doi：10.12452/j.fxcsxb.26011004

综 述

液相色谱-质谱数据处理算法的演进分析

谢志明，高树辉*

（中国人民公安大学 侦查学院，北京 100038）

摘 要：液相色谱-质谱联用技术(LC-MS)结合了液相色谱的高分离能力与质谱的高灵敏度鉴定能力，已成为生命科学、临床医学及环境科学等领域的核心分析工具。但其产出的高维、复杂且含有噪声的原始数据，需高效、稳健的数据处理算法，实现从原始信号到有效科学信息的转化。该文聚焦LC-MS数据处理流程中的三大关键任务——信号提取、时空对齐与物质鉴定，深入剖析其方法论演进与内在逻辑，并进一步揭示当前算法研究面临的泛化性不足、基准数据集缺失以及软件生态碎片化等共性挑战。针对上述痛点，提出物理引导的人工智能建模与标准化基准体系构建方向，强调算法开发需融合仪器机理与数据科学，以期构建高鲁棒、可解释、云端协同的新一代LC-MS数据处理技术。

关键词：液相色谱-质谱；数据处理；深度学习；特征提取；保留时间校正；物质鉴定

中图分类号：O657.63 **文献标识码：**A **文章编号：**1004-4957(2026)05-1108-15

Evolution and Pathways of Liquid Chromatography-Mass Spectrometry Data Processing Algorithms

XIE Zhi-ming，GAO Shu-hui*

(School of Investigation，People's Public Security University of China，Beijing 100038，China)

Abstract： Liquid chromatography-mass spectrometry (LC-MS) combines high separation ability with high sensitivity. It has become a core analysis tool. It is widely used in life science. It is also essential in clinical medicine and environmental science. However，LC-MS produces high-dimensional data. The raw data is complex and contains significant noise. Therefore，efficient data processing algorithms are necessary. These algorithms must be robust. The goal is to convert raw signals into effective scientific information. This paper focuses on the data processing workflow. It highlights three key tasks. The first task is signal extraction. The second task is spatiotemporal alignment. The third task is substance identification. This review analyzes the methodological evolution of these tasks. It also explores their internal logic deeply. Furthermore，the study reveals common challenges in current research. One major problem is the lack of algorithm generalization. Another issue is the absence of standardized benchmark datasets. The fragmentation of the software ecosystem is also a critical bottleneck. To address these pain points，specific directions are proposed. One direction is physics-guided artificial intelligence (AI) modeling. Another is the construction of standardized benchmark systems. Algorithm development must integrate instrument mechanisms with data science. This integration is crucial. It aims to build next-generation LC-MS technologies. These technologies should be high-robustness and interpretable. Cloud collaboration is also a key feature for the future.

Key words： liquid chromatography-mass spectrometry；data processing；deep learning；feature extraction；retention time alignment；substance identification

液相色谱-质谱联用技术(LC-MS)凭借其高灵敏度与宽动态范围，已成为生命科学等领域的核心分析方法。然而，随着硬件分辨率提升与数据非依赖性采集(DIA)等模式的革新，LC-MS数据呈指数级增长，复杂度剧增，尤其在非靶向筛查[1]等复杂体系分析中，对从海量高维噪声数据中精准提取信息提出了更高要求。以基于R语言的开源软件XCMS[2]和模块化数据处理软件MZmine[3]为代表的经典工具基于预设数学模型，在常规分析中表现稳健，但其固定参数模型在面对复杂基质、低信噪比信号及大

收稿日期：2026-01-10；修回日期：2026-02-14
*通讯作者：高树辉，博士，教授，研究方向：刑事科学技术，E-mail：gaoshuhui@ppsuc.edu.cn
网络首发日期：2026-03-27

规模保留时间漂移时，常面临灵敏度与特异性的失衡。当前，LC-MS 数据处理正经历从“规则驱动”向“数据驱动”的范式变革。机器学习与深度学习技术使算法能够从数据中直接学习特征，模仿专家判断，其应用已渗透至从峰解卷积到化合物结构推断的各个环节。然而，尽管新模型在特定数据集上表现优异，但由于缺乏统一基准测试与存在跨实验室数据变异，模型泛化能力面临严峻挑战。同时，大量算法仍停留于脚本阶段，缺乏集成化的软件生态，阻碍了其在实际场景中的落地应用。

本文聚焦LC-MS数据处理三大核心环节的技术演进：针对信号提取，探讨如何利用深度学习模型实现复杂噪声与基线的端到端处理；围绕色谱时空对齐，阐释其从局部峰形匹配向全局轨迹聚类的智能化发展；在物质鉴定方面，分析其从依赖标准品库匹配向基于裂解规则与生成式模型的结构推理跨越。进而批判性指出当前领域因基准缺失与生态碎片化引发的算法泛化危机，并提出通过物理引导机器学习与标准化评估体系，构建可解释、强鲁棒的新一代数据处理范式。旨在通过梳理技术演进逻辑与构建全链条认知框架，推动LC-MS分析从离散工具集成向自适应智能生态系统跃迁，为复杂体系的全景表征、痕量物发现及未知风险确证提供关键技术支撑，对提升分析通量、准确性与数据挖掘深度具有科学意义。

1　LC-MS数据处理算法基本原理

信号提取、时空对齐与物质鉴定是LC-MS三大核心任务，将高维、非结构化的连续离子流数据转化为结构化的化合物特征矩阵，这一过程在逻辑上对应为3个连续的逆问题求解步骤：降维与去噪、误差校正与一致性重建以及赋予化学语义。

1.1　信号提取

信号提取旨在将原始LC-MS数据中连续的离子强度信号转化为离散化学特征。每个特征由3个核心参数定义：质荷比(m/z)、保留时间(RT)和积分强度。m/z是质谱直接测量的物理量，作为区分化合物的第一维坐标，高分辨质谱可将其精确至小数点后4~5位，从而有效分辨质量近似的异构体。RT反映化合物的极性、疏水性等理化性质，是第二维坐标，但因实验条件波动易发生非线性漂移，需后续算法校正。积分强度(峰面积或峰高)代表组分的相对丰度，是比较样本间差异、筛选标志物的关键依据。原始LC-MS数据常含数百万数据点，其中多为噪声，直接分析计算成本高、信噪比低。信号提取的核心是在保留真实化学信号的前提下实现数据降维，剔除冗余噪声，并将离子流解析为对应化学实体的同位素包络。

早期算法多依赖于数学形态学假设，即预设色谱峰严格遵循高斯分布或墨西哥帽小波等特定函数形式，并借助二阶导数判定峰边界。这种方法在处理非理想峰形，如拖尾、共流出时容易产生误差。新一代算法引入了“感兴趣区域(ROI)”与“信号聚类”理论，AriumMS[4]利用ROI筛选机制锁定最小强度与丰度阈值实现了对冗余背景噪声的靶向剔除。MassCube[5]将策略深化至像素级信号聚类，辅以高斯滤波边缘检测，成功在像素层面解构紧密相邻的异构体信号，完成了对复杂基质中重叠峰的精细化解析(见图1)。XFlow[6]则提出了一种无参数的强度优先聚类方法，利用离子色谱图的潜在流行属性实现信噪分离，赋予了算法对多分辨率数据的自适应兼容能力。针对低丰度信号易被硬阈值过滤的问题，PASTAQ[7]提出了全谱量化策略。通过RT和m/z双维度的同步2D高斯核平滑，构建规则的网格化数据拓扑，有效避免了早期信号截断，显著提升了对低浓度化合物的检测灵敏度。

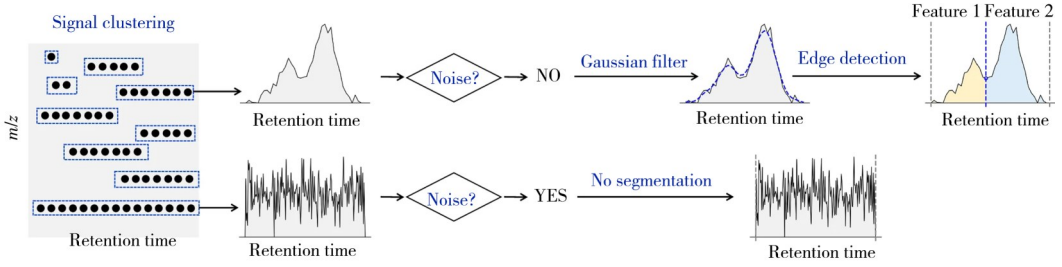

图1　MassCube对MS1信号进行聚类确定噪声并将特征分割为色谱峰[5]

Fig. 1　MassCube performs clustering on MS1 signals to identify noise and segment features into chromatographic peaks[5]

1.2 时空对齐

时空对齐是指在非靶向分析中，通过数学建模修正不同样本或批次间由于色谱系统物理波动引发的保留时间非线性漂移，从而在多维特征空间中重建特征对应关系的过程。其终极目标是将分散在不同测量文件中的离散信号，映射到一个具有统计可比性的主特征表中。在液相色谱-质谱分析中，保留时间是表征化合物物理化学性质的关键信息。但这种表征具有内在的不稳定性，一方面，色谱柱老化、流动相微扰或热力学波动导致的非线性伸缩变形，使得简单的线性平移校正失效；另一方面，在大规模队列研究中，批次效应进一步加剧了这一难题，批次间RT偏移常显著超越批次内变异，甚至出现保留顺序的颠倒，即化合物A和B在不同批次中的流出顺序发生互换。若不进行高精度的拓扑对齐，统计分析将面临严重的错位匹配风险，导致特征组分的假阳性或假阴性[8-9]。

早期色谱对齐策略多基于几何规整(如动态时间规整(DTW))，易因过度校正产生畸变。现代方法则引入统计约束以提升稳健性：PASTAQ通过最大化二维峰体积重叠实现高精度对齐；metabCombiner利用样条拟合与丰度排序完成跨色谱条件的特征匹配。递归式填补策略可在特征部分缺失时，回溯原始数据并在预期保留时间窗口强制积分，有效减少漏检[10-11]。针对靶向分析中的保留时间漂移，SmartPeak采用相对保留时间与二次混合整数规划框架，较传统绝对窗口法鲁棒性显著提升[12]。此外，针对多批次实验中常见的保留顺序交换问题，kmersAlignment算法通过将特征序列分解k-mers子序列进行局部比对，有效解决了非线性漂移带来的错位。

1.3 物质鉴定

物质鉴定是将具有*m/z*、RT等数字坐标的特征映射为具体化学结构的过程。现有标准品数据库覆盖有限，依赖匹配会遗漏大量未知但有价值的化合物。电喷雾电离产生的加合物与碎片若不加区分，还会导致单化合物被误判为多个，影响机理推断[13]。

传统方法基于MS^2谱图与数据库的余弦相似度匹配，难以鉴定未知物。DFBuilder[14]提出诊断性碎片过滤，依据特征性碎片或中性丢失筛选特定结构，突破数据库限制。nLossFinder[15]则将类似逻辑用于加合物筛查，通过追踪中性丢失模式发现未知修饰位点。针对同一化合物的冗余谱图干扰，MeRgeION与SLAW采用共识谱图生成机制，聚合多次扫描重构高信噪比指纹谱图，提升检索准确性。MassCube与Scaffold Element则通过保留时间一致性等关联，智能识别源内碎片，避免将其误判为独立杂质，这对保证复杂体系非靶向筛查结果的特异性至关重要。对于聚合物等同系物，HepParser等方法利用质量余数变换等数学周期律，实现了无标准品依赖的自动归类与可视化，为解析聚合反应提供了有效工具[16-17]。

2 LC-MS数据处理算法的研究进展

在算法演进进程中，LC-MS数据处理经历了从显式规则到数据驱动的范式迁移。基于规则的时代主要依赖小波变换、高斯拟合等数学模型实现特征提取(如XCMS)与色谱对齐(如DTW)。统计机器学习时代引入了稀疏约束与网络拓扑等统计模型，用于特征选择(Inxparse)、色谱对齐(G-Aligner)及物质鉴定(mWISE)。当前深度学习时代则以数据驱动的表示学习为核心，通过卷积网络进行端到端特征提取(EVA)、基于轨迹追踪实现对齐(Asari)，并融合生成式人工智能进行结构推理(MassKG)。如图2所示，这一发展路径体现了从“人工参数调优”到“数据自主感知”的方法论根本转变。

2.1 复杂基质中的信号提取

信号提取算法经历了从“模型驱动”向“数据驱动”进而向“人机协同”的智能化演进。传统基于显式数学模型的拟合方法(如以高斯曲线拟合为代表的centWave算法)常因模型假设与复杂真实色谱峰形失配，导致漏检或误判。为克服此局限，研究转向借鉴计算机视觉思路：一种路径将提取问题转化为图像分类任务，直接识别色谱图中的峰区域；更精细的路径则采用像素级分割模型(如U-Net)，对色谱信号进行像素级解析，输出每个数据点属于色谱峰的概率。针对深度学习模型对大规模标注数据的依赖，融合主动学习(Active learning)与人机交互的策略应运而生，该机制通过迭代筛选高不确定性样本触发专家介入，在有限标注成本下构建了面向少样本场景的可持续优化闭环，见图3。这一演进轨迹本质上标志着算法逻辑从封闭的数学近似，向开放式感知与交互反馈的根本转向，对于提升复

杂生物基质中痕量代谢物或环境非靶向筛查中未知污染物的检出效能具有决定性意义。

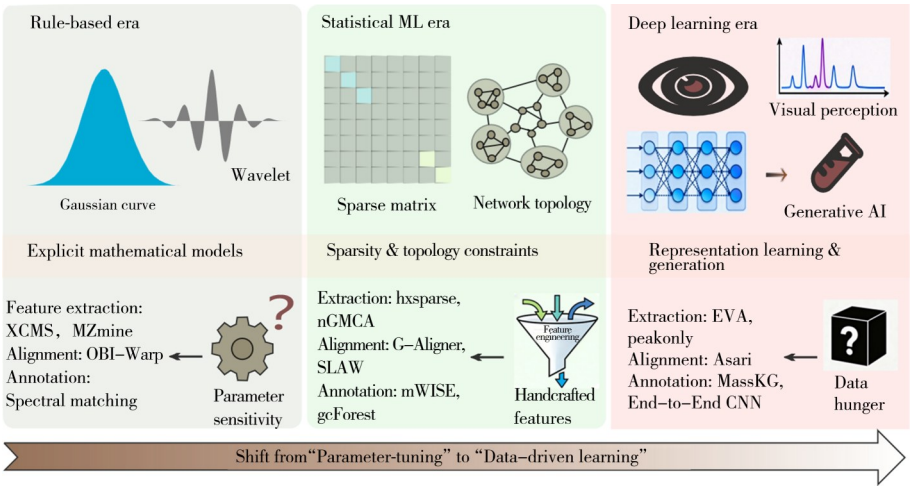

图2　液相色谱-质谱数据处理算法演进的全景概念图

Fig. 2　Panoramic conceptual diagram of the evolution of liquid chromatography-mass spectrometry data processing algorithms

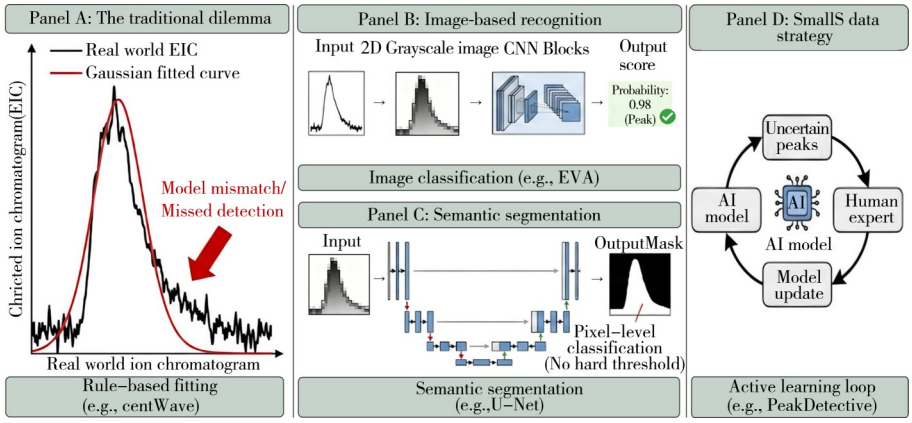

图3　信号提取机制的演变：从基于规则的数学拟合向基于深度学习的视觉感知及人机协同的主动学习转型

Fig. 3　Evolution of signal extraction mechanisms: Transformation from rule-based mathematical fitting to deep learning-based visual perception and human-machine collaborative active learning

2.1.1　传统算法的数学假设与局限　信号提取算法经历了从“规则驱动”到“数据驱动”的演进。早期XCMS系列算法是典型代表：matchedFilter[18]基于高斯峰形假设进行匹配滤波，centWave[19]则引入连续小波变换(CWT)以适应高分辨质谱中多变的峰宽，后者因其优异的时频局部化能力，也成为解析非平稳瞬态信号的有效工具。ADAP[20]算法进一步结合CWT与聚类，改善了共流出峰的解析。

这类算法虽具有数学清晰、可解释性强的优点，但其性能严重依赖预设模型(如高斯或小波)与参数调优。面对复杂基质中的非理想峰形或噪声，往往需要专家反复试参。然而，原始信号中往往掺杂着复杂的随机噪声与基线漂移，若预处理算法选择不当，极易导致微弱特征峰的淹没或畸变。针对这一跨领域的信号保真难题，本课题组前期在复杂一维光谱信号的去噪研究中发现，基于希尔伯特变换的滤波策略在处理非平稳信号时表现出优于传统快速傅里叶变换的边缘保持能力[21]。这一结论对于质谱信号处理同样具有启示意义：即针对不同信噪比和形态特征的原始离子流信号，建立自适应的差异化滤波机制，是最大限度保留痕量组分特征信息、降低“假阳性”干扰的前提。尽管IPO[22]和AutoTuner[23]等工具试图通过统计策略实现参数自动优化，在保留数学模型的同时提升鲁棒性，但其在处理跨批次或梯度剧烈波动的非线性数据时仍存在局限，这一方法学瓶颈最终催生了向数据驱动智能化算法的转型。

2.1.2　引入稀疏与解构的数学约束　为突破刚性规则对复杂信号的限制，稀疏约束与矩阵分解技术被引入以实现“盲源分离”。针对氢氘交换质谱等体系中谱图重叠与双峰分布难题，传统积分方法依赖人工，易误判。针对色谱峰重叠导致的信号模糊与定量误差，传统的时域处理往往难以奏效，而基于

频域的复原策略则展现出独特的数学优势。本课题组前期在模糊信号复原的研究中发现，基于频率域点扩散函数的参数估算策略能有效解决信号的盲复原难题[24]。该研究通过优化频域内的反卷积算法，显著抑制了复原过程中的振铃效应与噪声放大。这一数学思想虽源于图像信号处理，但其核心逻辑——即将时域上高度混叠的信号转化为频域特征进行解构——对于解析色谱-质谱中因色谱扩散效应导致的严重共流出与峰形畸变同样具有重要的借鉴价值，为复杂体系的信号分离提供了跨领域的数学思路。在此基础上，Hxsparse[25]采用最小绝对收缩和选择算子(LASSO)L1 正则化，利用其“尖峰”特性迫使无关变量归零，从超完备同位素字典中稀疏重构观测数据，抑制噪声过拟合，实现欠定条件下的盲源分离，显著降低特征提取误差(见图 4)。除了稀疏重构，基于进化理论的特征优选策略在高维数据降维中同样表现优异。本课题组在处理高维光谱数据时，创新性引入了竞争性自适应重加权采样结合连续投影算法[26]以及区间变量迭代空间收缩[27]策略。这些方法通过模拟“优胜劣汰”的迭代过程，成功从数百个连续波段中筛选出最具鉴别力的核心特征变量，在有效剔除共线性冗余与背景噪声的同时，将数据维度降低了 75%以上。这种减法思维对于解决 LC-MS 中同位素包络冗余及复杂基质干扰下的特征选择具有重要的借鉴价值。

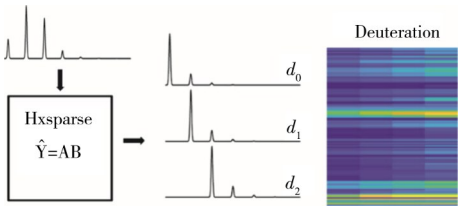

图 4 Hxsparse 基于线性氘化模型和 LASSO 正则化的氢交换特征提取算法质谱[25]

Fig. 4 Hxsparse: A feature extraction algorithm for hydrogen exchange mass spectrometry based on a linear deuteration model and LASSO regularization[25]

面对复杂生物基质或环境样本分析中常见的乘性噪声与非平稳特性，标准非负矩阵分解往往失效。nGMCA[28]通过引入非平稳噪声模型，无需预设峰形即可重构微量成分信号，在混合物分析中实现 97%的信号恢复率，展现了无模型数学解构优势。然而盲源分离在极低信噪比下可能产生“鬼影峰”，结果需对照原始谱图验证。

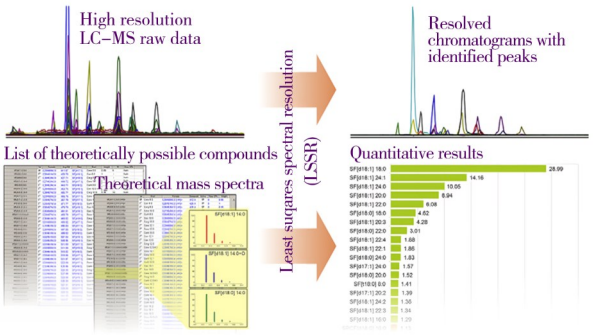

图 5 最小二乘谱分辨率算法扩展到高分辨率脂质组学数据[29]

Fig. 5 Extension of least squares spectral resolution algorithm to high-resolution lipidomics data[29]

针对特定分子体系的专用型算法日趋精细：LSSR[29]基于最小二乘光谱解析，自适应分离重叠同位素包络(见图 5)；ROIMCR[30]在感兴趣区域内通过多元曲线解析-交替最小二乘法直接解析共流出组分，避免了对齐误差。在此领域，覃佐剑等[31]曾系统总结了化学计量学方法在脂质组学数据解析中的应用，指出多元曲线解析(MCR)与平行因子分析(PARAFAC)等数学分离策略在解决复杂生物基质中色谱共流出难题时具有显著优势，这类基于统计学约束的经典方法为后续智能化算法的演进奠定了坚实的方法论基础。而在大分子分析方面，ProMex[32]聚合多电荷态信号，利用洗脱时间连续性约束提升大分子定量重现性；CRANE[33]将 LC-MS 数据映射为灰度图像，通过未抽取小波变换实现基线、噪声与背景的一步剔除。此外，针对结构近似化合物在常规检测中信号差异微弱的难题，数学变换增强策略展现了独特优势。本课题组提出了基于对数比的光谱差异模型[34]，发现通过对原始信号进行特定的非线性数学变换，可以显著放大微弱的物理化学差异信号，在低信噪比条件下成功实现了对混合重叠信号的高精度鉴别。这一思路表明，构建对特定差异敏感的数学特征空间，是提升 LC-MS 对痕量、高相似度物质鉴定能力的有效途径。

2.1.3 深度学习驱动的端到端特征识别 稀疏模型缓解了部分重叠问题，而深度学习的引入推动了从“计算峰”到“看见峰”的智能化转变。其核心在于摒弃人工特征与刚性假设，转而模仿专家视觉，从原始数据中直接学习特征。卷积神经网络(CNN)凭借其平移不变性与层级特征提取能力[35]，在此发挥了关键作用。与传统算法的“硬拟合”不同，CNN 通过卷积层捕捉信号边缘的微小梯度，池化层则使其对色谱峰漂移保持稳定响应。通过层级堆叠，网络将底层特征组合为高层抽象，在特征空间中构建了能包容峰形异质的非线性流形，从而摆脱了对固定峰宽或信噪比阈值的依赖，确立了基于形态语

义的端到端识别架构。

图像化识别是这一方向的代表。EVA[36]将一维色谱图转化为二维图像，利用CNN的平移不变性模拟视觉判别，依据峰的整体拓扑而非单点阈值进行识别(见图6)。在包含数万张图像的数据集上训练后，其分类准确率超过90%，并在不同仪器平台上表现出强鲁棒性。这种能力对化工监测极具价值：在高浓度主成分背景下，模型能通过形态语义直接捕获痕量杂质，为反应机理研究提供完整数据基础。然而，信号转图像过程会损失分辨率，且模型的可解释性仍面临挑战。

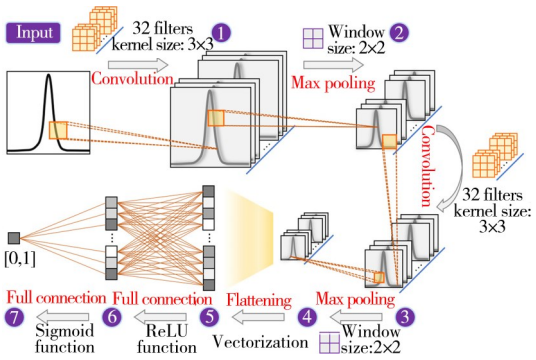

图6　为EVA构建的基于卷积神经网络的深度学习模型示意图[36]

Fig. 6　Illustration of the CNN-based deep learning model constructed for EVA[36]

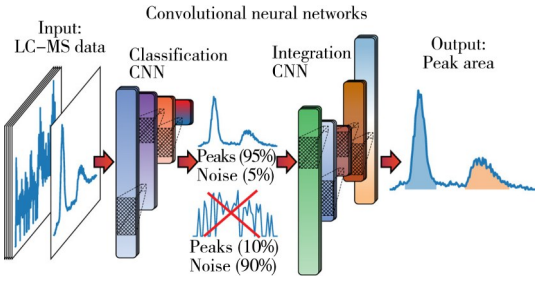

图7　深度学习用于高精度LC-MS数据峰值检测的peakonly算法[37]

Fig. 7　The peakonly algorithm for high-precision LC-MS data peak detection using deep learning[37]

(1)精细分割与检测。随后的研究不再满足于简单的“是/否”二分类，而是向像素级的精细度演进。peakonly[37]算法借鉴医学影像处理领域表现优异的U-Net架构，将峰检测任务重构为像素级分割问题(见图7)。迥异于传统算法的边界框回归逻辑，该方法建立在逐像素预测基础之上，实现了对色谱峰积分区域的精准拓扑界定。在高分辨LC-MS数据实测中，这种架构优势使其实现了97%的真阳性检测精度，核心优势在于完全摒弃了硬强度阈值，从而能够从高噪声背景中有效召回传统算法因固定阈值筛选而遗漏的低丰度信号。

(2)小样本学习。针对深度学习模型对海量标注数据的高度依赖，PeakDetective[38]结合无监督学习与主动学习，通过自编码器提取峰特征，基于信息熵筛选高不确定性样本交由专家标注，仅需不足百次标注即可建立高精度分类器，显著降低了模型构建初期的样本需求(见图8)。然而，该策略效果依赖于专家水平，若引入认知偏差可能导致错误放大。从传统高斯拟合到视觉识别，算法演进实现了从“让数据适应模型”到“让算法学会辨识信号”的转变。

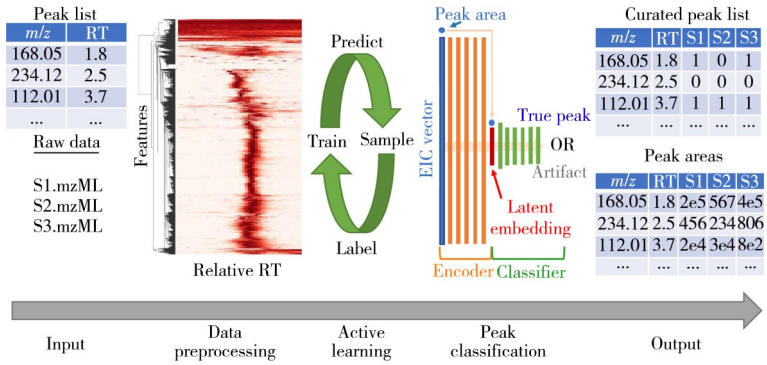

图8　PeakDetective工作流程[38]

Fig. 8　PeakDetective workflow[38]

2.2　时空对齐与校正

精准信号提取后，跨批次分析面临保留时间非线性漂移的挑战。时空对齐技术随之经历了从“几何规整”“化学融合”到“统计建模”的演进(见图9)。早期动态时间规整(DTW)通过拉伸时间轴匹配信号，但这种强制扭曲易掩盖真实的工艺波动与组分变化，导致化学信息丢失。为此，基于化学图论的方法(如G-Aligner)通过引入同位素模式等先验知识，构建拓扑约束网络，实现化学引导的对齐。近期无扭曲追踪策略(如Asari)则放弃直接校正，转而建立全局统计模型追踪特征复现规律，虽保持了时间轴真实，却需更高建模能力以避免因检测批次不同给实验结果带来的偏差。这一进程体现对齐任务已从信号处理提升至化学知识与统计推断融合的层次，对化工生产中的批次稳定性比对至关重要。

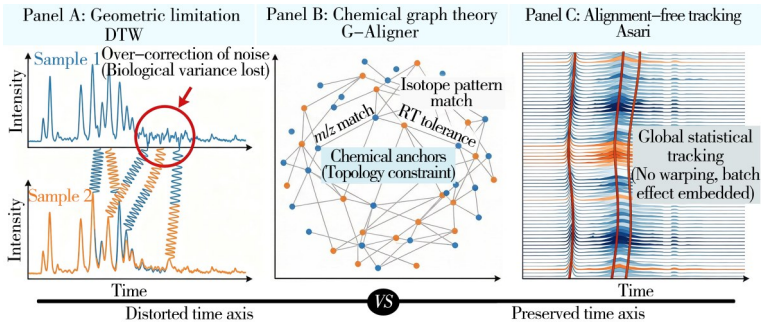

图9 时空对齐策略的革新：从导致波形扭曲的几何规整进化为基于化学锚点的图论优化与保留原始时间轴的全局统计追踪

Fig. 9 Innovation of spatiotemporal alignment strategies: Evolution from geometric normalization that causes waveform distortion to graph-theoretic optimization based on chemical anchors and global statistical tracking that preserves the original time axis

2.2.1 基于波形的非线性规整 在保留时间校正的早期，对齐被视为纯信号处理问题。以XCMS的OBI-Warp为代表的方法采用动态时间规整(DTW)，通过对总离子流图进行非线性扭曲实现波形匹配[39-40]。这类无标策略虽通用，却存在根本缺陷：仅依赖峰的几何位置，完全忽略其化学属性。这种对波形相似的单一依赖缺乏化学标准品作为参照，在低信噪比或严重漂移样本中易误对齐噪声，将真实组分差异误判为技术误差，从而掩盖关键痕量成分。

2.2.2 引入化学信息与图论约束 为突破几何校正的局限，新一代算法引入化学信息与数据拓扑约束对齐过程。G-Aligner[41]采用图论网络模型，将不同样本中的质谱特征视为网络“节点”，将潜在的同源组分关系视为连接“边”，从而把复杂的对齐任务转化为网络中的全局最优匹配问题。该算法结合精确质荷比与同位素分布等已知的化学特征信息，通过组合优化求取全局最优匹配，使复杂数据集上的对齐精度较传统方法提升9.8%~26.6%(见图10)。这标志着对齐逻辑从单纯的信号波形相似性转向了更深层的化学身份确认，能有效抑制保留时间波动导致的假阳性匹配。然而，在大规模样本(如催化剂筛选)场景下，该算法构建的拓扑网络易引发组合爆炸，算力需求呈指数增长，限制了其工业实时应用。

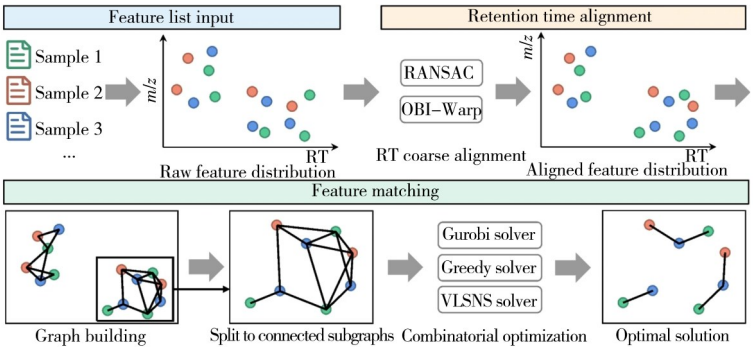

图10 G-Aligner通过三个主要步骤进行特征对齐：功能列表输入、保留时间对齐和功能匹配[41]

Fig. 10 G-Aligner performs feature alignment through three main steps: functional list input, retention time alignment, and functional matching[41]

针对对齐导致的特征缺失与数据稀疏，SLAW[11]采用递归填补机制(见图11)。该算法基于质量控制样本建立闭环反馈，动态校准提取参数并回溯原始数据，迭代重检索潜在微量组分，在修复漏检空缺的同时维持同位素模式与光谱拓扑的完整性，从而重构高密度、高置信度的特征矩阵。需注意的是，在应用于药物代谢动力学或临床微量组分等受监管的定量分析场景时，必须审慎设定信噪比过滤阈值。过度激进的递归填补策略极易引入虚假的低丰度特征，导致方法空白中出现积分背景噪声的“假阳性”风险，进而人为抬高统计学上的假发现率，严重损害分析方法的特异性与定量准确度。

2.2.3 全局视角的重构 尽管图论和递归填补提升了匹配精度，但其本质仍是对保留时间漂移的后修正。相比之下，Asari[42]算法引入了一种“无对齐”的特征追踪技术方案(见图12)，试图从底层逻辑上规避传统对齐步骤引入的系统性误差。

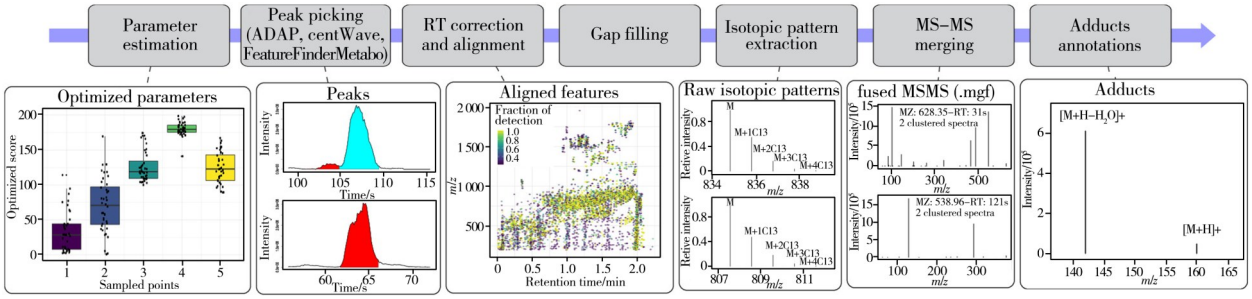

图 11　完整的 SLAW 处理流程[11]

Fig. 11　Complete SLAW processing workflow[11]

Asari 算法的核心是从局部波形对齐转向全局统计聚类。传统 DTW 算法易受噪声干扰，强行对齐会导致“过度校正”。Asari 则利用高分辨质谱的质量精度作为硬约束，将对齐任务重构为密度聚类问题：通过核密度估计构建全局图谱，将保留时间漂移视为统计分布参数，在严格定义的质量容差范围内实现特征归类。该算法的创新在于内置批次效应消除与非扭曲追踪。它通过建立统一特征索引，减少匹配错误，并避免扭曲时间轴，从而高效解决大规模数据中的特征碎片化问题，显著提升处理效率。在大规模代谢组学队列或长周期纵向研究中，该算法展现了卓越的跨批次、跨仪器平台特征追踪能力，能有效剔除仪器系统误差引入的假性波动，确保了生物学变异的真实表达。但其全局统计模型的鲁棒性依赖于足够样本量以保证概率密度的收敛，故在小样本检测时性能可能不如传统方法。

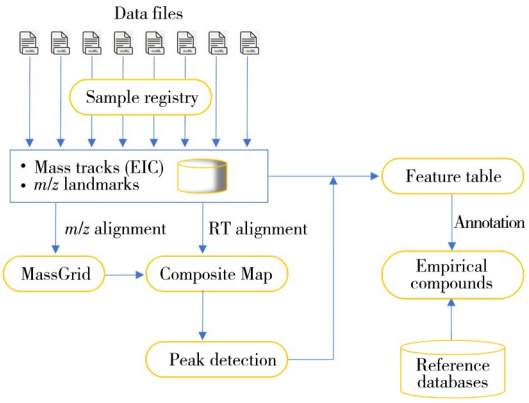

图 12　Asari 处理 mzML 文件生成特征表并匹配化合物[42]

Fig. 12　Asari processes mzML files to generate feature tables and match compounds[42]

2.3　物质注释与鉴定

通过对全局特征的追踪与时空对齐，可获得包含质荷比与保留时间坐标的统计峰表。然而，该表本质上是缺乏化学语义的数字矩阵，仅能反映信号变化，却无法揭示其对应的物质本质。因此，物质鉴定成为将数字坐标转化为具有特定理化属性与工艺指示意义的化学结构的关键环节，是实现从“几何校正”到“语义解读”跨越的核心步骤。传统方法受限于标准品覆盖范围与数据库匹配的多解性，面对海量未知信号，鉴定技术正经历变革：其范式正从依赖已知库的被动检索，经统计学习优化评分置信度，向利用生成式人工智能主动探索未知结构及预测理化属性演进。这一发展为未知代谢物与环境转化产物的快速确证、复杂体系的非靶向解析及高通量生物活性筛选提供了强大的数据智能支撑。

2.3.1　数据库依赖的瓶颈与统计修正　在传统的注释流程中，无论是 MetaboAnalyst[43]等综合平台，还是针对特定分子比如氧化磷脂 LPPtiger[44]、糖肽 LaCyTools[45]等的工具，其核心逻辑均为谱图匹配。这种被动检索模式面临两大痛点：一是高度依赖数据库的完备性，二是传统打分机制难以区分假阳性。

让评分回归“可信”。为了解决原始软件评分不可靠的难题，机器学习被引入作为质量控制的裁判。研究者不再盲目采信基于点积的原始分数，而是利用逻辑回归[46]构建监督模型（见图 13）。该模型将正交的多维特征——包括库匹配分数、同位素分布匹配度和质量误差——进行加权整合，将离散、标准不一的软件评分映射为 0~1 区间内连续的“鉴定概率”。这一概率化转换有效抑制了因参数设置差异导致的假阳性，为精细化工产

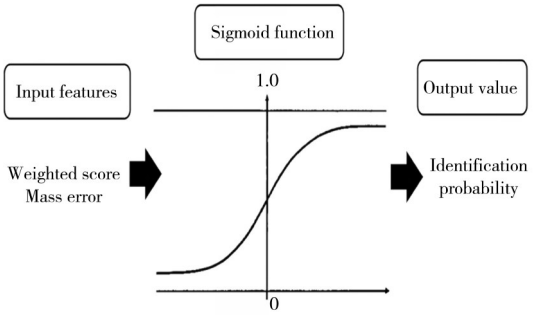

图 13　LC-MS/MS 逻辑回归分析数据：一种监督式机器学习方法[46]

Fig. 13　LC-MS/MS logistic regression analysis data: a supervised machine learning method[46]

品中关键杂质或活性成分的定性分析提供了标准化的置信度量尺。

面对新型人工合成的精神类致瘾物质标准品缺失的挑战，集成学习如 gcForest[47]展现出优势。其级联森林结构通过层级特征增强实现高效表征，降低了模型复杂度和对标注数据的依赖，为新型危害物筛查等少样本场景提供了实用方案。在特定分子解析方面，算法正融合化学机理向智能化发展：针对糖类或聚合物异构体，GlyKAn AZ[48-49]等工具利用反应网络拓扑规则进行结构评分与指认；对于修饰大分子，EpiProfile 2.0[50]通过提取离子色谱峰实现修饰模式的定量分析；在无标准品情况下，GlobalStd[51]依据质量距离推断潜在生化反应链，建立性质与机理间的关联。这些方法通过嵌入领域规则，有效提升了垂直场景下的解析精度。

2.3.2 图网络与上下文挖掘 为克服传统数据库匹配的孤立性局限，新一代方法开始利用数据内部的拓扑关联与化学反应上下文信息，以提升物质结构鉴定的全局一致性。

机器学习正深入数据拓扑结构，以“关系重构”破解未知物关联难题。传统注释常孤立看待特征，而 mWISE[52]等算法引入图扩散策略(见图 14)，利用化学反应网络拓扑，结合共流出与强度共变等统计相关性，在网络中传播与重排序注释置信度。这种基于“反应上下文”的方法通过系统信息收敛了匹配多解性，提升了鉴定准确性。但其效能受限于已知知识库(如 KEGG)的完备性，对未收录的副反应或非典型转化产物，可能因强行关联而产生误导。

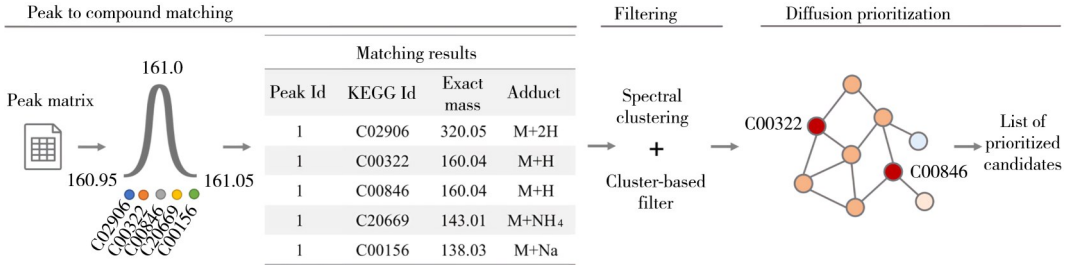

图 14 mWISE 算法的三个主要阶段方案：匹配阶段、滤波阶段、扩散优先级[52]

Fig. 14 Three main phase scheme of the mWISE algorithm: matching phase, filtering phase, and diffusion priority[52]

与此同时，传统统计与机器学习分类器的融合增强了标志物筛选效能。Marcišauskas 等[53]在卵巢癌研究中采用逻辑模型树(LMT)等算法，捕捉非线性特征组合，识别出被传统单变量检验掩盖的关键蛋白，构建出更鲁棒的诊断模型。

机器学习已深入物质鉴定及数据质量评估环节。AP3 模型[54]基于随机森林算法，通过学习分子片段理化特征预测质谱可检测性，有效甄别真实缺失与假阴性结果。保留时间预测方面，Lu 等[55]开发的 LsRP 工具结合序列特征向量与 SVR 回归，实现高精度 RT 预测，显著提升了靶向与全景分析的鉴定准确率。针对 Top-down 表征中的复杂信号，TopFD[56]引入神经网络评分系统，通过识别同位素包络形状区分信号与噪声，降低大分子解卷积假阳性。在非标记定量中，FFId[57]采用 SVM 分类器筛选跨运行对齐特征，有效减少缺失值导致的定量偏差，提升工艺评价可靠性。为增强模型可解释性，dffROI[58]提出特征融合架构，结合人工设计特征与 CNN 抽象特征，在保持高准确率的同时提升决策透明度。

2.3.3 生成与预测 然而，无论是统计优化还是上下文挖掘，均未完全突破对已知化合物数据库的依赖。生成式人工智能的介入，正推动物质鉴定从“被动检索”迈向“主动生成”的全新阶段。

上述方法优化了已知物的鉴定，但面对数据库中不存在的全新化合物传统算法依然束手无策。深度学习的引入彻底改变了这一局面，实现了从检索到生成的现实意义，生成式 AI 之所以能突破数据库的限制并非凭空臆造，而是基于表征学习将离散的化学结构映射到连续的高维潜在空间。不同于传统谱图匹配在有限已知库中的点对点检索，MassKG 等模型通过学习海量分子的 SMILES 序列掌握了化学键连接的深层语法规则，当遇到未知谱图时模型实质上是在特征空间的流形上进行插值与采样，寻找在化学结构上合理且预测谱图与观测数据最接近的理论分子。这种机制从数学上将离散的“检索问题”转化为了连续空间中的“优化生成问题”，赋予了算法探索未知化学空间的推演能力。

深度学习模型(如 DNN、CNN、RNN)[59]利用其高阶非线性函数拟合能力构建了质谱碎片模式与分子结构指纹之间的复杂映射关系，实证研究表明基于大规模谱图训练的深度模型能精准预测未知化合

物的分子指纹，其在候选化合物排序上的表现显著优于基于核方法的经典工具，为解析无库匹配谱图提供了数据驱动的新路径。

生成式模型带来了革命性突破。MassKG[60]系统基于循环神经网络(RNN)化学语言模型(图15)，不仅学习已知分子的结构规则，更能生成数据库中未记录的全新结构。在银杏叶提取物分析中，该系统成功识别出如MKGR15673等新型化合物，实现了从已知物检索到未知物生成的跨越。这对解构“化学暗物质”意义显著，特别是在非靶向代谢组学与环境暴露组学研究中，通过生成式模型，研究人员无需依赖合成标准品，即可逆向推测生物转化过程中的瞬态中间体或环境降解转化产物，为解析复杂的代谢通路与环境归趋提供了分子层面的直接依据。然而，也需警惕生成式模型的“幻觉”风险，其可能生成结构合理但不真实或难以稳定存在的“伪分子”。因此，AI生成结果目前应视为高置信度推测，必须结合保留时间预测、正交实验验证或热力学约束评估，方可转化为确凿的定性结论。

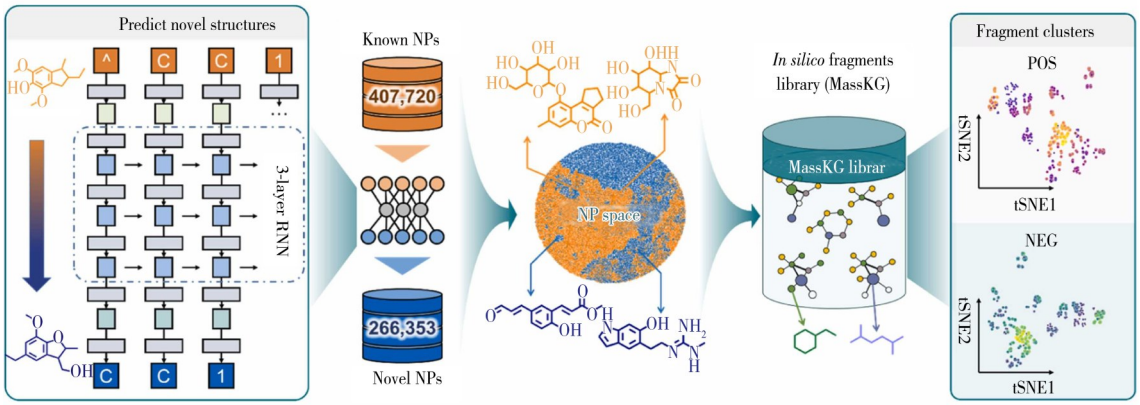

图15　基于RNN的化学语言模型结构用于预测新NP结构[60]

Fig. 15　RNN-based chemical language model architecture for predicting novel NP structures[60]

端到端的表型预测，另一种极端的进化路径是完全跳过繁琐的定性鉴定与定量积分步骤直接建立原始谱图与宏观理化属性之间的非线性映射。使用基于CNN的模型直接处理HPLC-MS/MS原始数据，实现了对癌症亚型的精准分类[61]，准确率达95%；在SONAR-MSI[62]研究中，通过将复杂的DIA数据转化为保留时空信息的伪图像，CNN模型实现了对灵芝物种100%的分类准确率。相比于传统机器学习依赖人工设计特征，深度神经网络能够直接从高维光谱数据中提取特征。例如，本课题组针对复杂信号中的噪声干扰问题，构建了一种基于一维深度残差收缩网络的智能识别模型[63]，该模型创新性地引入了注意力机制与软阈值化策略[64]，能够自适应地感知并消除一维信号中的非相关背景噪声。尽管该研究最初应用于高光谱数据，但鉴于其与质谱数据在“一维序列结构”与“非平稳噪声特征”上的高度同构性，该策略有力证明了深度学习可以绕过化学结构鉴定的“黑箱”，直接从高维原始信号中提取具有分类判别价值的抽象特征，实现对微小差异化学组分的精准区分(准确率>99%)。这种“端到端”的策略证明了在特定应用场景下，深度学习可以绕过化学结构鉴定的“黑盒”直接提取具有分类判别价值的高维抽象特征。这为临床疾病的快速筛查、中药材与食品的产地溯源及复杂生物制品的整体质量评价等场景，提供了一种无需解析具体组分即可实现快速智能化分级的实战手段。

LC-MS数据处理算法的演进呈现阶段性互补。传统方法(如XCMS)可解释性强但依赖人工调参；统计学习方法(如mWISE)对复杂信号更稳健却计算成本高；深度学习方法(如EVA)可实现端到端感知，但通用性与可解释性不足。3类方法分别适用于标准流程、大规模队列与痕量发现等场景，共同构成多层次工具集，也折射出该领域泛化性不足、标准缺失与生态碎片化的挑战。未来需融合物化已知信息与智能学习，建立标准化、可协作的数据处理体系(表1)。为便于读者系统把握这一技术演进脉络，表2详细汇总了上述各阶段主流算法的提出时间、核心原理、引用次数及开源代码库，为不同需求的研究者提供了从理论溯源到工具获取的全景式索引。

3　现存挑战

尽管深度学习在实验环境下准确率屡创新高(>95%)，但在标准方法验证与常规分析的底层生态中

仍由 XCMS 及 MZmine 等经典规则驱动工具占据主导。这种学术界模型高分与工业界实际转化滞后的现象，构成了鲜明的效能倒挂。基准数据集上的完美拟合往往掩盖了模型在跨实验室、跨仪器平台复杂场景下的泛化失效，这要求该领域内研究者必须从客观审视的视角，重新评估特定数据集上的过拟合风险以及工业级评价标准的缺位问题。如果将算法革新定义为 LC-MS 数据处理的技术上限，那么“软件工程化”的完备度就直接代表了实际应用的下限。智能化算法在特征解析上展现了范式级的突破，但从算法原型到工业级工具之间仍横亘着一道深邃的“工程化断层”。当前，大量前沿算法因滞留于独立脚本阶段，缺乏模块化集成与交互式界面支持，导致难以融入主流工作流，成为制约新技术规模化推广应用的关键技术瓶颈。

表 1 LC-MS 数据处理 3 大算法维度对比与工具矩阵

Table 1 Comparison and tool matrix of three algorithmic dimensions for LC-MS data processing

Comparison dimension	Traditional algorithm	Statistical machine learning	Deep learning
核心逻辑	假设信号符合已知数学模型(如高斯分布)，依靠硬阈值过滤	利用数据内在的稀疏性、拓扑结构或统计规律进行盲源分离	模仿人类视觉认知(CNN)或语言逻辑(RNN)，进行端到端学习
信号提取工具	XCMS(centWave)[2]利用小波变换匹配色谱峰 MZmine(ADAP)[3]基于连续小波变换与聚类 IPO[22]/AutoTuner[23]自动化优化上述参数	Hxsparse[25]利用 LASSO 稀疏约束解决严重重叠 nGMCA[26]非负矩阵分解处理非平稳噪声 LSSR[29]最小二乘法解析同位素	EVA[36]将 EIC 转化为图像，用 CNN 识别峰 PeakOnly[37]利用 U-Net 进行像素级语义分割 PeakDetective[38]主动学习+人机协同，解决小样本问题
时空对齐工具	OBI-Warp[2]基于动态时间规整(DTW)，非线性扭曲时间轴以匹配波形	G-Aligner[41]利用图论和化学锚点进行网络优化 Asari[42]无对齐策略，通过全局统计建立共识特征 SLAW[11]递归式填补，回溯原始数据	FFId[57]利用 SVM 筛选跨运行对齐特征 SONAR-MSI[62]跳过特征对齐，直接利用伪图像进行表型分类
物质鉴定工具	Spectral Matching[59]基于余弦相似度的数据库检索 MetaboAnalyst[43, 73-76]综合流程与富集分析	gcForest[47]集成学习用于新精神活性物质筛查 mWISE[52]利用代谢网络拓扑扩散推断上下文 AP3[54]随机森林预测多肽可检测性 LsRP(SVR)[55]利用 SVR 和氨基酸位置向量预测肽段保留时间，辅助精确鉴定	TopFD[56]神经网络评分用于自上而下的去噪 CSI：FingerID[59]深度模型预测分子指纹(优于核方法) MassKG[60]生成式 AI(RNN) 预测未知物结构
工程化痛点	参数复杂：灵敏度与特异性难以平衡，对用户经验要求极高	计算密集：图优化和矩阵分解在大规模队列上极其耗时	泛化困难：跨实验室、跨仪器的模型迁移能力差，存在过拟合
适用场景建议	标准化分析：需要明确解释来源的场景	复杂重叠/大队列：适用于解决生物或环境样本中的组分重叠与共流出难题；大规模临床组学研究中跨批次数据的一致性评价与质量监控	未知物探索/低信噪比：发现新化合物、处理高噪声背景的痕量分析

表 2 LC-MS 数据处理主流算法与工具资源汇总

Table 2 Summary of mainstream algorithms and tool resources for LC-MS data processing

Reference	Algorithm/Tool	Category	Publication year	Core principle	Citation count	Address/Code repository
[2]	XCMS(centWave)	基于规则	2008	利用连续小波变换(CWT)匹配色谱峰，非线性校正	909	Bioconductor(XCMS)
[2]	OBI-Warp	基于规则	2006	基于动态时间规整(DTW)，非线性扭曲时间轴以匹配波形	4132	(集成于 XCMS)
[3]	MZmine(ADAP)	基于规则	2010	基于 CWT 与信号聚类，引入“感兴趣区域”去噪	339	GitHub(mzmine/mzmine3)
[22]	IPO	基于规则	2015	自动化优化 XCMS/MZmine 的参数，提升稳定性	265	Bioconductor(ipo)
[11]	SLAW	统计机器学习	2021	递归式填补，回溯原始数据减少缺失值	34	GitHub(zamboni-lab/slaw)
[25]	Hxsparse	统计机器学习	2025	利用 LASSO 稀疏约束解决严重谱图重叠	20	GitHub(Hxsparse)
[41]	G-Aligner	统计机器学习	2023	利用图论和化学锚点进行网络优化对齐	3	GitHub(G-Aligner)
[42]	Asari	统计机器学习	2023	无对齐策略，通过全局统计建立共识特征	38	GitHub(shuzhao-li/asari)
[36]	EVA	深度学习	2021	将 EIC 转化为图像，利用卷积神经网络(CNN)识别峰	29	GitHub(EVA-Code)
[37]	PeakOnly	深度学习	2020	利用 U-Net 进行像素级语义分割，无需硬阈值	138	GitHub(AutoFlow-Omics/PeakOnly)
[38]	PeakDetective	深度学习	2023	主动学习+人机协同，解决小样本标注问题	8	GitHub(PeakDetective)
[60]	MassKG	深度学习	2024	生成式 AI(RNN)预测未知物结构，优于传统检索	9	GitHub(Wren-4/MassKG)

3.1　准确率虚高与泛化能力存在缺陷

当前深度学习模型主要在受控的干净数据集上训练，难以迁移至多尺度、多源异构的实际分析场景。Rehfeldt等[65]的研究表明，公共LC-MS数据因实验条件、仪器差异存在巨大系统性变异，远超真实的生物学变异或环境波动。这本质上是数据分布差异问题：虽然化合物物化规律P(Y|X)稳定，但仪器与条件差异导致输入数据分布P(X)显著偏移。当前主流深度学习模型的设计与泛化能力均基于独立同分布(i. i. d.)假设，默认训练数据与目标场景数据来自同一概率分布(P(X)不变)，面对跨场景分布失配时，所学特征无法正确映射，性能显著下降。迁移学习也因数据异质性过高，往往效果有限甚至产生负迁移。这提示，若不从数据采集标准化入手，仅增加算法复杂度无法解决跨场景适用性难题。

当前该领域缺乏公认的标准数据集，导致算法性能评估缺乏客观比较。研究者多使用自建测试集，使不同算法优劣难以量化。有基准测试[66]指出，不同工具对噪声、保留时间漂移等的敏感度存在差异，并无通用算法。因此，建立涵盖多仪器、多基质的标准化数据集至关重要。已有研究在此方面做出探索：MVAPACK[67]构建了模拟与实验相结合的双层基准集，首次在同等条件下评估了不同软件的抗干扰能力；Cho等[68]利用同位素标记样本构建特征明确的基准集，系统比较了多种采集策略的组分覆盖能力；ProteomicsML平台则整合了保留时间、碎片离子强度等多维度数据集[69]，推动形成了标准化处理流程。这些工作标志着该领域正从分散方法向标准化工程阶段演进。未来算法评估需超越单一数据集的准确率竞争，其跨平台、跨实验室的泛化与鲁棒性将成为核心评价标准。

3.2　软件生态与工程化

当前，大量深度学习模型仅以独立Python脚本形式发布于GitHub，虽满足开源要求，却为多数非编程背景的一线分析人员与临床研究者带来使用门槛。Sarpe等[70]曾指出，软件应具备模块化适应性以支持持续创新。然而许多优秀模型因缺乏此类设计而难以集成至现有流程。关键在于，代码开源不等同于工业适用，未来需依托KNIME、Galaxy或MZmine等可视化工作流平台，将算法封装为插件，真正降低使用壁垒。

为破解工具碎片化困局，需构建统一数据结构与生态。TidyMass[71]以对象导向的R包体系建立标杆，通过标准化数据对象实现全流程可追溯与跨平台复现，显著缓解格式兼容难题。类似地，rtmsEcho[72]为AEMS技术提供开源R包，实现多模态数据自动化处理，弥补了厂商软件在批量数据分析上的不足。MetaboAnalyst[43,73-76]则依托Web平台实现了从原始谱图处理到机理推断的全流程服务，通过云端模式消除环境依赖，大幅降低了非专业用户进行复杂分析的技术门槛。

传统命令行工具用户体验差，低代码方案通过简化开发有效改善了这一状况。Streamlit[77]框架允许开发者仅用Python即可快速构建含交互组件的Web应用，OpenMS WebApps项目借此将命令行工具重构为友好界面，显著降低了开发成本。在R生态中，基于Shiny的GlycoDash[78]为糖蛋白质组学提供了从数据处理到可视化的交互式链路。

云端与可视化协同[79]推动了“云端后端+交互前端”的新架构。MS-PyCloud[80]基于AWS无服务器架构提供弹性算力，解决了本地处理瓶颈；ili[81]则专注于质谱成像数据的3D可视化与交互探索。这种架构兼顾了大规模数据处理的效率与科研中对细节的交互把控。相关软件工具总结见表3。

表3　代表性LC-MS数据处理软件工具与平台概览

Table 3　Overview of representative LC-MS data processing software tools and platforms

Software/Tool	Core function	Architectural feature	Addressed engineering pain point
MVAPACK[67]	算法鲁棒性评估	基准测试套件	提供了含“真值”的标准化数据集，解决了算法评价标准缺失问题
ProteomicsML[69]	深度学习模型训练与评估	数据基础设施	汇集了高质量基准数据，促进了机器学习模型的标准化开发
TidyMass[71]	全流程分析(清洗/注释/统计)	R 语言生态系统	解决了模块间数据格式不兼容与流程可追溯性问题
rtmsEcho[72]	AEMS 数据自动化处理(MRM/全扫描)	R 语言包(基于 rtms 框架)	解决了厂商软件缺乏自动化批处理能力的问题，提升了高通量筛选效率
MetaboAnalyst 6. 0[73]	从光谱处理到多组学整合	Web 平台	降低了非计算专家的技术门槛，无需本地配置环境
Streamlit OpenMS WebApps[77]	交互式参数调整与可视化	低代码 Web 应用	将命令行脚本转化为可视化工具，提升了用户交互体验

(续表3)

Software/Tool	Core function	Architectural feature	Addressed engineering pain point
GlycoDash[78]	复杂大分子/聚合物数据清洗与可视化	R Shiny Web 应用(Docker化)	解决了 Skyline/LaCyTools 后处理阶段依赖手工操作的瓶颈，实现了交互式质控
PSpecteR[79]	数据可视化	R Shiny Web 应用(Docker化)	提供了交互式界面，降低了命令行工具的使用门槛，支持多工具整合

4 结 论

本文系统构建了液相色谱-质谱(LC-MS)数据处理分析框架，将算法演进划分为基于规则、统计机器学习与深度学习三阶段，并围绕信号提取、时空对齐与物质鉴定三大核心任务，明晰了从数学模型驱动到数据智能感知的发展路径，同时指出当前面临“算法泛化危机”“基准缺失”与“生态碎片化”等挑战。为推进实验室算法向工业级工具转化，未来的发展将聚焦于以下方面：(1)发展可解释人工智能(XAI)。在临床诊断、法医毒理等受监管领域，算法决策的可解释性是结果可信的基石。须构建内嵌可解释性机制(如注意力、显著图)的解析框架，将模型决策映射至质谱特征，弥合“黑箱”模型与化学直觉的鸿沟，实现从“AI替代”到“AI增强认知”的转变[82]；(2)构建数据高效范式与可信治理体系。针对标注数据稀缺难题，应发展基于自监督学习的质谱基础模型，挖掘海量未标注光谱价值[83]。针对临床与法医质谱数据的高敏感性，需推广以“区块链+人工智能”的双驱架构为代表的新型治理模式[84]，在不交换原始数据的前提下实现可信溯源与协同计算，为多中心数据合作提供安全合规的解决方案；(3)推进融合物理化学机制的AI与多模态融合。算法需深度融合分析化学已知理化参数(如保留规律、裂解机理)，将其作为正则化约束引导模型训练，确保输出符合物化逻辑[85]。同时，融合质谱数据与化学文本等多模态信息，构建具有推理能力的智能分析系统；(4)建立标准化基准与工程化生态。亟须构建跨仪器、跨基质的标准数据集，以客观评估算法、推动方法标准化[86]。软件生态应朝向模块化、云端化与交互式发展，通过低代码界面与高性能后端协同，降低使用门槛，提升分析流程的自动化与可复现性。

总之，LC-MS数据处理技术的智能化演进是一项系统性工程，须在算法可解释性、数据治理、知识融合及软件生态四维度协同突破，以增强模型跨场景鲁棒性，切实赋能生命健康、环境及新材料等领域的精准分析与科学发现。

参考文献：

[1] Su Q Z, Dong B, Li D, Zhong H N. *J. Instrum. Anal.*(苏启枝，董犇，李丹，钟怀宁. 分析测试学报), **2022**, 41(10): 1558-1567.

[2] Louail P, Brunius C, Garcia-Aloy M, Kumler W, Storz N, Stanstrup J, Treutler H, Vangeenderhuysen P, Witting M, Neumann S, Rainer J. *Anal. Chem.*, **2025**, 97(50): 27639-27645.

[3] Heuckeroth S, Damiani T, Smirnov A, Mokshyna O, Brungs C, Korf A, Smith J D, Stincone P, Dreolin N, Nothias L F, Hyötyläinen T, Orešič M, Karst U, Dorrestein P C, Petras D, Du X, van der Hooft J J J, Schmid R, Pluskal T. *Nat. Protoc.*, **2024**, 19(9): 2597-2641.

[4] Naumann L, Haun A, Höchsmann A, Mohr M, Novák M, Flottmann D, Neusüß C. *Anal. Bioanal. Chem.*, **2023**, 415(16): 3137-3154.

[5] Yu H X, Ding J, Shen T, Liu M, Li Y Y, Fiehn O. *Nat. Commun.*, **2025**, 16(1): 5487.

[6] Gutierrez M, Smith R. *PLoS One*, **2020**, 15(10): e0227659.

[7] Sánchez Brotons A, Eriksson J O, Kwiatkowski M, Wolters J C, Kema I P, Barcaru A, Kuipers F, Bakker S J L, Bischoff R, Suits F, Horvatovich P. *Anal. Chem.*, **2021**, 93(32): 11215-11224.

[8] Habra H, Kachman M, Bullock K, Clish C, Evans C R, Karnovsky A. *Anal. Chem.*, **2021**, 93(12): 5028-5036.

[9] Malinka F, Zareie A, Prochazka J, Sedlacek R, Novosadova V. *Bioinformatics*, **2022**, 38(15): 3759-3767.

[10] Liu Y, Zhang Y, Vennekens T, Lippens J L, Duijsens L, Bui-Thi D, Laukens K, de Vijlder T. *Anal. Chem.*, **2023**, 95(22): 8433-8442.

[11] Delabriere A, Warmer P, Brennsteiner V, Zamboni N. *Anal. Chem.*, **2021**, 93(45): 15024-15032.

[12] Kutuzova S, Colaianni P, Röst H, Sachsenberg T, Alka O, Kohlbacher O, Burla B, Torta F, Schrübbers L, Kristensen M, Nielsen L, Herrgård M J, McCloskey D. *Anal. Chem.*, **2020**, 92(24): 15968-15974.

[13] Seitzer P M, Searle B C. *J. Proteome Res.*, **2019**, 18(2): 791-796.

[14] Murray K J, Carlson E S, Stornetta A, Balskus E P, Villalta P W, Balbo S. *Anal. Chem.*, **2021**, 93(14): 5754-5762.

[15] Sousa P F M, Martella G, Åberg K M, Esfahani B, Motwani H V. *Toxics*, **2021**, 9(4): 78.
[16] Wang H, Wang Y, Hou M, Zhang C, Wang Y, Guo Z, Bu D, Li Y, Huang C, Sun S. *Front. Chem.*, **2021**, 9: 723149.
[17] Molenaar S R A, van de Put B, Desport J S, Samanipour S, Peters R A H, Pirok B W J. *Anal. Chem.*, **2022**, 94(14): 5599-5607.
[18] Smith C A, Want E J, O'Maille G, Abagyan R, Siuzdak G. *Anal. Chem.*, **2006**, 78(3): 779-787.
[19] Tautenhahn R, Böttcher C, Neumann S. *BMC Bioinf.*, **2008**, 9: 504.
[20] Myers O D, Sumner S J, Li S, Barnes S, Du X. *Anal. Chem.*, **2017**, 89(17): 8696-8703.
[21] Wu D, Gao S H, Lu Y F, Zhang N. *Appl. Laser*(吴迪, 高树辉, 陆一帆, 张宁. 应用激光), **2023**, 43(7): 102-108.
[22] Libiseller G, Dvorzak M, Kleb U, Griesberger E, Langes T, Bailer M, Riester-Röger T, Thallinger G G. *BMC Bioinf.*, **2015**, 16(1): 118.
[23] McLean C, Kujawinski E B. *Anal. Chem.*, **2020**, 92(8): 5724-5732.
[24] Zhu P W, Gao S H, Xie Z Y, Fu Y. *Comput. Sci.*(朱沛伍, 高树辉, 解朝玉, 傅裕. 计算机科学), **2024**, 51(4): 223-229.
[25] Shi Y, Hart J, Weis D D. *Anal. Chem.*, **2025**, 97(40): 21917-21924.
[26] Zhang Y F, Gao S H. *Appl. Chem. Ind.*(张宇帆, 高树辉. 应用化工), **2025**, 54(3): 667-675.
[27] Li C S, Gao S H, Li K K. *Spectrosc. Spectral Anal.*(李昌盛, 高树辉, 李开开. 光谱学与光谱分析), **2025**, 45(10): 2804-2815.
[28] Rapin J, Souloumiac A, Bobin J, Larue A, Junot C, Ouethrani M, Starck J L. *Signal Process.*, **2016**, 123: 75-83.
[29] Zeng Y X, Mjøs S A, David F P A, Schmid A W. *Anal. Chim. Acta*, **2016**, 914: 35-46.
[30] Dalmau N, Bedia C, Tauler R. *Anal. Chim. Acta*, **2018**, 1025: 80-91.
[31] Qin Z J, Xie Y, Wei F, Chen H. *J. Instrum. Anal.*(覃佐剑, 谢娅, 魏芳, 陈洪. 分析测试学报), **2020**, 39(3): 406-415.
[32] Park J, Piehowski P D, Wilkins C, Zhou M, Mendoza J, Fujimoto G M, Gibbons B C, Shaw J B, Shen Y, Shukla A K, Moore R J, Liu T, Petyuk V A, Tolić N, Paša-Tolić L, Smith R D, Payne S H, Kim S. *Nat. Methods*, **2017**, 14(9): 909-914.
[33] Seneviratne A J, Peters S, Clarke D, Dausmann M, Hecker M, Tully B, Hains P G, Zhong Q. *Bioinformatics*, **2021**, 37(24): 4719-4726.
[34] Li C S, Gao S H. *Microchem. J.*, **2025**, 214: 114125.
[35] Kantz E D, Tiwari S, Watrous J D, Cheng S, Jain M. *Anal. Chem.*, **2019**, 91(19): 12407-12413.
[36] Guo J, Shen S, Xing S, Chen Y, Chen F, Porter E M, Yu H X, Huan T. *Anal. Chem.*, **2021**, 93(36): 12181-12186.
[37] Melnikov A D, Tsentalovich Y P, Yanshole V V. *Anal. Chem.*, **2020**, 92(1): 588-592.
[38] Stancliffe E, Patti G J. *Anal. Chem.*, **2023**, 95(25): 9397-9403.
[39] Zhang M, Wang Y, Moore R, Upton R, Harrington P, Chen P. *J. Agric. Food Chem.*, **2022**, 70(17): 5450-5457.
[40] Byrdwell W C, Kalscheur K F. *Anal. Bioanal. Chem.*, **2024**, 416(25): 5527-5555.
[41] Wang R, Lu M, An S, Wang J, Yu C. *BMC Bioinf.*, **2023**, 24(1): 431.
[42] Li S, Siddiqa A, Thapa M, Chi Y, Zheng S. *Nat. Commun.*, **2023**, 14(1): 4113.
[43] Guo J, Huan T. *Anal. Chim. Acta*, **2020**, 1137: 37-46.
[44] Ni Z, Angelidou G, Hoffmann R, Fedorova M. *Sci. Rep.*, **2017**, 7(1): 15138.
[45] Jansen B C, Falck D, de Haan N, Hipgrave Ederveen A L, Razdorov G, Lauc G, Wuhrer M. *J. Proteome Res.*, **2016**, 15(7): 2198-2210.
[46] Chen C, Mondal K, Vervliet P, Covaci A, O'Brien E P, Rockne K J, Drummond J L, Hanley L. *Anal. Chem.*, **2023**, 95(12): 5205-5213.
[47] Yang Y, Liu D, Hua Z, Xu P, Wang Y, Di B, Liao J, Su M. *J. Chem. Inf. Model.*, **2023**, 63(3): 815-825.
[48] Dhingra A, Schaeffer Z, Majewska Nepomuceno N I, Au J, Ahn J. *BMC Bioinf.*, **2023**, 24(1): 259.
[49] Klein J, Carvalho L, Zaia J. *Bioinformatics*, **2018**, 34(20): 3511-3518.
[50] Yuan Z F, Sidoli S, Marchione D M, Simithy J, Janssen K A, Szurgot M R, Garcia B A. *J. Proteome Res.*, **2018**, 17(7): 2533-2541.
[51] Yu M, Olkowicz M, Pawliszyn J. *Anal. Chim. Acta*, **2019**, 1050: 16-24.
[52] Barranco-Altirriba M, Solà-Santos P, Picart-Armada S, Kanaan-Izquierdo S, Fonollosa J, Perera-Lluna A. *Anal. Chem.*, **2021**, 93(31): 10772-10778.
[53] Marcišauskas S, Ulfenborg B, Kristjansdottir B, Waldemarson S, Sundfeldt K. *J. Proteomics*, **2019**, 196: 57-68.
[54] Gao Z, Chang C, Yang J, Zhu Y, Fu Y. *Anal. Chem.*, **2019**, 91(13): 8705-8711.
[55] Lu W, Liu X, Liu S, Cao W, Zhang Y, Yang P. *Sci. Rep.*, **2017**, 7(1): 43959.
[56] Basharat A R, Zang Y, Sun L, Liu X. *Anal. Chem.*, **2023**, 95(21): 8189-8196.
[57] Weisser H, Choudhary J S. *J. Proteome Res.*, **2017**, 16(8): 2964-2974.

[58] Zhang H, Xu Z, Fan X, Wang Y, Yang Q, Sun J, Wen M, Kang X, Zhang Z, Lu H. *Anal. Chem.*, **2023**, 95(2): 612-620.
[59] Chau H Y K, Zhang X, Ressom H W. *Metabolites*, **2025**, 15(2): 132.
[60] Zhu B, Li Z, Jin Z, Zhong Y, Lv T, Ge Z, Li H, Wang T, Lin Y, Liu H, Ma T, Wang S, Liao J, Fan X. *Comput. Struct. Biotechnol. J.*, **2024**, 23: 3327-3341.
[61] Petrovsky D V, Kopylov A T, Rudnev V R, Stepanov A A, Kulikova L I, Malsagova K A, Kaysheva A L. *J. Pers. Med.*, **2021**, 11(12): 1288.
[62] Jin Z, Zhu B, Li Z, Li Z, Tang Y, Wang Y. *Anal. Chem.*, **2025**, 97(41): 22475-22481.
[63] Zhang H, Gao S H. *J. Instrum. Anal.*(张浩, 高树辉. 分析测试学报), **2023**, 42(7): 817-824.
[64] Gao S H, Zhang H. *J. People's Public Secur. Univ. China*: *Sci. Technol.*(高树辉, 张浩. 中国人民公安大学学报: 自然科学版), **2024**, 30(1): 1-7.
[65] Rehfeldt T G, Krawczyk K, Echers S G, Marcatili P, Palczynski P, Röttger R, Schwämmle V. *GigaScience*, **2023**, 12: giad096.
[66] Li Z, Lu Y, Guo Y, Cao H, Wang Q, Shui W. *Anal. Chim. Acta*, **2018**, 1029: 50-57.
[67] Jurich C P, Jeppesen M J, Sakallioglu I T, De Lima Leite A, Yesselman J D, Powers R. *Anal. Chem.*, **2024**, 96(32): 12943-12956.
[68] Cho K, Schwaiger-Haber M, Naser F J, Stancliffe E, Sindelar M, Patti G J. *Anal. Chim. Acta*, **2021**, 1149: 338210.
[69] Rehfeldt T G, Gabriels R, Bouwmeester R, Gessulat S, Neely B A, Palmblad M, Perez-Riverol Y, Schmidt T, Vizcaíno J A, Deutsch E W. *J. Proteome Res.*, **2023**, 22(2): 632-636.
[70] Sarpe V, Schriemer D C. *Curr. Opin. Biotechnol.*, **2017**, 43: 110-117.
[71] Shen X, Yan H, Wang C, Gao P, Johnson C H, Snyder M P. *Nat. Commun.*, **2022**, 13(1): 4365.
[72] Rimmer M A, Twarog N, Ranathunge T A, Wang J, Li Y, Chen T, Shelat A A, Yang L. *Anal. Chem.*, **2025**, 97(37): 20444-20452.
[73] Pang Z, Lu Y, Zhou G, Hui F, Xu L, Viau C, Spigelman A F, MacDonald P E, Wishart D S, Li S, Xia J. *Nucleic Acids Res.*, **2024**, 52(W1): W398-W406.
[74] Chong J, Yamamoto M, Xia J. *Metabolites*, **2019**, 9(3): 57.
[75] Pang Z, Chong J, Li S, Xia J. *Metabolites*, **2020**, 10(5): 186.
[76] Tian L, Li Z, Ma G, Zhang X, Tang Z, Wang S, Kang J, Liang D, Yu T. *Bioinformatics*, **2022**, 38(14): 3662-3664.
[77] Müller T D, Siraj A, Walter A, Kim J, Wein S, von Kleist J, Feroz A, Pilz M, Jeong K, Sing J C, Charkow J, Röst H L, Sachsenberg T. *J. Proteome Res.*, **2025**, 24(2): 940-948.
[78] Pongracz T, Gijze S, Hipgrave Ederveen A L, Derks R J E, Falck D. *Anal. Bioanal. Chem.*, **2025**, 417(10): 2003-2014.
[79] Degnan D J, Bramer L M, White A M, Zhou M, Bilbao A, McCue L A. *J. Proteome Res.*, **2021**, 20(4): 2014-2020.
[80] Hu Y, Schnaubelt M, Chen L, Zhang B, Hoang T, Lih T M, Zhang Z, Zhang H. *Anal. Chem.*, **2024**, 96(25): 10145-10151.
[81] Protsyuk I, Melnik A V, Nothias L F, Rappez L, Phapale P, Aksenov A A, Bouslimani A, Ryazanov S, Dorrestein P C, Alexandrov T. *Nat. Protoc.*, **2018**, 13(1): 134-154.
[82] Streun G L, Steuer A E, Ebert L C, Dobay A, Kraemer T. *Clin. Chem. Lab. Med.*, **2021**, 59(8): 1392-1399.
[83] Tang L, Mao Y H, Cai J, Liu H Q, Min H, An Y R, Liu S. *J. Instrum. Anal.*(唐磊, 茅晔辉, 蔡婧, 刘恒钦, 闵红, 安雅睿, 刘曙. 分析测试学报), **2025**, 44(6): 1227-1236.
[84] Gao S H, Wang G R. *J. Xi'an Jiaotong Univ.*: *Med. Sci.*(高树辉, 王贵容. 西安交通大学学报: 医学版), **2025**, 46(1): 1-11.
[85] Pace C L, Simmons J, Kelly R T, Muddiman D C. *J. Proteome Res.*, **2022**, 21(3): 713-720.
[86] Parker E J, Billane K C, Austen N, Cotton A, George R M, Hopkins D, Lake J A, Pitman J K, Prout J N, Walker H J, Williams A, Cameron D D. *Metabolites*, **2023**, 13(4): 463.

(责任编辑: 盛文彦)