大数据和人工智能技术用于计算机辅助药物设计的研究进展

杨艳伟[1#]，胡文元[2#]，林　志[1]，霍桂桃[1]，张　頔[1]，李双星[1]，张亚群[2]，闫振龙[2]，屈　哲[1*]，吕建军[2*]

1. 中国食品药品检定研究院，国家药物安全评价监测中心，北京市重点实验室，北京　100176

2. 益诺思生物技术南通有限公司，江苏 南通　226133

摘　要： 大数据和人工智能（AI）技术不仅可以对海量的生物医学数据进行准确和综合地分析，而且可以帮助构建药物设计领域的预测模型。随着算法和统计方法的发展，大数据和AI技术已应用于计算机辅助药物设计（CADD）。CADD可被用于有效克服药物设计领域的困难，从而高效地设计和开发新药。介绍了药物设计和发现过程中数据预处理和建模步骤、药物设计和发现过程中基于AI的建模方法、大数据和AI技术在CADD中的最新应用，以期为我国药物设计和开发提供参考。

关键词： 药物设计；大数据；人工智能；数据分析；预测模型

中图分类号： R914.2　　**文献标志码：** A　　**文章编号：** 1674-6376（2023）06-1369-07

DOI： 10.7501/j.issn.1674-6376.2023.06.025

Research progress in big data and artificial intelligence techniques for computer-aided drug design

YANG Yanwei[1], HU Wenyuan[2], LIN Zhi[1], HUO Guitao[1], ZHANG Di[1], LI Shuangxing[1], ZHANG Yaqun[2], YAN Zhenlong[2], QU Zhe[1], LYU Jianjun[2]

1. Beijing Key Laboratory, National Institutes for Food and Drug Control, National Center for Safety Evaluation of Drugs, Beijing 100176, China

2. Innostar Biotechnology Nantong Co., Ltd., Nantong 226133, China

Abstract: Big data and artificial intelligence (AI) techniques not only enable the accurate and comprehensive analysis of massive biomedical data, but also assist the development of predictive models in the field of drug design. With the development of computational and statistical methods, big data and AI techniques have been used for computer-aided drug design (CADD). CADD can be used to overcome troubles in the field of drug design, so as to effectively and efficiently design and develop new drugs. The steps of data pre-processing and development of models during drug design and development, AI-based modeling methods during drug design and development, recent applications of big data and AI-driven technologies in CADD were introduced, so as to provide some references for drug design and development in China.

Key words: drug design; big data; artificial intelligence; data analysis; predictive models

由于人类疾病的复杂性，药物设计和发现（涉及潜在靶点的识别和安全、有效药物的开发）过程复杂、费用高昂和耗时费力。计算机辅助药物设计（CADD）可以辅助进行药物设计和发现，一方面各种算法和统计方法可以用于有效地分析生物医学相关领域资料，以识别靶点和先导化合物。另一方面CADD可以进一步利用其他资料和计算机技术，确保药物的安全性和有效性，并帮助发现药物的潜在毒性和不良反应，从而顺利完成药物开发、安全性评价和上市申请[1-3]。目前，最常用的CADD方法包括基于结构的药物设计（SBDD）和基于配体的药物设计（LBDD）[4-7]。随着大数据在生物、化学

收稿日期： 2023-01-28

基金项目： 中国食品药品检定研究院学科带头人课题（2021X2）；江苏省新药一站式高效非临床评价公共服务平台建设项目（BM2021002）

#共同第一作者： 杨艳伟（1981—），男，副主任技师，研究方向为药物非临床安全性评价。E-mail：yangyanwei@nifdc.org.cn

胡文元（1990—），男，硕士，研究方向为药物非临床安全性评价。E-mail：wyhu@innostar.cn

***共同通信作者：** 屈　哲（1982—），女，副研究员，研究方向为药物非临床安全性评价。 E-mail：quzhe@nifdc.org.cn

吕建军（1970—），男，主任药师，研究方向为药物非临床前安全性评价。E-mail：jjlv@innostar.cn

和医药领域的发展，各种机器学习和算法在CADD领域得到了优化和应用。在最初的靶点识别阶段，可以使用生物信息学和反向对接等方法筛选与验证潜在靶点[8-10]，进一步使用组合库设计、虚拟筛选、药效团模型等方法发现先导化合物[11]。在先导化合物优化阶段，可以使用二维和三维定量构效关系（QSAR）的方法[4]。最后，在药物进入临床试验之前，还可以对药物的吸收、分布、代谢、排泄和毒性（ADMET）进行预测[12]。CADD不仅可以提高药物研发的成功率，降低药物研发费用，还可极大缩短药物的研发周期，是目前创新药物研发的核心技术之一[13-15]。本文介绍了药物设计和发现过程中数据预处理和建模步骤、药物设计和发现过程中基于人工智能（AI）的建模方法、大数据和AI技术在CADD中的最新应用，以期为我国药物设计和新药研发提供参考。

1 药物设计和发现过程中数据预处理和建模步骤

传统药物发现领域因生物医学数据的规模庞大和复杂性而存在局限性，而大数据和基于AI算法的计算和分析技术可以克服上述困难[16-18]。在CADD中对数据进行预处理对于正确理解和分析生化数据至关重要，最重要的是为预测模型的建立提供可靠的数据。对输入数据进行预处理后即可进行预测模型的建立。药物设计和发现过程中数据预处理和建模包括数据的收集与数据集划分、数据预处理、超参数调节和模型验证等步骤。

1.1 数据的收集与数据集划分

机器学习的建模必须有海量的数据支撑[19]。一般情况下，特征数不应少于样本数。建立回归模型，输出的是连续的数值；建立分类模型，输出的是类别标签。数据分为训练集和测试集，训练集用来建立模型，测试集由于测试模型的性能。对于回归任务，一般是将所有样本按输出值由高到低（或由低到高）排序，再按照一定比例，采用“逢n抽1”的方法抽取出训练集，剩余的样本组成测试集。通过这种方法可以使训练集和测试集中的输出值分布均匀。而对于分类任务，一般需要保证在训练集和测试集中各个类别之间的比例相等。

1.2 数据预处理

原始数据通常都不完整，可能会存在缺失、重复等，无法直接进行数据分析。为了提高数据分析的质量，需要对原始数据进行预处理。数据预处理没有标准的流程，通常因任务和数据集属性的不同而不同。在生物医学大数据分析中，考虑为1个包含n个样本和p个生物医学特征的数据矩阵。对数据进行统计预处理时，采用了缺失数据填补（missing data imputation）、离群值检测（outlier detection）和冗余特征剔除（redundant feature elimination）等方法。数据预处理的实现方法需要有效的算法来保证预测的准确性和效率。与数据预处理相关的算法包括神经网络（NN）和随机森林（RF）。

1.2.1 缺失数据填补 如果使用不充分的药物数据训练模型，对药物设计的预测结果可能不准确或不一致。因此，用处理分子描述符的填补模型填补缺失数据，可保证数据的完整性和唯一性。针对药物研发数据分析中缺失数据的填补方法较少的问题，Alchemite[20-22]提出了1种新型的神经网络缺失数据填补模型（Alchemite模型），Alchemite模型在缺失数据填补中表现出明显优势。研究证明，在试验数据不足时，Alchemite深度学习填补改进预测模型优于协同矩阵分解（CMF）、深度神经网络（DNN）或RF。Alchemite模型还可以估计结果预测的不确定性。QSAR首先构建1个RF模型，然后使用线性模型从初始模型的单个预测中寻找分析相关性。到目前为止，这种方法是集中在由一类蛋白质的活性组成的同质数据集上，例如激酶，以使用相关蛋白质中活性之间的线性相关性。贝叶斯矩阵分解（Bayesian matrix factorization）[22]结合了贝叶斯概率推理和矩阵分解，该方法还将化学描述符作为输入，可以处理含有数百万种化合物和数千万个数据点的数据集。贝叶斯矩阵分解本质上是一种线性方法，但由于空行和空列，可能会影响虚拟复合预测结果[22]。

1.2.2 离群值检测 药物研发数据集中的数据值，如QSAR模型，可以使用标准统计学方法根据相似度进行分组。基于标准化技术的异常化合物识别对QSAR模型有很大的影响[23]。如果离群值包括在内或将重要的数据值排除为离群值，则所构建的模型会导致错误的预测。为了获得可靠的预测结果，用于构建预测模型的分子数据集应覆盖化学空间，并检测分子数据集适用范围之外的新化合物[24]。因此，在建立预测模型之前，需要排除离群值。Alchemite算法也可以用来检测药物发现数据中潜在的离群值，因此，Alchemite软件程序既可以进行缺失数据填补，也可以进行离群值检测[22]。

1.2.3 冗余特征剔除 完成离群值检测之后，需要剔除数据中的冗余数据。当预测模型在数据集中

选择多个显著特征时，选择统计分析和生物学意义上高度相关的变量等冗余特征可能会导致模型分析的错误。为了正确理解预测模型，必须剔除冗余特征，以免影响模型的可靠性[25]。为了降低模型过拟合的风险并提高模型泛化能力，通常用尽可能少的特征建立模型，而不是用所有特征建模，特别是对于多元线性回归、逻辑回归、支持向量机（SVM）等容易造成过拟合的算法。药物研发数据常用的冗余特征剔除方法是利用RF算法的RGIFE软件程序。

1.3 超参数调节

机器学习算法可通过超参数调节来提升模型的性能[19]。该过程也是机器学习中最为繁琐的一步。传统的超参数调节方法有随机搜索法、单因素轮换法、网格搜索法等。网格搜索法最为常用。对于所有需要调节的超参数，首先在取值范围内按一定的步长取值，通过多层循环将全部的超参数组合代入模型，根据打分值确定最优组合。之后再将各个超参数的取值范围设置于最优组合附近，重复上述过程。然而，当需要调节的超参数数量增多时，网格搜索需要的计算量急剧增加。而且，这种方法不利于编程能力较弱的研究者使用。因此，需要开发高效的易于使用的自动调节算法超参数的方法迫在眉睫。

1.4 模型验证

对超参数的优化及模型可靠性都需要验证。一般将模型验证的方法分为训练集验证、测试集验证、验证集验证和交叉验证[19]。训练集验证是将训练集数据重新带入已经建立完毕的模型中，将预测值与真实值比对。测试集验证是在对原始数据处理之前，留出一部分数据，待模型建立完成之后，将测试集数据代入模型，通过比对预测结果和真实结果检验模型，作模型的最终评估结果。验证集验证与测试集验证类似，在训练模型前划分出一部分数据作为验证集，主要用来优化模型的超参数，而不作为最终的评价结果。交叉验证常见的有留一法（LOO）和最常用的k折交叉验证法。k折交叉验证法与验证集验证法相比更稳定，更不易造成过拟合，而且比LOO的效率高很多。因此，在模型的优化阶段，一般使用k折交叉验证法。

2 药物设计和发现过程中基于AI的建模方法

机器学习是统计学和计算机科学的交叉学科，是AI和数据科学的核心技术[26]，被广泛应用于药学[27]、化学[28]、生物学[29]、医学[30]等多个学科。构建预测模型的AI方法中值得关注的是连续结果的预测建立模型的回归方法、建立不同类别的预测模型的分类方法、根据2个特征之间的相似性或距离对特征进行分组的聚类方法和从高维数据中提取由显著特征组成的低维数据的降维方法。回归和分类属于监督学习（supervised learning），给定输入数据和目标结果，即可训练每个模型，以预测测试和验证过程的结果。聚类和降维属于无监督学习，主要是发现数据中隐藏的模式或分组，研究输入数据特征之间的相互作用。有监督学习算法包括SVM、RF、NN等；无监督学习的算法包括K均值聚类（K-means clustering）、主成分分析（PCA）等。

2.1 监督学习算法

2.1.1 SVM SVM算法最早由Vapnik等于1964年提出，并在20世纪90年代开始迅速发展，广泛应用于各个领域[31]。目前，SVM算法通常使用核函数将数据映射到1个足够高的维度，再寻找1个能将训练集数据正确划分为两类，并且几何间隔最大的超平面，将全体样本一分为二。当预测未知样本时，根据样本落在超平面的哪1侧得到预测结果。因此，基础的SVM模型是1个二分类模型。在数据量不太大时，SVM的效率很高，但当数据量增大到一定程度时，时间消耗急剧增加。另外，当建模使用的特征数过多时，容易造成过拟合。目前，有最佳描述符的生物学或化学结构适合用SVM分析，以进行QSAR预测。

2.1.2 RF RF是由Breiman[32]开发的集成的分类算法，其核心思路是利用随机采样和随机选取特征的方法构建大量多样化的决策树（DT），最后综合全部决策树的预测结果得到最终的预测结果。RF回归算法与分类算法相近，用同样的随机采样和随机选择特征的方法，构建大量回归树，最后综合全部回归树得到最终输出值。由于引入了随机采样和随机选取特征，RF模型可以允许使用大量的特征值进行建模，而且不容易产生过拟合现象。由于RF是基于决策树和回归树构建的模型，不需要对特征进行标准化处理。RF算法已被应用于使用细胞系的基因组信息、药物靶点和药理学信息将几种药物相关联。

2.1.3 NN NN最早于1974年被Werbos作为一般网络的特例被提出，近年随着计算机计算能力的提升被广泛应用于各个领域中[19]。NN是受动物大脑生物神经网络启发的算法，NN的基本处理单元聚合成层，其中第1层为输入层（即数据的特征值），最

后1层为输出层（即输出值），输入层与输出层之间的层均为隐含层。蛋白质数据通常被视为体素网格，基于网格的方法允许将网格体素投影到多通道蛋白质描述符中，类似于几何和基于能量的策略。因此，每个蛋白质体素都包含所有描述符的信息。蛋白质多通道网格已成功在三维卷积神经网络（3D-CNN）模型中进行处理，用于识别蛋白质结合位点和预测良好的蛋白结合剂。

2.2 无监督学习算法

2.2.1 K均值聚类 K均值聚类是根据数据特征将数据分类为K组的算法，是一种简单的迭代型聚类算法，采用距离作为相似性指标，从而发现给定数据集中的K个类，且每个类的中心是根据类中所有值的均值得到，每个类用聚类中心来描述。在药物发现研究中，K均值聚类可以为每个样本生成适当的分子描述符，计算化合物样本之间的相似性，并根据计算出的相似性对化合物特征进行分组。

2.2.2 PCA PCA是能够极大提升无监督特征学习速度的数据降维算法，其利用正交变换把线性相关变量表示的观测数据转换为少数几个由线性无关变量表示的数据（即主成分）[33-34]。PCA可用于构建具有分子描述符的QSAR模型，可以解释化合物样品如何对生物、化学或药物靶点产生影响。当额外的分子描述符被用于分析相同的生物靶点（例如受不同受体影响的蛋白质）时，PCA模型可预测生物活性。

3 大数据和AI技术在CADD中的最新应用

在药物设计的多个领域，大数据和AI技术已经成功地应用于CADD。在SBDD过程中有3个相关应用，即靶点蛋白结合位点的识别、基于结构虚拟筛选（SBVS）、药动学特性和毒性的预测以及计算机分子模拟技术。此外，基于分子数据的DT算法可以用来分析美国食品药品管理局（FDA）批准的药物的作用（如药物诱导的肝损伤），包括朴素贝叶斯分类在内的方法可以用来建立框架，来研究与生物、化学或药理学不同的化合物数据集相关的暴露等[33]。

3.1 靶点蛋白结合位点的识别

蛋白质结合位点是类药物分子结合并引发治疗反应的结构元件，结合位点的大规模识别仍然具有挑战性[34]。其原因包括蛋白质具有动态特性，构象的取样范围很广，而且通常只有小部分含有结合位点。可用构象数量的增加以及蛋白质构象的复杂性使得蛋白质数据分析更具挑战性[35]。目前已经开发了几种使用经典方法的工具，包括Fpocket[36]、SiteHound[37]和MetaPocket[38]。考虑到蛋白质表面的几何和势能因素，这些工具可以预测结合位点。不同的AI方法，如过采样和二进制分类（ENRI）、RF（P2Rank）和深度学习方法（DeepSite）[34,39]，已经成为提高结合口袋识别性能的潜在策略[40]。Kozlovskii和Popov开发了快速、准确的深度学习方法BiteNet[41]。BiteNet在预测能力和计算效率方面明显优于经典的结合位点预测方法。因此，基于深度学习的工具可以成功地用于以分子动力学轨迹作为输入来识别可成药构象，检测到的可成药构象有利于基于结构的药物设计程序。

用户在使用深度学习软件进行结合位点预测时，需要特别注意以下3个方面：（1）这类训练集不可避免地包括假阴性，可能遗漏一些结合位点特别是新的变构位点，因此，经典和深度学习方法联合使用可能避免遗漏；（2）建议考虑训练数据集所构建模型的适用领域，研究表明，不同方法的性能取决于所研究的蛋白质家族；（3）训练集忽略了蛋白质的柔性，因为训练集通常是由X射线刚性结构数据构成的，可以通过计算机生成蛋白质配体构象集合等数据增强技术来解决这一问题[42]。

3.2 SBVS

一旦识别了可成药蛋白构象，就可以从候选药物的化学空间中获得良好的结合剂，从而产生预期的疗效。那些强效结合剂被称为命中化合物（hit compound）。利用对接技术可以虚拟模拟候选药物和蛋白质结合位点之间的分子相互作用。SBVS时，需要对数据库中的大量配体根据其结合亲和力进行排序，并通过打分函数（SF）回归模型进行预测[43]。最近，新一代基于AI的SF已经被开发[44]，其表现优于传统SF，因为其能够从低水平特征的蛋白质-配体复合物中学习[45]。此外，基于AI的SF的灵活性允许定制训练数据集，以研究感兴趣的蛋白质家族[46]和额外的信息以提高预测性能[45]或多样化结果。然而，由于缺乏合格的验证实验，目前对于基于AI的SF在SBVS中的应用存在争议[45,47]。在对常用基准数据集的回顾性验证中，基于AI的SF在使用蛋白质-配体复合体信息训练和仅使用配体信息训练时均稳定地表现出良好的性能[45,47]。这些结果表明蛋白质结构信息对预测没有显著影响。考虑到基于AI的算法通常缺乏可解释性，为了确保方法的可靠性，需要进行偏差控制验证[47]。有研究通过AI计算机靶标虚拟筛选法，发现原本用于治疗

淋巴瘤的化疗药物普拉曲沙（Pralatrexate）可能是治疗新型冠状病毒感染的有效药物[48]。

3.3 药动学特性和毒性预测

药动学主要研究药物在生物体内的ADMET的规律。对于即将成药的化合物，良好的药动学特性至关重要。然而用传统的实验方法大批量测定化合物的药动学特性耗时费力、费用极高，而利用CADD方法先行对待检测化合物进行筛选可以极大地降低费用并提高效率。在药物发现领域，采用多种基于AI的方法来预测ADMET特性。例如通过SVM或贝叶斯方法构建的SwissADME网络工具可预测理化性质、描述符、类药性和ADMET特性。由于预测模型的质量取决于输入数据，因此需要大量高质量数据才能获得准确的预测结果。通过努力建立全面数据库和基准数据以及算法的开发，使用基于AI的建模，在药物发现领域进行更好的ADMET预测[49]。

3.4 计算机分子模拟技术

利用计算机图形学进行分子模拟的技术称为计算机分子模拟。通过计算机模拟手段进行分子对接、药物筛选、先导物的优化、定量构效关系和药效团模型等方法进行药物设计，可以揭示药物与受体靶标的作用机制，探索药物靶点的空间结构，最终目标是设计具有能选择性地与某一靶标结合的药物分子。同时，利用分子模拟技术来构造、显示、分析和储存复杂的分子模型，在三维空间中观测药物小分子的结构特征，更改小分子形状和方位，并探测小分子与受大分子靶点的作用机制，判断药物小分子与受体大分子结合的可能活性位点。还能对药物小分子的结构进行修正，提出改善药物的药效学和药动学性质的方案，在“三维空间”中实现直观、可视化的药物分子设计。AI利用大数据和机器学习方法，根据已有的药物研发数据自动设计出上百万种与特定靶标相关的小分子化合物，并根据药效、选择性、ADME等其他条件对化合物进行筛选。而后筛选出来的化合物被合成并且进行实验检测，然后实验数据会被反馈到AI系统中用于改善下一轮化合物的选择[50]。李群林[51]于2020年采用分子模拟技术研究了吡唑乙基苯甲酰胺衍生物作为食欲素受体1拮抗剂的作用，为以后的结构优化，设计、合成更为有效的选择性食欲素受体1拮抗剂提供了理论指导。

4 结语和展望

从标准化的数据采集、高效高质量的数据结构化，到数据平台中心提供的高效率的人工审核和逻辑核查，基于这样一整套智能化数据处理体系，可保证处理的海量数据的质量，并通过使用正确合理的统计分析方法来减少数据差错和控制风险，是获得高质量研究数据的重要前提和保证。利用大数据和AI方法，CADD能够更好地理解人类健康和疾病。生物医学大数据的有效和高效分析方法有助于确定与特定健康结果密切相关的重要靶点或特征。

虽然CADD技术已经成功应用于药物设计领域，但目前还不完善，在各个环节均需要加以改进。首先，目前的化合物数据虽然含有大量分子化合物，但是其中大部分在生物医药领域内是没有应用价值的。因此，高质量的虚拟筛选库不仅可以大大降低计算资源，而且可以提高筛选的成功率。其次，CADD中关键的分子模拟对接技术还有待改进和提高，如何建立一种快速的方法来考虑受体柔性是近年来虚拟高通量筛选技术面临的巨大挑战，而且“假阳性”也是虚拟高通量筛选的另一个重要问题，CADD的方法准确性仍有很大的改进空间。生化药物在结构和药效学上的复杂性导致研究中会出现越来越多的高维数据，这将推动药物设计的大数据和AI算法的不断创新和发展。因此，通过CADD技术确实是一条有效途径，而且随着超级计算机的出现，可以预见CADD在新药研发中将扮演越来越重要的角色。

本文介绍了大数据和AI技术在CADD中的研究进展，CADD可结合不同的数据预处理方法和AI算法，将显著改善药物设计、发现和开发的预测模型。随着大数据和AI算法快速发展和在数据预处理和建模方法领域的日臻完善，可以预见CADD将越来越多地用于我国的新药研发。

利益冲突 所有作者均声明不存在利益冲突

参考文献

[1] Grechishnikova D. Transformer neural network for protein-specific de novo drug generation as a machine translation problem [J]. Sci Rep, 2021, 11(1): 321.

[2] Lu R M, Hwang Y C, Liu I J, et al. Development of therapeutic antibodies for the treatment of diseases [J]. Biomed Sci, 2020, 27(1): 1.

[3] Emmerich C H, Gamboa L M, Hofmann M C J, et al. Improving target assessment in biomedical research: The GOT-IT recommendations [J]. Nat Rev Drug Discov, 2021, 20(1): 64-81.

[4] Andricopulo A D, Salum L B, Abraham D J. Structure-based drug design strategies in medicinal chemistry [J]. Curr Top Med Chem, 2009, 9(9): 771-790.

[5] Lavecchia A, Di Giovanni C. Virtual screening strategies in drug discovery: A critical review [J]. Curr Med Chem, 2013, 20(23): 2839-2860.

[6] Grinter S Z, Zhou X. Challenges, applications, and recent advances of protein-ligand docking in structure-based drug design [J]. Molecules, 2014, 19(7): 10150-10176.

[7] Lionta E, Spyrou G, Vassilatis D K, et al. Structure-based virtual screening for drug discovery: Principles, applications and recent advances [J]. Curr Top Med Chem, 2014, 14(16): 1923-1938.

[8] Schaduangrat N, Lampa S, Simeon S, et al. Towards reproducible computational drug discovery [J]. J Cheminform, 2020, 12(1): 9.

[9] Yang X, Wang Y F, Byrne R, et al. Concepts of artificial intelligence for computer-assisted drug discovery [J]. Chem Rev, 2019,119(18): 10520-10594.

[10] Chen Y P, Chen F. Identifying targets for drug discovery using bioinformatics [J]. Expert Opin Ther Targets, 2008, 12(4): 383-389.

[11] Kalyaanamoorthy S, Chen Y P. Structure-based drug design to augment hit discovery [J]. Drug Discov Today, 2011, 16(17/18): 831-839.

[12] Cavagnaro J A. Preclinical safety evaluation of biotechnology-derived pharmaceuticals [J]. Nat Rev Drug Discov, 2002, 1(6): 469-475.

[13] Hu Y H, Lin W C, Tsai C F, et al. An efficient data preprocessing approach for large scale medical data mining [J]. Technol Health Care, 2015, 23(2): 153-160.

[14] Car J, Sheikh A, Wicks P, et al. Beyond the hype of big data and artificial intelligence: Building foundations for knowledge and wisdom [J]. BMC Med, 2019, 17(1): 143.

[15] Katsila T, Spyroulias G A, Patrinos G P, et al. Computational approaches in target identification and drug discovery [J]. Comput Struct Biotechnol J, 2016, 14: 177-184.

[16] Miller J B. Big data and biomedical informatics: Preparing for the modernization of clinical neuropsychology [J]. Clin Neuropsychol, 2019, 33(2): 287-304.

[17] Suh D, Lee J W, Choi S, et al. Recent applications of deep learning methods on evolution-and contact-based protein structure prediction [J]. Int Mol Sci, 2021, 22(11): 6032.

[18] Mirza B, Wang W, Wang J, et al. Machine learning and integrative analysis of biomedical big data [J]. Genes (Basel), 2019, 10(2): 87.

[19] 张鹏翼 . 计算机辅助激酶抑制剂筛选的新策略及新工具的开发与构建 [D]. 兰州: 兰州大学, 2021: 6-9.
Zhang P Y. Development and construction of new strategy and modelling tools for computer-aided screening of kinase inhibitors [D]. Lanzhou: Lanzhou University, 2021: 6-9.

[20] Lazzarini N, Bacardit J. RGIFE: A ranked guided iterative feature elimination heuristic for the identification of biomarkers [J]. BMC Bioinform, 2017, 18(1): 322.

[21] Irwin B W J, Whitehead T M, Rowland S, et al. Deep imputation on large-scale drug discovery data [J]. Appl Al Lett, 2021, 2: e31.

[22] Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development [J]. Nat Rev Drug Discov, 2019, 18(6): 463-477.

[23] Tropsha A. Best practices for QSAR model development, validation, and exploitation [J]. Mol Inform, 2010, 29(6/7): 476-488.

[24] Yosipof A, Senderowitz H. Optimization of molecular representativeness [J]. J Chem Inform Model, 2014, 54 (6): 1567-1577.

[25] Zhang B T, Cao P. Classification of high dimensional biomedical data based on feature selection using redundant removal [J]. PLos One, 2019, 14(4): e0214406.

[26] Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects [J]. Science, 2015, 349 (6245): 255-260.

[27] Burbidge R, Trotter M, Buxton B, et al. Drug design by machine learning: support vector machines for pharmaceutical data analysis [J]. Comput Chem, 2001, 26 (1): 5-14.

[28] Butler K T, Davies D W, Cartwright H, et al. Machine learning for molecular and materials science [J]. Nature, 2018, 559(7715): 547-555.

[29] Libbrecht M W, Noble W S. Machine learning applications in genetics and genomics [J]. Nat Rev Genet, 2015, 16(6): 321-332.

[30] Deo R C. Machine learning in medicine [J]. Circulation, 2015, 132(20): 1920-1930.

[31] Vapnik V N. An overview of statistical learning theory [J]. IEEE Trans Neural Netw, 1999, 10(5): 988-999.

[32] Breiman L. Random forests [J]. Machine Learn, 2001, 45 (1): 5-32.

[33] Terranova N, Venkatakrishnan K, Benincosa L J. Application of machine learning in translational medicine: Current status and future opportunities [J]. AAPS J, 2021, 23(4): 74.

[34] Krivák R, Hoksza D. P2rank: Machine learning based tool for rapid and accurate prediction of ligand binding sites

from protein structure [J]. J Cheminform, 2018, 10(1): 39.

[35] Amaro R E, Baudry J, Chodera J, et al. Ensemble docking in drug discovery [J]. Biophys J, 2018, 114(10): 2271-2278.

[36] Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: An open source platform for ligand pocket detection [J]. BMC Bioinformatics, 2009, 10: 168.

[37] Hernandez M, Ghersi D, Sanchez R. SITEHOUND-web: A server for ligand binding site identification in protein structures [J]. Nucleic Acids Res, 2009, 37 (Web Server issue): W413-W416.

[38] Zhang Z M, Li Y, Lin B Y, et al. Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction [J]. Bioinformatics, 2011, 27(15): 2083-2088.

[39] Jiang M J, Li Z, Bian Y J, et al. A novel protein descriptor for the prediction of drug binding sites [J]. BMC Bioinformatics, 2019, 20(1): 478.

[40] Westbrook J, Feng Z, Chen L, et al. The protein data bank and structural genomics [J]. Nucleic Acids Res, 2003, 31(1): 489-491.

[41] Cimermancic P, Weinkam P, Rettenmaier T J, et al. CryptoSite: Expanding the druggable proteome by characterization and prediction of cryptic binding sites [J]. J Mol Biol, 2016, 428(4): 709-719.

[42] Son J, Kim D. Development of a graph convolutional neural network model for efficient prediction of protein-ligand binding affinities [J]. PLoS One, 2021, 16(4): e0249404.

[43] Ghislat G, Rahman T, Ballester P J. Recent progress on the prospective application of machine learning to structure-based virtual screening [J]. Curr Opin Chem Biol, 2021, 65: 28-34.

[44] Liu J, Wang R. Classification of current scoring functions [J]. J Chem Inf Model, 2015, 55(3): 475-482.

[45] Morrone J A, Weber J K, Huynh T, et al. Combining docking pose rank and structure with deep learning improves protein-ligand binding mode prediction over a baseline docking approach [J]. J Chem Inf Model, 2020, 60(9): 4170-4179.

[46] Bitencourt-Ferreira G, Duarte da Silva A, Filgueira de Azevedo W. Application of machine learning techniques to predict binding affinity for drug targets: A study of cyclin-dependent kinase 2 [J]. Curr Med Chem, 2021, 28 (2): 253-265.

[47] Sieg J, Flachsenberg F, Rarey M. In need of bias control: Evaluating chemical data for machine learning in structure-based virtual screening [J]. J Chem Inf Model, 2019, 59(3): 947-961.

[48] Zhang H P, Yang Y, Li J X, et al. A novel virtual screening procedure identifies pralatrexate as inhibitor of SARS-CoV-2 RdRp and it reduces viral replication *in vitro* [J]. PLoS Comput Biol, 2020, 16(12): e1008489.

[49] Wu Z, Ramsundar B, Feinberg E N, et al. MoleculeNet: A benchmark for molecular machine learning [J]. Chem Sci, 2017, 9(2): 513-530.

[50] 刘景陶, 柳耀花. 计算机分子模拟技术及人工智能在药物研发中的应用 [J]. 科技创新与应用, 2018, 2: 46-47.
Liu J T, Liu Y H. Application of molecular modeling techniques and artificial intelligence in drug research and development [J]. Technol Innov Appl, 2018, 2: 46-47.

[51] 李群林. 分子模拟技术研究吡唑乙基苯甲酰胺衍生物作为食欲素受体1拮抗剂 [D]. 上海: 上海应用科技大学, 2020: 1-32.
Li Q L. Molecular modeling technology studies of novel pyrazoylethylbenzamide derivative derivatives as selective orexin receptor 1 antagonists [D]. Shanghai: Shanghai Institute of Technology, 2020: 1-32.

[责任编辑 李红珠]