

人工智能与中医药前沿研究专栏

DOI: 10.16305/j.1007-1334.2026.z20250919002

机器学习算法在中药活性成分筛选中的应用

张能仪, 李园园, 王天明

上海中医药大学中药学院(上海 201203)

【摘要】 中药作为复杂的化学体系,其活性成分筛选长期面临成分多样性高、作用机制不明等挑战。近年来,机器学习算法技术的介入为中药现代化研究提供了新范式。系统梳理机器学习算法的分类并探讨其在中药活性成分筛选中的应用场景、技术挑战与优化策略。

【关键词】 人工智能;中药;活性成分;机器学习;大语言模型;算法

Application of machine learning algorithms in screening of active components in Chinese materia medica

ZHANG Nengyi, LI Yuanyuan, WANG Tianming

School of Pharmacy, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China

Abstract: Chinese materia medica (CMM), as a complex chemical system, has long faced challenges in active components screening, such as the high compositional diversity and unclear mechanisms of action. In recent years, the integration of machine learning algorithms has provided a new paradigm for the modernization of CMM research. This paper systematically reviews the classification of machine learning algorithms and explores their application scenarios, technical challenges, and optimization strategies in the screening of active components from CMM.

Keywords: artificial intelligence; Chinese materia medica; active components; machine learning; large language models; algorithms

中药是我国极具特色且珍贵的物质财富,是中医防治疾病的重要基础,然而,中药作为复杂的化学体系,其活性成分筛选长期面临多重挑战,包括复杂基质的严重干扰、药效物质基础界定模糊以及作用机制阐释不清等问题^[1]。传统的研究策略通常遵循“化学分离-生物筛选”的思路:首先对中药化学成分进行提取、分离与结构鉴定;随后,利用体外或体内生物活性评价方法及药理模型,逐一筛选单一成分的生物活性;最终旨在确定关键活性成分,并阐明中药的药效物质基础。^[2]该策略的核心逻辑在于确立药物与特定疾病之间单一的线性因果关系,即通过识别与特定靶蛋白具有高亲和力、高特异性

的相互作用成分,来揭示药物的作用机制^[3]。然而,这种策略存在明显局限:研究周期长、工作量大,且其还原论思路与中医药“辨证论治”所蕴含的整体观、系统观相悖,更难以契合中药多组分、多通路、多靶点协同作用的系统特性^[4]。因此,构建高效、系统的活性成分筛选方法,尤其是高通量筛选技术,对于克服传统方法的局限、深入解析中药复杂体系至关重要。

在这一背景下,机器学习(machine learning, ML)算法展现出强大的应用潜力。ML致力于构建能够通过经验自动改进的计算机系统^[5],凭借其强大的高通量数据处理能力、卓越的复杂模式识别与预测优势,能够通过多模态数据融合(如舌象、脉象)构建中医辨证逻辑,结合疗效反馈优化模型,实现“同病异治”的个性化分析,进一步借助知识图谱可解析方剂中药物间的“君臣佐使”结构关系,或通过人工智能系统模拟药物协同机制,辅助药物筛选

[基金项目] 国家自然科学基金项目(82404823);上海中医药大学科技发展项目(23KFL043)

[作者简介] 张能仪,女,本科,主要从事中药学研究工作

[通信作者] 王天明,高级实验师;E-mail:wtmtcm@126.com

与配伍优化,实现传统诊疗逻辑的数字化建模与复现^[6]。ML为中药现代化研究提供了全新的技术范式。该范式能够有效应对中药复杂体系的挑战,显著提升活性成分筛选的效率与精准度^[7],并有望为解析中药多组分、多靶点协同作用的整体效应提供新的视角和工具。

本研究系统梳理ML算法技术及其在中药活性成分筛选研究中的应用进展,深入分析其优势、挑战与未来趋势,以期为该领域的研究者提供新的方法学参考与思路启发,进而推动中药活性成分筛选研究的创新发展。

1 中药活性成分筛选的研究进展

近年来,对于生物色谱法、亲和超滤筛选法和计算机网络虚拟筛选等新技术在中药活性成分筛选的应用研究愈加广泛和深入。

1.1 生物色谱法 生物色谱法以生物活性材料作为固定相,通过色谱参数定量表征药物与生物大分子、靶体或细胞间的相互作用,从而模拟药物在体内的活性表达过程^[8]。该方法无需预先分离样品,兼具药理学相关性与结合特异性。例如,何丽君等^[9]采用人脐静脉内皮细胞(HUVEC)生物色谱模型筛选心叶淫羊藿乙酸乙酯提取物,通过比对色谱图谱识别潜在心血管活性成分,并借助多种分离技术最终获得7个单体化合物。然而,该技术仍存在固定相制备成本高、柱寿命短、商品化难度大以及难以模拟体内复杂代谢环境等局限,可能导致活性成分筛选不全^[10]。

1.2 亲和超滤筛选法 亲和超滤筛选法基于配体与受体间的特异性结合,借助超滤装置实现活性小分子的快速筛选,并结合液相色谱-质谱联用(LC-MS)完成结构鉴定^[11],具有筛选与鉴定一体化的特点,尤其适用于复杂体系中活性成分的发现。黄晨存等^[12]以凝血酶为靶标,采用亲和超滤-质谱技术对赤芍不同极性部位进行逆靶向筛选,成功识别出鞣花酸作为潜在抗凝血成分;随后的分子对接结果进一步证实其与凝血酶活性位点具有较高亲和力。

1.3 计算机网络虚拟筛选 计算机网络虚拟筛选技术借助数学算法与计算机建模、数据库构建及搜索能力,实现对中药药效物质基础的高通量研究^[13]。该技术通过整合前期研究成果,构建涵盖中药化学成分、药理作用及临床功效的专题数据库,并持续引入系统生物学等多学科进展,不断拓展与完善数据资源。在此基础上,虚拟筛选能够系统探

究中药的药效物质基础、药性理论及作用机制,不仅推动中药知识的数字化与系统化整合,也为活性成分筛选与药效机制解析提供了高效、集成的信息分析平台^[14]。

2 ML算法类型与选择

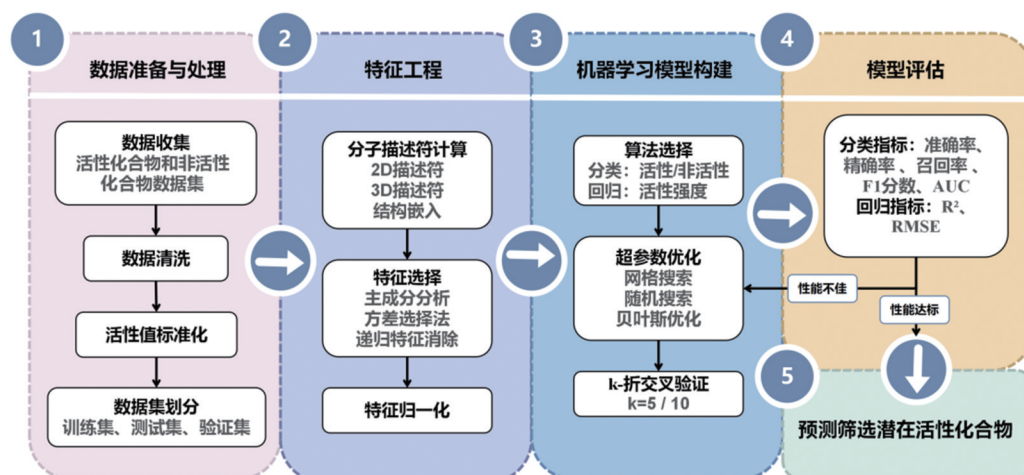
2.1 监督学习算法 监督学习算法利用带标签的训练数据集,学习从输入特征到目标输出的映射关系,以构建预测模型,并对未知样本进行精准预测或分类^[15]。监督学习算法可通过已知活性数据训练模型,实现对化合物生物活性的高效预测,在中药靶向筛选中具有高精度优势^[16-17],因而被广泛应用于虚拟筛选、药物-靶点相互作用预测及药物反应建模等领域^[18-19]。其应用于中药活性成分筛选的一般流程如图1所示。典型的监督学习算法包括支持向量机(SVM)、随机森林(RF)与XGBoost等。

2.1.1 SVM SVM是一种基于统计学习理论的监督学习算法,其核心思想是在特征空间中构造一个具有最大间隔的最优分类超平面,以实现不同类别样本的最大化分离^[20]。针对非线性分类问题,SVM通过引入核函数将原始数据映射到高维特征空间,从而在该空间中实现线性划分。SVM具有良好的泛化能力,该方法在小样本、非线性及高维数据处理中表现出优异的泛化能力,在药物活性成分筛选等领域得到广泛应用^[21]。

2.1.2 决策树 决策树通过树状结构递归划分特征空间,以根节点起始、内部节点做特征判断、叶节点输出分类或回归结果^[22]。其核心是依据信息增益等指标选择最优特征,构建直观可解释的决策路径,但易过拟合,需剪枝优化。为解决单一决策树容易过拟合的弱点,集成学习算法应运而生,如RF、XGBoost。

2.1.3 RF RF是一种集成学习算法,通过构建多棵决策树并集成其预测结果以提升模型性能与稳定性^[23]。该算法在每棵树的生成过程中采用双重随机策略:从原始数据集中进行有放回抽样,并在节点分裂时仅随机选取部分特征进行评价。这种机制使RF能够有效抑制过拟合,同时具备处理高维数据与缺失值的能力。RF可用于对大量化学成分进行分类,识别具有潜在活性的候选分子^[24]。

2.1.4 XGBoost XGBoost是一种高效的梯度提升集成学习算法,其通过迭代式构建多棵决策树以持续提升模型预测性能。在每一轮迭代中,该算法基于前一棵树的预测残差,通过优化可微目标函数来训



注:AUC为曲线下面积, R^2 为决定系数, RMSE为均方根误差。

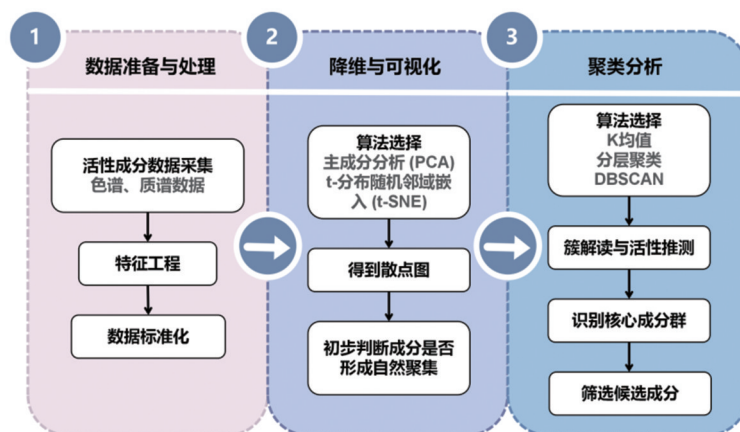
图1 监督学习算法和深度学习算法筛选中药活性成分一般流程

练新树,从而逐步降低模型误差、提高预测精度^[25-26]。

2.1.5 浅层神经网络 浅层神经网络是仅包含一个输入层、一个输出层和一个隐藏层的人工神经网络,是深度学习的基础结构^[27]。它通过模拟人脑神经元的连接方式,实现对非线性数据的拟合和特征提取,其中前馈神经网络(FNN)是最基础也是应用最广泛的网络结构,可用于虚拟筛选^[28]、靶点预

测^[29]和药物反应预测^[30]。

2.2 无监督学习算法 无监督学习算法通过对无标签数据集进行内在结构挖掘,实现模式识别与关联发现,尤其适用于从复杂数据中提取潜在信息^[15,31]。该技术在中药活性成分筛选中的一般流程如图2所示。其主要方法包括聚类分析与降维分析两大类。



注:DBSCAN为基于密度的噪声应用空间聚类。

图2 无监督学习算法筛选中药活性成分一般流程

2.2.1 聚类算法 聚类算法通过度量数据点间的相似性,将数据集划分为若干具有内部同质性的群组。常用算法包括K均值、分层聚类、基于密度的噪声应用空间聚类(DBSCAN)以及模糊聚类等方法^[32]。

2.2.2 降维算法 降维算法旨在将高维数据投影至低维空间,在保留关键结构信息的同时简化数据复杂性。主成分分析(PCA)和因子分析等方法已广泛应用于分子模拟及物理化学领域^[33],可用于解析

中药多靶点作用的分子网络机制^[34]以及区分化合物的成药潜力^[35]。

2.3 强化学习算法 强化学习算法的核心思想是智能体通过与环境进行持续交互,根据环境反馈的奖励信号来学习最优策略,以达成长期累积奖励最大化的目标^[36]。强化学习与深度学习、大语言模型等先进技术结合,已在药物发现的多个环节取得突破性进展^[37],为中医药的现代化研究提供了全新的技术工具。

2.4 深度学习算法 深度学习算法基于多层神经网络结构,能够自动提取数据中的多层次抽象特征^[38-39],在虚拟筛选及中药多组学数据整合与机制解析方面展现出显著优势^[40-41]。主动深度学习能够用有限的数据进行高效筛选,与传统方法相比,命中发现率提高了6倍^[42]。该技术应用于中药活性成分筛选的一般流程如图1所示。

2.4.1 卷积神经网络(CNN) CNN通过卷积层提取局部特征,擅长处理网格结构数据。CNN已成为药物筛选和发现的强大工具,可以处理复杂的分子和生物数据,从而能够更快、更准确地识别潜在候选药物、预测药物反应以及评估药物相互作用^[43-45]。

2.4.2 循环神经网络(RNN) RNN通常具有长短期记忆(LSTM)单元,可用于通过从现有化学数据库中学习来生成新型类药物化合物^[46-48]、预测药物靶标结合亲和力^[49-51]、高精度预测药物不良反应^[52]。

2.4.3 图神经网络(GNN) GNN处理图结构数据(如分子键连接、社交网络),可直接建模节点与边的关系。其常用于预测分子特性^[53]、结合亲和力和药物靶点相互作用^[54]、从头药物设计^[55]、药物筛选^[56-57]。

2.4.4 大语言模型 大语言模型是深度学习领域在自然语言处理方向的核心技术成果,指通过数百亿至数千亿级参数的深度神经网络构建,依托海量文本数据训练,具备自然语言理解、生成与推理能力的先进模型,已成为推动人工智能产业化应用的关键技术之一^[58]。以ChatGPT为代表的大语言模型正迅速变革着药物发现的传统流程^[59]。例如,ConPLex模型凭蛋白质序列和SMILES字符串,无需依赖复杂的三维结构计算,就能快速预测药物与靶标的相互作用^[60]。

2.5 算法的选择 算法的选择应紧密围绕研究目标、数据特征和可用计算资源来决定。首先,要明确研究任务类型,包括:分类任务,即预测一个类别标签(是否为活性成分);回归任务,即预测一个连续数值(半数抑制浓度得分);聚类任务,即在无标签数据中发现内在的群组结构。在确定任务类型后,根据数据规模与特征进行初选,对于样本数<10 000的小规模数据集,传统ML算法通常表现优异,且训练速度快、可解释性强;对于样本数>10 000的大规模数据集,深度学习算法能够自动学习复杂的特征表示,潜力巨大。接着尝试上述提到的主流算法,利用交叉验证和网格搜索对其进行优化。在

实践中,将多个表现良好的模型进行集成(如投票法、堆叠法),往往能获得比单一模型更准确的预测结果。

3 资源与工具

3.1 中药系统药理学数据库与分析平台(TCMSP) TCMSP(<https://tcmsp-e.com>)是目前应用最广泛的中药数据库之一,收录了中药化学成分、药物靶点、药代动力学参数(如口服生物利用度、药物相似性)等信息,为构建“成分-靶点-疾病”网络提供核心数据^[61]。在ML工作流中,其核心作用是为网络分析和特征构建提供化合物、靶点基础数据,是提取中药特异性特征的关键起点,可为后续模型训练提供高质量的基础关联数据支撑。

3.2 中药综合数据库(TCMID) TCMID(<http://www.megabionet.org/tcmid/>)侧重于中药的现代化与标准化研究,整合了中药、成分、靶点、疾病等相关信息,可将现代药理学和生物医学体系联系在一起,常作为TCMSP的补充^[62],用于交叉验证和扩充“成分-靶点-疾病-证候”关联网络,增强网络构建的覆盖度和可靠性。

3.3 本草组鉴(HERB) HERB(<http://herb.ac.cn/>)为集成了高通量实验数据和文献挖掘数据的中药数据库,提供了中药、活性成分、基因靶点、疾病等关键信息,数据全面,尤其适合进行深入的机制挖掘^[63]。其可为ML模型提供更丰富的标注数据和关系网络,支撑复杂的多任务学习与预测。

3.4 RDKit RDKit是一个开源的化学信息学工具包,可用于分子的读写、2D/3D结构处理、分子指纹和描述符计算、化合物相似性搜索及亚结构分析等,是进行分子特征提取和预处理的强大工具^[64]。在ML工作流中,RDKit是连接中药化学成分与ML模型的关键桥梁,其核心作用聚焦于分子表征环节,可将中药化学成分转化为ML模型可识别的数字特征(如拓扑指纹、物理化学描述符等),实现化学成分的量化表达,为后续模型训练提供标准化的输入特征。

3.5 DeepChem DeepChem(<https://deepchem.io/>)是一个专注于深度学习在药物发现、材料科学和量子化学领域应用的开源工具库。它内置了MoleculeNet基准数据集,包含多个已标注的分子性质数据集,方便模型训练与比较^[65]。其核心作用是为中药成分提供先进的分子表征方法(如GNN表示),并集成多种深度学习模型,直接支持化合物性

质预测、活性分类等端到端的 ML 建模任务。

3.6 Cytoscape Cytoscape (<https://cytoscape.org>) 是一款开源的网络可视化与分析工具,支持复杂网络的构建、拓扑结构分析、模块划分及可视化展示。在 ML workflows 中,可对 TCMSp、TCMID、HERB 等数据库构建的“成分-靶点-疾病-证候”网络进行结构优化与特征提取,如计算节点度、中介中心性、聚类系数等拓扑特征,这些特征可作为补充输入融入 ML 模型,帮助模型捕捉网络层面的关联规律,提升模型对中药多成分协同作用机制的解析能力。

3.7 NetworkX NetworkX (<https://networkx.org>) 是 Python 语言的网络分析工具库,专注于复杂网络的数学建模与算法实现。在 ML workflows 中,可实现多源中医药数据(如成分-靶点数据、靶点-疾病数据、证候-功效数据)的整合与结构化处理,通过编程化方式高效构建和分析“成分-靶点-疾病-证候”网络,其输出的标准化网络数据可直接对接后续 ML 模型的特征处理流程,尤其适用于大规模中医药网络数据的自动化分析场景。

3.8 Neo4j Neo4j (<https://neo4j.com>) 是基于图数据库的知识图谱构建与查询工具,具备高效的关联数据存储与复杂关系查询能力。在 ML workflows 中,可将分散的中医药多源数据(中药、成分、靶点、疾病、证候、功效等)整合构建为标准化的中医药知识图谱,通过图谱的语义关联推理功能挖掘潜在的“成分-靶点-疾病-证候”关联关系,为 ML 模型提供额外的特征信息和数据标注依据,同时可支撑模型预测结果的可解释性分析,实现“数据-模型-推理”的闭环。

4 ML 算法在中药活性成分筛选中的应用

4.1 基于配体的活性预测 基于配体的活性预测本质上是定量构效关系(QSAR)研究的延伸,其核心理论基础在于:分子的理化性质和结构特征与其生物活性之间存在内在联系。通过 ML 算法这种关联模式,能够实现对未知化合物活性的有效预测^[66]。王曦廷等^[67]采用梯度提升决策树(GBDT)与逻辑回归相结合的集成策略,以分子指纹作为特征描述符,构建了用于抗纤维化中药成分的虚拟筛选模型,经验证该模型对姜黄素、甘草酸等已知活性化合物的预测结果,与后续分子对接实验高度一致,证实了该计算模型可有效替代部分传统的实验筛选流程,提升早期药物发现的效率。Yang 等^[68]采用 4 种单分类器(包括 AdaBoost、K 最邻近法、分类树

和 RF)结合分子指纹描述符,构建了堆叠朴素贝叶斯(s-NB)模型,并借助虚拟筛选与细胞实验验证,发现两种化合物可同时抑制过氧化氢和连二亚硫酸钠诱导的神经毒性,该研究证实了 ML 在预测中药神经保护活性成分及阐释其配伍机制方面的有效性。Xiong 等^[69]提出了一种融合深度玻尔兹曼机(DBM)与偏最小二乘法(PLS)的新型分析模型,该模型首先利用 DBM 从特征空间中提取非线性特征,再将这些特征输入 PLS 进行多元线性回归分析,通过在麻杏石甘汤实验数据上的评估,该方法的预测准确率较现有主流方法平均提升了 10%。Huang 等^[70]以转录因子叉头框蛋白 O1(FoxO1)为靶点,采用朴素贝叶斯(NB)、RF 及 SVM 等 ML 算法,结合 MACCS 分子指纹与电拓扑状态指数作为特征,从甘草提取物中筛选得到 FoxO1 抑制剂欧甘草素 B;随后,通过同位素标记实验、免疫染色及荧光素酶报告基因检测等多种体内外药理学实验,进一步验证了该化合物的生物活性,从而强化了模型预测的可靠性。

4.2 多模态数据融合策略 中药活性成分筛选的现代研究正逐步由“单一模态分析”向“多模态融合”的研究范式转变。由于中药成分复杂、作用机制多元,单一维度的数据往往难以全面捕捉其活性特征。跨模态数据融合通过整合化学分析、多组学数据与药理实验等多源信息,借助 ML 算法的强大学习能力,能够显著提升活性成分筛选的准确性与效率,为中药创新研发提供新的路径。Guo 等^[71]将无监督学习策略应用于小儿消食片颗粒的研究,探索其对功能性消化不良的多向药理机制,利用基于细胞功能相似性的分层聚类,建立从化学聚类模块到细胞功能的联系,有助于将中药化合物分类为几个具有相似治疗功能的模块,以研究中药的多药理作用。Zhao 等^[72]开发了条形图指纹(BFD)结合谱效关系分析技术,以栀子根为模型,通过灰色关联分析(GRA)、偏最小二乘回归(PLSR)和高分辨峰分馏(HRPF)策略,筛选出 8 种护肝关键质量标志物(Q-markers),为中药现代化质量控制提供了新方法。Tian 等^[73]融合 LC-MS 成分鉴定、斑马鱼抗炎验证及双向门控循环单元-多注意力模型(BiGRU-MAR),筛选出安宫牛黄丸中牛磺胆酸等 7 种关键成分,构建 NIR 便携检测体系,预测精度达 0.98,实现“分钟级”快速质控。Pan 等^[74]采用集成学习策略,从化合物的整体拓扑结构、局部化学环境及药效团

特征 3 个层面进行系统性解析,并结合与抗衰老表型相关的特征重要性进行加权整合,构建了具有更强判别能力的分子指纹表征方法,该方法显著提升了分子指纹在抗衰老化合物筛选任务中的表征质量与实用性;基于此,从天然活性化合物、中药成分库及美国食品药品监督管理局批准药物数据库中筛选出一系列候选分子,并对其中排名最高的 6 种化合物进行了深入的功能验证,实验结果表明,所有 6 种化合物在热休克应激条件下均能显著延长模式生物线虫的寿命,其中 4 种化合物(百里醌、重楼皂苷 VI、美迪松和白花前胡素 C)更具备直接延长线虫健康寿命的能力,展现出良好的抗衰老活性。Wang 等^[75]提出一种 TCMCPi 模型,通过图卷积网络(GCN)提取化合物的“分子图结构模态”,通过 CNN 提取靶点的“序列/结构模态”,实现化合物与靶点双实体特征的跨模态融合;进一步模型创新将 CPI 网络转化为线图,引入“网络拓扑模态”,并借助 GAT 在学习线图节点特征时,将前述融合后的双实体特征与网络拓扑特征深度结合,形成“实体属性模态+网络结构模态”的多模态协同利用框架,模型的性能优于其他基准方法。

4.3 多技术联合应用 多技术联合应用于 ML 算法筛选活性药物,核心在于通过不同技术的互补性突破单一方法的局限,实现筛选效率、准确性和范围的多重提升,为药物研发的早期阶段提供更精准、高效指导。其中,网络药理学与 ML 算法的深度整合,为推进中医药作用机制的深度解析展现出潜力^[76-77]。网络药理学的核心优势在于能够系统阐明药物、靶点与疾病之间复杂的相互作用网络,这与中医药“多组分、多靶点、多通路”的整体调节特征高度契合。结合 ML 算法在处理海量异构数据和挖掘深层语义关系方面的能力,这种融合策略能够更深入地揭示中医药复杂机制背后的生物学基础,提供更为系统、动态的见解。

Ma 等^[78]则采用 ML 方法从血清代谢组学数据中识别差异代谢物,并基于网络药理学构建“成分-靶点-通路-疾病”多维网络,在成分-靶点相互作用预测模型的框架下,他们成功识别出药效成分及其作用靶点,建立了一种高效、准确的中药复方药效物质基础与作用机制解析方法,为阐明中药复方的复杂作用机制奠定了基础。有研究^[78]以中药复方白及五味子颗粒为对象,利用 ML 方法筛选出与疾病相关的差异代谢物及潜在生物标志物;进一步通

过构建“化合物-靶点-通路-疾病”多维网络,整合 Floyd-Warshall 算法识别网络中最短路径,从而定位关键调控节点;同时,采用 AdaBoost 模型,结合分子指纹特征,对化合物与靶点之间的相互作用进行高效预测;最终,通过分子对接等实验手段对筛选得到的有效成分及其作用靶点进行验证,系统揭示了该复方抗酒精性肝损伤的核心有效成分及可能的作用通路。

为探究百咳宁(BKN)颗粒治疗儿童流感的潜在机制,Gong 等^[79]首先从 TCMSP 和中国知网收集其活性成分,并基于 PubChem 与 PharmMapper 平台预测其作用靶点,同时,从 GSE34205(分析集)与 GSE42026(验证集)数据集中识别差异表达基因(DEGs),作为疾病相关靶标用于后续网络药理学分析;通过构建“BKN-成分-靶点”相互作用网络,筛选出 BKN 靶点与流感 DEGs 之间的重叠基因,进而构建蛋白质相互作用网络,并进行基因本体数据库功能注释与京都基因和基因组数据库通路富集分析;进一步将 BKN 相关的 DEGs(定义为 BKN-RCGs)进行聚类分析与 PCA,以评估其分类能力;在此基础上,采用加权基因共表达网络分析(WGCNA)识别与流感表型高度相关的核心模块及枢纽基因;随后,整合 RF、LASSO 回归与支持向量机-递归特征消除(SVM-RFE)3 种 ML 算法,筛选用于儿童流感诊断的关键标志物,并在外部数据集 GSE42026 中进行了独立验证;最后,通过分子对接与分子动力学模拟(MDS)评估活性成分与核心靶点之间的结合模式与稳定性。

Zheng 等^[80]建立了一种整合多组学数据与计算及实验验证的系统性研究策略,以深入解析解毒散根汤(JSD)的抗结直肠癌(CRC)作用机制,首先采用超高效液相色谱-串联质谱(UPLC-MS/MS)技术,从 JSD 及其在大鼠体内的血浆样本中,共筛选出 18 种原型成分与 8 种潜在代谢物;在此基础上,结合网络药理学与 ML 方法,进一步从中确定了 3 种最具潜力的生物活性成分;通过 RNA-seq 技术和生物信息学方法阐明其作用靶点,同时评估了其临床意义;随后,通过分子对接与分子动力学模拟,从计算层面预测了最优的“成分-靶点”相互作用对;表面等离子共振(SPR)与微量热泳动(MST)实验则直接证实了这些成分-靶点存在特异性结合;进一步的分子生物学实验揭示,白藜芦醇可能通过调控相关信号通路诱导 CRC 细胞凋亡,而金雀异黄素则通过下调 β -连

环蛋白表达,从而抑制 CRC 细胞增殖;此研究不仅证实 JSD 在体外对 CRC 细胞具有显著的促凋亡作用,更开发并实践了一套从活性成分筛选到靶点机制验证的完整研究框架,为中药复方的现代化机制研究提供了新模式。

5 技术挑战与优化策略

5.1 数据质量和数量问题 高质量、大规模数据是支撑 ML 算法有效性能的基础^[81]。在中药活性成分筛选研究中,数据来源的异质性(实验误差、文献矛盾、数据库滞后^[82])进一步破坏了模型的独立同分布假设,使得模型在跨数据集验证时泛化能力显著下降。此外,与化学药物相比,中药相关数据规模整体有限,尤其缺乏针对特定疾病或靶点的高质量标注数据集^[83],多数研究样本量远未达到 ML 模型对大规模训练数据的基本要求,这一现状严重制约了 ML 模型的训练效果与泛化能力,影响预测结果的准确性和可靠性。

为应对上述挑战,未来研究需整合多源数据——包括高通量实验数据、临床观察记录及中医古籍知识等,构建更全面、系统的中药活性成分信息库。同时,应建立严格的数据质量控制体系,涵盖数据清洗、验证与标准化等环节,以提升数据的一致性与可用性。通过多源高质量数据的融合,可为 ML 模型提供更充分的训练基础,从而显著提升其在中药活性成分筛选中的性能与可靠性。

5.2 特征选择和提取的复杂性 中药活性成分具有复杂的化学结构和多样的物理化学性质,如何从中选取并构建有效特征成为 ML 应用中的关键挑战^[84]。目前常用的分子描述符多达数百种,涵盖二维指纹、三维结构参数及理化性质等多种类型,当前特征选择陷入“盲目堆砌”误区,与目标活性无关的特征仍被保留,严重干扰模型学习。同时,中药成分与生物活性间多呈非线性关系,难以通过简单特征组合实现准确描述^[85]。此外,许多与活性相关的关键特征尚未被充分发掘与定义,进一步增加了特征工程的难度。不当的特征选择易导致模型出现过拟合或欠拟合,从而降低预测性能。

针对中药活性成分特征选择和提取的复杂性,需要进一步开发更有效的特征工程方法。结合领域知识和先进的数据分析技术,系统探索能够更准确表征成分-活性关系的特征表达体系。

5.3 模型的可解释性问题 ML 模型,特别是深度学习等复杂模型,常被视为“不可解释模型”,其决

策逻辑与内在机制往往缺乏透明性^[86]。在中药活性成分筛选中,模型的可解释性尤为重要^[87],因为只有明晰模型的预测依据,才能深入理解活性成分的作用机制,进而推动中药研究的深化与药物开发。然而,当前多数 ML 模型在预测结果的解释方面仍存在明显局限^[88]。复杂模型在有限数据条件下易产生“虚假可解释性”——模型通过记忆噪声弥补知识空白,提供看似合理却无生物学基础的解释。常用的事后解释方法[如沙普利加法解释工具(SHAP)、黑箱模型预测结果解释工具(LIME)]仅能描述数学关联,无法建立机制性联系。这一不足制约了该类方法在中药筛选领域的深入应用与推广。

为提升模型的可解释性,需进一步发展可解释 ML 技术。一方面,应针对现有模型开发解释性工具,如基于特征重要性分析的局部解释模型^[89]等,帮助理解模型的决策过程。另一方面,需探索构建具有内在可解释性的 ML 模型,使其输出能够直接关联中药活性成分的结构与功能特性,从而为研究人员提供更直观、可信的预测分析支持。

5.4 中药复杂体系的特殊性 现有研究仍沿袭“单成分-单靶点”的西药研究范式,无法表征中药多组分协同的本质特征^[90]。模型预测在表征中药成分间相互作用方面存在三重局限,即缺乏协同/拮抗效应的数学框架、忽略时空动态过程、未能整合炮制配伍等工艺参数。如何将这些因素纳入 ML 模型中,也是需要解决的问题。

考虑到中药复杂体系的特殊性,未来需要构建能够更好反映中药活性成分之间相互作用以及中药炮制、配伍等因素影响的 ML 模型,使模型能够更真实地模拟中药在实际应用中的情况,提高中药活性成分筛选的准确性和有效性,减少能耗,以期“精准中药”的创新提供思路^[91-92]。

6 小结

ML 算法在中药活性成分筛选中展现出巨大的潜力,为中药现代化研究提供了有力的技术支持。通过对多种 ML 算法的应用,能够有效地预测中药活性成分与靶点的相互作用、生物活性,揭示结构特征与活性的关系,以及筛选中药复方中的活性成分。然而,目前在应用过程中仍面临数据质量和数量、特征选择和提取、模型可解释性以及中药复杂体系特殊性等方面的挑战。未来,通过融合多源数据、开发更有效的特征工程方法、提高模型可解释性、构建适应中药复杂体系的模型以及与实验研究

的深度结合等措施,有望进一步提升 ML 算法在中药活性成分筛选中的应用效果,加速中药活性成分的研究和开发,推动中药现代化进程,为人类健康事业做出更大贡献。

参考文献:

- [1] 巩亦萍,顾志荣,郭引兄,等. 中药药效物质基础研究方法与思路的进展与展望[J]. 中成药, 2025, 47(4): 1211-1216.
- [2] 许晴,李智,万梅绪,等. 中药活性成分筛选新技术研究进展[J]. 药物评价研究, 2021, 44(7): 1541-1547.
- [3] 刘晓光,李梅. 基于深度学习的药物-靶标相互作用预测研究综述[J]. 智能系统学报, 2024, 19(3): 494-524.
- [4] 丁梦,谭旺晓,张笑,等. 基于靶点的中药活性成分发现技术研究进展[J]. 中国中药杂志, 2025, 50(13): 3645-3656.
- [5] SUN J, DONG Q X, WANG S W, et al. Artificial intelligence in psychiatry research, diagnosis, and therapy [J]. Asian J Psychiatr, 2023, 87: 103705.
- [6] 张君宇,方格,陈家旭,等. 基于机器学习方法构建气滞证及亚型证辨证决策网络模型[J]. 辽宁中医药大学学报, 2025, 27(7): 175-180.
- [7] MENG X, LANG Y, LI X, et al. Progress in the application of AI in the standardization of traditional Chinese medicine: A review based on machine learning and deep learning [J]. Pharmac Res Modern Chin Med, 2025, 16: 100639.
- [8] TANG Y P, XU D Q, YUE S J, et al. Modern research thoughts and methods on bio-active components of TCM formulae [J]. Chin J Nat Med, 2022, 20(7): 481-493.
- [9] 何丽君,顾承真,马黄璜,等. 基于色谱技术对淫羊藿成分及指纹图谱研究[J]. 海峡药学, 2023, 35(2): 61-65.
- [10] 文星星,沈萍,张涛,等. 生物色谱法在中药研究中的应用进展[J]. 工业微生物, 2023, 53(6): 81-83.
- [11] HE Y, ZHAO X, YU M, et al. Affinity ultrafiltration mass spectrometry for screening active ingredients in traditional Chinese medicine: A review of the past decade (2014-2024) [J]. Molecules, 2025, 30(3): 608.
- [12] 黄晨存,严晟,赵元,等. 赤芍抗凝血活性成分的筛选[J]. 中国药物化学杂志, 2023, 33(3): 181-190.
- [13] 赵东方,何荣景,侯旭东,等. 源于天然的酶抑制剂高效发现及评价新技术:进展与展望[J]. 上海中医药大学学报, 2021, 35(1): 1-11, 19.
- [14] 赵珊,张文青,张欣悦,等. 基于药物靶标识别中药活性成分的研究方法及应用[J]. 药物评价研究, 2023, 46(7): 1578-1586.
- [15] PATEL L, SHUKLA T, HUANG X, et al. Machine learning methods in drug discovery [J]. Molecules, 2020, 25(22): 5277.
- [16] 刘利梅,陈晓晋,孙世伟,等. 深度学习在药物活性预测研究中的应用[J]. 生物化学与生物物理进展, 2022, 49(8): 1498-1519.
- [17] DAINA A, ZOETE V. Testing the predictive power of reverse screening to infer drug targets, with the help of machine learning [J]. Commun Chem, 2024, 7(1): 105.
- [18] OBAIDO G, MIENYE I D, EGBELOWO O F, et al. Supervised machine learning in drug discovery and development: Algorithms, applications, challenges, and prospects [J]. Machine Learn Applic, 2024, 17: 100576.
- [19] JIN K, RONG C, CHANG J. Application of machine learning algorithms in drug screening [J]. J Complex Health Sci, 2023, 6(2): 35-38.
- [20] CERVANTES J, GARCIA-LAMONT F, RODRIGUEZ-MAZAHUA L, et al. A comprehensive survey on support vector machine classification: Applications, challenges and trends [J]. Neurocom, 2020, 408: 189-215.
- [21] RODRIGUEZ-PEREZ R, BAJORATH J. Evolution of support vector machine and regression modeling in chemoinformatics and drug discovery [J]. J Comput Aided Mol Des, 2022, 36(5): 355-362.
- [22] LU Y, YE T, ZHENG J. Decision tree algorithm in machine learning [C]// IEEE. Proceedings of 2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA). Dalian: IEEE, 2022: 1014-1017.
- [23] SALMAN H A, KALAKECH A, STEITI A. Random forest algorithm overview [J]. Babylon J Machine Learn, 2024, 2024: 69-79.
- [24] LIND A P, ANDERSON P C. Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties [J]. PLoS One, 2019, 14(7): e0219774.
- [25] BENTEJAC C, CSORGO A, MARTINEZ-MUNOZ G. A comparative analysis of gradient boosting algorithms [J]. Art Intel Rev, 2021, 54(3): 1937-1967.
- [26] STAUFFER R, ZEILEIS A, MAYR G J. Long-term foehn reconstruction combining unsupervised and supervised learning [J]. Int J Clim, 2024, 44(16): 5890-5901.
- [27] YANG Y, ZHOU D X. Optimal rates of approximation by shallow reLUk neural networks and applications to nonparametric regression [J]. Construc Approx, 2024, 62: 329-360.
- [28] HADIBY S, ALI Y M B. FNN based-virtual screening using 2D pharmacophore fingerprint for activity prediction in drug discovery [J]. Int J Comput Intel App, 2022, 21(3): 2250019.
- [29] LIANG L, LIU Y, KANG B, et al. Large-scale comparison of machine learning algorithms for target prediction of natural products [J]. Brief Bioinform, 2022, 23(5): bbac359.
- [30] AHMED H, SHEDEED H A, HAMAD S, et al. Feedforward neural network in cancer treatment response prediction [C]// ICAICV. Proceedings of The International Conference on Artificial Intelligence and Computer Vision. Cham: Springer Nature Switzerland, 2023: 119-128.
- [31] GLIELMO A, HUSIC B E, RODRIGUEZ A, et al. Unsupervised learning methods for molecular simulation data [J]. Chem Rev, 2021, 121(16): 9722-9758.
- [32] NAEEM S, ALI A, ANAM S, et al. An unsupervised machine learning algorithms: Comprehensive review [J]. Int J Comput Dig Sys, 2023, 13(1): 911-921.

- [33] ECKHARDT C M, MADJAROVA S J, WILLIAMS R J, et al. Unsupervised machine learning methods and emerging applications in healthcare [J]. *Knee Surg Sports Traumatol Arthrosc*, 2023, 31(2): 376-381.
- [34] 江启煜, 曾慧妍. 基于人工智能和组学数据驱动的中药潜在机制新型分析预测方法[J]. *中国组织工程研究*, 2025, 29(35): 7552-7561.
- [35] LEE K, JANG J, SEO S, et al. Drug-likeness scoring based on unsupervised learning[J]. *Chem Sci*, 2022, 13(2): 554-565.
- [36] 秦智慧, 李宁, 刘晓彤, 等. 无模型强化学习研究综述[J]. *计算机科学*, 2021, 48(3): 180-187.
- [37] SHEIKHOLESLAMI M, MAZROUEI N, GHEISARI Y, et al. DrugGen enhances drug discovery with large language models and reinforcement learning[J]. *Sci Rep*, 2025, 15(1): 13445.
- [38] ALZUBAIDI L, ZHANG J, HUMAIDI A J, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions[J]. *J Big Data*, 2021, 8(1): 53.
- [39] AHMED S F, ALAM M S B, HASSAN M, et al. Deep learning modelling techniques: current progress, applications, advantages, and challenges[J]. *Art Intel Rev*, 2023, 56(11): 13521-13617.
- [40] KIMBER T B, CHEN Y, VOLKAMER A. Deep learning in virtual screening: Recent applications and developments[J]. *Int J Mol Sci*, 2021, 22(9): 4435.
- [41] ZHANG P, ZHANG D, ZHOU W, et al. Network pharmacology: towards the artificial intelligence-based precision traditional Chinese medicine[J]. *Brief Bioinform*, 2024, 25(1): bbad518.
- [42] TILBORG D V, GRISONI F. Traversing chemical space with active deep learning for low-data drug discovery[J]. *Nat Comput Sci*, 2024, 4(10): 786-796.
- [43] NGUYEN-VO T H, DO T T T, NGUYEN B P. Multitask learning on graph convolutional residual neural networks for screening of multitarget anticancer compounds [J]. *J Chem Inform Model*, 2024, 64(18): 6957-6968.
- [44] DAS P, MAZUMDER D H. MLCNNF: a multi-label convolutional neural network framework for predicting adverse COVID drug reactions from the chemical structure [J]. *IEEE Trans Comput Biol Bioinform*, 2025, 22(1): 285-297.
- [45] SUN M, ZHAO S, GILVARY C, et al. Graph convolutional networks for computational drug development and discovery [J]. *Brief Bioinform*, 2020, 21(3): 919-935.
- [46] XIA W, XIAO J, BIAN H, et al. Deep learning-based construction of a drug-like compound database and its application in virtual screening of HsDHODH inhibitors[J]. *Methods*, 2024, 225: 44-51.
- [47] ZOU J, ZHAO L, SHI S. Generation of focused drug molecule library using recurrent neural network[J]. *J Mol Model*, 2023, 29(12): 361.
- [48] LIU X, YE K, VAN VLIJMEN H W T, et al. An exploration strategy improves the diversity of de novo ligands using deep reinforcement learning: a case for the adenosine A_{2A} receptor[J]. *J Cheminform*, 2019, 11(1): 35.
- [49] CARPENTER K, PILOZZI A, HUANG X. A pilot study of multi-input recurrent neural networks for drug-kinase binding prediction [J]. *Molecules*, 2020, 25(15): 3372.
- [50] ZHANG H, LIU X, CHENG W, et al. Prediction of drug-target binding affinity based on deep learning models[J]. *Comput Biol Med*, 2024, 174: 108435.
- [51] WANG X, LIU J, ZHANG C, et al. SSGraphCPI: A novel model for predicting compound-protein interactions based on deep learning[J]. *Int J Mol Sci*, 2022, 23(7): 3780.
- [52] NARGIS T, SHAHA S A N, HAZAREENA A. An approach in prediction of medicine side effects using RNN networks [C]// IEEE. *Proceedings of 2024 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*. Wuhan: IEEE, 2024: 416-421.
- [53] CAI H, ZHANG H, ZHAO D, et al. FP-GNN: a versatile deep learning architecture for enhanced molecular property prediction [J]. *Brief Bioinform*, 2022, 23(6): 1-12.
- [54] ZHANG S, JIN Y, LIU T, et al. SS-GNN: A simple-structured graph neural network for affinity prediction [J]. *ACS Omega*, 2023, 8(25): 22496-22507.
- [55] XIONG J, XIONG Z, CHEN K, et al. Graph neural networks for automated de novo drug design[J]. *Drug Discov Today*, 2021, 26(6): 1382-1393.
- [56] WANG Y, SU Y, ZHAO K, et al. A deep learning drug screening framework for integrating local-global characteristics: A novel attempt for limited data [J]. *Heliyon*, 2024, 10(14): e34244.
- [57] BOONPALIT K, CHUNTAKARUK H, KINCHAGAWAT J, et al. Pre-training strategy for antiviral drug screening with low-data graph neural network: A case study in HIV-1 K103N reverse transcriptase[J]. *J Comput Chem*, 2025, 46(1): e27514.
- [58] NAVEED H, KHAN A U, QIU S, et al. A comprehensive overview of large language models[J]. *ACM Trans Int Sys Tech*, 2025, 16(5): 1-72.
- [59] ZHENG Y, KOH H Y, JU J, et al. Large language models for drug discovery and development [J]. *Patterns*, 2025, 6(10): 101346.
- [60] CHOI Y S, HAN S K, KIM J, et al. ConPlex: a server for the evolutionary conservation analysis of protein complex structures [J]. *Nucleic Acids Res*, 2010, 38(Web Server issue): W450-W456.
- [61] 马璐, 田国祥, 耿辉, 等. 中药系统药理数据库 TCMSP 及其分析应用简介[J]. *中国循证心血管医学杂志*, 2020, 12(12): 1413-1416.
- [62] XUE T, GONG S, XIE T H, et al. Development status and trend of traditional Chinese medicine database [J]. *TMR Mod Herb Med*, 2022, 5: 19.
- [63] GAO K, LIU L, LEI S, et al. HERB 2.0: an updated database integrating clinical and experimental evidence for traditional Chinese medicine [J]. *Nucleic Acids Res*, 2025, 53(D1):

- D1404-D1414.
- [64] RYZHKOV F V, RYZHKOVA Y E, ELINSON M N. Python tools for structural tasks in chemistry[J]. *Mol Divers*, 2025, 29(4): 3733-3752.
- [65] KORSHUNOVA M, GINSBURG B, TROPSHA A, et al. OpenChem: a deep learning toolkit for computational chemistry and drug design[J]. *J Chem Inform Model*, 2021, 61(1): 7-13.
- [66] CARRACEDO-REBOREDO P, LINARES-BLANCO J, RODRIGUEZ-FERNANDEZ N, et al. A review on machine learning approaches and trends in drug discovery [J]. *Comput Struct Biotechnol J*, 2021, 19: 4538-4558.
- [67] 王曦廷,李彧,张澜,等. 基于机器学习的抗纤维化中药化合物筛选研究[J]. *北京中医药大学学报*, 2019, 42(1): 30-36.
- [68] YANG S, SHEN Y, LU W, et al. Evaluation and identification of the neuroprotective compounds of Xiaoxuming Decoction by machine learning: a novel mode to explore the combination rules in traditional Chinese medicine prescription[J]. *Biomed Res Int*, 2019, 2019(1): 6847685.
- [69] XIONG W P, ZHU Y M, ZENG Q X, et al. Dose-effect relationship analysis of TCM based on deep Boltzmann machine and partial least squares [J]. *Math Biosci Eng*, 2023, 20(8): 14395-14413.
- [70] HUANG Z Y, WANG L J, WANG J J, et al. Hispaglabridin B, a constituent of liquorice identified by a bioinformatics and machine learning approach, relieves protein-energy wasting by inhibiting forkhead box O1[J]. *Br J Pharmacol*, 2019, 176(2): 267-281.
- [71] GUO F, TANG X, ZHANG W, et al. Exploration of the mechanism of traditional Chinese medicine by AI approach using unsupervised machine learning for cellular functional similarity of compounds in heterogeneous networks, XiaoErFuPi granules as an example[J]. *Pharmacol Res*, 2020, 160: 105077.
- [72] ZHAO Y, MAO S, LI Y, et al. Rapid screening of active ingredients in traditional Chinese medicine based on bar-form-diagram fingerprint combined with spectrum-effect relationship [J]. *J Ethnopharmacol*, 2025, 351: 120137.
- [73] TIAN M, MA X, NIE L, et al. Screening of bioactive compounds and deep learning-driven quality control of Angong Niuhuang Pills [J]. *J Ethnopharmacol*, 2025, 351: 120095.
- [74] PAN Y, CAI H, YE F, et al. ElixirSeeker: A machine learning framework utilizing fusion molecular fingerprints for the discovery of lifespan-extending compounds [J]. *Aging Cell*, 2025, 24(8): e70116.
- [75] WANG H, LIU G, XU M. A novel method for predicting the interaction between Chinese herbal compounds and targets [C]// SPIE. Proceedings of Third International Conference on Biomedical and Intelligent Systems (IC-BIS 2024). Nanchang: SPIE, 2024: 677-683.
- [76] JOSHI C P, BALDI A, KUMAR N, et al. Harnessing network pharmacology in drug discovery: an integrated approach [J]. *Naunyn Schmiedebergs Arch Pharmacol*, 2025, 398(5): 4689-4703.
- [77] NOOR F, ASIF M, ASHFAQ U A, et al. Machine learning for synergistic network pharmacology: a comprehensive overview[J]. *Brief Bioinform*, 2023, 24(3): bbad120.
- [78] MA J, WEI P, XU X, et al. Machine learning-assisted analysis of serum metabolomics and network pharmacology reveals the effective compound from herbal formula against alcoholic liver injury[J]. *Chin Med*, 2025, 20(1): 48.
- [79] GONG Z, CHE Q, HU M, et al. Integrating multi-dimensional data to reveal the mechanisms and molecular targets of baikenning granules for treatment of pediatric influenza [J]. *Front Mol Biosci*, 2025, 12: 1637980.
- [80] ZHENG X, SHI C, XIE Y, et al. Bioactive components of Jiedu Sangen decoction against colorectal cancer: A novel and comprehensive research strategy for natural drug development [J]. *Phytomedicine*, 2025, 142: 156795.
- [81] MOHAMMED S, BUDACH L, FEUERPFIL M, et al. The effects of data quality on machine learning performance on tabular data[J]. *Inform Sys*, 2025, 132: 102549.
- [82] 张泽丹,王斌,张雯,等. 中医药真实世界数据质量评估方法进展[J]. *科技导报*, 2024, 42(21): 13-19.
- [83] 陈健,陈启龙. 网络药理学在中医药研究中的现状及思考[J]. *上海中医药大学学报*, 2021, 35(5): 1-6.
- [84] ZENG J, JIA X. Systems theory-driven framework for AI integration into the holistic material basis research of traditional Chinese medicine[J]. *Engineering*, 2024, 40: 28-50.
- [85] VO A H, VAN VLEET T R, GUPTA R R, et al. An overview of machine learning and big data for drug toxicity evaluation [J]. *Chem Res Toxicol*, 2019, 33(1): 20-37.
- [86] WINTER P M, EDER S, WEISSENBOCK J, et al. Trusted artificial intelligence: Towards certification of machine learning applications[J]. *arXiv*, 2021, 2103: 16910.
- [87] MURDOCH W J, SINGH C, KUMBIER K, et al. Definitions, methods, and applications in interpretable machine learning [J]. *Proc Natl Acad Sci U S A*, 2019, 116(44): 22071-22080.
- [88] RUDIN C, CHEN C, CHEN Z, et al. Interpretable machine learning: Fundamental principles and 10 grand challenges [J]. *Stat Surv*, 2022, 16: 1-85.
- [89] STIGLIC G, KOCBBEK P, FIJACKO N, et al. Interpretability of machine learning-based prediction models in healthcare [J]. *WIREs Data Mining Knowl Discov*, 2020, 10(5): e1379.
- [90] SONG Z, CHEN G, CHEN C Y C. AI empowering traditional Chinese medicine?[J]. *Chem Sci*, 2024, 15(41): 16844-16886.
- [91] 苗明三,谢逸轩. 论精准医学视角下的“精准中药”[J]. *上海中医药杂志*, 2023, 57(5): 31-36.
- [92] 王志强,薛启隆,刘新龙,等. 基于数值模拟的中药制药减碳绿色优化策略[J]. *上海中医药杂志*, 2025, 59(9): 1-8.

编辑:黄博韬

收稿日期:2025-09-19