科学通报 2025 年 第 70 卷 第 4-5 期: 481 ~ 491
人工智能在化学研究中的应用与展望专题 进 展
《中国科学》杂志社 SCIENCE CHINA PRESS

人工智能赋能色谱技术研究

林京龙, 莫凡洋*

北京大学材料科学与工程学院, 北京 100871

* 联系人, E-mail: fmo@pku.edu.cn

2024-02-20 收稿, 2024-04-22 修回, 2024-07-03 接受, 2024-07-17 网络版发表
国家自然科学基金(22071004, 21933001, 22150013)资助

摘要 本文首先介绍了人工智能赋能色谱技术(AI4Chromatography)研究的整体流程, 涵盖数据收集、特征工程、模型搭建和可解释性分析等关键环节. 在数据收集部分, 介绍了常见的色谱数据来源, 包括实验、文献和数据库; 在特征工程部分, 阐述了分子表示和实验条件编码的方法; 在模型搭建部分, 介绍了常用的机器学习方法及模型框架; 在可解释性分析部分, 介绍了特征重要性分析和相关性分析等重要研究手段. 随后, 通过详尽讨论AI算法在色谱分析中的各类问题, 特别是定量结构保留关系(quantitative structure-retention relationship, QSRR)研究的应用案例, 向读者展示了利用AI解决色谱问题的全方位视角. 文章最后, 从数据获取、模型构建、知识嵌入与发现三个维度, 探讨了AI4Chromatography研究面临的挑战与未来展望.

关键词 AI4Chromatography, 定量结构保留关系(QSRR), 色谱技术, 机器学习

色谱技术自20世纪初提出以来, 已成为化合物分离与分析的核心方法. 其基于不同组分在流动相与固定相间的亲和力差异实现有效分离. 从最初的纸色谱到如今的高效液相色谱(high performance liquid chromatography, HPLC)和气相色谱(gas chromatography, GC), 色谱技术的每一次进步都极大促进了化学、生物学及环境科学等多领域的研究发展. 同时, 人工智能(artificial intelligence, AI)尤其是机器学习(machine learning, ML), 在化学领域展现出强大的数据处理和分析能力, 被广泛应用于逆合成分析[1,2]、反应产率预测[3]、化学动力学机理解析[4]等方面.

色谱分析的核心挑战在于准确预测和确定色谱条件. 传统方法通常依赖于经验判断及反复试验, 通过试错过程逐步积累分析知识. 然而, AI的引入为这一领域带来了创新性解决方案, 因为它拥有基于数据学习的模型预测能力, 可以快速进行色谱条件的虚拟筛选, 从而有效降低试错的频率和成本. 此外, AI在提升分析的准确性和效率方面展现了明显的优势, 特别是在处理传统色谱技术难以应对的复杂样本时, 其价值更加显著. 在色谱技术领域, ML的应用主要集中在数据的高效处理和精确解析上, 包括优化色谱峰的识别、提升分离效率以及精确预测色谱条件等方面. 色谱技术的高分辨率与AI的数据处理能力的结合, 不仅开辟了实现更快、更准确、成本效益更高分析方法的新路径, 也满足了日益增长的复杂化合物分析需求.

保留值是由分子与色谱柱固定相及流动相之间的相互作用决定的, 因此可以依据分子结构和实验条件来预测保留值, 从而辅助分子鉴定和色谱条件优化. 近十年来, 基于ML的保留值预测模型得到广泛报道, 这些模型通常被称为定量结构保留关系(quantitative structure-retention relationship, QSRR)[5]模型, 开发QSRR模型是AI4Chromatography研究的核心所在. 本文将重点介绍AI4Chromatography研究工作流, 并通过详述AI算法在各种色谱问题中的典型应用案例, 帮助

引用格式: 林京龙, 莫凡洋. 人工智能赋能色谱技术研究. 科学通报, 2025, 70: 481–491
Lin J, Mo F. AI-enabled chromatography research (in Chinese). Chin Sci Bull, 2025, 70: 481–491, doi: 10.1360/TB-2024-0184

读者深入了解这一领域.

1 AI4Chromatography研究的工作流

在AI4Chromatography研究中, 研究要素涵盖数据、计算支持及硬件自动化支持三个方面. 其研究流程如图1所示. 首先进行色谱数据的收集, 然后对分子及色谱实验条件进行详尽的特征工程处理. 接着, 使用机器学习框架构建并训练模型. 最后, 通过特征重要性分析和相关性分析等方法进行模型的可解释性研究, 以提出新的化学见解.

数据主要来自数据库、文献和实验, 研究人员通常需利用爬虫技术和自然语言处理(natural language processing, NLP)技术批量、快速获取目标数据. 计算支持涉及ML算法选择、量子化学计算和分析软件等, 旨在从数据中提炼假设. 硬件自动化则要求小型化和并行化, 如高通量实验装置和流动化学平台, 以获取高质量数据.

在处理色谱问题时, 合理的特征工程至关重要. 其目的是从原始数据中提取有意义的信息, 并转换成适合计算机算法学习和预测的格式, 这一过程需色谱领域专家参与. 特征工程主要包括两个方面: 分子表示和实验条件编码. 分子作为非结构化数据, 可采用SMLES[6]或分子指纹如MACCS[7]、Morgan[7]等进行表示. 在色谱分析中, 重点应放在计算那些影响分子与色谱柱相互作用的分子描述符(molecular descriptor, MD)上, 诸如分子的大小、极性、电荷等特征. 这些描述符对于理解和预测分子在色谱中的行为至关重要. 此外, 采用分子图表示学习[8~11]的方法也是一个有效的选择, 能够省去传统的描述符计算步骤. 色谱实验条件(如温度、流速、流动相、固定相)对分析结果具有重要影响, 常采用独热编码(one-hot encoding, OHE)等技术进行表示. 特征计算完成后, 可以考虑应用主成分分析(principal component analysis, PCA)等降维技术以降低模型过拟合风险.

特征工程后, 可以搭建模型并进行训练和测试, 常用的人工智能模型包括各种集成算法(例如随机森林(random forest, RF)、极限梯度提升(extreme gradient boosting, XGB)、轻量梯度提升机(light gradient boosting machine, LGB))和深度学习算法(例如人工神经网络(artificial neural network, ANN)、卷积神经网络(convolutional neural network, CNN)、图神经网络(graph neural network, GNN)). 这些算法可以通过不同的框架实现, 包括Scikit-learn(sklearn)、PyTorch、华为的MindSpore以及百度的飞桨(PaddlePaddle). 而后, 需要对模型进行可解释性研究, 这包括分析特征的重要性、评估特征与标签间的相关性, 以及应用符号回归(symbolic regression, SR)等方法. 特征的重要性可以通过SHAP(shapley additive explanations)值分析等技术来确定, 而特征与标签间的相关性可以通过计算Spearman系数等方法来评估.

2 AI在色谱研究中的应用

2.1 AI在TLC中的应用

薄层色谱法(thin layer chromatography, TLC)是一种分离混合物中各组分的有效技术. 在此方法中, 待分

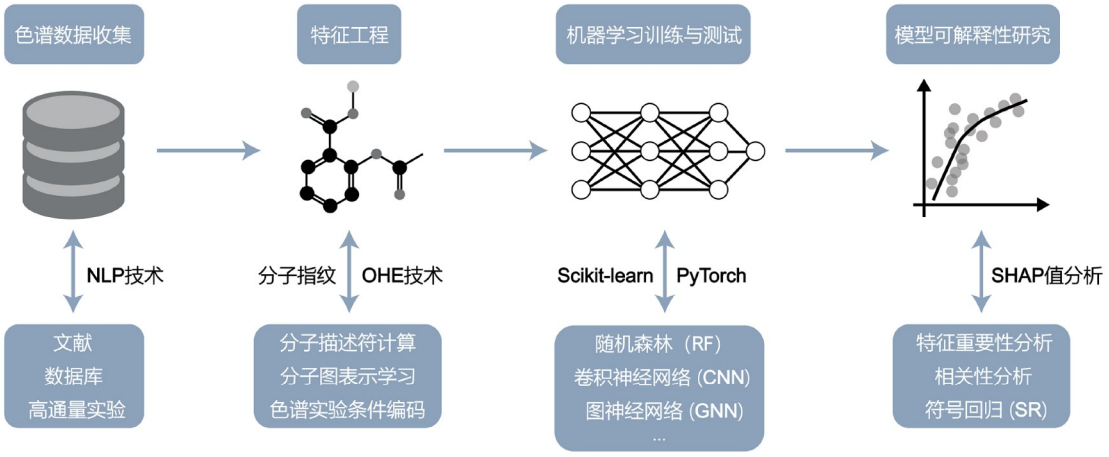

图 1 AI4Chromatography研究流程示意图
Figure 1 Schematic of the AI4Chromatography research workflow

离组分在流动相(即溶剂)的作用下，沿固定相表面移动. 由于组分与固定相的亲和力差异, 各组分的移动速度不同, 进而实现分离.

TLC不仅快速、经济, 而且操作简便, 是有机化学实验室常用的初步筛选工具. 它主要用于筛选和确定最佳溶剂系统, 为后续的柱层析提供依据. 我们课题组结合自动化TLC平台和ML技术，开发了一种预测不同溶剂体系下延迟因子(retardation factor, R_f)的方法[12](图2). 采用的Ensemble模型在测试集上预测的R^2值高达0.961，在训练集未见化合物上预测的R^2为0.887，显示出优异的预测准确性. 特征重要性分析表明, 分子的拓扑极性表面积(TPSA)对R_f值的影响最为显著，这个结论是完全基于数据驱动获得的. 该模型能够精准预测有机化合物在不同溶剂组合下的R_f值曲线，为纯化条件的选择提供依据, 有效提高TLC数据的分析效率.

ML结合TLC不仅在纯化条件预测方面发挥作用, 还广泛应用于食品快速检测领域. Tan等人[13]将TLC与表面增强拉曼散射(SERS)技术相结合，运用PCA进行降维，并通过支持向量机(support vector machine, SVM)进行定量分析, 较传统的偏最小二乘(partial least squares, PLS)法展现出了更佳的预测性能. 此方法适用于海鲜中组胺的快速、灵敏、定量检测，尤其是对变质金枪鱼样本的分析. 同样，Hu等人[14]应用TLC与SERS结合的技术，通过SVM实现了火锅底料中罂粟皮

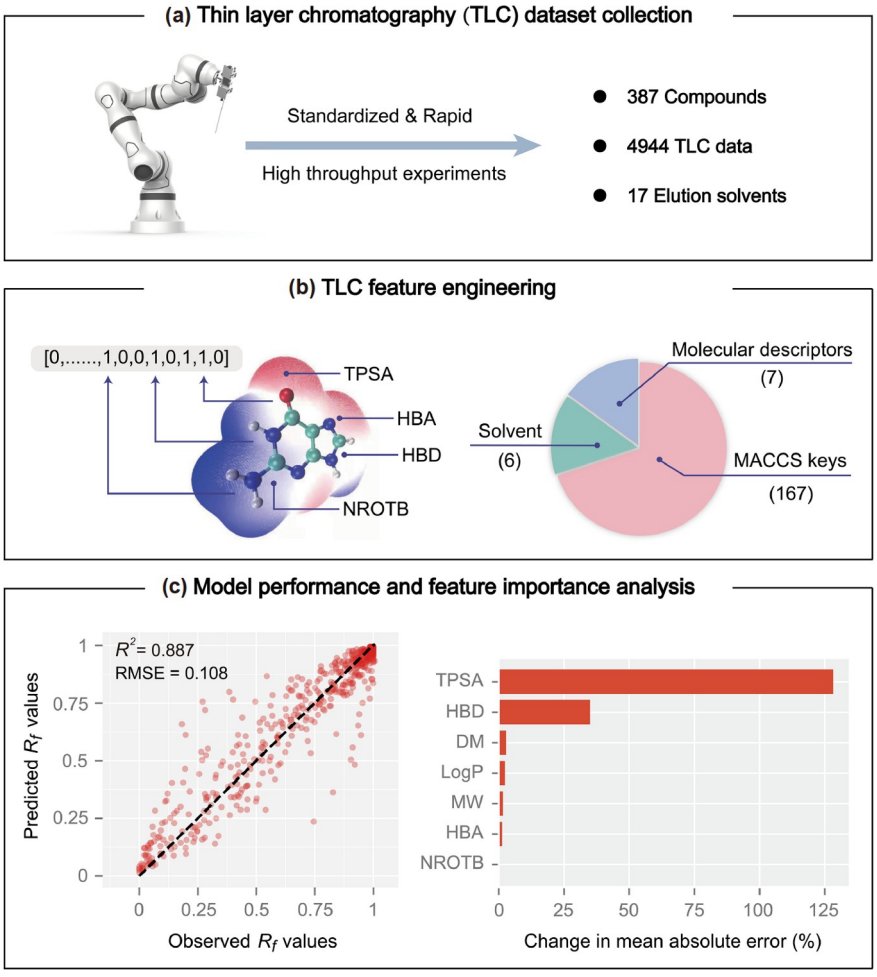

图 2 结合机器人技术和ML模型进行R_f值预测. (a) 开发的自动化TLC平台有助于高通量实验从而产生标准化的TLC数据; (b) 使用分子指纹、分子描述符和溶剂特征来表示TLC 过程; (c) Ensemble模型在训练集未见化合物上预测的R^2为0.887, 通过特征重要性分析发现对R_f值影响最大的分子描述符为TPSA

Figure 2 Integration of robotics and ML models for R_f value prediction. (a) The developed automated TLC platform facilitates high-throughput experiments, generating standardized TLC data. (b) TLC processes are represented using molecular fingerprints, molecular descriptors, and solvent features. (c) The ensemble model achieves an R^2 of 0.887 on previously unseen compounds in the training set, with feature importance analysis identifying TPSA as the most influential molecular descriptor affecting R_f values

的快速检测. 这些研究展示了TLC+ML在提高食品检测效率和准确性方面的巨大潜力.

2.2 AI在液相色谱中的应用

液相色谱法(LC)是利用液态流动相在固定相中的流动来分析和分离化合物的技术, 适用于极性和非极性化合物的鉴定与定量.

Choi等人[15]采用ANN预测丹磺酰化代谢物的液相色谱保留时间(retention time, RT). 通过MORDRED分子描述符计算器进行特征提取, 并运用方差过滤等方法进行降维. 测试集的R^2达到0.975, 相关性分析显示化合物的Log*P*与LC-RT存在高度正相关(Pearson相关系数为0.836), 验证了LogP是预测LC-RT的重要分子描述符[16~18]. 该LC-RT预测模型集成于用户友好的图形用户界面(GUI)中, 被用于鉴定尿液样本中未知RT的丹磺酰化代谢物. 另一研究中Xu等人[19]利用液相色谱-四极杆-飞行时间质谱(LC-Q-TOF-MS)和178种化学标准物质数据, 结合不同类型的分析柱(C18, phenylhexyl, pentafluorophenyl, cyano)开发了基于SVM和RF的QSRR模型, 用于识别食品包装浸出物. 该模型能够有效地识别和排除假阳性结果, 为食品安全监管提供了有力工具.

HPLC是一种先进的色谱技术, 用于分离、鉴定和定量混合物中的各种组分, 被广泛应用于手性分离[20]等领域. 在正相HPLC(NP-HPLC)预测的研究当中, Pérez-Baeza等人[21]运用ANN预测特定手性固定相下的对映体分辨率(*Rs*), 均方误差为0.047. 该研究选取了手性碳参数、分子拓扑参数(源自ChemSpider数据库)及LogP等特征, 并通过特征重要性分析发现表面张力与Rs正相关, 而-NHR基团与Rs负相关. 我们课题组则采用分位数几何增强图神经网络(QGeoGNN)预测HPLC中手性对映体的RT[22](图3), 该模型可推荐手性对映体的最优色谱分离条件, 减少人工试错. 通过NLP技术, 我们从644篇关于不对称催化的文章中提取了25847个分子的数据, 包括SMILES、实验信息、HPLC柱信息及RT. QGeoGNN通过转换分子的三维构象至原子-键图(G)和键-角图(H), 提供了对手性对映体区分的优异能力, 其性能超越传统ML方法. 模型中还嵌入了色谱学领域知识, 以支持多柱通用预测并为对映体分离概率预测提供了坚实基础.

在反相HPLC(RP-HPLC)预测的研究当中, D'Archivio通过ANN预测了16种氨基酸邻苯二甲醛衍生物的保留行为, 考虑了溶剂浓度梯度(*Φ*梯度)、pH梯度以及pH/*Φ*梯度等不同梯度洗脱模式[23]. 通过融合梯度分布描述符和氨基酸标识符作为输入, 实现了在不同溶剂浓度和pH条件下RT的预测, 其中*Φ*梯度、pH梯度、pH/*Φ*梯度的平均预测误差分别为1.1%、1.4%、2.5%. 在另外的研究中, Fedorova等人利用迁移学习(transfer learning, TL)开发了RP-HPLC-RT预测模型[24], 预训练模型的平均绝对误差(mean absolute error, MAE)为0.58 min. 该研究使用METLIN SMRT数据集中的80038个分子信息, 这些分子均在Zorbax Extend-C18反相柱上分离. 将分子以SMILES形式表示并计算one-hot矩阵, 以供一维卷积神经网络(1D-CNN)预训练. 在TL过程中, 通过冻结第一个卷积层的权重(保留通用特征)并对其余层进行微调, 使TL模型在五个外部数据集上的性能均超过了从零开始训练的模型.

亲水相互作用液相色谱(hydrophilic interaction liquid chromatography, HILIC)是专门用于分离极性化合物的液相色谱技术, 特别适合分离那些在RP-HPLC中难以分离的物质. 其工作原理虽与NP-HPLC相似, 但采用了类似RP-HPLC的流动相. Cao等人运用RF方法发展了HILIC的QSRR模型, MAE为0.52 min[16]. Torigoe等人[25]也应用RF技术开发了一种基于HILIC/阴离子交换/高分辨率串联质谱(HILIC/AEX/HRMS/MS)的QSRR模型, MAE为0.80 min, 该模型被成功应用于人血浆非靶向代谢组学分析, 识别出62种已知代谢物并推断出216种未知极性代谢物. Taraji等人[26]通过PLS预测了五种不同固定相下HILIC的RT, 并利用遗传算法(genetic algorithm, GA)进行MD的筛选. Yang等人[27]采用GNN结合TL技术对HILIC的RT进行预测, 首先利用约306K分子的伪标记数据集进行预训练, 随后通过TL对目标色谱的实验数据进行模型微调, 测试集最低MAE为0.64 min.

2.3 AI在气相色谱中的应用

GC是一种用于分析和分离易挥发化合物的色谱技术, 其通过控制气体流动相在固定相(即色谱柱)内的流动实现分离. 该技术在石油化工[28]、代谢组学[29,30]、食品分析[31,32]、农业[33,34] 等多个领域有着广泛应用, 对于挥发性和半挥发性化合物的鉴定与定量分析具有重要意义.

在GC预测研究领域中, 保留指数(retention index, RI)是一个核心参数, 它反映了特定固定相(stationary

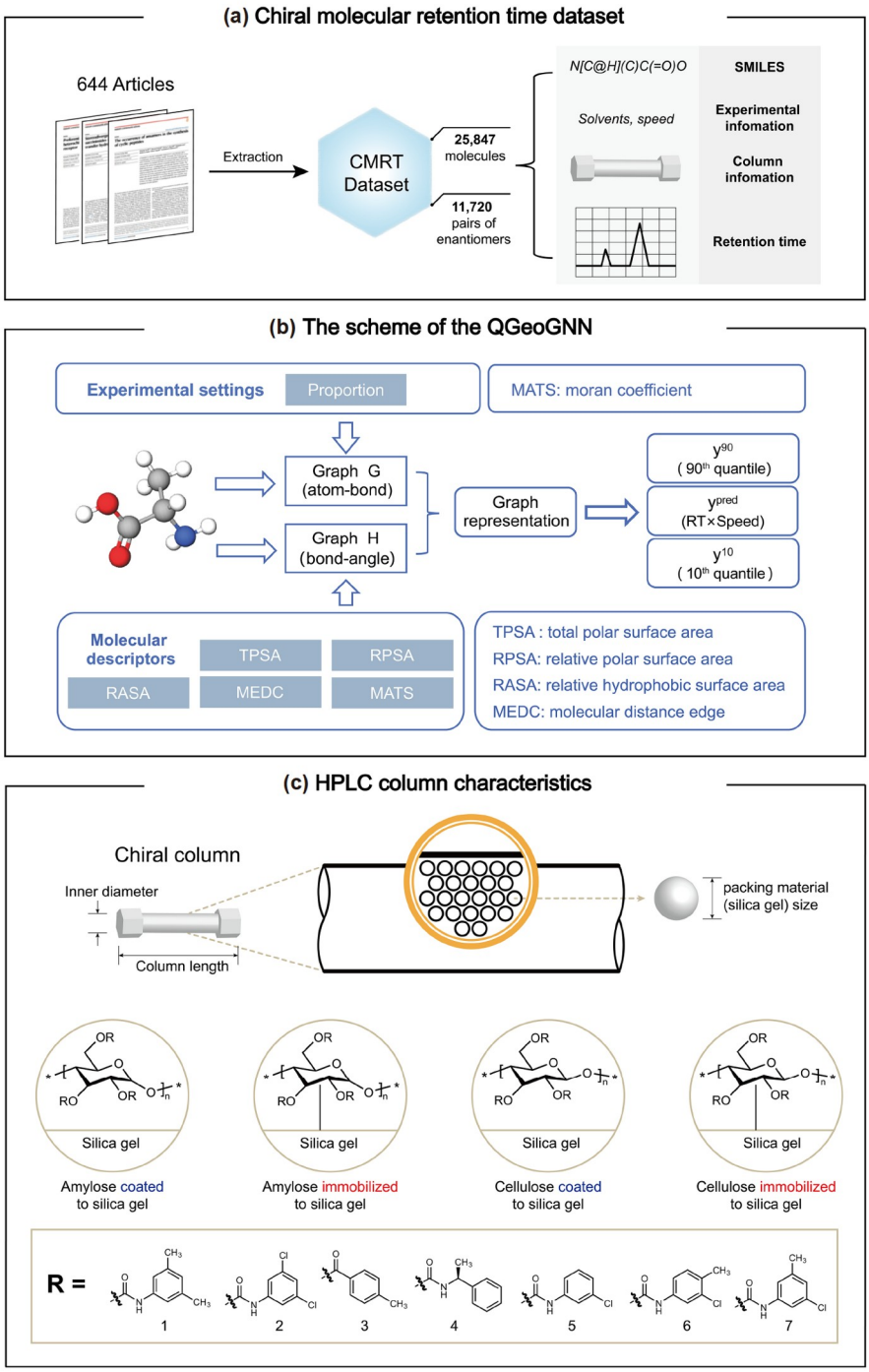

图 3 使用QGeoGNN预测HPLC的RT. (a) 通过NLP批量提取644篇论文的RT数据, 以及分子信息和色谱条件; (b) 将分子表示为原子-键图和键-角图, 并且嵌入实验条件和MD, 用以全面表示分子的三维信息, 从而实现对手性对映体的区分. (c) 手性固定相填充物的粒径、基体、取代基和连接类型(固定化或涂覆)都会影响HPLC柱的手性识别能力, 需在ML建模过程中考虑. 内径和柱长也会影响手性识别能力, 但在商业HPLC柱中这些参数是保持不变的

Figure 3 Predicting HPLC retention time using QGeoGNN. (a) RT data, molecular information, and chromatographic conditions were extracted from 644 papers using NLP techniques. (b) Molecules were represented as atom-bond and bond-angle graphs, with experimental conditions and molecular descriptors embedded to comprehensively represent the three-dimensional molecular information, enabling the distinction of chiral enantiomers. (c) The chiral recognition ability of HPLC columns is influenced by the particle size, matrix, substituents, and linkage type (immobilized or coated) of the chiral stationary phase, which must be considered in ML modeling. Although internal diameter and column length also affect chiral recognition, these parameters are consistent in commercial HPLC columns

phase, SP)对分子的保留能力[35]. RI的独特之处在于其对色谱条件的独立性, 它主要受SP和分子结构影响[36], 因此能够跨越不同GC条件及仪器应用, 这样的特点使其成为化合物鉴别中的关键参数. 目前, NIST 20数据库收录了大约14万种化合物的RI数据, 是最大的RI数据库. Matyushin等人在RI预测方面进行了开创性的研究. 他们首先利用CNN针对特定非极性固定相的RI进行预测[37], 将分子的SMILES转化为one-hot矩阵, 再通过卷积方法缩减特征图尺寸, 并通过平均池化后输入全连接层以实现RI预测(图4). 此外, 他们还发展了集成1D-CNN、2D-CNN、残差网络及XGB等技术的多模态机器学习方法[38]. 在后续的工作中, 他们利用ANN与CNN相结合的技术, 成功预测了极性及中极性固定相(如聚乙二醇、DB-WAX、DB-624、DB-210、DB-1701、OV-17)的RI, 测试集的R^2值可达0.989[39]. Veselinović等人[40]则开发了基于图不变量和SMILES预测RI的新策略, 并通过蒙特卡罗方法调整描述符权重以找到最佳描述符组合, 最终测试集的R^2值达到0.937. 然而, 尽管这些RI预测有助于了解化合物的保留特性, 但却无法用于推荐待分离化合物体系(如位置异构体系/顺反异构体系)的最优色谱条件. 相较之下, 直接预测GC的RT能够更精确地反映特定色谱条件下的分析结果, 对色谱条件的快速虚拟筛选具有重要意义, 但相关研究尚处于起步阶段.

气相色谱-质谱联用(gas chromatography-mass spectrometry, GC-MS)被广泛应用于代谢组学研究, 特别是在挥发性有机化合物(VOC)的检测方面[30]. Jirayupat等人[41]利用GC-MS谱图构建二维质谱图(2D MS map), 并通过逻辑回归算法分析, 成功鉴定了呼吸样本中以ppb级别存在的肺癌标志物. Qiu等人[42]开发的Me-

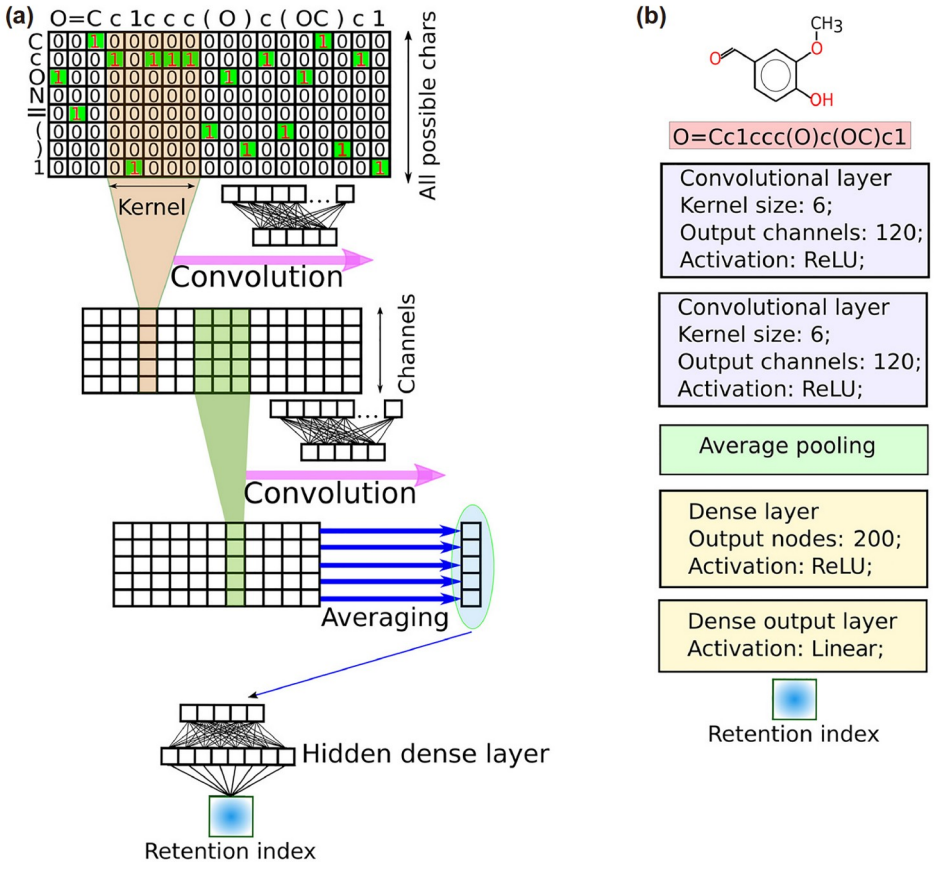

图 4 (网络版彩色)RI预测模型的架构[37]. (a) 模型示意图. 首先将分子的SMILES表示为one-hot矩阵, 其次使用CNN进行特征提取, 池化后的特征输入到全连接层进行RI预测. (b) 模型参数设定方案

Figure 4 (Color online) Architecture of the RI prediction model. (a) Schematic of the model: Molecular SMILES strings are first converted into one-hot matrices, followed by feature extraction using a CNN. The pooled features are then fed into a fully connected layer for RI prediction. (b) Configuration of model parameters

tExpert系统结合了多种技术，包括计算机衍生化、代谢物相似性评估、Kovats-RI预测及子结构预测(采用PLS方法)，为代谢物的快速鉴定提供了有效手段. Fan等人[43]提出的基于伪孪生卷积神经网络(pseudo-siamese convolutional neural network, PSCNN)的自动解析方法，利用NIST质谱库中超过26万个的EI-MS谱图，能够预测化合物的选择性区域(PSCNN1)和洗脱区域(PSCNN2).

2.4 AI在其他色谱中的应用

疏水相互作用色谱(hydrophobic interaction chromatography，HIC)是一种基于分析物与色谱介质疏水性区域的相互作用进行分离的技术. 固定相通常含有疏水基团，流动相则为水溶性缓冲液. 通过调整流动相的离子强度或加入有机溶剂，可控制分析物与固定相的相互作用，进而调节分析物的RT. Jain等人[44]利用RF从单克隆抗体序列直接预测氨基酸侧链的表面暴露度，模型的MAE为4.6%. 该模型能实时评估抗体的疏水性，从而辅助药物发现过程中抗体优先级的确定，并指导抗体的合理工程设计以降低疏水性.

凝胶渗透色谱(gel permeation chromatography, GPC)是一种专门用于测定高分子物质分子量分布的技术. 其工作原理依赖于分子尺寸的分离，而非分子间的相互作用力. Nagy等人[45]开发了两种ANN模型，成功地从GPC数据中提取共聚物的详细信息，包括分子量和组成.

离子交换色谱技术(ion exchange chromatography, IEC)基于分析物离子与色谱柱填料上离子交换基团之间的亲和力差异进行分离，被广泛应用于蛋白质、肽、核酸等生物大分子的分离与纯化. Giese等人[46]应用线性回归(linear regression，LR)、ANN等模型预测亲水性强阴离子交换色谱中肽的RT，并通过特征重要性分析发现带电残基是RT的主要影响因素. Nikita等人[47]采用强化学习(reinforcement learning，RL)优化阳离子交换色谱的工艺流速，设定奖励函数后，RL能推荐最优流速以实现产量最大化.

3 总结与展望

本文首先概述了AI4Chromatography研究的流程，涵盖了研究要素、特征工程、模型构建、可解释性分析及模型验证等关键步骤. 接着，通过详细阐述AI算法在不同色谱问题中的应用实例(尤其是QSRR研究)，为读者提供了对色谱问题AI解决方案的全面理解. 尽管AI4Chromatography研究已得到了初步的发展，但仍然存在着许多问题亟待解决，下面将从3个方面分析AI4-Chromatography研究的挑战与未来发展方向:

(1) 数据困境. 高质量、大数据是AI4Chromatography研究的根本. 目前，大部分QSRR研究依赖于数据库和文献，但部分数据库的不开源性和文献中色谱条件描述的不一致性，加之数据分布的不均匀性，给ML研究带来了困难. 因此，未来应当着力推进实验室自动化技术，以实现高通量的标准化数据采集. 例如，我们课题组已在自动化TLC分析平台结合AI应用方面进行了初步研究[12]. 同时，亦应致力于开发课题组色谱数据共享平台，建设AI4Chromatography研究社区和开源色谱数据库，以形成健康的研究生态环境，汇聚众智.

(2) AI4Chromatography模型的发展方向. AI4Chromatography研究已逐渐发展为硬编码与软编码相结合的方式. 硬编码即在模型中直接应用预设的固定特征，如MD计算就属于此范畴. 目前，主流特征筛选算法包括卡方/方差过滤、PCA、GA、蒙特卡罗方法等，未来需开发更高效的特征筛选算法. 软编码则指模型训练过程中自动从数据学习和提取特征的方法，例如CNN自动提取SMILES特征、GNN提取图特征等. 当前，随着ChatGPT的发展，基于注意力机制的Transformer模型已经吸引了化学研究者的关注[48~52]，它是一种软编码解决方案，在处理序列数据方面展现出显著优势. 然而，利用Transformer的AI4Chromatography研究尚少，这将是未来重点发展方向之一，目标是开发适用于色谱领域的大模型. 此外，未来的主要发展方向还应包括半监督学习(如伪标记[53]技术等)、多模态学习、迁移学习，这些方法可最大限度地利用有限的色谱数据，并实现模型在不同色谱系统中的有效应用.

(3) 应大力发展色谱知识嵌入和知识发现技术. 目前的AI4Chromatography算法普遍缺少对物理概念的整合，主要依赖于从数据中识别模式或规律. 这种方法可能导致得出的结果违背物理定律，进而影响模型的收敛性和稳定性. 因此，如何将色谱领域知识嵌入到AI算法中，构建物理合理、数学精确、计算高效的模型，是未来的重大挑战. 例如，我们课题组构建了一个物理约束神经网络(physics-informed neural network, PINN)，将TLC的单调规则纳入损失函数，同时还构建了一个硬约束，通过sigmoid激活函数将输出限制为(0,1)，使得与R_f值的现实世界行为保持一致[12].

另外, AI4Chromatography研究的最终目标是通过QSRR模型学习新知识和规则, 以促进科学发现. 然而, 神经网络因其“黑箱”特性而经常受到诟病, 其模型可解释性方面的研究仍然不足. 未来研究的重点应是发展基于AI算法的知识发现技术, 从数据中挖掘尚未被发现的知识, 以突破现有的认知限制. 其中主要的策略包括稀疏回归、符号数学等. 例如, Jiang等人[54]利用SISSO方法获得了谱构效关系的简明数学公式, 该公式具有较强的可解释性, 实现了跨不同系统的知识迁移.

我们对未来的愿景是: 通过实验室自动化快速获取色谱数据, 研究人员利用这些高质量的数据构建具有知识嵌入的AI算法, 并通过知识发现技术, 发现新的色谱知识和规则, 进而进一步指导算法设计, 形成闭环. AI4Chromatography研究领域蕴含着丰富的探索潜力, 我们期望本文分享的洞见与观点能激发研究人员的灵感, 共同推动该领域的进步与发展.

参考文献

1 Klucznik T, Mikulak-Klucznik B, McCormack M P, et al. Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. Chem, 2018, 4: 522–532
2 Segler M H S, Preuss M, Waller M P. Planning chemical syntheses with deep neural networks and symbolic AI. Nature, 2018, 555: 604–610
3 Ahneman D T, Estrada J G, Lin S, et al. Predicting reaction performance in C–N cross-coupling using machine learning. Science, 2018, 360: 186–190
4 Burés J, Larrosa I. Organic reaction mechanism classification using machine learning. Nature, 2023, 613: 689–695
5 Taraji M, Haddad P R, Amos R I J, et al. Chemometric-assisted method development in hydrophilic interaction liquid chromatography: A review. Anal Chim Acta, 2018, 1000: 20–40
6 Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci, 1988, 28: 31–36
7 Cereto-Massagué A, Ojeda M J, Valls C, et al. Molecular fingerprint similarity search in virtual screening. Methods, 2015, 71: 58–63
8 Kearnes S, McCloskey K, Berndl M, et al. Molecular graph convolutions: Moving beyond fingerprints. J Comput Aided Mol Des, 2016, 30: 595–608
9 Schütt K T, Sauceda H E, Kindermans P J, et al. SchNet—A deep learning architecture for molecules and materials. J Chem Phys, 2018, 148: 241722
10 Fang X, Liu L, Lei J, et al. Geometry-enhanced molecular representation learning for property prediction. Nat Mach Intell, 2022, 4: 127–134
11 Li S W, Xu L C, Zhang C, et al. Reaction performance prediction with an extrapolative and interpretable graph model based on chemical knowledge. Nat Commun, 2023, 14: 3569
12 Xu H, Lin J, Liu Q, et al. High-throughput discovery of chemical structure-polarity relationships combining automation and machine-learning techniques. Chem, 2022, 8: 3202–3214
13 Tan A, Zhao Y, Sivashanmugan K, et al. Quantitative TLC-SERS detection of histamine in seafood with support vector machine analysis. Food Control, 2019, 103: 111–118
14 Hu X, Fang G, Han A, et al. Rapid detection of *Pericarpium papaveris* in hot pot condiments using thin-layer chromatography and surface enhanced Raman spectroscopy combined with a support vector machine. Anal Methods, 2017, 9: 2177–2182
15 Choi E, Yoo W J, Jang H Y, et al. Machine learning liquid chromatography retention time prediction model augments the dansylation strategy for metabolite analysis of urine samples. J Chromatogr A, 2023, 1705: 464167
16 Cao M, Fraser K, Huege J, et al. Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics. Metabolomics, 2015, 11: 696–706
17 Park H, Lee J M, Kim J Y, et al. Prediction of liquid chromatography retention times of erectile dysfunction drugs and analogues using chemometric approaches. J Liquid Chromatogr Relat Technol, 2017, 40: 790–797
18 Jang I, Lee J, Lee J, et al. LC–MS/MS software for screening unknown erectile dysfunction drugs and analogues: Artificial neural network classification, peak-count scoring, simple similarity search, and hybrid similarity search algorithms. Anal Chem, 2019, 91: 9119–9128
19 Xu Z, Chughtai H, Tian L, et al. Development of quantitative structure-retention relationship models to improve the identification of leachables in food packaging using non-targeted analysis. Talanta, 2023, 253: 123861
20 Okamoto Y, Ikai T. Chiral HPLC for efficient resolution of enantiomers. Chem Soc Rev, 2008, 37: 2593–2608
21 Pérez-Baeza M, Martín-Biosca Y, Escuder-Gilabert L, et al. Artificial neural networks to model the enantioresolution of structurally unrelated neutral and basic compounds with cellulose tris(3,5-dimethylphenylcarbamate) chiral stationary phase and aqueous-acetonitrile mobile phases. J

Chromatogr A, 2022, 1672: 463048
22 Xu H, Lin J, Zhang D, et al. Retention time prediction for chromatographic enantioseparation by quantile geometry-enhanced graph neural network. Nat Commun, 2023, 14: 3095
23 D'Archivio A A. Artificial neural network prediction of retention of amino acids in reversed-phase hplc under application of linear organic modifier gradients and/or pH gradients. Molecules, 2019, 24: 632
24 Fedorova E S, Matyushin D D, Plyushchenko I V, et al. Deep learning for retention time prediction in reversed-phase liquid chromatography. J Chromatogr A, 2022, 1664: 462792
25 Torigoe T, Takahashi M, Heravizadeh O, et al. Predicting retention time in unified-hydrophilic-interaction/anion-exchange liquid chromatography high-resolution tandem mass spectrometry (Unified-HILIC/AEX/HRMS/MS) for comprehensive structural annotation of polar metabolome. Anal Chem, 2024, 96: 1275–1283
26 Taraji M, Haddad P R, Amos R I J, et al. Prediction of retention in hydrophilic interaction liquid chromatography using solute molecular descriptors based on chemical structures. J Chromatogr A, 2017, 1486: 59–67
27 Yang Q, Ji H, Fan X, et al. Retention time prediction in hydrophilic interaction liquid chromatography with graph neural network and transfer learning. J Chromatogr A, 2021, 1656: 462536
28 Pollo B J, Alexandrino G L, Augusto F, et al. The impact of comprehensive two-dimensional gas chromatography on oil & gas analysis: Recent advances and applications in petroleum industry. TrAC Trends Anal Chem, 2018, 105: 202–217
29 Dunn W B, Broadhurst D, Begley P, et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat Protoc, 2011, 6: 1060–1083
30 Lubes G, Goodarzi M. GC–MS based metabolomics used for the identification of cancer volatile organic compounds as biomarkers. J Pharm Biomed Anal, 2018, 147: 313–322
31 Lopez P, van Sisseren M, De Marco S, et al. A straightforward method to determine flavouring substances in food by GC–MS. Food Chem, 2015, 174: 407–416
32 Pereira V L, Fernandes J O, Cunha S C. Comparative assessment of three cleanup procedures after QuEChERS extraction for determination of trichothecenes (type A and type B) in processed cereal-based baby foods by GC–MS. Food Chem, 2015, 182: 143–149
33 Sugitate K, Saka M, Serino T, et al. Matrix behavior during sample preparation using metabolomics analysis approach for pesticide residue analysis by GC-MS in agricultural products. J Agric Food Chem, 2012, 60: 10226–10234
34 Sequeiros A, Labidi J. Characterization and determination of the S/G ratio via Py-GC/MS of agricultural and industrial residues. Industrial Crops Products, 2017, 97: 469–476
35 Vigdergauz M S, Martynov A A. Some applications of the gas chromatographic linear retention index. Chromatographia, 1971, 4: 463–467
36 Tarjan G, Nyiredy S, Györ M, et al. Thirtieth anniversary of the retention index according to Kováts in gas-liquid chromatography. J Chromatogr A, 1989, 472: 1–92
37 Matyushin D D, Sholokhova A Y, Buryak A K. A deep convolutional neural network for the estimation of gas chromatographic retention indices. J Chromatogr A, 2019, 1607: 460395
38 Matyushin D D, Buryak A K. Gas Chromatographic retention index prediction using multimodal machine learning. IEEE Access, 2020, 8: 223140–223155
39 Matyushin D D, Sholokhova A Y, Buryak A K. Deep learning based prediction of gas chromatographic retention indices for a wide variety of polar and mid-polar liquid stationary phases. Int J Mol Sci, 2021, 22: 9194
40 Veselinović A M, Velimorović D, Kaličanin B, et al. Prediction of gas chromatographic retention indices based on Monte Carlo method. Talanta, 2017, 168: 257–262
41 Jirayupat C, Nagashima K, Hosomi T, et al. Image processing and machine learning for automated identification of chemo-/biomarkers in chromatography–mass spectrometry. Anal Chem, 2021, 93: 14708–14715
42 Qiu F, Lei Z, Sumner L W. MetExpert: An expert system to enhance gas chromatography–mass spectrometry-based metabolite identifications. Anal Chim Acta, 2018, 1037: 316–326
43 Fan Y, Yu C, Lu H, et al. Deep learning-based method for automatic resolution of gas chromatography-mass spectrometry data from complex samples. J Chromatogr A, 2023, 1690: 463768
44 Jain T, Boland T, Lilov A, et al. Prediction of delayed retention of antibodies in hydrophobic interaction chromatography from sequence using machine learning. Bioinformatics, 2017, 33: 3758–3766
45 Nagy T, Róth G, Benedek M, et al. Enhanced copolymer characterization for polyethers using gel permeation chromatography combined with artificial neural networks. Anal Chem, 2023, 95: 10504–10511
46 Giese S H, Ishihama Y, Rappsilber J. Peptide retention in hydrophilic strong anion exchange chromatography is driven by charged and aromatic residues. Anal Chem, 2018, 90: 4635–4640

47 Nikita S, Tiwari A, Sonawat D, et al. Reinforcement learning based optimization of process chromatography for continuous processing of biopharmaceuticals. Chem Eng Sci, 2021, 230: 116171
48 Schwaller P, Laino T, Gaudin T, et al. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. ACS Cent Sci, 2019, 5: 1572–1583
49 Chithrananda S, Grand G, Ramsundar B. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. 2020
50 Ross J, Belgodere B, Chenthamarakshan V, et al. Large-scale chemical language representations capture molecular structure and properties. Nat Mach Intell, 2022, 4: 1256–1264
51 Frey N C, Soklaski R, Axelrod S, et al. Neural scaling of deep chemical models. Nat Mach Intell, 2023, 5: 1297–1305
52 Kang Y, Park H, Smit B, et al. A multi-modal pre-training transformer for universal transfer learning in metal–organic frameworks. Nat Mach Intell, 2023, 5: 309–318
53 Arazo E, Ortego D, Albert P, et al. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: 2020 International Joint Conference on Neural Networks (IJCNN). New York: IEEE, 2020. doi: 10.1109/ijcnn48605.2020.9207304
54 Wang X, Jiang S, Hu W, et al. Quantitatively determining surface–adsorbate properties from vibrational spectroscopy with interpretable machine learning. J Am Chem Soc, 2022, 144: 16069–16076

Summary for “人工智能赋能色谱技术研究”

AI-enabled chromatography research

Jinglong Lin & Fanyang Mo*

School of Materials Science and Engineering, Peking University, Beijing 100871, China
* Corresponding author, E-mail: fmo@pku.edu.cn

Chromatographic techniques, developed since the early 20th century, are fundamental for compound separation and analysis, evolving from paper chromatography to high-performance liquid chromatography (HPLC) and gas chromatography (GC). These advancements have significantly propelled research in chemistry, biology, and environmental science. Concurrently, artificial intelligence (AI) and machine learning (ML) have demonstrated robust data processing and analysis capabilities in the chemical domain, being extensively applied in retrosynthetic analysis, reaction yield prediction, and elucidation of chemical kinetics.

The primary challenge in chromatography lies in accurately predicting and determining chromatographic conditions, traditionally dependent on empirical judgment and iterative experimentation. AI introduces innovative solutions with data-driven model prediction capabilities, enabling rapid virtual screening of conditions, thereby reducing trial-and-error frequency and cost. AI’s advantages in improving analytical accuracy and efficiency are particularly notable when handling complex samples that traditional techniques struggle with.

This review details the AI4Chromatography research workflow, encompassing data collection, feature engineering, model building, and interpretability. Key applications include predicting retention factors in thin-layer chromatography (TLC), retention times in HPLC, and retention indices in GC. The development of quantitative structure-retention relationship (QSRR) models, pivotal to AI4Chromatography, is highlighted. The workflow involves collecting data from databases, literature, and experiments, using natural language processing (NLP) and web scraping for rapid data acquisition.

Feature engineering, crucial for meaningful data extraction, involves molecular representation and experimental condition encoding. Common AI models include random forest (RF), extreme gradient boosting (XGB), light gradient boosting machine (LGB), artificial neural networks (ANN), convolutional neural networks (CNN), and graph neural networks (GNN), implemented via frameworks such as Scikit-learn, PyTorch, MindSpore, and PaddlePaddle. Each model offers unique strengths in handling different chromatographic data and can be tailored to specific analytical needs.

Applications in TLC involve predicting retardation factors using ensemble models with high accuracy. In HPLC, models predict retention times and enantiomeric resolutions, aiding in optimal condition selection. For GC, retention indices are predicted using CNN and ANN, facilitating the identification of volatile compounds. These AI-driven models have significantly enhanced the efficiency and accuracy of chromatographic analyses, enabling reliable results with less experimental effort.

Despite initial advancements, several challenges remain. Firstly, high-quality, large datasets are essential, yet inconsistencies in database openness and chromatographic condition descriptions pose difficulties. Secondly, AI4Chromatography research must evolve to integrate both hard-coded features and soft-coded features, such as those derived from Transformer models. Future directions include semi-supervised learning, multimodal learning, and transfer learning to maximize data utility. Thirdly, embedding chromatographic knowledge into AI algorithms is crucial for building physically accurate, mathematically precise, and computationally efficient models. This includes developing physics-informed neural networks (PINNs) and advancing knowledge discovery techniques to overcome the “black box” nature of neural networks.

The vision for the future is to rapidly acquire chromatographic data through automation, build AI models with embedded knowledge, and discover new chromatographic insights, forming a virtuous cycle. The AI4Chromatography field holds immense potential, and this review aims to inspire further progress and innovation.

doi: 10.1360/TB-2024-0184