人工智能技术在中药药理学中的研究进展

张　楠[1,2], 王晓云[2], 韩　波[1*], 何　谷[2*]

(1. 成都中医药大学药学院, 四川 成都 611137;　2. 四川大学华西医院, 四川 成都 610041)

摘要: 人工智能 (artificial intelligence, AI) 技术在各领域的应用日益广泛, 特别是在处理和解析海量数据方面, 其强大的能力为许多科学研究带来了突破。在中医药研究中, AI通过其卓越的学习和数据处理能力, 在提升研究的系统性、高效性和准确性方面展现出显著优势。中医药作为具有悠久历史和丰富理论体系的学科, 其研究过程中需要整合大量复杂的信息, 而AI的引入则为此提供了重要支持。AI技术尤其是机器学习 (machine learning, ML) 和深度学习 (deep learning, DL) 技术, 可以解析复杂的生物和化学数据, 推动中药药理学的新发现。通过这些技术, 研究者能够系统分析中药成分的多靶点机制, 并优化配方疗效。AI与多组学数据结合, 以及在细胞表型分析中的应用, 有助于精准识别药物靶点并探索新机制。此外, 融合AI的网络药理学整合实验、计算和临床数据, 解析药物多靶点机制, 增强中药配方疗效。AI加速了活性化合物靶点识别及效果分析。大语言模型的发展也在中医药知识图谱构建和文献分析中发挥重要作用, 利用自然语言处理技术, 从海量文献中提取有价值信息, 构建系统知识结构。AI技术的引入推动了中医药研究的现代化, 还在国际化和精准医疗的发展中发挥着关键作用。AI技术不仅提升了中医药研究的整体水平, 还为多学科的交叉合作与创新提供了坚实的基础。在未来的发展中, 随着AI技术的不断进步, 可以预见中医药将在全球范围内发挥更大的作用。这一过程不仅是中医药现代化的重要标志, 更是科学与传统智慧结合的体现, 必将推动整个医学领域的进步和发展。

关键词: 人工智能; 中医药; 机器学习; 深度学习; 大语言模型

中图分类号: R966　　**文献标识码**: A　　**文章编号**: 0513-4870(2025)03-0550-09

Research progress of artificial intelligence technology in pharmacology of traditional Chinese medicine

ZHANG Nan[1,2], WANG Xiao-yun[2], HAN Bo[1*], HE Gu[2*]

(*1. School of Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China;　2. West China Hospital, Sichuan University, Chengdu 610041, China*)

Abstract: Artificial intelligence (AI) technology is increasingly applied across various fields, particularly in handling and analyzing large volumes of data, providing breakthroughs for numerous scientific studies. In Chinese medicine research, AI demonstrates significant advantages by enhancing the systematic, efficient, and accurate nature of studies through its exceptional learning and data processing capabilities. As a discipline with a long-standing history and rich theoretical framework, Chinese medicine research requires the integration of complex information, for which AI provides crucial support. AI technologies, especially machine learning and deep learning, can decipher complex biological and chemical data, advancing new discoveries in Chinese medicine pharmacology. Researchers can systematically analyze the multi-target mechanisms of Chinese medicine components and optimize formulation efficacy through these technologies. The combination of AI with multi-

收稿日期: 2024-10-31;　修回日期: 2025-01-27.

基金项目: 国家自然科学基金资助项目 (82104373); 四川省自然科学基金 (2024NSFSC1842); 成都中医药大学杏林学者学科人才科研提升计划 (QJRC2023020).

*通讯作者 E-mail: hegu@scu.edu.cn; hanbo@cdutcm.edu.cn

DOI: 10.16438/j.0513-4870.2024-1078

omics data and its application in cell phenotype analysis aids in accurately identifying drug targets and exploring new mechanisms. Additionally, AI-integrated network pharmacology combines experimental, computational, and clinical data to analyze multi-target drug mechanisms, enhancing the efficacy of TCM formule. AI accelerates the target identification of active compounds as well as dissecting the pharmacological effects. The development of large language models also plays a crucial role in constructing Chinese medicine knowledge graphs and literature analysis, extracting valuable information from extensive literature using natural language processing to build a systematic knowledge structure. The introduction of AI technology has propelled the modernization of Chinese medicine research and has a pivotal role in the development of internationalization and precision medicine. AI not only enhances the overall level of Chinese medicine research but also provides a solid foundation for interdisciplinary collaboration and innovation. With the continuous advancement of AI technology, Chinese medicine is anticipated to have a greater influence and role globally. This process is not only a significant marker of the modernization of Chinese medicine but also a reflection of the integration of science and traditional wisdom, which will undoubtedly drive progress and development in the entire medical field.

Key words: artificial intelligence; Chinese medicine; machine learning; deep learning; large language model

人工智能 (artificial intelligence, AI) 是计算机科学的一个分支, 旨在打造具有人类智能的计算系统, 使其能够执行通常需要人类智能才能完成的任务。AI起源于20世纪50年代, 并在近年来取得了突破性进展, 成为一门广泛应用于各个领域的核心技术。AI涵盖了许多子领域, 包括但不限于机器学习 (machine learning, ML)、深度学习 (deep learning, DL)、自然语言处理、计算机视觉等。AI技术的进步得益于大数据、计算能力的提升和创新算法的发展。随着AI技术的快速发展与广泛应用, 以及医药相关数据的不断积累, AI已广泛服务于包括药物开发及医疗健康相关的各个领域。AI在药理学研究领域的应用同样展示了其巨大的潜力和广阔的前景。基于ML和DL的技术正广泛应用于药物发现、药物作用机制解析及药物反应预测等多个方面 (图1)。首先, 在药物筛选阶段, AI技术通过虚拟筛选和分子模拟大幅提升了新药发现的效率。通过构建分子对接模型, AI可以快速筛选出与目标蛋白高度亲和的化合物, 预测其药理活性, 从而大幅

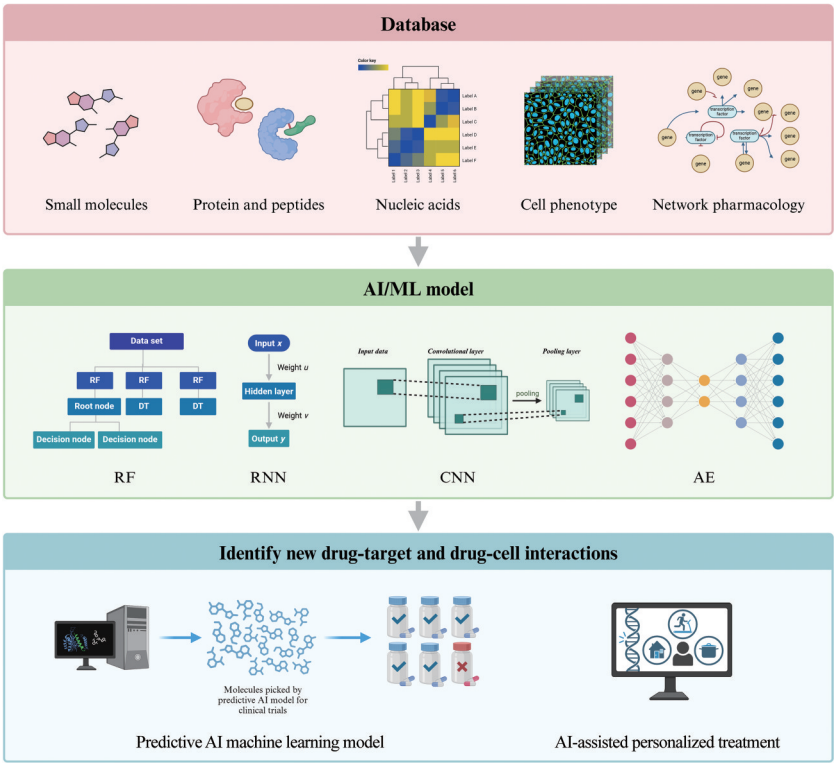

Figure 1　Traditional Chinese medicine assisted by artificial intelligence (AI). RF: Random forest; RNN: Recurrent neural network; CNN: Convolutional neural network; AE: Autoencoder

缩短实验周期。其次, AI技术在药物作用机制研究中也起到了关键作用。利用网络药理学方法, AI可以构建药物−靶标−疾病的多层次网络, 分析中药复方的多成分、多靶点的复杂作用机制。此外, AI还可以通过多组学数据分析 (基因组、转录组、蛋白质组和代谢组数据) 揭示药物在不同生物层面的作用机制。深度学习模型通过对这些组学数据进行整合和分析, 能够识别出潜在的生物标志物, 预测患者对药物的反应, 从而实现精准医疗。最后, AI技术在药物毒性评估和药物安全性监控中也发挥了重要作用。通过构建毒性预测模型, AI可以提前预测药物的潜在不良反应, 减少药物研发的风险。中药药理研究因其复杂的多成分、多靶点和多途径特性, 传统方法往往难以全面揭示其中的作用机制和药效关系。AI技术的引入能够通过处理大量数据、建立复杂模型以及优化研究流程, 有效提高研究效率和准确性, 从而推动新药开发和精准用药, 加速中药现代化和国际化进程。理解并应用AI技术, 将为中药药理学研究带来全新的方法和视角, 并将继续在生物医药领域发挥重要作用。

1 常用的AI算法

1.1 ML算法介绍

ML是一种通过数据驱动的方法, 使计算机系统能够自主学习和改进的关键技术。通过使用统计方法, ML算法可以从大量数据中提取特征, 进行模式识别, 并做出预测。常见的ML方法包括监督学习、无监督学习和强化学习。监督学习利用标注数据训练模型, 用于分类和回归任务; 无监督学习则处理未标注数据, 用于聚类和降维; 强化学习通过奖励机制不断优化行为决策。

1.1.1 监督学习算法 监督学习算法是ML中的一个重要分支, 主要用于预测和分类任务。其常见类型包括线性回归 (linear regression)、逻辑回归 (logistic regression)、支持向量机 (support vector machine, SVM)、K最近邻 (K-nearest neighbors, KNN)、决策树 (decision tree, DT)、随机森林 (random forest, RF) 和梯度提升树 (gradient boosting trees, GBT), 每种算法各有特点[1]。

线性回归是一种用于预测连续目标变量的基础算法。它假设目标变量与自变量之间存在线性关系, 通过最小化误差平方和来找到一个最佳拟合直线。其优点在于模型简单易懂、计算效率高, 且结果具有高度可解释性。然而, 它仅能捕捉线性关系, 无法处理复杂的非线性问题, 同时对异常值较为敏感, 并且需要对输入数据进行标准化处理。

逻辑回归主要用于二分类问题, 它通过逻辑函数将线性回归的输出映射到0～1之间的概率值, 以实现分类。逻辑回归的优点在于具有良好的可解释性, 系数明确表示特征对分类决策的影响, 且计算复杂度低, 适合大规模数据处理。

SVM通过找到能够最大化类别之间间隔的超平面来实现分类。SVM能够有效处理高维数据, 适用于线性或非线性分类, 具有处理小样本和高维数据的优势。但其对缺失数据敏感, 训练时间较长, 特别是对于大规模数据集, 且选择和调优核函数需要经验和尝试[2]。

KNN是一种简单且直观的分类算法。它通过计算待分类数据点与训练数据集中所有点的距离, 选取最接近的K个邻居进行投票决策。KNN的优点是实现非常简单, 不需要训练过程, 适用于多类分类问题。然而, KNN的计算复杂度较高, 对于大数据集, 查询速度慢, 且对数据分布和尺度非常敏感, 因此需要进行特征标准化。

DT通过树状结构进行分类和回归。这种算法递归地对数据进行分割, 生成一棵树, 每个节点表示对某个特征的测试, 叶子节点代表分类结果或回归值。决策树的优点在于简单直观, 易于理解和可视化, 还可以处理数值型和分类型数据, 不需要特征标准化和处理缺失值。但决策树容易过拟合, 需要采用剪枝策略, 并且对噪声和异常值敏感。

RF是一种基于决策树的集成算法。通过生成多个决策树, 并结合各个决策树的结果进行分类或回归。每棵树都是从原始数据集中随机抽样, 并从部分特征中随机选取特征进行分裂, 最终结果通过多数投票或平均值确定。随机森林能有效减少过拟合, 提高预测精度, 适用于高维数据和多类别问题, 对缺失数据和噪声具有较好的鲁棒性。然而, 面对非常大的数据量时, 训练和预测时间较长, 且单独结果缺乏解释性。随机森林常用于信用评分和基因表达数据分析等[3]。

GBT是一种集成学习方法, 通过逐步改进组合多个弱学习器 (通常是决策树) 来最小化总损失函数。GBT因其强大的预测性能而广泛应用于回归和分类任务。其优点在于灵活且能处理各种数据类型, 对数据中的非线性关系具有很强的捕捉能力, 但其训练时间较长且容易过拟合, 需要进行正则化[4]。

1.1.2 无监督学习算法 无监督学习算法是ML中的重要类别, 主要用于从未标记数据中发现隐藏的模式和结构, 常见的类型包括K均值聚类 (K-means clustering)、层次聚类 (hierarchical clustering)、主成分分析 (principal component analysis, PCA) 和独立成分分析 (independent component analysis, ICA)。

K均值聚类是一种用于发现数据集中K个簇的算法。K均值聚类的优点是实现简单, 计算速度快, 且易于解释和理解, 但其对初始簇中心的选择敏感, 可能收敛到局部最优解, 且需要预先指定簇的数量, 对outliers和噪声数据也较为敏感[5]。

层次聚类是一种建立聚类层次结构的方法, 可以是自底向上 (凝聚层次聚类) 或自顶向下 (分裂层次聚类), 通过计算样本之间的距离或者相似度进行聚类。层次聚类的优点是不需要预先指定簇的数量, 生成的树状图 (dendrogram) 可以直观地展示聚类过程和数据结构, 能发现不同层次的聚类结构。但其计算复杂度高, 适合小规模数据集, 对噪声和outliers敏感, 不具备全局最优性, 应用包括基因表达数据分析[6]。

PCA是一种经典的降维技术, 通过找到数据中方差最大的方向, 将高维数据投影到低维空间, 尽可能保留数据的主要特征。PCA可以有效降维, 减少数据维度并提高计算效率, 揭示数据的内在结构和主要矛盾[7]。

ICA是一种降维和信号分离方法, 用于寻找统计独立成分的信号, 常用于盲源分离。ICA的优点在于能够有效分离混合信号, 对非高斯信号具有良好的分离能力, 能发现底层独立的源信号。但其对信号的独立性假设依赖较强, 计算复杂度高, 对噪声和outliers敏感[8]。

1.1.3　强化学习算法　强化学习是一种通过与环境交互来学习最优策略的ML方法, 目标是使智能体在给定环境中通过试错获得最大累积奖励。强化学习算法主要包括3种类型: Q学习、策略梯度和深度Q网络。Q学习是一种无模型的算法, 通过学习动作的Q值函数来选择最佳动作, 以最大化累积奖励, 尽管它实现简单, 但在处理高维状态空间时表现较差且收敛较慢。策略梯度算法直接优化策略函数, 能处理连续和离散动作空间, 适应复杂任务, 然而计算复杂度高, 对超参数敏感, 需要大量数据。深度Q网络将Q学习与深度神经网络结合, 能够有效处理高维状态, 通过经验回放和固定目标网络进行稳定训练, 提高学习效率, 但对样本和计算资源的需求较高, 训练过程可能不稳定。

1.2　DL模型介绍

DL是ML的一个重要分支, 其灵感来源于对人脑神经网络细胞的研究, 通过构建多层神经网络模型来模拟人脑处理信息的方式。与传统ML方法不同, DL能够自动从海量数据中学习和提取特征, 具有强大的数据处理和模式识别能力。DL的一个显著特点是其层次化的结构, 每一层神经网络都会对输入数据进行抽象和表征, 从而在深层次上捕捉数据的复杂模式。

DL模型通过多层神经网络来对输入数据进行处理。每层神经网络由一定数量的神经元组成, 这些神经元通过权重连接并接收输入, 将其转换为加权和, 再通过激活函数输出。通过反向传播算法 (backpropagation), 模型会计算预测误差, 并根据该误差调整神经元的权重, 逐步优化模型参数, 以实现最优性能。DL在图像和语音识别、自然语言处理等任务中表现出色, 经典的DL架构有卷积神经网络 (convolutional neural network, CNN)、递归神经网络 (recurrent neural network, RNN)、生成对抗网络 (generative adversarial network, GAN) 和自编码器 (autoencoder, AE)。

1.2.1　CNN　CNN是DL中最常见的一类模型, 特别适用于处理图像数据。CNN通过卷积层 (convolutional layer) 提取图像的局部特征, 并通过池化层 (pooling layer) 进行特征降维, 最终通过全连接层 (fully connected layer) 进行分类或回归。卷积操作能够有效捕捉图像中的空间层次结构, 使其在计算机视觉领域广泛应用, 如图像分类、目标检测和图像生成[9]。

1.2.2　RNN　RNN在处理序列数据 (如时间序列、文本) 方面表现出色。RNN通过循环连接, 使得网络能够记忆和利用序列中的前后文信息, 适用于动态变化的数据。经典的RNN存在长时依赖问题, 而长短期记忆网络 (long short-term memory, LSTM) 和门控循环单元 (gated recurrent unit, GRU) 则通过特殊的门控机制解决了这一问题, 因此广泛用于自然语言处理 (NLP)、语音识别和时间序列预测等领域[10]。

1.2.3　GAN　生成对抗网络通过两个网络——生成器 (generator) 和判别器 (discriminator) 的对抗过程来生成新的数据。生成器试图生成逼真的数据以骗过判别器, 而判别器则学习区分真实数据和生成数据。通过这种对抗训练, GAN能够生成高质量的图像、视频和音频等数据, 广泛用于图像生成、风格迁移和数据增强等任务。GAN中的DCGAN、WGAN和StyleGAN是一些著名的变种[11]。

1.2.4　AE　AE是一种无监督学习模型, 通过将输入数据编码为低维表示, 再解码重构回原始数据, 从而实现数据压缩和特征提取。其结构包括编码器 (encoder) 和解码器 (decoder) 两个部分, 编码器将数据映射到低维隐藏层, 解码器则将隐藏层信息还原为高维数据。变分自编码器 (variational autoencoder, VAE) 进一步结合了概率模型, 用于生成和优化数据分布, 常用于降维、图像生成和异常检测[12]。

DL作为AI的重要分支, 借助多层神经网络强大的特征提取和模式识别能力, 已在各个领域取得了显著成果。从卷积神经网络到递归神经网络, 再到生成

对抗网络、自动编码器及变压器模型, 每种模型都有其独特的优势和适用场景。理解和掌握这些DL模型, 可以为解决复杂的数据分析和预测问题提供强有力的工具, 推动AI技术的进一步发展和应用。

2 用于中药药效研究的ML与DL方法

2.1 药物与靶点相互作用

2.1.1 基于结构的虚拟筛选与AI相结合 依托于计算机分子建模, 基于结构的虚拟筛选方法可以预测小分子与目标靶标之间的结合亲和力, 优先选出可能靶向目标蛋白质的小分子, 降低实验成本, 加速药物的发现过程。分子对接是一种典型的广泛使用的基于结构的技术, 可以量化评估药物与靶标结合位点之间的相互作用力, 计算结合自由能 (Δ*G*), 较低的Δ*G*预示着更大的结合亲和力。为了计算Δ*G*, 分子对接程序引入了评分函数, 包含了药物和蛋白质的物理化学参数。不同药物及目标蛋白质参数的差异以及结合亲和力的不同产生差异化的评分函数, 用于有效筛选靶向目标蛋白的小分子。但是经典的评分函数也存在着固有的局限性, 这种仅仅基于结合亲和力和结构而假定药物与靶标之间存在功能关系的方式, 会产生较高的假阳性预测[13]。

ML可有助于改善评分函数, 提高对接的精确性。通过ML对评分函数进行训练, 识别对结合亲和力贡献最大的数据集的关键特征, 推断出药物与靶标之间的关系, 可以一定程度上减少假定功能关系的影响。DL方法也极大地优化了分子对接的准确性, 可以更准确地预测药物与靶标的结合姿势和药物活性, 实现靶标的快速筛选。如Wallach等[14]利用CNN开发了AtomNet, AtomNet是一种基于结构的深度卷积神经网络, 可以预测药物发现中小分子的生物活性。AtomNet将3D网格放置在靶标和小分子共复合物上提取基本结构特征, 如原子类型计数和更复杂的结构蛋白配体相互作用指纹 (SPLIF), 最后将3D网格展开为1D浮点向量。Ragoza等[15]描述了一种CNN评分函数, 可自动学习与结合相关的蛋白质−配体相互作用的关键特征, 并在区分正确和不正确的绑定姿势以及已知的binders和非binders方面优于AutoDock Vina评分函数。但这些CNN驱动的DL模型无法提供有关药物−靶标相互作用强度的信息。Stepniewska-Dziubinska等[16]开发的Pafnucy也是一种基于3D卷积的CNN, 可以评估药物−靶标复合物的结合亲和力。

2.1.2 细胞表型数据的ML 在许多可公开访问的专用数据存储库中 (如PubChem、TCGA等), 存在高通量的生物表型数据集, 使研究人员可以快速访问及深度探索分析。针对这些大规模的数据集开发ML和DL模型, 可以高效预测药物与表型的关联。Kadurin等[17]开发了一种具有7层架构的对抗性自动编码器 (AAE), 分析6 252种化合物的NCI-60细胞系测定的全剂量反应数据, 开发DL模型。而后对PubChem数据库中超过7 200万种化合物进行处理, 选择具有潜在抗癌特性的候选分子。高通量成像检测可以捕获药物处理后细胞形态的变化, 再使用图像处理软件 (如CellProfiler) 提取形态特征信息, 转化为针对特定药物的指纹图谱, 用以进行ML以预测化合物活性[18]。当前药物作用机制的研究方法成本高昂且通量较低, Yu等[19]提出了一种通过分析线粒体表型变化鉴定作用机制的方法。他们开发了DL模型MitoReID, 可用于监测时间分辨率的线粒体图像, 并成功根据线粒体表型鉴定了6种药物的作用机制, 为靶点识别提供了一种自动化且具有成本效益的替代方案, 提升大规模药物发现和再利用的效率。

多向药理学 (polypharmacology) 是实现新药开发的一个途径, 中药多成分多靶点的特征也符合多向药理学中药物对多靶点的作用。利用多向药理学实现靶标反卷积有助于识别新的疾病相关靶标, 帮助合理设计更有效低度的药物组合。Gujral等[20]提出了一种使用弹性网正则化的集合方法, 结合mRNA表达谱和先前表征的大量激酶抑制剂的数据, 以识别与细胞迁移相关的癌细胞类型特异性激酶。

2.2 协同多组学研究

近年来, 随着各类组学研究技术的不断发展与丰富, 积累了大量的与中医药相关的组学数据, 包括转录组学、蛋白质组学、代谢组学、空间转录组学以及与中医药密切相关的单细胞组学等。这些数据涵盖了海量的信息, 尤其涉及一些微观水平药物分子、细胞、基因、生物大分子之间的复杂关系。利用AI技术发掘数据网络之间的关系, 可以高效系统地分析复杂疾病的病理机制, 深度发掘中药复杂药效成分与疾病证候之间的关系, 让传统中医药理论在AI与大数据时代迎来新的发展机遇。

新一代测序 (NGS) 技术的出现使得大规模转录谱的生成变得经济高效。大规模转录组学数据可以帮助阐明药物的作用机制。通常可以利用一些ML的算法, 如统计分析与回归, 构建共表达网络[21]。如Li等[22]基于ML算法构建网络平衡模型, 将基于共表达模式与网络拓扑特征结合, 构建冷热综合征的分子网络, 筛选确定了一系列与冷热综合征相关的生物标志物。癌症基因组图谱 (TCGA) 以及大量的公共数据库中海量的患者样本基于表达图谱信息, 也可以与ML方法一起应用于阐明疾病靶点和开发新的治疗方法。与中医

药相关的单细胞转录组数据, 也为病理机制研究与药物开发提供了丰富的资源。单细胞 RNA 测序技术 (scRNA-Seq) 可生成大量的单个细胞的转录图谱, 对于识别细胞类型簇、根据轨迹拓扑推断细胞群排列、突出体细胞克隆结构以及表征复杂疾病中的细胞异质性至关重要, 也为基于 DL 的细胞水平网络分析及细胞特异性的调控网络提供了充分的数据资源。单细胞图形神经网络 (scGNN) 是一种利用带有多模态自动编码器的 GNN 来构建和聚合细胞间关系的强大框架, 可应用于 scRNA-Seq 分析, 有效表示基因表达与细胞之间的关系[23]。

2.3　基于 AI 的中药组效关系研究

中药整合多成分、多靶点、多途径调控机体, 与现代医学中联合用药的治病方式, 均能对复杂疾病起到“增效减毒”的治疗效果。传统中药的配伍规则往往来源于经验的积累, 利用 AI 算法框架来评估复杂中药组合, 可以更高效识别和验证有效的药物组合。

在研究药物组效的方法中, 包括基于网络的方法和利用 ML 模型预测方法。基于 AI 算法的预测模型, 用于药物组合的快速筛选, 弥补了传统方法的不足。DeepSynergy 是一种基于 DL 的新型方法, 基于输入层神经元输入两种药物的化学性质和细胞系的基因表达作为特征向量, 最后通过隐藏层的深度协同网络转换输出预测的协同分数, 评估药物组合的协同作用[24]。在 DeepSynergy 的基础上还衍生优化出许多药物组合的预测模型。这些基于 DL 的预测模型, 适用于处理大量高维且有噪声的数据, 通过内部优化自动调整, 更好地捕捉复杂的数据关系。DL 模型 DrugCell 将传统的人工神经网络 (ANN) 与可见神经网络 (VNN) 结合, 能够模拟人类肿瘤细胞对不同化合物的反应, 智能预测出有效的药物组合[25]。TCMFP 是一种中医、AI 和网络科学算法相结合的中药方剂预测方法, 该方法集成了基于网络目标重要性的草药评分 (Hscore)、基于实证学习的配对评分 (Pscore) 和基于智能优化和遗传算法的草药配方预测评分 (FmapScore), 以有效地筛选疾病的最佳方剂组合[26]。Dai 等[27]也报道了一种基于网络药理学方法筛选治疗亨廷顿病的新型中医方剂, 并利用基于 SVM、CoMFA 和 CoMSIA 的定量构效关系 (QSAR) 模型, 预测候选药物的生物活性。

现有的基于 AI 算法的预测模型仍需要依赖于药效物质的化学结构及理化性质相似性等数据, 较为局限于两种药物之间的组合, 预测中药复杂多成分组合的协同效应仍然面临重大挑战。

2.4　整合 AI 的中药网络药理学

近年来, 网络药理学逐渐成为以中药为代表的多靶点药物发现的重要方法。它通过系统生物学和网络理论解析药物作用机制及其多靶点交互, 结合实验、计算和临床进展, 旨在加深对药物作用机制的深入理解。这一研究模式侧重于在网络视角下识别活性化合物的靶点, 以及化合物对多靶点的综合作用。

中药的多成分多靶点组合产生复杂的效应, 通过网络拓扑结构识别靶标, 使用“化合物–蛋白/基因–疾病”网络模型, 探索中药成分的药理机制, 分析药效分子在高通量环境中的调控, 并评估药物组合效果。这一过程包括识别与疾病和化合物相关的靶标, 构建蛋白质–蛋白质相互作用网络, 提取核心基因, 并进行网络验证以评估活性化合物与疾病靶标的相互作用[28]。

尽管网络药理学在多成分、多靶点配方的有效性评估及治疗策略探索方面潜力巨大, 但面临挑战, 如中药靶标识别技术进步有限, 以及中药多成分产生的组合效应高度复杂。为解决这些问题, 应利用计算算法和 AI 技术来克服中药网络药理学的局限性。

AI 技术吸引了各个科学领域的研究者, AI 算法的使用可以简化新药发现过程, 促进数据驱动的决策, 加速药物发现过程, 降低错误率。在中药网络药理学中, AI 算法可以辅助预测潜在疾病靶标、解析复杂相互作用, 并预测组合药物的协同效应, 提升预测性能。

2.4.1　ML 改进中药网络药理学中多靶点药物及机制的解析　SVM 是监督学习方法, 可用于分类和回归分析, 经常用于网络药理学中, 以高效筛选确定针对复杂疾病的有效药效分子。Dai 等[27]在中医理论整体观的指导下, 使用基于网络药理学的方法, 分析亨廷顿病的相关蛋白靶标网络, 筛选高结合能力的药物分子。随后使用 SVM 模型进行分析验证。这项研究将 ML 算法与网络药理学相结合, 获得了适用于治疗亨廷顿病的新型中药配方以及反应蛋白, 以整体概念处理复杂疾病, 以便将来探索更多的多靶点药物。

RF 是一种功能强大且用途广泛的监督式 ML 算法, 它可以增长和组合多个去相关的决策树以创建“森林”。在网络药理学中, 特别是对于活性分子和目标蛋白之间相互作用的验证, RF 算法可以提高评分函数的性能。Wu 等[29]基于 FDA 批准药物的药物靶点关系, 使用 RF 方法构建了靶点预测模型, 揭示了 22 种化合物具有多靶点效应, 分析了乌头 (Aconiti Lateralis Praeparata) 的协同活性。类似方法的应用可以帮助全面系统寻找中药的潜在靶点, 同时也为复杂疾病的治疗提供更多选择。

2.4.2　DL 扩展网络药理学中化合物与靶点研究　DL 是 ML 的一个子类型, DL 模型使用不同的算法, 算法

的主干由神经网络架构, 由多层堆叠组成, 允许不同层之间的处理输入。MLP、CNN、DNN和RNN等都有自己的优点和缺点, 已被广泛应用于药物的发现。

多层感知器 (MLP) 模型是一类完全连接的前馈ANN, 模拟神经细胞在人脑中的工作方式。MLP的架构由一系列层、神经元及连接组成。可以用来分析特定化合物与靶蛋白的稳定性、评估化合物的药物相似性等方面[30]。MLP是一种稳健且相对通用的模型, 可以扩展网络药理学中化合物−靶点相互作用的研究。除了MLP之外, 还有多种类型的ANN可以应用于网络药理学中多靶点的发现及药物的虚拟筛选。如由ANN扩展演变而来的DNN, 在输入和输出节点之间包含多个层, 被应用于评估小分子化合物的生物活性, 也有研究利用DNN模型基于单细胞测序数据进行细胞类型特异性基因调控网络的重建[31,32]。CNN模型是一种常规版的MLP, 可与网络药理学结合, 用于预测小分子生物活性, 以及具有生物活性的化合物与靶蛋白之间的亲和力。RNN是一种特殊类型的ANN, 在网络药理学研究中, 可以用于虚拟筛选和新的化合物库的生成, 可有助于多靶点药物的研究[33]。

网络药理学和AI技术的结合为中药药理研究带来了显著的优势。首先, 中药成分复杂, 多成分作用多靶点形成复杂的组合效应, 这一特点使得传统的研究方法难以全面解析其作用机制。网络药理学通过系统性分析中药成分与多种靶点的关系, 全面揭示其药理作用机制。AI技术则增强了这种分析能力, 通过高效的数据处理和模式识别, 加速了研究进程。其次, AI技术在大规模数据处理和分析中的优势, 使得研究者能够快速筛选出具有潜力的中药成分及其靶点, 大大提高了研究效率和精确度。最后, AI技术还可以用于预测中药成分的毒性和安全性, 降低药物研发中的风险, 使中药研究更具科学性和可预测性。

总之, 结合AI技术, 网络药理学不仅提高了多靶点药物发现的效率, 还为中药现代化和精准医疗的发展提供了新思路和强大工具。通过多学科的交叉与合作, 研究人员将能够更好地理解和利用中药的药理优势, 推动中药研究的国际化和现代化发展。

3 大语言模型在中医药研究中的应用与前景

近年来, 随着AI技术及自然语言处理技术的飞速发展, 大语言模型 (large language model, LLM) 的发展与应用也快速崛起。LLM是基于DL的自然语言处理技术, 通过极大规模语料的无监督式预训练, 利用神经网络模拟人类语言生成的过程。中医理论及实践应用涉及大量的文本数据, 涵盖海量的专业术语及病症描述, 可以作为LLM模型训练的资源库。通过对LLM模型的训练, 将传统中医理论与现代科学技术相融合, 可以将中医药数据进行数字化、智能化管理与应用, 提高中医药领域自然语言处理的准确性及实用性, 帮助医生更好地把握患者的病情, 提供更加准确优化的诊疗方案。

随着聊天生成预训练转换器 (chat generative pre-trained transformer, Chat GPT) 等市面上流行的LLM模型的发布, 中医药领域也有诸多聚焦于中医药智能诊断、养生咨询、知识问答以及辅助诊断等类型的LLM问世, 如岐黄问道、ShenNong-TCM、黄帝Huang-Di、数智岐黄、本草智库等。这些模型在提升问诊流程、改善智能化诊断准确性、优化人机交互体验、传播中医药知识等多方面展现出优异的表现。但也因应用场景复杂、模型训练数据资源不足、核心算法适配性局限等原因, 使中医药LLM面临诸多挑战。

LLM不仅在辅助支持临床诊断、优化提升临床诊疗效率及准确性, 以及推广中医药学相关知识方面具有丰富的应用场景。LLM在深入挖掘医学文献、辅助药物研发等科研场景中同样存在巨大的潜力[34]。中药及方剂复杂的药理机制解析, 以及大量中医古籍与前沿进展的整理, 使中药研究人员面临繁重的科研压力。随着LLM模型的发展, 其在中医药文献的学习分析及整理方面的应用, 可以提供更全面准确的中药知识库, 帮助科研人员从繁杂的数据中快速获取有用信息。中药成分复杂多样, 作用机制纷杂不明, LLM模型同样可以应用到中药化学成分及其作用靶点的预测与分析中。同时, 一些模型还具备预测药物之间不良相互作用的功能, 在一定程度上增强了药物研发的安全性[35]。不仅如此, LLM模型也可以应用至中药制剂的设计与优化中, 利用ML算法对中药制剂的制备工艺进行模拟和优化。

LLM模型作为一种AI技术, 赋予了中医药学的传承与发展更多的可能性, 但也面临一些问题与挑战。如模型训练数据更新不及时, 可能导致模型的时效性问题; 传统中医理论中不同流派的标准差异, 导致模型难以掌握真正的语义, 产生不确定的输出结果; 模型使用过程涉及的合法性、知识产权以及与社会价值观不一致的问题。因此, 在未来的研究与应用过程中, 还需中医药科研工作者与AI领域的研究人员共同努力, 解决模型在医学领域应用中存在的问题与挑战, 不断地优化和完善模型, 以适应不同场景的应用需求。

4 总结与展望

中医药在临床实践中积累形成了独特的理论体系, 但这些理论的形成和发展受限于微观水平的对治疗分子机制及效应原理的认知及科学解释, 这些都极

大限制了中医药的现代化创新发展。整合现代科学技术与理论, 对传统中医药的治疗原理进行深度阐释, 可以更好地促进其与现代医学发展相融合。在中医药研究中, AI技术同样面临多重技术挑战和局限性, 中医药数据往往存在不一致性和不完整性, 这可能影响模型训练的准确性。此外, AI模型的复杂性及其“黑箱”性质使其可解释性不足, 而临床应用需要全面理解模型的决策机制以确保安全性和有效性, 这将限制其在医学研究中的接受度和应用范围。因此, 提升数据质量和增强模型解释能力是未来中医药AI应用的关键方向。

作者贡献: 张楠、王晓云负责文献调研及论文撰写; 何谷、韩波负责论文选题和拟定论文框架的指导、把关, 以及论文的修改和定稿。

利益冲突: 所有作者声明不存在利益冲突。

References

[1] Boateng EY, Otoo J, Abaye D. Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: a review [J]. J Data Anal Inf Process, 2020, 8: 341-357.

[2] Hearst MA, Dumais ST, Osuna E, et al. Support vector machines [J]. IEEE Intell Syst, 1998, 13: 18-28.

[3] HO TK. Random decision forests [C]//Proceedings of 3rd International Conference on Document Analysis and Recognition. Montreal: Institute of Electrical and Electronics Engineers, 1995: 278-282.

[4] Friedman JH. Greedy function approximation: a gradient boosting machine [J]. Ann Stat, 2001, 29: 1189-1232.

[5] Ikotun AM, Ezugwu AE, Abualigah L, et al. K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data [J]. Inf Sci, 2023, 622: 178-210.

[6] Patel S, Sihmar S, Jatain A. A study of hierarchical clustering algorithms [C]//2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom): Institute of Electrical and Electronics Engineers, 2015: 537-541.

[7] Lever J, Krzywinski M, Altman N. Principal component analysis [J]. Nat Methods, 2017, 14: 641-642.

[8] Hyvärinen A, Oja E. Independent component analysis: algorithms and applications [J]. Neural Netw, 2000, 13: 411-430.

[9] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proc IEEE, 1998, 86: 2278-2324.

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

[11] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks [J]. Commun ACM, 2020, 63: 139-144.

[12] Ilnicka A, Schneider G. Designing molecules with autoencoder networks [J]. Nat Comput Sci, 2023, 3: 922-933.

[13] Issa NT, Stathias V, Schürer S, et al. Machine and deep learning approaches for cancer drug repurposing [J]. Semin Cancer Biol, 2021, 68: 132-142.

[14] Wallach I, Dzamba M, Heifets A. AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery [J]. Math Z, 2015, 47: 34-46.

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

[16] 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: 3666-3674.

[17] 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.

[18] Scheeder C, Heigwer F, Boutros M. Machine learning and image-based profiling in drug discovery [J]. Curr Opin Syst Biol, 2018, 10: 43-52.

[19] Yu M, Li W, Yu Y, et al. Deep learning large-scale drug discovery and repurposing [J]. Nat Comput Sci, 2024, 4: 600-614.

[20] Gujral TS, Peshkin L, Kirschner MW. Exploiting polypharmacology for drug target deconvolution [J]. Proc Natl Acad Sci U S A, 2014, 111: 5048-5053.

[21] Dong KF, Huo MQ, Sun HY, et al. Mechanism of *Astragalus membranaceus* in the treatment of laryngeal cancer based on gene co-expression network and molecular docking [J]. Sci Rep, 2020, 10: 11184.

[22] Li R, Ma T, Gu J, et al. Imbalanced network biomarkers for traditional Chinese medicine syndrome in gastritis patients [J]. Sci Rep, 2013, 3: 1543.

[23] Wang J, Ma A, Chang Y, et al. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses [J]. Nat Commun, 2021, 12: 1882.

[24] Preuer K, Lewis RPI, Hochreiter S, et al. DeepSynergy: predicting anti-cancer drug synergy with deep learning [J]. Bioinformatics, 2018, 34: 1538-1546.

[25] Kuenzi BM, Park J, Fong SH, et al. Predicting drug response and synergy using a deep learning model of human cancer cells [J]. Cancer Cell, 2020, 38: 672-684.e6.

[26] Niu Q, Li H, Tong L, et al. TCMFP: a novel herbal formula prediction method based on network target's score integrated with semi-supervised learning genetic algorithms [J]. Brief Bioinform, 2023, 24: bbad102.

[27] Dai W, Chen HY, Chen CY. A network pharmacology-based

approach to investigate the novel TCM formula against Huntington's disease and validated by support vector machine model [J]. Evid Based Complement Alternat Med, 2018, 2018: 6020197.

[28] Zhang P, Zhang D, Zhou W, et al. Network pharmacology: towards the artificial intelligence-based precision traditional Chinese medicine [J]. Brief Bioinform, 2023, 25: bbad518.

[29] Wu L, Gao X, Wang L, et al. Prediction of multi-target of Aconiti Lateralis Radix Praeparata and its network pharmacology [J]. China J Chin Mater Med (中国中药杂志), 2011, 36: 2907-2910.

[30] Stokes A, Hum W, Zaslavsky J. A minimal-input multilayer perceptron for predicting drug-drug interactions without knowledge of drug structure [J]. STEM Fellowship J, 2020, 6: 19-23.

[31] Grebner C, Matter H, Kofink D, et al. Application of deep neural network models in drug discovery programs [J]. Chem Med Chem, 2021, 16: 3772-3786.

[32] Chen J, Cheong C, Lan L, et al. DeepDRIM: a deep neural network to reconstruct cell-type-specific gene regulatory network using single-cell RNA-seq data [J]. Brief Bioinform, 2021, 22: bbab325.

[33] Noor F, Asif M, Ashfaq UA, et al. Machine learning for synergistic network pharmacology: a comprehensive overview [J]. Brief Bioinform, 2023, 24: bbad120.

[34] Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records [J]. NPJ Digit Med, 2022, 5: 194.

[35] Lv Q, Chen G, He H, et al. TCMBank-the largest TCM database provides deep learning-based Chinese-Western medicine exclusion prediction [J]. Signal Transduct Target Ther, 2023, 8: 127.