大多数基于传统机器学习和基于传统化学与生物学特征的预测模型。并且 Tsubaki 在模型中用注意力机制捕捉化合物-蛋白质的结合位点，可视化地表示出蛋白质活性口袋处的氨基酸序列。

简化分子线性输入规范（simplified molecular-input line-entry system，SMILES）序列也是一种常用的化合物分子表示方式，是将分子中的非氢原子和相关化学键按照一定的顺序排列并表示成序列[25]。Zhang 等[26]将 BindingDB 数据库中的化合物 SMILES 序列和蛋白质氨基酸序列分别视为自然语言，用擅长处理自然语言的 Word2Vec[27]将两组序列各转换为低维向量，然后输入分类器中，发现模型能很好地定性预测化合物与蛋白质之间的相互作用。

上海科技大学白芳课题组提出以 GCN 为基础的 DeepPROTACs 模型，预测所设计的 PROTACs 分子对靶蛋白的降解效果[28]。模型的输入信息包含靶蛋白-PROTAC-E3 酶三元复合物中的靶蛋白活性口袋、靶蛋白配体、E3 酶口袋、E3 酶配体以及连接子 5 个部分，使用图卷积操作分别提取前 4 个部分的相应特征信息，使用长短期记忆（long short-term memory，LSTM）提取 PROTACs 连接子的相应特征信息。最后将特征信息整合，输入到全连接层，用训练好的模型预测 PROTACs 分子的降解效果。

1.2　回归模型

除了利用分类模型定性地判断化合物与靶点蛋白是否有良好的相互作用并筛选先导化合物，药物化学家利用回归模型可以预测化合物与靶点蛋白之间亲和力的具体数值，也可以预测化合物的理化性质以及 ADMET 性质的具体数值，将预测值作为筛选先导化合物的判断依据。下面将按照输入方式和用途对回归模型进行分类归纳。

1.2.1　原子对计数

RF-Score 是预测化合物-蛋白质亲和力的经典模型。RF-Score 采用蛋白质-化合物的原子对计数，即定义 4 种蛋白质原子（C、N、O 和 S）和 9 种化合物原子（C、N、O、S、P、F、Cl、Br 和 I），在一定原子间距离阈值内统计蛋白质-化合物原子对出现的次数，将统计结果与标签值（亲和力）一起作为输入，训练模型，最终得到预测复合物亲和力的定量预测模型[29]。后来基于 RF-Score 也构建了一系列衍生模型，2021 年 Sanchez-Cruz 等[30]提出“扩展连接交互特性”模型——ECIF。ECIF 不仅定义蛋白质和化合物中原子的类型，还用多个参数定义原子的状态，即原子的显性化合价、连接的重原子数、连接的氢原子数、是否处于芳香环结构中以及是否处于环状结构中，更详细地表征出原子对中原子的具体情况。ECIF 预测化合物-蛋白质亲和力的准确性优于 RF-Score。

1.2.2　化合物和蛋白质的一维序列

化合物的一维 SMILES 序列和蛋白质的一维氨基酸序列输入到回归模型中预测化合物-蛋白质的亲和力。Ozturk 等[31]提出基于一维 CNN 的 DeepDTA 模型，将化合物和蛋白质的序列信息作为输入信息训练模型，预测复合物的亲和力。Hua 等[32]构建了一种多功能、鲁棒性好的亲和力预测模型 MFR-DTA，解决了序列表示特征效果差以及难以验证化合物-靶点结合区域预测结果的问题。他们设计 BioMLP/CNN 块提取生物序列特征，将单个元素特征和全局特征进行整合，用 Elem-feature 融合块细化提取的特征，构建 Mix-Decoder 块提取化合物-靶点相互作用信息同时预测它们的结合区域。

1.2.3　化合物的二维分子图和蛋白质相关信息

化合物的二维分子图作为回归模型的输入方式用于亲和力预测。Nguyen 等[33]构建基于 GCN 的 GraphDTA 用于药物-靶点的亲和力预测。化合物的二维分子图通过图卷积操作提取相关特征，而蛋白质的氨基酸序列通过卷积操作提取相关特征。用提取的特征训练网络模型，利用训练好的网络模型预测化合物-靶点的亲和力。Yang 等[34]构建的 MGraphDTA 包含多尺度图神经网络

(MGNN)和多尺度卷积神经网络(MCNN)两个模块。MGNN提取化合物的分子图特征,同时将化合物的局部和整体结构特征信息进行整合;MCNN提取蛋白质的氨基酸序列特征,最后用模型预测复合物的亲和力。且为了可视化网络模型,采用梯度加权亲和激活映射(Grad-AAM)的视觉解释方法,从化学角度分析模型。

Krasoulis等[35]将化合物的原子特征和蛋白质活性口袋的表面特征作为输入,构建基于GCN的深度神经虚拟筛选模型DENVIS,实现可扩展和高通量虚拟筛选。化合物的原子特征由图同构网络(GIN)提取,蛋白质活性口袋的表面特征使用特殊的GCN——混合模型网络(MoNet)提取。因为MoNet对流形几何采取卷积运算,通过这种方式提取蛋白质活性口袋表面的几何特征。

1.2.4 *化合物和蛋白质的三维结构* 一维卷积操作处理序列信息,二维卷积操作处理二维图(像素图或者分子图),而三维卷积操作实现对化合物和蛋白质的三维结构进行特征提取。3D-CNN是将复合物置于三维网格中,通过卷积的方式提取特征并将特征信息输入网络。Ragoza等[36]和Koes等[37]构建多个3D-CNN模型预测复合物的亲和力,同时筛选能与靶点蛋白结合的化合物三维构象。Jimenez等[38]构建的K-DEEP也是一种用于亲和力预测的3D-CNN模型。

1.2.5 *回归模型预测ADMET性质的具体数值* 在先导化合物的研究中,不仅关注化合物与靶点蛋白质之间的相互作用,还关注化合物的吸收、分布、代谢、排泄和毒性(ADMET)性质。较差的ADMET性质是导致药物研发后期失败的原因,因此利用回归模型预测ADMET性质也是目前的研究热点。拜耳公司利用化合物在体外/体内的实验数据和合适的化合物描述方式训练深度神经网络,预测化合物的ADMET性质,助力药物的后期研发[39]。默克公司利用大量分子描述符描述化合物,根据化合物的毒性数据训练深度神经网络并得到模型DeepTox,在Tox21预测毒性挑战中有出色的发挥[40]。中南大学曹东升课题组开发便于使用的、基于多任务图注意力框架的网络模型ADMETlab 2.0预测化合物的ADMET性质[41]。向ADMETlab 2.0的输入端输入化合物的SMILES序列或者化合物结构,能快速预测出化合物分子ADMET性质的具体数值。

1.3 *药物重定位*

筛选模型不仅可以从海量的化合物中筛选出先导化合物,也可以从上市药物或通过临床Ⅰ期的候选药物中筛选出针对其他疾病相关靶点的先导化合物。从上市药物和候选药物中得到先导化合物的方法称为药物重定位(又被称为“老药新用”)。药物重定位最明显的作用是在面对突然暴发的传染病且无药可用时,能快速得到有治疗效果且毒性较小的化合物。2020年新型冠状病毒肆虐全球,由此引起的疫情造成全球死亡人数剧增。为了快速得到能治疗新型冠状病毒感染且安全的化合物,药物化学家利用人工智能筛选模型从“老药”中筛选对新型冠状病毒有良好的抑制作用的先导化合物[42-43]。例如,亚马逊上海AI实验室联合国内多个研究团队构建了药物重定位知识图谱(DRKG)以及用于药物重定位的机器学习,进行药物重定位研究[44]。利用先进的深度图学习方法学习DRKG中表示药物特征的低维向量,并预测药物与新型冠状病毒靶点结合的可能性,将药物重定位的预测问题转换为药物和新型冠状病毒之间的置信度评估问题。

2 生成模型

筛选模型筛选得到的先导化合物可以直接购买,但是购买的化合物分子存在结构新颖性低,不容易突破专利壁垒等问题。而利用生成模型生成新颖结构的化合物分子,拓展化合物化学空间,避开结构专利壁垒,因此生成模型成为人工智能用于先导化合物发现研究的主要模型[45]。生成模型的构建流程见图4,其中生成模型中数据获取和数据特征提取与筛选模型一致。

图4 生成模型的构建流程图

在生成模型的评价指标中，Inception score（IS）和 Fréchet inception distance（FID）用来评价生成模型的性能。IS 反映生成化合物分子的多样性，值越大表明生成的化合物分子具有更好的多样性。FID 反映生成的化合物分子与真实的化合物分子在特征空间的距离，值越小表明生成的化合物分子的相关特征越接近于真实的化合物分子。可合成性和类药性主要评价具体生成的化合物分子，可合成性反映合成化合物分子的难易程度，类药性反映化合物分子是否具有成为药物的可能。

生成模型大致可以分为 4 类：循环神经网络（recurrent neural network，RNN）、强化学习（reinforcement learning，RL）、自编码器（autoencoder）和生成对抗网络（generative adversarial network，GAN）[46-47]。生成模型的输入信息类型有化合物的 SMILES 序列、化合物的二维分子图以及化合物和蛋白质的三维结构信息。

2.1　化合物的 SMILES 序列

化合物的 SMILES 序列可以输入到 RNN、RL 和自编码器等生成模型中，用于生成结构新颖的化合物分子。RNN 会对前面输入的信息进行记忆并应用于当前的计算中，因此擅长处理自然语言以及序列信息，其处理序列信息的机制见图 5。林克江课题组将 SMILES 序列输入到基于 LSTM 的 RNN，审查和调整生成的 SMILES 序列，最终得到 1 个结构新颖且药理活性较好的 Pim1 激酶抑制剂和 2 个 CDK4 抑制剂[48]。Segler 等[49]也将 SMILES 序列输入到 RNN，用于生成大量具有合适物理化学性质的新结构化合物分子 SMILES 序列。

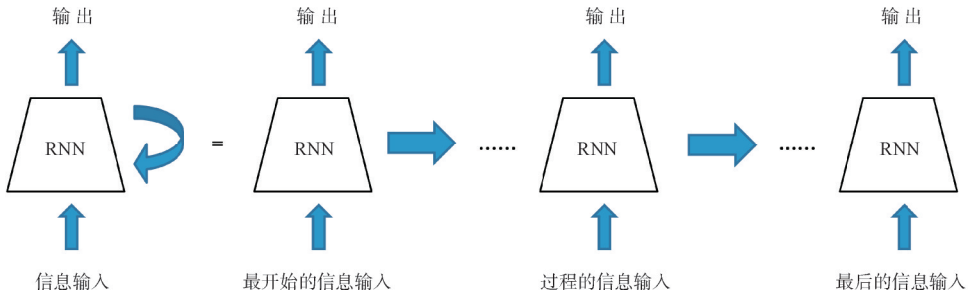

图 5　RNN 处理序列信息的示意图

RL 模型在获取样本数据的同时完成模型的更新并利用当前的模型指导下一步行动，使得模型不断迭代重复直到收敛。阿斯利康公司 Olivecrona 等[50]将 SMILES 序列输入进 RL-RNN 模型，生成满足需求的新结构化合物分子的 SMILES 序列。Popova 等[51]将 SMILES 序列输入到基于结构进化的强化学习模型 ReLeaSE 中，生成得到满足目标需求的化合物分子。英矽智能（Insilico Medicine）在 2019 年提出生成张量强化学习模型 GENTRL，向模型输入化合物的 SMILES 序列以生成靶向盘状结构域受体 1（DDR1）的新颖化合物。21 d 利用模型生成结构新颖的 DDR1 抑制剂分子并完成合成，46 d 就完成临床试验前的全部研究，得到合适的化合物进入临床研究[52]。由此可见，人工智能模型可以解决药物研发过程中耗时长，研发成本高等问题。

中山大学杨跃东课题组利用生成模型设计靶向含溴结构域蛋白 4（BRD4）的 PROTACs 分子[53]。将 PROTACs 分子的 SMILES 序列作为输入，使模型生成结构新颖的 PROTACs 分子，分别根据分子的药代动力学属性、新颖性、结构缺陷、分子动力学模拟结果和分子可合成性等条件对新生成 PROTACs 分子进行筛选。最终筛选得到 6 个生成的 PROTACs 分子，合成并测试化合物分子对 BRD4 的降解活性，发现其中 3 个分子有良好的降解活性。

自编码器采用编码—解码的方式训练模型，编码部分是将输入序列表示为带有语义的向量，解码部分是将编码生成的向量解码成目标文本序列，所以自编码器本质上是一个自然语言模型。自编码器的编码-解码过程示意图见图 6。变分自编码器（variational autoencoders，VAE）用于生成全新结构分子的 SMILES 序列并预测分子的相关性质[54]。Polykovskiy 等[55]描述了一种新颖的生成模型——纠缠条件对抗自编码器 ECAAE，可以基于各种要求（如化合物活性、溶解度和可合成性等）

生成结构新颖的化合物分子SMILES序列。作者使用ECAAE生成与类风湿性关节炎、银屑病和白癜风等疾病有关的Janus激酶3(JAK3)抑制剂,并且体外测试结果表明生成的化合物有良好的药理活性和选择性。

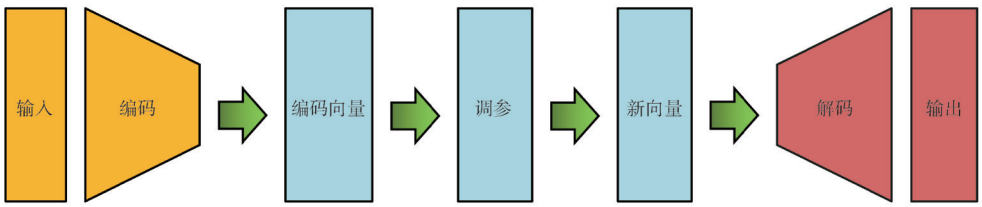

图6 自编码器的编码-解码过程示意图

Wang等[56]提出了一种以SMILES序列为输入方式的人工智能模型算法ChemistGA,该算法将传统的启发式遗传算法与深度学习算法相结合,使用基于Transformer的反应预测算法作为遗传算法与深度学习算法杂交的核心。ChemistGA模型保留了遗传算法的优势,同时极大地提高了生成含有期望特性以及可合成性的分子的比例。

SMILES序列虽然广泛应用于生成模型并能成功得到新颖的化合物分子,但存在一定的不足。生成模型不会关注SMILES序列的语法规则,导致大多新生成的序列不能表示为分子,需要对新生成的序列进行化学语法和原子排序的鉴别或者人工审查,才能得到正确的SMILES序列。

2.2 化合物的二维分子图

二维分子图作为输入方式,使生成模型完全生成有效的化合物分子,避免输入SMILES序列所引起的语法问题。因此越来越多的生成模型通过输入二维分子图训练模型,并生成结构新颖的化合物二维分子图。

Khemchandani等[57]构建基于二维分子图的强化学习生成模型——DeepGraphMolGen,采用图卷积策略网络(GCPNs)生成化合物分子结构。GCPNs能预测给定分子状态的下一个动作,在分子生成过程中会受到网络其他指标的指导,最终生成有效的分子。

GAN是可以生成二维分子图的生成模型,包含生成器和判别器。生成器生成与原始数据相似的数据信息,判别器判别给定的数据信息是否真实。经过交替优化训练,使GAN模型性能整体得到提升,最终GAN模型生成表示化合物结构的二维分子图。GAN模型生成二维分子图的流程见图7。Maziarka等[58]提出GAN衍生的生成模型——Mol-CycleGAN,通过向模型输入二维分子图使模型生成与原始化合物结构相似且性质优良的化合物。

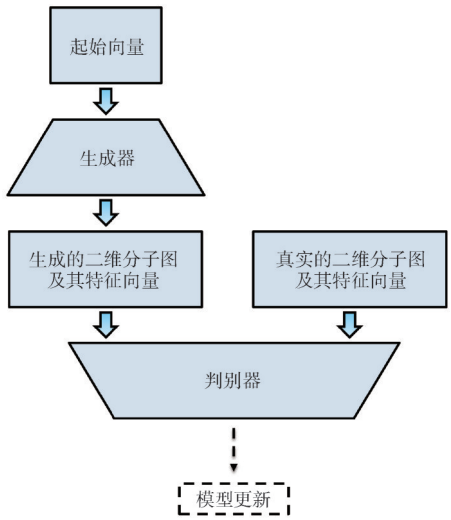

图7 GAN模型生成二维分子图的流程示意图

2.3 化合物和蛋白质的三维结构

上述生成模型都只是以化合物的SMILES序列(一维)信息和分子图(二维)信息作为输入和输出方式,没有关注化合物的三维结构信息以及靶点蛋白质的相关信息。化合物产生药理活性的本质是化合物与靶点蛋白在三维空间中结合从而改变靶点蛋白的生理活性,因此化合物和靶点蛋白的三维信息对于生成新颖结构的化合物分子起到重要作用。浙江大学侯廷军课题组提出一种以蛋白质-配体复合物的三维结合构象作为输入、基于自编码器的分子生成模型RELATION[59]。该模型以复合物的三维网格构象作为输入,采用结构域分离网络促进化合物和蛋白质几何特征之间的信

息交换，将提取到的特征转移到潜在空间中表示，引入药效团约束和基于BO的采样优化，高效生成与靶点蛋白具有良好亲和力以及具有较好药效团特征的新结构分子。利用该模型设计了新颖、有效、多样且亲和力高的蛋白激酶AKT1和CDK2的潜在抑制剂。

Ragoza等[60]提出了一种新的基于条件变分自编码器和生成对抗网络（CVAE-GAN）的生成模型，通过采样蛋白质-化合物结合区域相互作用的条件分布，首先生成化合物分子的原子密度图，然后在原子密度图的基础上构建有效的分子构象，实现直接在受体蛋白质的口袋结构中生成与蛋白质结合的三维结构分子。

2.4　设计满足多目标需求的化合物分子

在先导化合物发现研究中，不仅只关注化合物的药理活性，还兼顾化合物的其他性质，例如可合成性、类药性和ADMET性质等。因此可以根据输入的化合物信息构建能兼顾并优化多个分子性质的生成模型，得到满足条件的化合物分子。侯廷军课题组提出了一种集知识蒸馏、条件Transformer和强化学习于一体的多约束分子生成模型MCMG，用于生成同时具有理想理化性质和药理活性的新颖结构化合物分子[61]。生成模型框架由一个条件Transformer模块，一个实现知识蒸馏功能的RNN模块和一个RL微调模块构成。MCMG以SMILES序列作为输入，用条件Transformer训练分子生成模型，并将结构-性质关系纳入有偏生成过程中。然后利用知识蒸馏和强化学习等方式降低模型的复杂度并对模型进行微调，使模型生成的分子具有更好的结构多样性。Li等[62]提出基于多目标要求的条件图生成模型，向模型中输入二维分子图，直接得到分子结构，并且有效输出率高于SMILES序列。

基于多目标要求的生成模型是目前人工智能模型应用于先导化合物发现研究的热点，但是如何选取合适的指标来描述化合物的目标性质和调节指标对模型的影响程度是构建多目标要求的生成模型需要重点解决的问题。

通过生成模型生成的满足多目标要求的先导化合物往往需要通过合成获得，而化合物结构的新颖性会影响化合物合成进度。目前，有多种人工智能模型可以预测化合物的合成路线和合成条件，提高化合物的合成效率[63-65]。

3　总结与展望

合适的先导化合物对于药物的整个研发周期具有重要意义，而高通量筛选和组合化学等传统方法不适合大规模筛选先导化合物，也不太适合设计结构新颖的先导化合物。人工智能助力于先导化合物的研究并取得令人满意的结果。本文总结了近几年用于先导化合物发现研究的筛选模型和生成模型，并按照输入信息类型的不同对每类模型进行整理，突出模型的特点和实际应用。

在先导化合物的研究中，筛选模型和生成模型具有各自的优势。筛选模型可以快速从化合物库中找到合适的先导化合物进行后续研究。更重要的是，筛选模型可以用于“药物重定位”研究，快速得到对重大卫生传染病有治疗效果且安全的药物。例如2020年新型冠状病毒感染疫情暴发，筛选模型快速地从“老药”中得到对新冠疫情有治疗作用的化合物进行研究，推进针对新型冠状病毒的药物研发。生成模型能生成结构新颖且抑制活性良好的化合物分子，并且可以设计满足多目标要求的化合物分子，比筛选模型得到的先导化合物更适合进行后续研究。就化合物新颖性而言，生成模型更具有研究意义。

从利用人工智能模型筛选或者生成满足单一需求的化合物，到利用生成模型设计并得到同时满足药理活性和ADMET性质等多个目标要求的化合物分子，体现了人工智能模型在先导化合物研究中的重要性。在设计得到满足多目标要求化合物分子的研究中，需要选择合适的描述指标体现生成分子的相关性质。例如在药物的ADMET性质中，每一个性质都包含多种描述指标，如何选择合适的指标表示化合物分子的相关性质，是设计满足多目标要求化合物分子的研究重点。目前，研究人员尝试使用知识蒸馏等方法寻找合适的描述指标[61,66]。另外，选择合适的指标后，如何调整不同指标对模型的影响程度以得到最好的生成模型也是研究的重点。这两方面的研究可能是未来人工智能在先导化合物发现研究领域的主要研究方向。

目前，利用人工智能模型设计得到的化合物进入了临床研究并取得令人满意的结果。前文介

绍英矽智能利用人工智能模型设计并合成出结构新颖的DDR1抑制剂,在短时间内完成了临床试验前的全部研究[52]。ISM001-055是英矽智能利用人工智能模型得到的另一个化合物,用于治疗与特定靶点有关的特发性肺纤维化(IPF),目前已完成临床Ⅰ期的健康志愿者给药试验。英矽智能仅耗时18个月就完成从发现靶点到确定ISM001-055为临床前候选药物的相关研究,所投入的研发成本仅为260万美元,而传统的药物研发或者基于计算机辅助药物设计的药物研发很难达到如此成就。例如靶向突变的K-鼠类肉瘤病毒癌基因(KRas)共价抑制剂索托拉西布:2013年发现KRas酶$G^{12}C$的靶点并且得到能共价结合的亲电片段,但是经过5年多的临床前研究并花费数亿美元,直到2018年才得到能进入临床研究的AMG-510,即后来的索托拉西布[67]。比较ISM001-055和索托拉西布的临床前研究所需的研发时间和研发成本就可以明显发现人工智能可以加快药物研发的进程和降低药物研发的成本。因此,希望越来越多利用人工智能模型设计的先导化合物能进入临床研究,促进药物研发。

References

[1] Smietana K, Siatkowski M, Møller M. Trends in clinical success rates[J]. *Nat Rev Drug Discov*, 2016, **15**(6): 379-380.

[2] DiMasi JA, Grabowski HG, Hansen RW. The cost of drug development[J]. *N Engl J Med*, 2015, **372**(20): 1972.

[3] Schirle M, Jenkins JL. Identifying compound efficacy targets in phenotypic drug discovery[J]. *Drug Discov Today*, 2016, **21**(1): 82-89.

[4] Mak KK, Pichika MR. Artificial intelligence in drug development: present status and future prospects[J]. *Drug Discov Today*, 2019, **24**(3): 773-780.

[5] Chen HM, Engkvist O, Wang YH, *et al*. The rise of deep learning in drug discovery[J]. *Drug Discov Today*, 2018, **23**(6): 1241-1250.

[6] Rifaioglu AS, Atas H, Martin MJ, *et al*. Recent applications of deep learning and machine intelligence on *in silico* drug discovery: methods, tools and databases[J]. *Brief Bioinform*, 2019, **20**(5): 1878-1912.

[7] Jumper J, Evans R, Pritzel A, *et al*. Highly accurate protein structure prediction with AlphaFold[J]. *Nature*, 2021, **596**(7873): 583-589.

[8] Baek M, DiMaio F, Anishchenko I, *et al*. Accurate prediction of protein structures and interactions using a three-track neural network[J]. *Science*, 2021, **373**(6557): 871-876.

[9] Raies A, Tulodziecka E, Stainer J, *et al*. DrugnomeAI is an ensemble machine-learning framework for predicting druggability of candidate drug targets[J]. *Commun Biol*, 2022, **5**(1): 1291.

[10] Luo FL, Wang MH, Liu Y, *et al*. DeepPhos: prediction of protein phosphorylation sites with deep learning[J]. *Bioinformatics*, 2019, **35**(16): 2766-2773.

[11] Du HY, Jiang DJ, Gao JB, *et al*. Proteome-wide profiling of the covalent-druggable cysteines with a structure-based deep graph learning network[J]. *Research*, 2022, **2022**: 9873564.

[12] Patel L, Shukla T, Huang XZ, *et al*. Machine learning methods in drug discovery[J]. *Molecules*, 2020, **25**(22): 5277.

[13] Xie QQ, Zhong L, Pan YL, *et al*. Combined SVM-based and docking-based virtual screening for retrieving novel inhibitors of c-Met[J]. *Eur J Med Chem*, 2011, **46**(9): 3675-3680.

[14] Chen JJF, Visco DP Jr. Developing an *in silico* pipeline for faster drug candidate discovery: virtual high throughput screening with the Signature molecular descriptor using support vector machine models[J]. *Chem Eng Sci*, 2017, **159**: 31-42.

[15] Chen JJF, Visco DP Jr. Identifying novel factor XIIa inhibitors with PCA-GA-SVM developed vHTS models[J]. *Eur J Med Chem*, 2017, **140**: 31-41.

[16] Singh H, Singh S, Singla D, *et al*. QSAR based model for discriminating EGFR inhibitors and non-inhibitors using Random forest[J]. *Biol Direct*, 2015, **10**: 10.

[17] Prathipati P, Ma NL, Keller TH. Global Bayesian models for the prioritization of antitubercular agents[J]. *J Chem Inf Model*, 2008, **48**(12): 2362-2370.

[18] Tuerkova A, Bongers BJ, Norinder U, *et al*. Identifying novel inhibitors for hepatic organic anion transporting polypeptides by machine learning-based virtual screening[J]. *J Chem Inf Model*, 2022, **62**(24): 6323-6335.

[19] Xu YQ, Chen PP, Lin XH, *et al*. Discovery of CDK4 inhibitors by convolutional neural networks[J]. *Future Med Chem*, 2019, **11**(3): 165-177.

[20] Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, *et al*. Convolutional networks on graphs for learning molecular fingerprints [C]//Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. New York: ACM, 2015: 2224-2232.

[21] Peng JJ, Wang YX, Guan JJ, *et al*. An end-to-end heterogeneous graph representation learning-based framework for drug-target interaction prediction[J]. *Brief Bioinform*, 2021, **22**(5): bbaa430.

[22] Shao KH, Zhang YH, Wen YQ, *et al*. DTI-HETA: prediction of drug-target interactions based on GCN and GAT on heterogeneous graph[J]. *Brief Bioinform*, 2022, **23**(3): bbac109.

[23] Wu YF, Gao M, Zeng M, *et al*. BridgeDPI: a novel Graph Neural Network for predicting drug-protein interactions[J]. *Bioinfor-*

matics, 2022, **38**(9): 2571-2578.

[24] Tsubaki M, Tomii K, Jun SS. Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences[J]. *Bioinformatics*, 2019, **35**(2): 309-318.

[25] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules[J]. *J Chem Inf Comput Sci*, 1988, **28**(1): 31-36.

[26] Zhang YF, Wang XG, Kaushik AC, *et al*. SPVec: a Word2Vec-inspired feature representation method for drug-target interaction prediction[J]. *Front Chem*, 2019, **7**: 895.

[27] Xu YQ, Yao HQ, Lin KJ. An overview of neural networks for drug discovery and the inputs used[J]. *Expert Opin Drug Discov*, 2018, **13**(12): 1091-1102.

[28] Li FL, Hu QY, Zhang XL, *et al*. DeepPROTACs is a deep learning-based targeted degradation predictor for PROTACs[J]. *Nat Commun*, 2022, **13**(1): 7133.

[29] Ballester PJ, Mitchell JB. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking[J]. *Bioinformatics*, 2010, **26**(9): 1169-1175.

[30] Sánchez-Cruz N, Medina-Franco JL, Mestres J, *et al*. Extended connectivity interaction features: improving binding affinity prediction through chemical description[J]. *Bioinformatics*, 2021, **37**(10): 1376-1382.

[31] Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction[J]. *Bioinformatics*, 2018, **34**(17): i821-i829.

[32] Hua Y, Song XN, Feng ZH, *et al*. MFR-DTA: a multi-functional and robust model for predicting drug-target binding affinity and region[J]. *Bioinformatics*, 2023, **39**(2): btad056.

[33] Nguyen T, Le H, Quinn TP, *et al*. GraphDTA: predicting drug-target binding affinity with graph neural networks[J]. *Bioinformatics*, 2021, **37**(8): 1140-1147.

[34] Yang ZD, Zhong WH, Zhao L, *et al*. MGraphDTA: deep multiscale graph neural network for explainable drug-target binding affinity prediction[J]. *Chem Sci*, 2022, **13**(3): 816-833.

[35] Krasoulis A, Antonopoulos N, Pitsikalis V, *et al*. DENVIS: scalable and high-throughput virtual screening using graph neural networks with atomic and surface protein pocket features[J]. *J Chem Inf Model*, 2022, **62**(19): 4642-4659.

[36] Ragoza M, Hochuli J, Idrobo E, *et al*. Protein-ligand scoring with convolutional neural networks[J]. *J Chem Inf Model*, 2017, **57**(4): 942-957.

[37] Hochuli J, Helbling A, Skaist T, *et al*. Visualizing convolutional neural network protein-ligand scoring[J]. *J Mol Graph Model*, 2018, **84**: 96-108.

[38] Jiménez J, Škalič M, Martínez-Rosell G, *et al*. K_{DEEP}: protein-ligand absolute binding affinity prediction via 3D-convolutional neural networks[J]. *J Chem Inf Model*, 2018, **58**(2): 287-296.

[39] Göller AH, Kuhnke L, Montanari F, *et al*. Bayer's in silico ADMET platform: a journey of machine learning over the past two decades[J]. *Drug Discov Today*, 2020, **25**(9): 1702-1709.

[40] Mayr A, Klambauer G, Unterthiner T, *et al*. DeepTox: toxicity prediction using deep learning[J]. *Front Environ Sci*, 2016, **3**: 80.

[41] Xiong GL, Wu ZX, Yi JC, *et al*. ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties[J]. *Nucleic Acids Res*, 2021, **49**(W1): W5-W14.

[42] Lu L, Qin JL, Chen JD, *et al*. Recent computational drug repositioning strategies against SARS-CoV-2[J]. *Comput Struct Biotechnol J*, 2022, **20**: 5713-5728.

[43] Su XR, Hu L, You ZH, *et al*. A deep learning method for repurposing antiviral drugs against new viruses via multi-view non-negative matrix factorization and its application to SARS-CoV-2 [J]. *Brief Bioinform*, 2022, **23**(1): bbab526.

[44] Zeng XX, Song X, Ma TF, *et al*. Repurpose open data to discover therapeutics for COVID-19 using deep learning[J]. *J Proteome Res*, 2020, **19**(11): 4624-4636.

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

[46] Xu YJ, Lin KJ, Wang SW, *et al*. Deep learning for molecular generation[J]. *Future Med Chem*, 2019, **11**(6): 567-597.

[47] Tong XC, Liu XH, Tan XQ, *et al*. Generative models for *de novo* drug design[J]. *J Med Chem*, 2021, **64**(19): 14011-14027.

[48] Li XY, Xu YQ, Yao HQ, *et al*. Chemical space exploration based on recurrent neural networks: applications in discovering kinase inhibitors[J]. *J Cheminform*, 2020, **12**(1): 42.

[49] Segler MHS, Kogej T, Tyrchan C, *et al*. Generating focused molecule libraries for drug discovery with recurrent neural networks [J]. *ACS Cent Sci*, 2018, **4**(1): 120-131.

[50] Olivecrona M, Blaschke T, Engkvist O, *et al*. Molecular de-novo design through deep reinforcement learning[J]. *J Cheminform*, 2017, **9**(1): 48.

[51] Popova M, Isayev O, Tropsha A. Deep reinforcement learning for *de novo* drug design[J]. *Sci Adv*, 2018, **4**(7): eaap7885.

[52] Zhavoronkov A, Ivanenkov YA, Aliper A, *et al*. Deep learning enables rapid identification of potent DDR1 kinase inhibitors [J]. *Nat Biotechnol*, 2019, **37**(9): 1038-1040.

[53] Zheng SJ, Tan YH, Wang ZY, *et al*. Accelerated rational PROTAC design via deep learning and molecular simulations[J]. *Nat Mach Intell*, 2022, **4**(9): 739-748.

[54] Gómez-Bombarelli R, Wei JN, Duvenaud D, *et al*. Automatic chemical design using a data-driven continuous representation of molecules[J]. *ACS Cent Sci*, 2018, **4**(2): 268-276.

[55] Polykovskiy D, Zhebrak A, Vetrov D, *et al*. Entangled conditional adversarial autoencoder for *de novo* drug discovery[J]. *Mol Pharm*, 2018, **15**(10): 4398-4405.

[56] Wang JK, Wang XR, Sun HY, *et al*. ChemistGA: a chemical synthesizable accessible molecular generation algorithm for

real-world drug discovery[J]. *J Med Chem*, 2022, **65**(18): 12482-12496.

[57] 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.

[58] Maziarka Ł, Pocha A, Kaczmarczyk J, *et al*. Mol-CycleGAN: a generative model for molecular optimization[J]. *J Cheminform*, 2020, **12**(1): 2.

[59] Wang MY, Hsieh CY, Wang JK, *et al*. RELATION: a deep generative model for structure-based de novo drug design[J]. *J Med Chem*, 2022, **65**(13): 9478-9492.

[60] Ragoza M, Masuda T, Koes DR. Generating 3D molecules conditional on receptor binding sites with deep generative models [J]. *Chem Sci*, 2022, **13**(9): 2701-2713.

[61] Wang JK, Hsieh CY, Wang MY, *et al*. Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning[J]. *Nat Mach Intell*, 2021, **3**(10): 914-922.

[62] Li YB, Zhang LR, Liu ZM. Multi-objective *de novo* drug design with conditional graph generative model[J]. *J Cheminform*, 2018, **10**(1): 33.

[63] Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI[J]. *Nature*, 2018, **555**(7698): 604-610.

[64] Ishida S, Terayama K, Kojima R, *et al*. Prediction and interpretable visualization of retrosynthetic reactions using graph convolutional networks[J]. *J Chem Inf Model*, 2019, **59**(12): 5026-5033.

[65] Schwaller P, Petraglia R, Zullo V, *et al*. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy[J]. *Chem Sci*, 2020, **11**(12): 3316-3325.

[66] Yuan WN, Chen GX, Chen CYC. FusionDTA: attention-based feature polymerizer and knowledge distillation for drug-target binding affinity prediction[J]. *Brief Bioinform*, 2022, **23**(1): bbab506.

[67] Goebel L, Müller MP, Goody RS, *et al*. KRasG^{12}C inhibitors in clinical trials: a short historical perspective[J]. *RSC Med Chem*, 2020, **11**(7): 760-770.

[**专家介绍**] 李宣仪,博士后,中国药科大学药学院药物化学系助理研究员。研究方向:(1)利用人工智能技术设计小分子活性化合物,理解大分子与小分子相互作用,探索新药发现的新途径。(2)利用计算机辅助药物设计手段研究重大疾病的相关靶标结构与活性关系,发现具有潜力的先导化合物及创新药物。主持国家自然科学基金青年项目,江苏省自然科学基金青年项目,中国博士后面上项目等课题项目。任职至今以第一作者在 *Journal of Cheminformatics*, *Organic Letters*, *Expert Opinion on Therapeutic Patents* 等期刊杂志发表论文数篇。

[**专家介绍**] 林克江,教授,博士生导师。获中国药科大学药物化学硕士、博士学位,美国加州大学尔湾分校访问学者。从事新药研究与开发的教学与科研工作,专注于利用人工智能及计算机辅助药物设计手段,设计并发现具有自主知识产权的全新药物。利用深度学习设计小分子活性化合物、理解大分子与小分子相互作用,探索新药发现的新途径。利用计算机辅助药物设计手段研究重大疾病的相关靶标结构与活性关系,发现具有潜力的先导化合物及创新药物。主持及参加了多项省、国家级自然科学基金项目、重大新药创制等科研项目,获得了多项药物专利授权及先导化合物发现新技术专利授权,并在国内、国际期刊上发表学术论文数十篇。