机器学习在创新药物研发中的应用进展

周 玥 张心苑 毛雪石

（中国医学科学院药物研究所信息中心 北京 100050）

〔摘要〕 介绍人工智能和药物设计基本概念、主要算法、技术和实际应用，探讨机器学习在创新药物研究中的应用，从分类回归、靶蛋白结构预测、活性位点识别和结合亲和力角度，详述基于机器学习策略的虚拟筛选技术在创新药物研发中的应用和挑战，对该技术发展进行展望。

〔关键词〕 人工智能；机器学习；药物设计；虚拟筛选

〔中图分类号〕R-056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673-6036.2020.08.005

Application Progress of Machine Learning in the Innovative Research and Development of Drugs *ZHOU Yue, ZHANG Xinyuan, MAO Xueshi, Network and Information Center, Institute of Materia Medica, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100050, China*

〔**Abstract**〕 The paper introduces the basic concepts, main algorithms, techniques and practical applications of Artificial Intelligence (AI) and drug design, discusses the application of Machine Learning (ML) in new drug research. It expounds the application and challenges of virtual screening technology based on ML strategy in the innovative research and development of drugs in detail from the perspectives of classification regression, target protein structure prediction, active site identification and binding affinity, and prospects the development of this technology.

〔**Keywords**〕 Artificial Intelligence (AI); Machine Learning (ML); drug design; virtual screening

1 引言

传统药物研发面临研发周期长、经费投入大、临床批准成功率低等方面的挑战，同时药物研发人员需要处理和分析海量信息[1]。随着计算机软硬件的进步，人工智能理论的发展和药理学数据的积累，人工智能技术的重要分支机器学习作为一种强大的数据挖掘工具已经应用于药物设计各个领域，如靶点识别、药物设计和结构优化、药物重新利用、性质评估和临床试验等[2]。本文将从机器学习重要算法、药物设计基本理论和机器学习在基于配体和受体虚拟筛选中的应用几个方面进行阐述。

〔修回日期〕 2020-01-10

〔作者简介〕 周玥，助理研究员，发表论文12篇；通讯作者：毛雪石，副研究员，发表论文15篇。

2 人工智能及其在药物设计领域应用发展历程

2.1 人工智能

人工智能概念始于1930年艾伦·图灵的通用图灵机并在1956年达特茅斯会议上由约翰·麦卡锡正式提出。作为一个交叉学科，人工智能整合计算机、数学、心理学和语言学等众多学科知识，已应用于文字语言处理、图像影像分析及自主智能领域[2]。从诞生至今人工智能共经历3个发展高峰期。20世纪50和60年代，逻辑推理和启发式搜索

概念的出现使人机交互成为可能。20世纪80年代，前馈神经网络和反向传播算法的创立成功实现人工智能在化学和分子生物学领域的首次探索，完成基于序列信息的蛋白质二级结构预测。2012年至今，深层网络模型的成熟使人工智能应用拓展到医学图像分析和自动驾驶车辆等领域[1-2]。

2.2 人工智能在药物设计领域应用

聚焦药物设计领域，药物化学家基于定量构效关系Hansh模型，逐渐开始应用人工智能方法以评估和预测化学与生物效应的核心问题[3]。20世纪90年代，神经网络、支持向量机和随机森林等方法已开始应用于抗癌药物筛选、蛋白序列设计和药物设计[4-5]。21世纪以来，人工智能在先导化合物优化、活性和毒性预测等领域取得成功[6]。基于人工智能在药物研发领域的快速发展，制药公司开始与人工智能公司开展合作，促进该领域的进一步发展[7]。

3 机器学习在药物设计中的应用

3.1 药物设计中的主要机器学习方法

3.1.1 概述 人工智能在药物设计中的应用即学习和解释与药物相关的大数据以发现新药物算法，以更加综合和自动的方式结合机器学习的发展[8]。与传统方法相比，基于机器学习的药物设计方法不依赖于基础原理和理论进步，而是更加注重从庞大生物医学大数据中提取新知识。

3.1.2 分类 目前药物研发使用最多的机器学习方法大致可分为5类：监督学习（监督、半监督、非监督）、主动学习、强化学习、迁移学习和多任务学习[2]。（1）监督学习。根据已知的输入和输出数据关系训练模型，以预测新样本数据分类和数值结果，主要用于药物疗效和ADMET预测等[1]。具体而言，可以对给定化合物库中的分子进行活性和非活性标记，通过分析分子特征与生物活性之间的关系预测新分子生物活性。（2）非监督学习。通过识别输入数据中的隐藏模式或固有结构来进行聚类和特征查找，已应用于疾病靶点的发现[1,9]。（3）主动学习。通过关注结构空间新颖性及最大可能化学空间领域来辅助选择过程，同时识别具有针对特定疾病靶标的潜在活性新型化合物[7,10]。（4）强化学习。在某种程度上模仿奖励驱动的学习方式，通过奖励和惩罚模式来设计和优化系统，主要用于从头药物设计领域[1-2]。（5）迁移学习和多任务学习。已应用于药物设计领域[2]。

3.1.3 具体实现算法 包括用于预测活性的回归算法，用于分类的随机森林、朴素贝叶斯和聚类算法，以及用于图像识别和结构创建的人工神经网络和深度学习等[1-2,11]。相对于传统学习方法，深度学习可以自动从输入数据中学习特征，通过多层特征提取将简单特征转换为复杂特征。目前比较流行的深度学习算法主要有深度神经网络、卷积神经网络、循环神经网络、深度自动编码器神经网络和生成对抗网络，已在生物活性预测、全新药物设计与合成及生物图像分析领域展现出巨大优势[1-2,11]。

3.2 机器学习应用于药物设计基本步骤

3.2.1 遵循药物研发过程 药物研发过程主要包括潜在药物靶标发现与验证、苗头化合物发现、先导化合物结构优化和候选化合物确认以及临床前与临床研究等[6]。机器学习在药物设计中的应用是一个顺序过程，包括研究问题的提出、机器学习方法结构设计、数据准备、模型训练与评估，以及结果理解和解释等[1-2]。

3.2.2 具体步骤 一是提出研究问题。确认特定问题属于回归预测活性任务、分类区分活性和非活性任务或产生新分子的结构性任务。二是根据问题和数据类型及数量选择合适算法并设置合理初始值。对于回归预测任务多使用逻辑回归方法；分类区分任务则较多使用支持向量机、随机森林和人工神经网络等算法；而对于生成性任务，深度学习网络则更为适用，如深度玻尔兹曼机和深度信念网络等。三是数据收集准备，初始数据的代表性、质量和数量对人工智能模型质量至关重要。为最大程度地提高可预测性，用于训练的数据需准确、合理且完整。四是模型训练和评估。通过训练搜寻一组参数以达到减小误差的目的。基于上述机器学习步

骤，科研人员可以解决药物设计中绝大多数的问题。但是目前机器学习在药物领域的应用还处于早期阶段，应重视结果的可解释性和可重复性，否则将制约机器学习在该领域的进一步发展。

4 机器学习在虚拟筛选中的应用

4.1 原理与步骤

4.1.1 原理 虚拟筛选已成为药物研发过程中一种重要的技术手段，通过该方法可对大批量化合物进行有效搜索，获得针对潜在靶标的苗头或先导化合物。虚拟筛选技术虽然克服了传统高通量筛选在时间及资源消耗上的不足，但其仅是高通量筛选的补充，减少后期筛选化合物数量，仍必须与实验相结合[12]。

4.1.2 步骤 常规虚拟筛选流程主要包括3大步骤。首先，基于自创、开源或商用数据库构建初始化合物库，依据类药性和假阳性评价标准过滤无法成药的化合物，进而构建筛选化合物库。其次，依据靶点结构是否已知，选择基于结构或配体的虚拟筛选技术，根据筛选条件获得理论上具有活性的化合物。最后，通过体外实验验证获得苗头或先导化合物。在上述过程中涉及大量的参数拟合、模型评价等工作，这正是机器学习优势所在，此外还可以提升整体运算速度，这些优点促使研究人员开始使用机器学习来完成虚拟筛选相关工作。

4.2 基于机器学习的虚拟筛选及其应用

4.2.1 基于机器学习的虚拟筛选 将机器学习算法和策略有机融合到基于结构和配体的虚拟筛选技术之中，可以自主完成相关模型构建及参数拟合工作，提高整体虚拟筛选完成速度、准确度和客观性，近年来越来越受到科研人员青睐。应用机器学习开展虚拟筛选工作，首先要构建化学基因数据库，根据筛选条件获得数据集。其次要根据k倍交叉验证方法和最小化结构风险原则以合理方式将数据集分为训练集和测试集。之后训练模型并依据混淆矩阵评价模型性能。最终将训练好的模型应用于虚拟筛选[12-14]。

4.2.2 应用模型 基于配体的虚拟筛选方法包括相似性搜索、化合物分类和回归活性预测3大技术。应用于配体虚拟筛选的机器学习方法以分类器为主。具有代表性的模型主要有：朴素贝叶斯、k最近邻居、支持向量机、随机森林和人工神经网络等。朴素贝叶斯模型适用于虚拟筛选分类和获取特异性结合于靶点的分子骨架[15]。k最近邻居模型对于预测多靶点结合活性等多任务学习具有明显优势[12]。支持向量机则可用于化合物分类和合成可及性或水溶性等化合物属性值预测[16]。随机森林可以改善定量构效关系数据预测，也可用于对接打分函数以及预测蛋白质-配体结合亲和力研究[14]。人工神经网络常应用于潜在药物靶标识别、化合物分类、定量构效关系以及蛋白质-配体结合亲和力等研究[6]。

4.2.3 基于结构的药物设计 这是一个复杂过程，主要涉及靶点结构预测、活性位点识别、配体和受体相互作用识别、对接打分函数和结合亲和力计算等[2]。靶点结构预测方面，机器学习已用于靶标蛋白质同源性检测，扭转角，二级结构、理化性质及翻译后修饰预测，区分活性和非活性构象以及模型评估等[17-18]。预测蛋白质二级结构的软件包主要有：基于分类器的ASAP和refineD以及基于深度学习网络算法的MUFOLD-SS等。活性位点与相互作用识别方面，机器学习可以基于卷积神经网络预测结合位点或联合决策树与人工神经网络识别别构位点[19-20]。关于靶标蛋白结合位点识别的经典方法和机器学习策略可参考相关综述[21-23]。针对活性位点识别的相关软件包有：基于3D卷积神经网络DeepSite和基于随机森林算法的P2Rank。对接打分函数和结合亲和力计算方面，主要是通过结构分类、回归模型和深度学习算法来预测[24-25]。此外深度学习网络还可应用于化合物的反向找靶[25]。Khamis和Colwell[26-27]详尽阐述有关机器学习在分子对接中的主要应用和该领域研究成果，以及深度神经网络面临的挑战。目前关于亲和力计算的软件包主要有：OnionNet、gnina、K_{DEEP}、DeepAffinity、DeepConv-DTI和GraphDTA等。

5 结语

创新药物设计虽然克服了传统药物设计中研发周期长和经费投入大等问题，但仍面临着海量数据挖掘与分析的难题。人工智能凭借其技术优势逐步应用于药物设计领域，虚拟筛选技术作为发现先导化合物的重要来源已成为药物设计中的重要组成部分。机器学习应用于虚拟筛选，可有效提升大量模型构建和参数拟合工作效率，从而获得更为理想的先导化合物或潜在药物分子。不同机器学习模型适用于虚拟筛选的不同问题，目前比较成熟的应用主要集中于基于配体虚拟筛选中的活性预测与化合物分类，基于受体虚拟筛选的新位点识别与结合亲和力计算。人工智能在创新药物研发中的应用还涉及先导化合物优化、全新药物设计和化合物性质预测等。目前越来越多的制药公司或人工智能公司开始与科研院所合作，共同促进人工智能在药物研发中的应用与发展。中国医学科学院药物研究所已与元气制药合作创建协和知药人工智能实验室。联合创建实验室将发挥各自优势，有利于医药事业发展，对我国创新药物研发领域起到推动作用，为生物医药领域做出一定贡献。

参考文献

1 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.

2 Yang X, Wang Y, Byrne R, et al. Concepts of Artificial Intelligence for Computer - assisted Drug Discovery [J]. Chem Rev, 2019, 119 (18): 10520 - 10594.

3 Miller E, Hansch C. Structure - activity Analysis of Tetrahydrofolate Analogs Using Sutstituent Constants and Regression Analysis [J]. J Pharm Sci, 1967, 56 (1): 92 - 97.

4 Weinstein JN, Kohn KW, Grever MR, et al. Neural Computing in Cancer Drug Development: predicting mechanism of action [J]. Science, 1992, 258 (5081): 447 - 451.

5 Schneider G, Wrede P. The Rational Design of Amino Acid Sequences by Artificial Neural Networks and Simulated Molecular Evolution: de novo design of an idealized leader peptidase cleavage site [J]. Biophys J, 1994, 66 (2 Pt 1): 335 - 344.

6 Zhong F, Xing J, Li X, et al. Artificial Intelligence in Drug Design [J]. Sci China Life Sci, 2018, 61 (10): 1191 - 1204.

7 Mak KK, Pichika MR. Artificial Intelligence in Drug Development: present status and future prospects [J]. Drug Discov Today, 2019, 24 (3): 773 - 780.

8 Duch W, Swaminathan K, Meller J. Artificial Intelligence Approaches for Rational Drug Design and Discovery [J]. Curr Pharm Des, 2007, 13 (14): 1497 - 1508.

9 Young JD, Cai C, Lu X. Unsupervised Deep Learning Reveals Prognostically Relevant Subtypes of Glioblastoma [J]. BMC Bioinformatics, 2017, 18 (Suppl 11): 381.

10 Reker D, Schneider G. Active - learning Strategies in Computer - assisted Drug Discovery [J]. Drug Discov Today, 2015, 20 (4): 458 - 465.

11 Chen H, Engkvist O, Wang Y, et al. The Rise of Deep Learning in Drug Discovery [J]. Drug Discov Today, 2018, 23 (6): 1241 - 1250.

12 Carpenter KA, Huang X. Machine Learning - based Virtual Screening and Its Applications to Alzheimer's Drug Discovery: a review [J]. Curr Pharm Des, 2018, 24 (28): 3347 - 3358.

13 Carpenter KA, Cohen DS, Jarrell JT, et al. Deep Learning and Virtual Drug Screening [J]. Future Med Chem, 2018, 10 (21): 2557 - 2567.

14 Lavecchia A. Machine - learning Approaches in Drug Discovery: methods and applications [J]. Drug Discov Today, 2015, 20 (3): 318 - 331.

15 Bender A. Bayesian Methods in Virtual Screening and Chemical Biology [J]. Methods Mol Biol, 2011 (672): 175 - 196.

16 Shi Z, Ma XH, Qin C, et al. Combinatorial Support Vector Machines Approach for Virtual Screening of Selective Multi - target Serotonin Reuptake Inhibitors from Large Compound Libraries [J]. J Mol Graph Model, 2012 (32): 49 - 66.

17 Kumari P, Nath A, Chaube R. Identification of Human Drug Targets Using Machine - learning Algorithms [J]. Comput Biol Med, 2015 (56): 175 - 181.

18 Mayr A, Klambauer G, Unterthiner T, et al. Large - scale Comparison of Machine Learning Methods for Drug Target Prediction on ChEMBL [J]. Chem Sci, 2018, 9 (24): 5441 - 5451.

（下转第47页）

4.2.5 增加患者权力和疾病预警功能 调查发现信息安全问题和电子病历共享无提示预警功能是大学生拒绝电子病历及其共享的重要原因。随着计算机网络发展，网上获取信息更加便捷，越来越多的患者担心其电子病历信息泄露，对此电子病历系统可增加患者权力，使其有权隐藏个人隐私，对不同隐私进行分级保护[9]。充分利用电子病历共享，增加电子病历疾病预警功能，使人们亲身体验到推动共享所带来的更多优势。

5 结语

随着时代的进步和网络信息的快速发展，越来越多的医院开始采用电子病历，但由于人们对电子病历缺乏了解以及电子病历自身发展的不完善，其共享进程缓慢。本研究调查大学生对电子病历共享的认知现状，分析我国电子病历共享发展现状，发现大学生对此整体上缺乏了解，对电子病历共享时的信息安全不信任。近年来各个医院患者数量不断增加，传统的纸质病历已无法满足现代医疗需求，因此快速推进电子病历及其共享发展不容搁置[10]。一方面需要国家法律及政策的支持，另一方面也需要各个医院的大力配合及医务工作人员的实施，此外更需要公众接受和认可，相信在多方共同努力下能更快、更好地推动电子病历及其共享发展。

参考文献

1 曹赛颖. 电子病历档案信息共享研究 [D]. 合肥：安徽大学，2019.

2 穆芳洁. 国内外电子病历的发展概况及思考 [J]. 中国病案，2014，15 (9)：40-42.

3 朱晓卓. 论电子病历真实性的法律保障 [J]. 医学与法学，2015，7 (3)：41-45.

4 相悦丽，李莹，尹永奎，等. 医疗纠纷中电子病历作为证据相关问题的探讨 [J]. 中国卫生事业管理，2017，34 (3)：211-212.

5 马竞. 电子病案的优势及其信息管理系统的构建 [J]. 江苏卫生事业管理，2018，29 (2)：207-209.

6 王璞，蒋海泥，江文佳，等. 实现电子病历资源共享的障碍与对策探索 [J]. 中国医院，2017，21 (2)：52-53.

7 赵瑞. 电子病历共享研究 [D]. 郑州：郑州大学，2011.

8 余冬. 电子病历在医院信息化管理中的应用 [J]. 电子技术与软件工程，2017 (17)：259.

9 雷文瑾. 病人可控的电子66病历访问控制与隐私保护研究 [D]. 北京：北京交通大学，2017.

10 刘亚静. 网络信息化背景下医疗档案信息共享存在的问题及对策讨论 [J]. 河北医学，2014，20 (5)：878-880.

(上接第28页)

19 Jiang M, Li Z, Bian Y, et al. A Novel Protein Descriptor for the Prediction of Drug Binding Sites [J]. BMC Bioinformatics, 2019, 20 (1): 478.

20 Zhou H, Dong Z, Tao P. Recognition of Protein Allosteric States and Residues: machine learning approaches [J]. J Comput Chem, 2018, 39 (20): 1481-1490.

21 Macari G, Toti D, Polticelli F. Computational Methods and Tools for Binding Site Recognition between Proteins and Small Molecules: from classical geometrical approaches to modern machine learning strategies [J]. J Comput Aided Mol Des, 2019, 33 (10): 887-903.

22 Xiong Y, Zhu X, Kihara D. In Silico Drug Discovery and Design [M]. London: Future Science Ltd, 2013: 204-220.

23 Chen R, Liu X, Jin S, et al. Machine Learning for Drug-target Interaction Prediction [J]. Molecules, 2018, 23 (9): 2208.

24 Sunseri J, Ragoza M, Collins J, et al. A D3R Prospective Evaluation of Machine Learning for Protein-ligand Scoring [J]. J Comput Aided Mol Des, 2016, 30 (9): 761-771.

25 Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P. Development and Evaluation of a Deep Learning Model for Protein-ligand Binding Affinity Prediction [J]. Bioinformatics, 2018, 34 (21): 3666-3674.

26 Khamis MA, Gomaa W, Ahmed WF. Machine Learning in Computational Docking [J]. Artif Intell Med, 2015, 63 (3): 135-152.

27 Colwell LJ. Statistical and Machine Learning Approaches to Predicting Protein-ligand Interactions [J]. Curr Opin Struct Biol, 2018 (49): 123-128.