广东农业科学 2022，49（11）：96-109
Guangdong Agricultural Sciences DOI:10.16768/j.issn.1004-874X.2022.11.011

黄文洁，吴绍文，刘蕊，孔谦，晏石娟 . 基于质谱的代谢组学数据分析技术研究进展[J]. 广东农业科学，2022，49（11）：96-109.

基于质谱的代谢组学数据分析技术研究进展

黄文洁[1]，吴绍文[1]，刘　蕊[2]，孔　谦[1]，晏石娟[1]

（1. 广东省农业科学院农业生物基因研究中心 / 广东省农作物种质资源保存与利用重点实验室，广东 广州　510640；2. 梅州市农林科学院果树研究所，广东 梅州　514071）

摘　要：代谢组学是系统生物学研究的重要组成部分，是一种对特定条件下生物体内所有内源性小分子代谢物进行全面定性和定量分析的技术。质谱和核磁共振系统的不断更新迭代推进了代谢组学技术的迅猛发展，其中质谱技术因其能同时检测出数千个生物流体、细胞和组织中的代谢物，且所需前处理步骤简单，已发展为当前代谢组学研究中应用最广泛的技术，开发基于质谱的代谢组学数据分析方法也因此成为过去 10 年代谢组学研究的热点领域。对基于 GC-MS 和 LC-MS 的代谢组学数据预处理、代谢组学数据统计分析、代谢途径富集分析，以及未知代谢物鉴定 4 个方向取得的研究进展进行系统总结，详细介绍常用的数据分析策略和分析软件；并重点综述了包括基于数据库、分子网络算法、人工智能算法等未知代谢物鉴定的前沿方法，最后展望了基于质谱的代谢组学数据分析的未来发展方向，在已知生化反应和分子网络分析的基础上再整合代谢物合成的遗传位点等信息，有望进一步提高代谢物的鉴定数量和准确度。全面综述基于质谱的代谢组学数据分析技术，将为开发新的代谢组学分析方法和挖掘代谢组学数据的生物学意义提供重要的参考和思路。

关键词：代谢组；质谱；数据预处理；数据库；代谢物鉴定；分子网络算法

中图分类号：Q94-3　　**文献标志码：**A　　**文章编号：**1004-874X（2022）11-0096-14

Progress in Mass Spectrometry-based Metabolomics Data Analysis Techniques

HUANG Wenjie[1], WU Shaowen[1], LIU Rui[2], KONG Qian[1], YAN Shijuan[1]

（1. *Agro-biological Gene Research Center, Guangdong Academy of Agricultural Sciences / Guangdong Key Laboratory for Crop Germplasm Resources Preservation and Utilization, Guangzhou* 510640, *China;* 2. *Institute of Fruit Tree Research, Meizhou Academy of Agricultural and Forestry Sciences, Meizhou* 514071, *China*）

Abstract: Metabolomics technique, as an important part of systems biology, aims to identify and quantify all endogenous small molecule metabolites in organisms at certain condition. The continuous iteration of mass spectrometry and nuclear magnetic resonance system facilitate great progress in metabolomics technologies. Among them, mass spectrometry and related metabolomic techniques have been the most widely used due to their ability to detect thousands of metabolites in biological fluids, cells and tissues simultaneously, without complex pre-processing steps for sample preparation. Therefore, the development of tools for mass spectrometry-based metabolomics data analysis has been a hot topic in metabolomics

收稿日期：2022-09-09

基金项目：高水平农科院建设 - 科技创新战略专项（202205，R2020PY-JX019）；2022 年梅州特色现代农业产业人才振兴计划“揭榜挂帅”项目 - 梅州柚木质化阻断技术研究项目；广东省农业科学院“十四五”农业优势产业学科团队建设项目（202114TD）

作者简介：黄文洁（1987—），女，硕士，助理研究员，研究方向为代谢组学技术研发与创新应用，E-mail：huangwenjie@agrogene.ac.cn

通信作者：晏石娟（1983—），女，博士，研究员，研究方向为作物品质控制与多组学技术，E-mail：shijuan@agrogene.ac.cn

research in the past decade. In this review, we systematically summarized the research progress in four main aspects of gas/liquid chromatography tandem mass spectrometry (GC/LC–MS)–based metabolomics data analysis, including metabolomics data preprocessing, statistical analysis of metabolomics data, metabolic pathway enrichment analysis, and identification of unknown metabolites. We mainly introduced the commonly used analysis strategies and software related with MS–based metabolomic data analysis; and highlighted the cutting–edge innovation about molecular networking–, artificial intelligence–and databases–based metabolite identification. Finally we gave a brief future perspective about MS–based metabolomic data analysis, and believe that new developed strategies, which integrate the known biochemical reactions, molecular networking tools, and genetic loci information regulating the metabolite biosynthesis, will promote the number and accuracy of identified metabolites. This review will provide new ideas for deeper exploration of new methods for metabolomic data analysis and biological significance from metabolomic data.

Key words: metabolomics; mass spectrometry; data preprocessing; database; metabolite identification; molecular network algorithm

自1999年Nichilson等[1]首次提出“代谢组学”的概念后，代谢组学得到不断发展，成为继基因组学、转录组学和蛋白质组学技术之后的又一门新兴组学技术。代谢组学研究旨在通过核磁共振技术（Nuclear magnetic resonance, NMR）、质谱技术（Mass spectrometry, MS）等分析手段对生物体内特定条件下的所有内源性代谢物（<1 000 u的小分子）进行全面定性和定量分析[2-3]。内源代谢物通常是生物反应的中间产物或最终产物，处于不断变化的过程，因此，代谢组学比其他组学方法更能直接地反映细胞、组织或生物体的表型信息。

质谱分析技术可以实现对生物流体、细胞和组织中数千个代谢物的高通量检测，具有分析速度快、灵敏度高、检测代谢物种类覆盖范围广等优点，且随着高分辨质谱技术的迅速发展，高精确度的离子质量更有助于提高代谢物的鉴定能力，因此，该技术已成为代谢组研究中不可或缺的工具。其中，气相色谱–质谱联用（Gas chromatography–mass spectrometry, GC–MS）、液相色谱–质谱联用（Liquid chromatography–mass spectrometry, LC–MS）技术是目前代谢组学研究中应用最广泛的质谱分析技术[4]。GC–MS适用于热稳定、易挥发或经衍生化后具有挥发性的代谢物，如氨基酸、糖类、有机酸和脂肪酸等初级代谢物[5-8]，且不受复杂样品的基质效应干扰，在定性分析方面具有通用的质谱数据库。LC–MS具有更全面和强大的分析能力，结合不同的离子源、电离模式和色谱柱等条件进行分析，可以在不需要复杂的样品预处理的情况下分离和鉴定样品中更多种类的代谢物，适用于热不稳定、不易挥发、相对分子质量较大的物质，如脂质、类黄酮、生物碱、类胡萝卜素、苯丙素类等代谢物[6,9]。近年来，基于质谱的代谢组学研究被广泛用于解决生物学研究中的重要问题，包括解析复杂生物合成途径的代谢调控，探索控制农作物重要性状形成的分子机制，解析包括进化和驯化综合征在内的植物遗传学，以及对生物或非生物应激的代谢反应等[10]。

基于质谱的代谢组学分析技术包括代谢组学样本前处理、质谱数据采集、代谢组学数据预处理、代谢组学数据统计分析、代谢途径富集分析以及未知代谢物鉴定等主要步骤。如何通过数据分析方法从采集到的质谱原始数据中提取代谢物离子、获得代谢物的含量信息、提高代谢物鉴定效率，找出具有生物学意义的信息是代谢组学研究的关键环节[11]。前期我们围绕基于质谱的代谢组学技术发展历程、工作流程以及其在植物、肠道微生物研究中的应用进展进行了系统的阐述[12-15]。本文将重点围绕基于质谱的代谢组学数据分析技术展开综述，包括数据分析策略、数据分析软件和算法、数据库构建等方面。

1 质谱原始数据的预处理

原始质谱数据包含质荷比（Mass–to–charge ratios, m/z）、保留时间（Retention time, RT）和峰强度（Peak intensity）等多维数据[16]，涵盖了样本中实际代谢物的信息、试剂中杂质和仪器残留等噪音的质谱碎片特征。因此，对原始质谱数据进行预处理以获取准确、可靠的代谢物特征的信息，消除随机误差(噪音)和其他干扰因素的影响，

能够保障下游数据分析和信息挖掘的准确性[17]。质谱数据预处理主要包括峰提取、峰对齐和归一化等主要步骤。目前代谢组学研究群体常用的一些质谱数据预处理分析软件及其功能介绍见表1。

表 1　质谱数据预处理常用软件

Table 1　Common software for mass spectrometry data pre-processing

软件 Software	质谱数据类型 MS data type	功能简介 Function description	网址 Website
XCMS	LC-MS, GC-MS	保留时间对齐、滤噪、峰提取和峰鉴定	http://bioconductor.org/packages/release/bioc/html/xcms.html
MetAlign	LC-MS, GC-MS	数据格式转换，精准分子量计算，基线校正，滤噪、峰提取、峰对齐	http://www.wageningenur.nl/en/show/MetAlign-1.htm
MZmine 2	LC-MS	支持高分辨质谱数据可视化分析，包括滤噪、峰识别、峰提取、峰对齐，使用在线数据库进行代谢物鉴定，归一化和统计分析	http://mzmine.sourceforge.net
OpenMS	LC-MS	一个高通量质谱数据分析平台。包括数据格式转换、峰提取、峰对齐、代谢物鉴定和统计分析等	http://www.openms.de
MS-DIAL	LC-MS	一款专门用于解决 DIA 质谱数据解卷积的软件，具有滤噪、峰提取、峰对齐、归一化等功能	http://prime.psc.riken.jp/
AMDIS	GC-MS	GC-MS 质谱数据预处理最受欢迎的软件之一，是一款功能强大的解卷积软件，具有峰滤噪、峰提取和谱图检索匹配等功能	http://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:amdis
eRah	GC-MS	一种基于盲源分离的多变量技术的色谱解卷积方法，包括峰提取、峰对齐、利用谱库实现代谢物自动识别的功能	http://CRAN.R-project.org/package=erah.
Metabolite Detector	GC-MS	具有解卷积、峰提取、峰鉴定等功能，可通过计算未知化合物的 Kovats 保留指数 (Kovats Index, KI) 与正构烷烃混合物的保留指数进行比对实现对未知化合物的准确定性	http://metabolitedetector.tu-bs.de.

注：AMDIS: 自动质谱退卷积定性系统。

Note: AMDIS: Automated mass spectral deconvolution and identification system.

1.1　LC/GC-MS 数据预处理

XCMS 是 LC-MS 数据预处理最常用的软件之一，它是基于 R 语言开发的，针对不同类型的质谱数据建立了不同的特征峰检测和峰对齐算法，也适合于 GC-MS 数据预处理。XCMS 可以实现质谱数据过滤、峰识别、峰提取、峰对齐和定量等功能，但在分析大规模样品时比较耗时。此外，XCMS 可以与其他 R 包如 ggplot2、prcomp 和 heatmap2 等，整合进行多变量统计分析、聚类分析等[18]。最新开发的 XCMS Online 是 XCMS 的网页版本，支持多种实验方案数据分析，还可进行单变量分析、多变量分析等统计分析以及代谢途径富集分析[19]。Lommen[20] 开发了 MetAlign 软件，可用于 GC-MS 和 LC-MS 数据预处理，支持数据格式自动转换、计算精确的质量数、基线校正、峰提取、滤噪及超 1 000 个数据集的峰对齐，该软件的缺点也是在大规模数据分析中比较耗时。此外，还有不少软件可同时兼顾 GC-MS 和 LC-MS 质谱数据的预处理，如 Normalyzer[21]、RUV-2[22]、NOREVA[23] 软件可适用于 GC-MS 和 LC-MS 质谱数据的归一化处理；MetTailor[24]、MetDIA[25]、TracMass 2[26]、MetFlow[27]、IP4M[28]、WiPP(Workflow for improved peak picking)[29] 等软件适用于滤噪、峰提取和峰对齐。

1.2　LC-MS 数据预处理

Pluskal 等[30] 开发了一个基于 Java 的开源 LC-MS 质谱数据分析工具 MZmine2，它可以实现数据的批处理和结果可视化，具有代谢组数据的峰提取、滤噪、解卷积、谱峰匹配和标准化等预处理功能。Röst 等[31] 开发了一个基于 C++ 编写的开源软件 OpenMS，提供了 185 个工具和现成的工作流程用于 LC-MS 质谱数据处理、可视化和定量分析，该软件为用户提供了高度灵活和专业的软件环境，以减少数据处理过程中出现潜在的错误分析。Tsugawa 等[32] 开发了一款专门用于解决数据非依赖采集（Data independent acquisition, DIA）LC-MS 数据解卷积的软件 MS-DIAL，该软件兼具有滤噪、峰提取、峰对齐、归一化等功能。Delabriere 等[33] 开发了一款用于分析大规模代谢组学和脂质组学 LC-MS 数据的软件 SLAW，该软件具有程序参数自动优化、峰提取、峰对齐、缺值填充、MS2 离子碎片信息提取和同位素模式识别等功能。Guo 等[34] 开发了一个多功能的代谢组数据分析 R 语言包 JPA，提供全面系统的代谢物特征提取和注释功能，其不仅可以直接从 LC-MS

原始数据中提取代谢物特征信息，而且还从其他数据处理软件（如XCMS、MS-DIAL、MZmine 2）处理的结果中对代谢物特征信息做进一步的提取。由于背景噪音、重复峰或污染会导致前处理后的数据存在假阳性色谱峰，因此还有一些功能相对专一的软件如ApLCMS[35]、质谱特征列表优化器（Mass Spectral Feature List Optimizer, MS-FLO）[36]、CPVA[37]被开发并应用于消除假阳性色谱峰，其中近期报道的Peakonly是一种基于卷积神经网络（CNN）的深度机器学习算法平台，用于检测LC-MS原始质谱数据的真实阳性色谱峰，该算法在检测或排除低强度噪声峰值方面具有较高的灵活性，对真实阳性色谱峰的检测精度非常高[38]。还有不少软件被开发用于LC-MS质谱数据的滤噪、峰提取、峰对齐等，如peakPantheR[39]、DecoID[40]、Galaxy-M[41]、SPICA[42]、MET-COFEI[43]等。

1.3 GC-MS数据预处理

由于LC-MS和GC-MS两种方法在电离模式、色谱分离、数据采集等方面都存在显著差异，因此，也有一些专门适用于GC-MS质谱数据分析的软件。AMDIS是GC-MS质谱数据预处理最常用软件之一。该软件可以有效克服GC-MS定性分析中基质效应和共洗脱效应的干扰，是一款功能强大的解卷积软件，自动完成峰滤噪、峰提取并利用GC-MS数据库完成谱图检索匹配[44]。Hiller等[45]开发了一款针对GC-MS数据开展有效峰提取和峰鉴定的软件MetaboliteDetector，该软件提供了一个交互式用户界面，以便没有经验的用户也可以轻易使用；同时，该软件还通过计算未知化合物的KI与正构烷烃混合物的保留指数进行比对实现对未知化合物的准确定性。Ni等[46]开发了一个基于质谱碎片离子分层聚类的解卷积算法平台ADAP-GC，具有峰提取、峰对齐等一系列数据处理功能，并且随着该平台的更新，最新版本ADAP-GC 4.0对代谢物峰检测的灵敏度、准确性和稳定性方面都有所提升[47]。Domingo-almenara等[48]开发了一个集成的R语言方法包eRah，它包含了一种基于盲源分离(blind source separation, BSS)的多变量技术的色谱解卷积方法，具有样品峰提取、峰对齐、定量和利用谱图数据库实现代谢物的自动识别的功能。最近报道的QPMASS软件，可以适用于大批量的GC-MS数据分析的软件，实现样品分组、峰提取、峰对齐、定量离子选择、缺失值过滤和填充等功能，使峰鉴定的假阳性和假阴性误差大大降低，其误差小于5%[49]。由于质谱检测的代谢物的相对强度或浓度存在数量级的差异，为了消除极限值数据在统计分析过程中忽略具有重要生物学意义但含量较低的代谢物的情况，在数据分析过程中需要减少极限值造成的误差。因此，在数据预处理后需要对数据进行归一化处理（normalization）或标度化（scaling）和数据转换等进一步的处理，系列软件被专门开发用于数据归一化处理，包括归一化自动编码器(Normalization Autoencoder, NormAE)[50]、MetTailor[24]、Normalyzer[21]、EigenMS[51]、MSPrep[52]等。此外还有很多软件在被开发用于GC-MS质谱数据的滤噪、峰提取、峰对齐等，如TagFinder[53]、MetaQuant[54]、PyMS[55]、MetaMS[56]、Maui-VIA[57]、GC2MS[58]等。

2 代谢组学数据分析

2.1 统计分析

采集的质谱原始数据通过滤噪、解卷积、峰识别、峰提取、峰对齐，归一化和缺失值填充等预处理后，形成的数据矩阵可用于进一步的数据统计分析。代谢组学数据统计分析主要分为单变量统计分析和多维统计分析，单变量统计分析包括相关性分析，例如皮尔森相关性系数、斯皮尔曼相关性系数、方差分析（ANOVA）和*t*-test检验分析等；多维统计分析又可分为非监督模式识别方法和监督模式识别方法两大类，非监督模式识别方法包括主成分分析（Principal component analysis, PCA）、自组织投影（Self-organizingmap, SOM）、聚类分析（Hierarchical cluster analysis，HCA），监督模式识别方法包括偏最小二乘法（Partial least squares, PLS）、偏最小二乘法－显著性分析联合法（Partial least squares-discriminant analysis, PLS-DA）、人工神经网络（Artificial neural network, ANN）、线性判别分析法（Linear discrimination analysis, LDA）、随机森林（Random forest, RF）和支持向量机法（Support vector machine, SVM）等[4, 16]，其中PCA和PLS-DA是目前代谢数据分析中使用最广泛的方法。SIMCA-P是一个功能强大、可实现多

元变量统计分析的商业软件，将数据转换成可视化信息，并应用于鉴定生物标志物和寻找差异代谢物等[59]。CV-ANOVA是基于交叉验证预测残差建立PLS和OPLS模型并进行方差分析，其优势是可以将交互验证的结果以统计学意义的p值展现出来，但该方法对于小样本集的检验效果较差[60]。MetabR使用线性混合模型对数据进行归一化处理然后采用方差分析ANOVA检验分析效果[61]。相比之下，种群模型分析-随机森林（Model population analysis-random forest, MPA-RF）是将随机森林与种群模型分析相结合，用于选择差异代谢物信息[62]。Metabomxtr通过建立混合分析模型处理代谢物缺失值的问题[63]。许多通用的统计软件能够执行常规的统计分析功能，但也有不少软件将其他代谢组学数据分析功能整合到同一个工作流程中，如Metabololyzer[64]、metaP-Server[65]、MSPrep[52]等。

2.2 代谢途径富集分析

富集分析是通过超几何分布检验（Hypergeometric test）或Fisher精确概率法建立统计模型分析数据中差异代谢物在各个生物通路中的富集情况，以此来帮助识别和解释其生物学功能。Xia等[66]开发了第一个小分子富集分析软件MSEA（Metabolite set enrichment analysis），它通过识别和解释代谢产物浓度变化模式来帮助研究人员注释代谢物的生物学意义，该方法的关键是通过构建分布于各个代谢途径上的1 000种具有相关性的代谢物数据库进行富集分析，MSEA可为代谢组学研究提供过表达分析（Over representation analysis, ORA）、单样本分析（Single sample profiling, SSP)和定量富集分析（Quantitative enrichment analysis, QEA）3种不同的富集分析。由于MSEA分析过程中常常对重叠代谢物集的权重分配不当而导致假阳性率较高，因此Deng等[67]提出了一种偏最小二乘扩展模型，用于解决重叠代谢物集的富集分析假阳性高的问题，称为ogPLS分析（Overlapping group PLS），将ogPLS模型的权重向量分解为代谢通路特异性子向量，从而再重新分配重叠代谢物的权重。以上两种方法相比，ogPLS方法具有较高的准确率、较低的假阳性率和更好的稳定性，适用于重叠代谢物集分析。Moreno等[68]开发了一个基于ChEBI（Chemical entities of biological interest）实体小分子数据库进行富集分析的工具BiNChE，该工具提供基于ChEBI角色实体（ChEBI Role Ontology）或ChEBI结构实体（ChEBI Structural Ontology）的简单的加权和片段分析，有助于探索代谢组学或其他系统生物学研究背景下产生的大量小分子，分析结果以交互式图形展示，并可导出为高分辨率图像或网络格式图片。MetaboAnalyst4.0经过近10年的发展已经成为代谢组学分析中使用最广泛的平台（30万用户）之一，支持LC-MS原始质谱数据预处理、数据归一化、统计分析、代谢通路富集分析等，旨在实现代谢组学的高通量分析，并缩小从原始数据到生物学见解之间的距离[69]。

2.3 代谢物鉴定

2.3.1 *基于数据库检索的代谢物鉴定* 代谢物鉴定是基于质谱的代谢组学研究中最具挑战性的步骤，代谢物鉴定的准确性在很大程度上取决于准确质量数、质谱谱图、离子碎裂模式、保留时间等信息。基于数据库检索的代谢物鉴定方法是最传统的方法，代谢物鉴定的常用数据库见表2。

NIST数据库是谱库检索中应用最广泛的质谱谱图数据库之一，可以用于谱库检索以识别GC-MS和LC-MS质谱中的未知化合物。NIST数据库包含有多个碰撞能级采集的二级（MS/MS）质谱图、不同加合离子的质谱图、化合物名称、分子式和CAS号等信息[70]。HMDB包含关于人体小分子代谢物的详细信息，截至2022年9月该数据库包含220 945个水溶性和脂溶性代谢物信息，同时还有DrugBank、T3DB、SMPDB和Food DB 4个子数据库可应用于药物、药物代谢物、毒素、环境污染物、人类疾病相关代谢物和食物方面的研究[71]。GNPS是一个利用分子网络构建天然产物数据库，具有代谢组学数据分析功能，其涵盖了Massbank、HMDB、NIST等第三方数据库的信息，以及实验室采集的化合物谱图和全球多个科研社团提供的质谱数据库，实现MS/MS质谱数据共享功能[72]。METLIN是另一个被广泛使用的高分辨质谱数据库，涵盖了不同碰撞能级和正/负模式条件下采集的MS/MS图谱，可以找到代谢产物的碎片离子、其来自标准品及其稳定同位素标记的类似物生成的谱图，在未知物的鉴定过程中起着关键作用[73]。MassBank数据库包含了来自不同实验室、不同仪器型号以及

表 2　代谢物鉴定常用数据库

Table 2　Common databases for metabolites identification

数据库 Database	质谱数据类型 MS data type	功能简介 Function description	网址 Website
NIST	LC–MS, GC–MS	应用最广泛的质谱数据库，包括化合物的二级质谱图、多离子加合物谱图、化合物名称、分子式、CAS 号等信息	http://www.sisweb.com/software/ms/nist.htm
HMDB	LC–MS, GC–MS	人类代谢组数据库，包括化合物分子量、分子式、化学性质、代谢通路和二级质谱图等信息	http://www.hmdb.ca/
GNPS	LC–MS, GC–MS	全球天然产物社会分子网络平台，利用分子网络构建天然产物数据库，具有代谢组学数据分析功能	http://gnps.ucsd.edu/
METLIN	LC–MS	高分辨质谱数据库，具有不同碰撞能级以及正 / 负模式条件下采集的代谢物标准品二级质谱图，精准分子量、分子式、结构式和质谱碎片信息等	http://metlin.scripps.edu/
GMD	GC–MS	植物代谢组学数据库，包含大量的植物代谢物的 GC–MS 谱图，特别是衍生化后的代谢物，还记录了部分代谢产物在植物中的浓度信息	http://gmd.mpimp–golm.mpg.de/
LipidMaps	LC–MS	一个包含生物相关脂质结构和注释的数据库，截至 2022 年 9 月包含 47 718 种独特的脂质结构，是世界上最大的脂质公共数据库	http://www.lipidmaps.org/
KEGG	LC–MS, GC–MS	最重要的生物信息学数据库之一，包含基因、蛋白、化学物质和酶 4 个部分，涵盖生物各类代谢物的代谢途径以及各个途径之间关系的信息	http://www.genome.jp/kegg/

注：NIST：美国国家标准与技术研究院数据库；HMDB：人类代谢组数据库；GNPS：全球天然产物社会分子网络；GMD：Golm 代谢物数据库；KEGG：京都基因和基因组百科全书。

Note: NIST: The National Institute of Standards and Technology; HMDB: Human Metabolome Database; GNPS: Global Natural Products Social Molecular Networking; GMD: The Golm Metabolome Database; KEGG: Kyoto Encyclopedia of Genes and Genomes.

不同质谱参数条件下采集的多级质谱数据用于代谢物鉴定代谢物，该数据库可以通过化学名称、质量数、质荷比 m /z 和分子式进行搜索，截至 2022 年 9 月数据库涵盖了 15 075 个代谢物的 90 190 个质谱数据，其中有 68 941 个二级质谱图，对化合物鉴定非常有用[74]。GMD 是一个植物代谢物数据库，含有大量的植物代谢产物的 GC–MS 图谱（特别是衍生化后的），用户可以导入样品的 GC–MS 数据进行搜索比对和鉴定。该数据库仅收录植物的代谢组，并含有部分代谢产物在植物中的浓度信息[75]。ReSpect 是另一个植物代谢物数据库，包括文献记录以及真实标准品的 MS/MS 数据[76]。针对脂类物质，Lipid Maps 是一个包含生物相关脂质结构和注释的数据库，截至 2022 年 9 月包含了 47 718 种独特的脂质结构，是世界上最大的脂质公共数据库。支持通过脂质类别、常用名、系统命名、分子量、InChIKey 命名或 Lipid Map 编号进行检索[77]。此外，还有一些常见的基于化合物谱库（谱图）信息建立的数据库，如 MetaboLights[78]、PubChem[79]、mzCloud[80]、Fiehn[81]、MoNA[82]、LipidIMMS Analyzer[83]等。

尽管从上述数据库中通过图谱匹配可以鉴定非常多的代谢物，但仍有许多代谢物由于缺乏标准品 MS/MS 图谱而难以鉴定出来。因此，基于代谢途径而开发的数据库应运而生。KEGG 是最重要的生物信息学数据库之一，涵盖了代谢通路和整合代谢、基因和蛋白通路的信息。截至 2022 年 10 月 9 日，KEGG 数据库含有 558 条代谢通路和 18 991 个代谢产物和化学结构信息，通过对生物代谢物分子的相互作用和反应网络实现对代谢物的注释[84]。MetaCyc 是一个包含了初级和次级代谢物途径的数据库，其中收集了来自 3 000 多种生物近 2 800 个代谢通路[85]。PlantCyc 9.5 数据库（https://plantcyc.org/databases/plantcyc/9.5）提供超过 350 种植物和 800 条代谢通路信息，包含代谢通路、催化的酶和基因，以及各种植物代谢物，同时整合了各种植物代谢通路数据库，包括 MetaCyc 数据库中所有的植物代谢通路。WikiPathways 包含 30 多个物种的代谢通路，如水稻（*Oryza sativa*）、玉米（*Zea mays*）等[86]。

2.3.2　基于分子网络技术的代谢物鉴定　2012 年，Watrous 等[87]首次开发了分子网络方法用于代谢物鉴定，是基于质谱的代谢组学数据分析的一个突破性进展，这一方法通过 MS/MS 谱图对比，构建以谱图为节点、谱图相似性为边线的网络，从而进行代谢物的注释。分子网络方法能够有效地利用已有数据，如 GNPS 中集成的大规模代谢组学、分子网络数据集，从而增强对代谢物的注释能力[72]。目前，已有许多先进的分子网络工具被开发并应用于 LC–MS/MS 数据分析和代谢物的注释。例如，在对复杂生物基质进行分析时，首先指认出提取物中的已知化合物（即去重

复）被认为是未知代谢物鉴定的关键步骤。Allard等[88]提出了一种分子网络和天然产物模拟MS/MS碎片数据库(in-silico MS/MS database, ISDB)相结合的去重复策略，并使用这一策略分析了植物和真菌提取物，结果表明模拟MS/MS碎片数据库能够有效地帮助分子网络中节点的注释。基于结构相似性的分子网络也被应用于提高模拟碎片峰预测的准确性，从而增强其注释能力[89]。

针对天然产物的鉴定，Mohimani等[90-91]使用去重复的策略开发了DEREPLICATOR和DEREPLICATOR+算法。这两种算法中，DEREPLICATOR通过将分子网络用于多肽匹配谱图的搜索，实现了已知多肽天然产物新变体的可变去重复，并允许对网络中的谱图所代表的多肽结构相关性提出假设。经测试，在GNPS分子网络平台中搜索近1亿个串联质谱后，DEREPLICATOR能够鉴定的多肽天然产物及其新变体的数量相比于以往的去重复策略有数量级的提升[90]。由于这一方法只能鉴定多肽天然产物，作者又开发了DEREPLICATOR+算法，将上述策略拓展于聚酮化合物、萜烯、苯类、生物碱、类黄酮等天然产物的鉴定，在GNPS分子网络平台中搜索近2亿个串联质谱的结果显示DEREPLICATOR+能够鉴定的分子数相比于以往的方法提高了5倍[91]。

上述分子网络以及结合ISDB的方法使用已知标准品或者模拟得到的碎片离子谱图库来鉴定代谢物，然而碎片离子谱图包含的与生物化学特征相关的信息却被忽略了。为此，Van der Hooft等[92]研究组开发了MS2LDA，一种无监督的分析方法，这一方法通过在碎片数据中提取生物化学相关的分子亚结构，并作为共同出现的分子片段和中性丢失碎片峰的集合（Mass2Motifs），然后使用分子共享的亚结构进行分组，再根据这些亚结构来推定新的结构注释。使用MS2LDA分析4种啤酒提取物的结果表明，在没有训练数据的情况下，使用30个结构表征的Mass2Motifs能够注释的分子数为传统库匹配方式的3倍。为了整合分子网络、生物化学特征和模拟碎片峰等多种来源的结构信息，以增强从不同数据集中提取化学信息的能力，Ernst等[93]开发了MolNetEnhancer以提供代谢组学数据的更全面的化学概述，并阐明每个碎片峰的结构细节，4个植物和细菌的研究案例显示MolNetEnhancer能够通过组合多个独立的分析流程来帮助研究者解读代谢组学数据。

此外，一些新的策略被整合到分子网络分析中。例如，通过整合高分辨率同位素模式分析和碎片峰树（Fragmentation trees），SIRIUS 4能够完成大型MS/MS数据集的分子结构评估，并通过分子网络传播注释[94]。结合贝叶斯统计和Gibbs采样，Ludwig等[95]建立了一种不依赖数据库的分子式注释方法ZODIAC，通过构建一个相对更小的相关化合物网络，其运算速度提升了25倍。基于代谢反应网络的递归算法，Shen等[96]开发了一种使用MS/MS谱图来表征初始种子代谢物，并利用其实验得到的MS/MS谱图作为替代谱图来注释其反应配对的领近代谢物的方法MetDNA。Beauxis等[97]则整合MS/MS谱图、GNPS中的分子网络、化学反应库和MS/MS谱图预测等信息开发了MetWork。一个比较大的进步是基于特征的分子网络方法（Feature-based molecular networking, FBMN）的开发，相比于传统的方法，FBMN整合了相对定量和离子淌度数据，从而实现了对同分异构体的分辨和分析[98]。另外，Tripathi等[99]提出了一种从碎片峰谱图预测分子指纹的分层组织策略Qemistree，这一方法可使用描述样本信息的元数据和化学本体来表示质谱数据，通过将分子关系表示为树，实现了使用基于系统发育的工具来分析代谢组学数据。

近年来，色谱和一级质谱信息（MS1）也被用于分子网络分析，以进一步开发高效的代谢物注释和鉴定方法。例如，Chen等[100]开发了一种全局网络优化方法NetID来注释非靶向代谢组学数据，这一方法根据对应于相关化学分子增减的MS1质量差异和MS/MS谱图的相似性来进行网络的全局优化。将此方法应用于酵母和小鼠数据的分析，作者鉴定到5种以前未识别的代谢产物。另外，在电离过程中，分子通常会形成具有不同碎裂行为的多种离子，而在传统的分子网络中这些离子的碎片峰通常不相连，导致相同类别的化合物的分子网络冗余且不相连。为了克服这一瓶颈，Schmid等[101]开发了一种离子识别分子网络算法（Ion identity molecular networking, IIMN），将色谱峰形状的相关性分析整合到分子网络中，以连接和折叠同一分子的不同离子种类。此外，Senan等[102]还建立了一种复杂生物样品和纯化

合物共洗脱曲线的相似性网络结合计算得到的加合物形成的自然频率，对冗余的 MS1 特征进行注释，从而为单个化合物提供准确注释的方法 CliqueMS。近期，Zhou 等[103]更是进一步开发了知识引导的多层网络算法（Knowledge-guided multi-layer network, KGMN），KGMN 使用基于知识的代谢反应网络、知识引导的 MS/MS 相似性网络和全局峰相关网络，实现了未知代谢物的有效注释。总的来说，通过多种实验数据、计算方法和分子网络算法的整合，实现了相对有效和准确的代谢物注释，具有广泛的应用前景。常见的用于分子网络鉴定的软件见表 3。

表 3　基于分子网络的代谢物鉴定相关软件

Table 3　Software for molecular networking-based metabolite identification

软件 Software	功能简介 Function description	参考文献 Reference
GNPS	一个社区范围内组织和共享原始、处理或注释的 MS/MS 质谱数据的开放访问知识库	[72]
ISDB	一种分子网络和天然产物模拟 MS/MS 谱图数据库相结合的去重复策略	[88]
DEREPLICATOR	一种新的去重复策略，通过将分子网络用于多肽匹配谱图的搜索，实现了已知多肽天然产物新变体的可变去重复，并允许对网络中的谱图所代表的多肽结构相关性提出假设	[90]
DEREPLICATOR +	一个将 DEREPLICATOR 的分析策略拓展到聚酮化合物、萜烯、苯类、生物碱、类黄酮等天然产物鉴定的分析工具	[91]
MS2LDA	一种无监督的分析方法，通过在碎片数据中提取生物化学相关的分子亚结构，作为共同出现的分子片段和中性丢失碎片峰的集合，并使用分子共享的亚结构构建分子网络，再根据这些亚结构来推定新的结构注释	[92]
MolNetEnhancer	一个整合分子网络、生物化学特征和模拟碎片峰等多种来源的结构信息的软件，能够提供代谢组学数据的全面的化学概述，并阐明每个碎片峰的结构细节	[93]
SIRIUS 4	一个通过整合高分辨率同位素模式分析和碎片峰树，完成大型 MS/MS 数据集的分子结构评估的工具，并能够通过分子网络传播注释	[95]
MetDNA	一种使用 MS/MS 谱图来表征初始种子代谢物，并利用其实验得到的 MS/MS 谱图作为替代谱图来注释其反应配对的领近代谢物的方法	[96]
MetWork	一种整合 MS/MS 谱图、GNPS 中的分子网络、化学反应库和 MS/MS 谱图预测等信息的方法	[97]
FBMN	一个基于特征的分子网络方法，整合了相对定量和离子淌度数据，从而实现了对同分异构体的分辨和分析	[98]
Qemistree	一个用于构建 MS/MS 特征树的计算工具，以树型结构整合描述特征分子网络、样本信息的元数据和其他化学注释信息，实现了使用基于系统发育的工具来分析代谢组学数据	[99]
NetID	一种全局网络优化方法来注释非靶向代谢组学数据，根据对应于相关化学分子增减的 MS1 质量差异和 MS/MS 谱图的相似性来进行分子网络的全局优化	[100]
IIMN	一种离子识别分子网络算法，将色谱峰形状的相关性分析整合到分子网络中，以连接和折叠同一分子的不同离子种类	[101]
CliqueMS	一种结合复杂生物样品和纯化合物的共洗脱曲线相似性网络，以及计算得到的加合物形成的自然频率，对冗余的 MS1 特征进行注释，从而为单个化合物提供准确注释的方法	[102]

2.3.3　*基于其他技术的代谢物鉴定*　尽管基于或整合分子网络的工具在代谢组学数据分析方面非常流行、通用且高效，但构建的网络依赖于分析参数，且没有保留对谱图相似性的全局分析。因此，一些基于降维和机器学习的分组方法被应用于质谱数据的分析，以提供分子网络中无法获得的信息，有望进一步提高代谢物注释的能力。例如，Bittremieux 等[104]提出了一种快速谱图相似性搜索方法 Falcon，能够对数百万 MS/MS 谱图进行有效的聚类和分组。传统的计算方法通常使用谱图相似性作为分子结构相似性的度量，两种指标的相关性制约了分析方法的有效性。为了解决这一问题，Huber 等[105]开发了一种孪生神经网络算法 MS2DeepScore，这一方法实现了根据两个化学结构的 MS/MS 谱图来预测其结构相似性。Falcon 和 MS2DeepScore 是两种大规模 MS/MS 谱图比较和分析的强大工具，被认为在代谢组学数据分析和注释方面具有较大潜力。此外，机器学习算法也被应用于预测色谱保留时间，以增强其在代谢物注释中的可用性。García 等[106]将多种机器学习算法应用于预测色谱保留时间并整合到代谢物注释的流程中，以获得候选注释的 Z-cores，实验测试结果显示 68% 的正确注释出现在按质量过滤并按 Z-cores 排序的前 3 个候选分子中，表明其对支持代谢物注释的有效性。而针对模拟谱图无法区分正确和错误注释的问题，Hoffmann

等[107]近期开发了一种模拟谱图数据库的生成、注释和置信度评分相结合的方法COSMIC（Confidence of small molecule identifications），这一方法库搜索的注释错误率更低，并实现了多个未知结构的天然胆汁酸的准确注释。

除了上述数据分析方法，全面、自动化和可重复的代谢组学分析流程对于准确有效的化合物注释也至关重要。为此，Shen等[108]首先开发了一个基于LC-MS数据进行自动化合物注释的R包metID。metID结合了所有主要数据库的信息，是一个灵活、简单、强大的工具，可以安装在所有平台上。使用metID分析一个已发表案例数据的结果显示其不仅完成了发表论文中所有的463个代谢物的注释，还注释了479个新的代谢物[98]。基于这一工具，作者又进一步开发了面向对象的计算框架TidyMass，实现了基于LC-MS的非靶向代谢组学数据处理和分析的可追溯、可共享和可重复[109]。另外，Yu等[110]提出了一种自动化的、全面且无统计模型的工作流程PMDDA（Paired mass distance-dependent analysis），这一流程根据MS1的特征进行全面的MS/MS数据采集，实现了更多化合物的注释。

总的来说，目前已经研发了很多新的代谢组学数据分析和注释工具，大大促进了这一领域的发展。同时，大量的新型工具也使得用户很难判断其适用性，尽管开发者通常会将其研发的工具与其他方法进行比较，但目前仍然缺少标准化的测试数据集来进行关键的性能评估和比较。建立适用于评估分析工具通用性、有效性和重复性的大量、随机的数据集，不仅能够帮助用户选择其需要的工具，也能促进方法开发的标准化，是未来值得探索的重要方向。

3 结语与展望

基于GC-MS和LC-MS等质谱系统的代谢组学数据分析主要包括质谱数据预处理、代谢组学数据统计分析、代谢途径富集分析以及代谢物鉴定等步骤。过去十来年许多关于质谱数据预处理、多维变量统计分析、代谢途径分析和代谢物数据库的分析软件被相继开发和成功应用。特别是近年来计算代谢组学方法迅猛发展，极大地推动了代谢组学数据分析流程的自动化和规范化，为大规模代谢组学数据的充分挖掘打下坚实基础。而分子网络、机器学习等前沿方法也大大提高了代谢物的注释和鉴定能力，显著提升了代谢物特征信息提取的准确性以及代谢物鉴定的覆盖范围。然而，由于生物样本的复杂性以及现有质谱分析技术的局限性，使得代谢组学所能检测并注释的代谢物数量远远少于生物体内源代谢物的数量，难以满足现代研究发展的要求。因此，在未来的研究中，首先需要进一步发展高通量、高分辨率和高灵敏度的先进质谱数据采集技术，提高对低丰度代谢物的检测能力，以实现对代谢组学原始数据更充分地采集，从而构建更全面的代谢物质谱数据库。其次，深入开发更强大的计算代谢组学分析工具对于提高代谢物的鉴定和注释能力也至关重要。已有研究显示代谢物的生物化学特征、反应网络等信息，能够明显提升分子网络技术对代谢物的注释能力；同一条代谢途径常常受到相同遗传位点的调控，因此，在已知生化反应和分子网络分析的基础上再整合代谢物合成的遗传位点等信息，也有望进一步提高代谢物的鉴定数和准确度。此外，不断优化代谢物注释的算法必将极大地促进代谢组学研究的发展，也是未来代谢组学数据分析研究的重点。

我们相信，随着高分辨质谱仪的更新迭代和不同代谢组数据分析方法的相继开发，定会极大提高基于质谱的代谢组学技术分析能力，主要表现为代谢物的分析效率、鉴定数量、灵敏度和精准度得到不断提升。在农业领域，基于质谱的代谢组学分析将助力于农业生物复杂性状形成的机制探索，农业生物重要代谢途径的解析、农作物生长发育与胁迫应答的代谢调控网络研究，以及转基因安全评估等不同学科领域。

参考文献（References）：

[1] NICHOLSON J K, LINDON J C, HOLMES E. ‘Metabonomics’: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data[J]. *Xenobiotica*, 1999, 29(11): 1181-1189. DOI: 10.1080/004982599238047.

[2] MA A, QI X. Mining plant metabolomes: Methods, applications, and perspectives[J]. *Plant Communications*, 2021, 2(5): 100238. DOI: 10.1016/j.xplc.2021.100238.

[3] 顾渝娟,吴振先. 代谢组学在植物研究中的应用[J]. 广东农业科学, 2012, 39(4): 105-107. DOI: 10.16768/j.issn.1004-874X.2012.04.003.
GU Y J, WU Z X. Application of metabolomics in plant research[J]. *Guangdong Agricultural Sciences*, 2012, 39(4): 105-107. DOI:

10.16768/j.issn.1004-874X.2012.04.003.

[4] 梁丹丹,李忆涛,郑晓皎,陈天璐. 代谢组学全功能软件研究进展[J]. 上海交通大学学报(医学版), 2018, 38(7): 805-810. DOI: 10.3969/j.issn.1674-8115.2018.07.017.
LIANG D D, LI Y T, ZHENG X J, CHEN T L. Advance in full-functional software of metabolomics[J]. *Journal of Shanghai Jiao Tong University (Medical Science)*, 2018, 38(7): 805-810. DOI: 10.3969/j.issn.1674-8115.2018.07.017.

[5] FERNIE A R, TRETHEWEY R, N, KROTZKY A J, WILLMITZER L. Metabolite profiling: From diagnostics to systems biology[J]. *Nature Reviews Molecular Cell Biology*, 2004, 5(9): 763-769. DOI: 10.1038/nrm1451.

[6] FERNIE A R, TOHGE T. The genetics of plant metabolism[J]. *Annual Review of Genetics*, 2017, 51: 287-310. DOI: 10.1146/annurev-genet-120116-024640.

[7] 刘丽娜, 傅曼琴, 徐玉娟, 吴继军, 余元善, 温靖. 基于GC-MS技术分析不同贮藏年份陈皮挥发性成分差异[J]. 广东农业科学, 2020, 47(9): 114-120. DOI: 10.16768/j.issn.1004-874X.2020.09.015.
LIU L N, FU M Q, XU Y J, WU J J, YU Y S, WEN J. Analysis of differences in volatile components of pericarpium citri reticulatae in different storage years based on GC-MS[J]. *Guangdong Agricultural Sciences*, 2020, 47(9): 114-120. DOI: 10.16768/j.issn.1004-874X.2020.09.015.

[8] 李辉, 易恒洁, 彭邦星 赵忠海. 贵妃鸡肌肉脂肪酸的GC-MS分析[J]. 广东农业科学, 2014, 41(1): 96-97. DOI: 10.16768/j.issn.1004-874X.2014.01.024.
LI H, YI H J, PENG B X, ZHAO Z H. Analysis on components of fatty acid in muscle of Royal chicken by GC-MS[J]. *Guangdong Agricultural Sciences*, 2014, 41(1): 96-97. DOI: 10.16768/j.issn.1004-874X.2014.01.024.

[9] 傅秀敏, 唐劲驰, 杨子银. 茶叶类胡萝卜素合成、代谢调控研究进展[J]. 广东农业科学, 2021, 48(5): 18-27. DOI: 10.16768/j.issn.1004-874X.2021.05.003.
FU X M, TANG J C, YANG Z Y. Research progress in biosynthesis and metabolism regulation of carotenoids in tea plants[J]. *Guangdong Agricultural Sciences*, 2021, 48(5): 18-27. DOI: 10.16768/j.issn.1004-874X.2021.05.003.

[10] RAZZAQ A, SADIA B, RAZA A, KHALID HAMEED M, SALEEM F. Metabolomics: A way forward for crop improvement[J]. *Metabolites*, 2019, 9(12): 303. DOI: 10.3390/metabo9120303.

[11] ZEKI Ö C, EYLEM C C, REÇBER T KıR S, NEMUTLU E. Integration of GC-MS and LC-MS for untargeted metabolomics profiling[J]. *Journal of Pharmaceutical and Biomedical Analysis*, 2020, 190: 113509. DOI: 10.1016/j.jpba.2020.113509.

[12] 殷志斌,黄文洁,伍欣宙,晏石娟. 空间分辨代谢组学进展和挑战[J]. 生物技术通报, 2021, 37(1): 32-51. DOI:10.13560/j.cnki.biotech.bull.1985.2020-1374.
YIN Z B, HUANG W J, WU X Z, YAN S J. Spatially resolved metabolomics:progress and challenges[J]. *Biotechnology Bulletin*, 2021, 37(1): 32-51. DOI: 10.13560/j.cnki.biotech.bull.1985.2020-1374.

[13] 黄小丹, 陈梦雨, 黄文洁, 张名位, 晏石娟. 基于代谢组学的植物多酚及其肠道健康效应研究进展[J]. 生物技术通报, 2021, 37(1): 123-136. DOI: 10.13560/j.cnki.biotech.bull.1985.2020-1409.
HUANG X D, CHEN M Y, HUANG W J, ZHANG M W, YAN S J. Progress based on metabolomics: plant polyphenols and their gut health benefit[J]. *Biotechnology Bulletin*, 2021, 37(1): 123-136. DOI: 10.13560/j.cnki.biotech.bull.1985. 2020-1409.

[14] YAN S, BHAWAL R, YIN Z, THANNHAUSER T W, ZHANG S. Recent advances in proteomics and metabolomics in plants[J]. *Molecular Horticulture*, 2022, 2: 17. DOI: 10.1186/s43897-022-00038-9.

[15] YAN S, HUANG J, CHEN Z, JIANG Z, LI X, CHEN Z. Metabolomics in gut microbiota: applications and challenges[J]. *Science Bulletin*, 2016, 61(15): 1151-1153. DOI: 10.1007/s11434-016-1142-7.

[16] PEREZ DE SOUZA L, NAAKE T, TOHGE T, FERNIE A R. From chromatogram to analyte to metabolite: How to pick horses for courses from the massive web resources for mass spectral plant metabolomics[J]. *Giga Science*, 2017, 6: 1-20. DOI: 10.1093/gigascience/gix037.

[17] 杨军, 刘心昱, 许国旺. 基于质谱数据的计算代谢组学方法学研究进展[J]. 中国科学: 化学, 2022, 52(9): 1580-1591. DOI: 10.1360/ssc-2022-0084.
YANG J, LIU X Y, XU G W. New advances in mass spectrometry data-based computational metabolomics methods[J]. *Scientia Sinica Chimica*, 2022, 52(9): 1580-1591. DOI: 10.1360/ssc-2022-0084.

[18] SMITH C A, WANT E J, O' MAILLE G, ABAGYAN R, SIUZDAK G. XCMS_ processing mass_spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification[J]. *Analytical Chemistry*, 2006, 78(3): 779-787. DOI: 10.1021/ac051437y.

[19] TAUTENHAHN R, PATTI G J, RINEHART D, SIUZDAK G. XCMS Online: A web-based platform to process untargeted metabolomic data[J]. *Analytical Chemistry*, 2012, 84(11): 5035-5039. DOI: 10.1021/ac300698c.

[20] LOMMEN A. MetAlign: Interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing[J]. *Analytical Chemistry*, 2009, 81(8): 3079-3086. DOI: 10.1021/ac900036d.

[21] CHAWADE A, ALEXANDERSSON E, LEVANDER F. Normalyzer: A tool for rapid evaluation of normalization methods for omics data sets[J]. *Journal of Proteome Research*, 2014, 13(6): 2114-3120. DOI: 10.1021/pr401264n.

[22] DE LIVERA A M, DIAS D A, DE SOUZA D, RUPASINGHE T, PYKE J, TULL D, ROESSNER U, MCCONVILLE M, SPEED T P. Normalizing and integrating metabolomics data[J]. *Analytical Chemistry*, 2012, 84(24): 10768-10776. DOI: 10.1021/ac302748b.

[23] YANG Q, WANG Y, ZHANG Y, LI F, XIA W, ZHOU Y, QIU Y, LI H, ZHU F. NOREVA: Enhanced normalization and evaluation of time-course and multi-class metabolomic data[J]. *Nucleic Acids Research*, 2020, 48(W1): W436-W448. DOI: 10.1093/nar/gkaa258.

[24] CHEN G, CUI L, TEO G S, ONG C N, TAN C S, CHOI H. MetTailor: Dynamic block summary and intensity normalization for robust analysis of mass spectrometry data in metabolomics[J]. *Bioinformatics*, 2015, 31(22): 3645-3652. DOI: 10.1093/bioinformatics/btv434.

[25] LI H, CAI Y, GUO Y, CHEN F, ZHU Z J. MetDIA: Targeted metabolite extraction of multiplexed MS/MS spectra generated by data-independent acquisition[J]. *Analytical Chemistry*, 2016, 88(17): 8757-8764. DOI: 10.1021/acs.analchem.6b02122.

[26] TENGSTRAND E, LINDBERG J, ÅBERG K M. TracMass 2-A modular suite of tools for processing chromatography-full scan mass spectrometry data[J]. *Analytical Chemistry*, 2014, 86(7): 3435-

3442. DOI: 10.1021/ac403905h.

[27] SHEN X, ZHU Z J. MetFlow: An interactive and integrated workflow for metabolomics data cleaning and differential metabolite discovery [J]. *Bioinformatics*, 2019, 35(16): 2870–2872. DOI: 10.1093/bioinformatics/bty1066.

[28] LIANG D, LIU Q, ZHOU K, JIA W, XIE G, CHEN T. IP4M: An integrated platform for mass spectrometry–based metabolomics data mining [J]. *BMC Bioinformatics*, 2020, 21(1): 444. DOI: 10.1186/s12859–020–03786–x.

[29] BORGSMÜLLER N, GLOAGUEN Y, OPIALLA T, BLANC E, SICARD E, ROYER A L, LE BIZEC B, DURAND S, MIGNÉ C, PÉTÉRA M, PUJOS–GUILLOT E, GIACOMONI F, GUITTON Y, BEULE D, KIRWAN J. WiPP: Workflow for improved peak picking for gas chromatography–mass spectrometry (GC–MS) data [J]. *Metabolites*, 2019, 9(9): 171. DOI: 10.3390/metabo9090171.

[30] PLUSKAL T, CASTILLO S, VILLAR–BRIONES A, ORESIC M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry–based molecular profile data [J]. *BMC Bioinformatics*, 2010, 11: 395. DOI: 10.1186/1471–2105–11–395.

[31] RÖST H L, SACHSENBERG T, AICHE S, BIELOW C, WEISSER H, AICHELER F, ANDREOTTI S, EHRLICH H C, GUTENBRUNNER P, KENAR E, LIANG X, NAHNSEN S, NILSE L, PFEUFFER J, ROSENBERGER G, RURIK M, SCHMITT U, VEIT J, WALZER M, WOJNAR D, WOLSKI W E, SCHILLING O, CHOUDHARY J S, MALMSTRÖM L, AEBERSOLD R, REINERT K, KOHLBACHER O. OpenMS: A flexible open–source software platform for mass spectrometry data analysis [J]. *Nature Methods*, 2016, 13(9): 741–748. DOI: 10.1038/nmeth.3959.

[32] TSUGAWA H, CAJKA T, KIND T, MA Y, HIGGINS B, IKEDA K, KANAZAWA M, VANDERGHEYNST J, FIEHN O, ARITA M. MS–DIAL: Data independent MS/MS deconvolution for comprehensive metabolome analysis [J]. *Nature Methods*, 2015, 12(6): 523–526. DOI: 10.1038/nmeth.3393.

[33] DELABRIERE A, WARMER P, BRENNSTEINER V, ZAMBONI N. SLAW: A scalable and self–optimizing processing workflow for untargeted LC–MS [J]. *Analytical Chemistry*, 2021, 93(45): 15024–15032. DOI: 10.1021/acs.analchem.1c02687.

[34] GUO J, SHEN S, LIU M, WANG C, LOW B, CHEN Y, HU Y, XING S, YU H, GAO Y, FANG M, HUAN T. JPA: Joint metabolic feature extraction increases the depth of chemical coverage for LC–MS–based metabolomics and exposomics [J]. *Metabolites*, 2022, 12(3): 212. DOI: 10.3390/metabo12030212.

[35] YU T, PARK Y, JOHNSON J M, JONES D P. apLCMS––Adaptive processing of high–resolution LC/MS data [J]. *Bioinformatics*, 2009, 25(15): 1930–1936. DOI: 10.1093/bioinformatics/btp291.

[36] DEFELICE B C, MEHTA S S, SAMRA S, ČAJKA T, WANCEWICZ B, FAHRMANN J F, FIEHN O. Mass spectral feature list optimizer (MS–FLO): A tool to minimize false positive peak reports in untargeted liquid chromatography – mass spectroscopy (LC–MS) data processing [J]. *Analytical Chemistry*, 2017, 89(6): 3250–3255. DOI: 10.1021/acs.analchem. 6b04372.

[37] LUAN H, JIANG X, JI F, LAN Z, CAI Z, ZHANG W. CPVA: A web–based metabolomic tool for chromatographic peak visualization and annotation [J]. *Bioinformatics*, 2020, 36(12): 3913–3915. DOI: 10.1093/bioinformatics/btaa200.

[38] MELNIKOV A, D, TSENTALOVICH Y P, YANSHOLE V V. Deep learning for the precise peak detection in high–resolution LC–MS data [J]. *Analytical Chemistry*, 2020, 92(1): 588–592. DOI: 10.1021/acs.analchem. 9b04811.

[39] WOLFER A M, CORREIA G D S, SANDS C J, CAMUZEAUX S, YUEN A H Y, CHEKMENEVA E, TAKÁTS Z, PEARCE J T M, LEWIS M R. peakPantheR, An R package for large–scale targeted extraction and integration of annotated metabolic features in LC–MS profiling datasets [J]. *Bioinformatics*, 2021, 37(42): 4886–4888. DOI: 10.1093/bioinformatics/btab433.

[40] STANCLIFFE E, SCHWAIGER–HABER M, SINDELAR M, PATTI G J. DecoID improves identification rates in metabolomics through database–assisted MS/MS deconvolution [J]. *Nature Methods*, 2021, 18(7): 779–787. DOI: 10.1038/s41592–021–01195–3.

[41] DAVIDSON R L, WEBER R J M, LIU H, SHARMA–OATES A, VIANT M R. Galaxy–M: A Galaxy workflow for processing and analyzing direct infusion and liquid chromatography mass spectrometry–based metabolomics data [J]. *Gigascience*, 2016, 5: 10. DOI: 10.1186/s13742–016–0115–8.

[42] MAK T D, LAIAKIS E C, GOUDARZI M, FORNACE A J JR. Selective paired ion contrast analysis: a novel algorithm for analyzing postprocessed LC–MS metabolomics data possessing high experimental noise [J]. *Analytical Chemistry*, 2015, 87(6): 3177–3186. DOI: 10.1021/ac504012a.

[43] ZHANG W, CHANG J, LEI Z, HUHMAN D, SUMNER L W, ZHAO P X. MET–COFEA: A liquid chromatography/mass spectrometry data processing platform for metabolite compound feature extraction and annotation [J]. *Analytical Chemistry*, 2014, 86(13): 6245–6253. DOI: 10.1021/ac501162k.

[44] STEIN S E. An integrated method for spectrum extraction and compound identification from gas chromatography mass spectrometry data [J]. *Journal of the American Society for Mass Spectrometry*, 1999, 10(8): 770–781. DOI: 10.1016/S1044–0305(99)00047–1.

[45] HILLER K, HANGEBRAUK J, JÄGER C, SPURA J, SCHREIBER K, SCHOMBURG D. MetaboliteDetector: Comprehensive analysis tool for targeted and nontargeted GC–MS based metabolome analysis [J]. *Analytical Chemistry*, 2009, 81(9): 3429–3439. DOI: 10.1021/ac802689c.

[46] NI Y, QIU Y, JIANG W, SUTTLEMYRE K, SU M, ZHANG W, JIA W, DU X. ADAP–GC 2.0: Deconvolution of coeluting metabolites from GC/TOF–MS data for metabolomics studies [J]. *Analytical Chemistry*, 2012, 84(15): 6619–6629. DOI: 10.1021/ac300898h.

[47] SMIRNOV A, QIU Y, JIA W, WALKER D I, JONES D P, DU X. ADAP–GC 4.0: Application of clustering–assisted multivariate curve resolution to spectral deconvolution of gas chromatography – mass spectrometry metabolomics data [J]. *Analytical Chemistry*, 2019, 91(14): 9069–9077. DOI: 10.1021/acs.analchem.9b01424.

[48] DOMINGO–ALMENARA X, BREZMES J, VINAIXA M, SAMINO S, RAMIREZ N, RAMON–KRAUEL M, LERIN C, DÍAZ M, IBÁÑEZ L, CORREIG X, PERERA–LLUNA A, YANES O. eRah: A computational tool integrating spectral deconvolution and alignment with quantification and identification of metabolites in GC/MS–based metabolomics [J]. *Analytical Chemistry*, 2016, 88(19): 9821–9829. DOI: 10.1021/acs.analchem. 6b02927.

[49] DUAN L, MA A, MENG X, SHEN G A, QI X. QPMASS: A parallel peak alignment and quantification software for the analysis of large–scale gas chromatography–mass spectrometry (GC–MS)–based

metabolomics datasets [J]. *Journal of Chromatography* A, 2020, 1620: 460999. DOI: 10.1016/j.chroma.2020.460999.

[50] RONG Z, TAN Q, CAO L, ZHANG L, DENG K, HUANG Y, ZHU Z J, LI Z, LI K. NormAE: Deep adversarial learning model to remove batch effects in liquid chromatography mass spectrometry-based metabolomics data [J]. *Analytical Chemistry*, 2020, 92(7): 5082-5090. DOI: 10.1021/acs.analchem.9b05460.

[51] KARPIEVITCH Y V, NIKOLIC S B, WILSON R, SHARMAN J E, EDWARDS L M. Metabolomics data normalization with EigenMS [J]. *PLoS ONE*, 2014, 9(12): e116221. DOI: 10.1371/journal.pone.0116221.

[52] HUGHES G, CRUICKSHANK-QUINN C, REISDORPH R, LUTZ S, PETRACHE I, REISDORPH N, BOWLER R, KECHRIS K. MSPrep--Summarization, normalization and diagnostics for processing of mass spectrometry-based metabolomic data [J]. *Bioinformatics*, 2014, 30(1): 133-134. DOI: 10.1093/bioinformatics/btt589.

[53] LUEDEMANN A, STRASSBURG K, ERBAN A, KOPKA J. TagFinder for the quantitative analysis of gas chromatography-mass spectrometry (GC-MS)-based metabolite profiling experiments [J]. *Bioinformatics*, 2008, 24(5): 732-737. DOI: 10.1093/bioinformatics/btn023.

[54] BUNK B, KUCKLICK M, JONAS R, MÜNCH R, SCHOBERT M, JAHN D, HILLER K. MetaQuant: A tool for the automatic quantification of GC/MS-based metabolome data [J]. *Bioinformatics*, 2006, 22(23): 2962-2965. DOI: 10.1093/bioinformatics/ btl526.

[55] O' CALLAGHAN S, DE SOUZA D P, ISAAC A, WANG Q, HODKINSON L, OLSHANSKY M, ERWIN T, APPELBE B, TULL D L, ROESSNER U, BACIC A, MCCONVILLE M J, LIKIĆ V A. MS: A Python toolkit for processing of gas chromatography-mass spectrometry (GC-MS) data. Application and comparative study of selected tools [J]. *BMC Bioinformatics*, 2012, 13: 115. DOI: 10.1186/1471-2105-13-115.

[56] WEHRENS R, WEINGART G, MATTIVI F. metaMS: An open-source pipeline for GC-MS-based untargeted metabolomics [J]. *Journal of Chromatography* B, 2014, 996: 109-116. DOI: 10.1016/j.jchromb.2014.02.051.

[57] KUICH P H J L, HOFFMANN N, KEMPA S. Maui-VIA: A user-friendly software for visual identification, alignment, correction, and quantification of gas chromatography-mass spectrometry data [J]. *Frontiers in Bioengineering and Biotechnology*, 2015, 2: 84. DOI: 10.3389/fbioe.2014.00084.

[58] TIAN T F, WANG S Y, KUO T C, TAN C E, CHEN G Y, KUO C H, CHEN C S, CHAN C C, LIN O A, TSENG Y J. Web server for peak detection, baseline correction, and alignment in two-dimensional gas chromatography mass spectrometry-based metabolomics data [J]. *Analytical Chemistry*, 2016, 88(21): 10395-10403. DOI: 10.1021/acs.analchem.6b00755.

[59] WHEELOCK Å M, WHEELOCK C E. Trials and tribulations of 'omics data analysis: assessing quality of SIMCA-based multivariate models using examples from pulmonary medicine [J]. *Molecular Biosystems*, 2013, 9(11): 2589-2596. DOI: 10.1039/c3mb70194h.

[60] ERIKSSON L, TRYGG J, WOLD S. CV-ANOVA for significance testing of PLS and OPLS® models [J]. *Journal of Chemometrics*, 2008, 22(11-12): 594-600. DOI: 10.1002/cem.1187.

[61] ERNEST B, GOODING J R, CAMPAGNA S R, SAXTON A M, VOY B H. MetabR: An R script for linear model analysis of quantitative metabolomic data [J]. *BMC Research Notes*, 2012, 5: 596. DOI:0.1186/1756-0500-5-596.

[62] HUANG J H, YAN J, WU Q H, DUARTE FERRO M, YI L Z, LU H M, XU Q S, LIANG Y Z. Selective of informative metabolites using random forests based on model population analysis [J]. *Talanta*, 2013, 117: 549-555. DOI: 10.1016/j.talanta. 2013.07.070.

[63] NODZENSKI M, MUEHLBAUER M J, BAIN J R, REISETTER A C, LOWE JR W L , SCHOLTENS D M. Metabomxtr: An R package for mixture-model analysis of non-targeted metabolomics data [J]. *Bioinformatics*, 2014, 30(22): 3287-3288. DOI: 10.1093/bioinformatics/btu509.

[64] MAK T D, LAIAKIS E C, GOUDARZI M, FORNACE JR A J. MetaboLyzer: A novel statistical workflow for analyzing Postprocessed LC-MS metabolomics data [J]. *Analytical Chemistry*, 2014, 86(1): 506-513. DOI: 10.1021/ac402477z.

[65] KASTENMÜLLER G, RÖMISCH-MARGL W, WÄGELE B, ALTMAIER E, SUHRE K. metaP-server: A web-based metabolomics data analysis tool [J]. *Journal of Biomedicine and Biotechnology*, 2011, 2011: 839862. DOI: 10.1155/2011/ 839862.

[66] XIA J, WISHART D S. MSEA: A web-based tool to identify biologically meaningful patterns in quantitative metabolomic data [J]. *Nucleic Acids Research*, 2010, 38: W71-W77. DOI: 10.1093/nar/gkq329.

[67] DENG L, MA L, CHENG K K, XU X, RAFTERY D, DONG J. Sparse PLS-based method for overlapping metabolite set enrichment analysis [J]. *Journal of Proteome Research*, 2021, 20(6): 3204-3213. DOI: 10.1021/acs.jproteome.1c00064.

[68] MORENO P, BEISKEN S, HARSHA B, MUTHUKRISHNAN V, TUDOSE I, DEKKER A, DORNFELDT S, TARUTTIS F, GROSSE I, HASTINGS J, NEUMANN S, STEINBECK C. BiNChE: A web tool and library for chemical enrichment analysis based on the ChEBI ontology [J]. *BMC Bioinformatics*, 2015, 16(1): 56. DOI: 10.1186/s12859-015-0486-3.

[69] CHONG J, WISHART D S, XIA J. Using MetaboAnalyst 4.0 for comprehensive and integrative metabolomics data analysis [J]. *Current Protocols in Bioinformatics*, 2019, 68(1): e86. DOI: 10.1002/cpbi.86.

[70] MURRAY-RUST P, RZEPA H S, STEWART J J P, ZHANG Y. A global resource for computational chemistry [J]. *Journal of Molecular Modeling*, 2005, 11(6): 532-541. DOI: 10.1007/s00894-005-0278-1.

[71] WISHART D S, JEWISON T, GUO A C, WILSON M, KNOX C, LIU Y. DJOUMBOU Y, MANDAL R, AZIAT F, DONG E, BOUATRA S, SINELNIKOV I, ARNDT D, XIA J, LIU P, YALLOU F, BJORNDAHL T, PEREZ-PINEIRO R, EISNER R, ALLEN F, NEVEU V, GREINER R, SCALBERT A. HMDB 3.0-The human metabolome database in 2013 [J]. *Nucleic Acids Research*, 2013, 41(Database issue): 801-807. DOI: 10.1093/nar/gks1065.

[72] WANG M, CARVER J J, PHELAN V V, SANCHEZ L M, GARG N, PENG Y, NGUYEN D D, WATROUS J, KAPONO C A, LUZZATTO-KNAAN T, PORTO C, BOUSLIMANI A, MELNIK A V, MEEHAN M J, LIU W T, CRÜSEMANN M, BOUDREAU P D, ESQUENAZI E, SANDOVAL-CALDERÔN M, KERSTEN R D, BANDEIRA N. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking [J]. *Nature Biotechnology*, 2016, 34(8): 828-837. DOI: 10.1038/nbt.3597.

[73] GUIJAS C, MONTENEGRO-BURKE J R, DOMINGO-ALMENARA X, PALERMO A, WARTH B, HERMANN G, KOELLENSPERGER G, HUAN T, URITBOONTHAI W, AISPORNA A E, WOLAN

D W, SPILKER M E, BENTON H P, SIUZDAK G. METLIN: A technology platform for identifying knowns and unknowns [J]. *Analytical Chemistry*, 2018, 90(5): 3156–3164. DOI: 10.1021/acs.analchem.7b04424.

[74] HORAI H, ARITA M, KANAYA S, NIHEI Y, IKEDA T, SUWA K, OJIMA Y, TANAKA K, TANAKA S, AOSHIMA K, ODA Y, KAKAZU Y, KUSANO M, TOHGE T, MATSUDA F, SAWADA Y, HIRAI M Y, NAKANISHI H, IKEDA K, AKIMOTO N, MAOKA T, TAKAHASHI H, ARA T, SAKURAI N, SUZUKI H, SHIBATA D, NEUMANN S, IIDA T, TANAKA K, FUNATSU K, MATSUURA F, SOGA T, TAGUCHI R, SAITO K, NISHIOKA T. MassBank: A public repository for sharing mass spectral data for life sciences [J]. *Journal of Mass Spectrometry*, 2010, 45(7): 703–714. DOI: 10.1002/jms.1777.

[75] KOPKA J, SCHAUER N, KRUEGER S, BIRKEMEYER C, USADEL B, BERGMÜLLER E, DÖRMANN P, WECKWERTH W, GIBON Y, STITT M, WILLMITZER L, FERNIE A R, .STEINHAUSER D. GMD@CSB.DB: the Golm Metabolome Database [J]. *Bioinformatics*, 2005, 21(8): 1635–1638. DOI: 10.1093/bioinformatics/bti236.

[76] SAWADA Y, NAKABAYASHI R, YAMADA Y, SUZUKI M, SATO M, SAKATA A, AKIYAMA K, SAKURAI T, MATSUDA F, AOKI T, HIRAI M Y, SAITO K. RIKEN tan dem mass spectral database (ReSpect) for phytochemicals: a plant–specific MS/MS–based data resource and database [J]. *Phytochemistry*, 2012, 82: 38–45. DOI: 10.1016/j.phytochem.2012.07.007.

[77] COTTER D, MAER A, GUDA C, SAUNDERS B, SUBRAMANIAM S. LMPD: LIPID MAPS proteome database [J]. *Nucleic Acids Research*, 2006, 34: 507–510. DOI: 10.1093/nar/gkj122.

[78] KALE N S, HAUG K, CONESA P, JAYSEELAN K, MORENO P, ROCCA–SERRA P, NAINALA V C, SPICER R A, WILLIAMS M, LI X, SALEK R M, GRIFFIN J L, STEINBECK C. MetaboLights: An open–access database repository for metabolomics data. [J]. *Current Protocols in Bioinformatics*, 2016, 53: 14.13.1–14.13.18. DOI: 10.1002/0471250953.bi1413s53.

[79] KIM S, CHEN J, CHENG T, GINDULYTE A, HE J, HE S, LI Q, SHOEMAKER B A, THIESSEN P A, YU B, ZASLAVSKY L, ZHANG J, BOLTON E E. PubChem in 2021: New data content and improved web interfaces [J]. *Nucleic Acids Researchearch*, 2021, 49(D1): D1388–D1395. DOI: 10.1093/nar/gkaa971.

[80] https://www.mzcloud.org/

[81] KIND T, WOHLGEMUTH G, LEE D Y, LU Y, PALAZOGLU M, SHAHBAZ S, FIEHN O. FiehnLib: Mass spectral and retention index libraries for metabolomics based on quadrupole and time–of–flight gas chromatography/mass spectrometry [J]. *Analytical Chemistry*, 2009, 81(24): 10038–10048. DOI: 10.1021/ac9019522.

[82] https://mona.fiehnlab.ucdavis.edu/

[83] ZHOU Z, SHEN X, CHEN X, TU J, XIONG X, ZHU Z J. LipidIMMS Analyzer: Integrating multi–dimensional information to support lipid identification in ion mobility–mass spectrometry based lipidomics [J]. *Bioinformatics*, 2019, 35(4): 698–700. DOI: 10.1093/bioinformatics/bty661.

[84] OGATA H, GOTO S, SATO K, FUJIBUCHI W, BONO H, KANEHISA M. KEGG: Kyoto encyclopedia of genes and genomes [J]. *Nucleic Acids Research*, 1999, 27(1): 29–34. DOI: 10.1093/nar/27.1.29.

[85] KARP P D, RILEY M, PALEY S M, PELLEGRINI–TOOLE A. The MetaCyc Database [J]. *Nucleic Acids Research*, 2002, 30(1): 59–61. DOI: 10.1093/nar/30.1.59.

[86] MARTENS M, AMMAR A, RIUTTA A, WAAGMEESTER A, SLENTER D N, HANSPERS K, MILLER R A, DIGLES D, LOPES E N, EHRHART F, DUPUIS L J, WINCKERS L A, COORT S L, WILLIGHAGEN E L, EVELO C T, PICO A R, KUTMON M. WikiPathways: Connecting communities [J]. *Nucleic Acids Research*, 2021, 49(D1): D613–D621. DOI: 10.1093/nar/gkaa1024.

[87] WATROUS J, ROACH P, ALEXANDROV T, HEATH B S, YANG J Y, KERSTEN R D, VAN DER VOORT M, POGLIANO K, GROSS H, RAAIJMAKERS J M, MOORE B S, LASKIN J, BANDEIRA N, DORRESTEIN P C. Mass spectral molecular networking of living microbial colonies [J]. *PNAS*, 2012, 109(26): E1743–E1752. DOI: 10.1073/pnas.1203689109.

[88] ALLARD P M, PÉRESSE T, BISSON J, GINDRO K, MARCOURT L, PHAM V C, ROUSSI F, LITAUDON M, WOLFENDER J L. Integration of molecular networking and in–silico MS/MS fragmentation for natural products dereplication [J]. *Analytical Chemistry*, 2016, 88(6): 3317–3323. DOI: 10.1021/acs.analchem.5b04804.

[89] DA SILVA R R, WANG M, NOTHIAS L F, VAN DER HOOFT J, CARABALLO–RODRÍGUEZ A M, FOX E, BALUNAS M J, KLASSEN J L, LOPES N P, DORRESTEIN P C. Propagating annotations of molecular networks using in silico fragmentation [J]. *PLoS Computational Biology*, 2018, 14(4): e1006089. DOI: 10.1371/journal.pcbi.1006089.

[90] MOHIMANI H, GUREVICH A, MIKHEENKO A, GARG N, NOTHIAS L F, NINOMIYA A, TAKADA K, DORRESTEIN P C, PEVZNER P A. Dereplication of peptidic natural products through database search of mass spectra [J]. *Nature Chemical Biology*, 2017, 13(1): 1–10. DOI: 10.1038/nchembio.2219.

[91] MOHIMANI H, GUREVICH A, SHLEMOV A, MIKHEENKO A, KOROBEYNIKOV A, CAO L, SHCHERBIN E, NOTHIAS L F, DORRESTEIN P C, PEVZNER P A. Dereplication of microbial metabolites through database search of mass spectra [J]. *Nature Communications*, 2018, 9(1): 4035. DOI: 10.1038/s41467–018–06082–8.

[92] VAN DER HOOFT J J, WANDY J, BARRETT M P, BURGESS K E, ROGERS S. Topic modeling for untargeted substructure exploration in metabolomics [J]. *PNAS*, 2016, 113(48): 13738–13743. DOI: 10.1073/pnas.1608041113.

[93] ERNST M, KANG K B, CARABALLO–RODRÍGUEZ A M, NOTHIAS L F, WANDY J, CHEN C, WANG M, ROGERS S, MEDEMA M H, DORRESTEIN P C, VAN DER HOOFT J. MolNetEnhancer: Enhanced molecular networks by integrating metabolome mining and annotation tools [J]. Metabolites, 2019, 9(7): 144. DOI: 10.3390/metabo9070144.

[94] DÜHRKOP K, FLEISCHAUER M, LUDWIG M, AKSENOV A A, MELNIK A V, MEUSEL M, DORRESTEIN P C, ROUSU J, BÖCKER S. SIRIUS 4: A rapid tool for turning tandem mass spectra into metabolite structure information [J]. *Nature Methods*, 2019, 16(4): 299–302. DOI: 10.1038/s41592–019–0344–8.

[95] LUDWIG M, NOTHIAS L F, DÜHRKOP K, KOESTER I, FLEISCHAUER M, HOFFMANN M A, PETRAS D, VARGAS F, MORSY M, ALUWIHARE L, DORRESTEIN P C, BÖCKERD S. Database–independent molecular formula annotation using Gibbs sampling through ZODIAC [J]. *Nature Machine Intelligence*, 2020, 2(10): 629–641. DOI: 10.1038/s42256–020–00234–6.

[96] SHEN X, WANG R, XIONG X, YIN Y, CAI Y, MA Z, LIU N, ZHU Z J. Metabolic reaction network–based recursive metabolite annotation for

untargeted metabolomics [J]. *Nature Communications*, 2019, 10(1): 1516. DOI: 10.1038/s41467- 019-09550-x.

[97] BEAUXIS Y, GENTA-JOUVE G. Metwork: A web server for natural products anticipation [J]. *Bioinformatics*, 2019, 35(10): 1795-1796. DOI: 10.1093/bioinformatics/bty864

[98] NOTHIAS L F, PETRAS D, SCHMID R, DÜHRKOP K, RAINER J, SARVEPALLI A, PROTSYUK I, ERNST M, TSUGAWA H, FLEISCHAUER M, AICHELER F, AKSENOV A A, ALKA O, ALLARD P M, BARSCH A, CACHET X, CARABALLO-RODRIGUEZ A M, DA SILVA R R, DANG T, GARG N, DORRESTEIN P C. Feature-based molecular networking in the GNPS analysis environment [J]. *Nature Methods*, 2020, 17(9): 905-908. DOI: 10.1038/s41592-020-0933-6.

[99] TRIPATHI A, VÁZQUEZ-BAEZA Y, GAUGLITZ J M, WANG M, DÜHRKOP K, NOTHIAS-ESPOSITO M, ACHARYA D D, ERNST M, VAN DER HOOFT J, ZHU Q, MCDONALD D, BREJNROD A, D, GONZALEZ A, HANDELSMAN J, FLEISCHAUER M, LUDWIG M, BÖCKER S, NOTHIAS L F, KNIGHT R, DORRESTEIN P C. Chemically informed analyses of metabolomics mass spectrometry data with Qemistree [J]. *Nature Chemical Biology*, 2021, 17(2): 146-151. DOI: 10.1038/s41589-020-00677-3.

[100] CHEN L, LU W, WANG L, XING X, CHEN Z, TENG X, ZENG X, MUSCARELLA A D, SHEN Y, COWAN A, MCREYNOLDS M R, KENNEDY B J, LATO A M, CAMPAGNA S R, SINGH M, RABINOWITZ J D. Metabolite discovery through global annotation of untargeted metabolomics data [J]. *Nature Methods*, 2021, 18(11): 1377-1385. DOI: 10.1038/s41592-021-01303-3.

[101] SCHMID R, PETRAS D, NOTHIAS L F, WANG M, ARON A T, JAGELS A, TSUGAWA H, RAINER J, GARCIA-ALOY M, DÜHRKOP K, KORF A, PLUSKAL T, KAMENÍK Z, JARMUSCH A K, CARABALLO-RODRÍGUEZ A, M, WELDON K C, NOTHIAS-ESPOSITO M, AKSENOV A A, BAUERMEISTER A, ALBARRACIN ORIO A, ... DORRESTEIN P C. Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment [J]. *Nature Communications*, 2021, 12(1): 3832. DOI: 10.1038/s41467-021-23953-9.

[102] SENAN O, AGUILAR-MOGAS A, NAVARRO M, CAPELLADES J, NOON L, BURKS D, YANES O, GUIMERÀ R, .SALES-PARDO M. CliqueMS: A computational tool for annotating in-source metabolite ions from LC-MS untargeted metabolomics data based on a coelution similarity network [J]. *Bioinformatics*, 2019, 35(20): 4089-4097. DOI: 10.1093/bioinformatics/btz207.

[103] ZHOU Z, LOU M, ZHANG H, YIN Y, CAI Y, ZHU Z J. Metabolite annotation from knowns to unknowns through knowledge-guided multi-layer metabolic network [J]. *Nature Communications*, 2022, 13(1): 6656. DOI: 10.1101/2022.06.02. 494523.

[104] BITTREMIEUX W, LAUKENS K, NOBLE W S, DORRESTEIN P C. DORRESTEIN. Large-scale tandem mass spectrum clustering using fast nearest neighbor searching [J]. *Rapid Communications in Mass Spectrometry*, 2021, e9153. DOI: 10.1002/rcm.9153.

[105] HUBER F, VAN DER BURG S, VAN DER HOOFT J, RIDDER L. MS2DeepScore: A novel deep learning similarity measure to compare tandem mass spectra [J]. *Journal of Cheminformatics*, 2021, 13(1): 84. DOI: 10.1186/s13321-021-00558-4.

[106] GARCÍA C A, GIL-DE-LA-FUENTE A, BARBAS C, OTERO A. Probabilistic metabolite annotation using retention time prediction and meta-learned projections [J]. *Journal of Cheminformatics*, 2022, 14(1): 33. DOI: 10.1186/s13321-022-00613-8.

[107] HOFFMANN M A, NOTHIAS L F, LUDWIG M, FLEISCHAUER M, GENTRY E C, WITTING M, DORRESTEIN P C, DÜHRKOP K, BÖCKER S. High-confidence structural annotation of metabolites absent from spectral libraries [J]. *Nature Biotechnology*, 2022, 40(3): 411-421. DOI: 10.1038/s41587-021-01045-9.

[108] SHEN X, WU S, LIANG L, CHEN S, CONTREPOIS K, ZHU Z J, SNYDER M. metID: An R package for automatable compound annotation for LC2MS-based data [J]. *Bioinformatics*, 2021, 38(2), 568-569. DOI: 10.1093/bioinformatics/btab583.

[109] SHEN X, YAN H, WANG C, GAO P, JOHNSON C H, SNYDER M P. TidyMass an object-oriented reproducible analysis framework for LC-MS data [J]. *Nature Communications*, 2022, 13(1): 4365. DOI: 10.1038/s41467-022-32155-w.

[110] YU M, DOLIOS G, PETRICK L. Reproducible untargeted metabolomics workfow for exhaustive MS2 data acquisition of MS1 features [J]. *Journal of Cheminformatics*, 2022, 14(1): 6. DOI: 10.1186/s13321-022-00586-8.

（责任编辑　邹移光）

黄文洁，硕士，助理研究员。研究方向包括代谢组学技术研发与创新应用、作物品质鉴评与农产品品质识别新技术研发。作为核心成员先后承担了国家自然科学基金－广东省联合基金子课题、国家转基因重大专项调增课题任务、国家自然科学基金－青年基金、广东省自然科学基金等科研项目共13项，共发表学术论文42篇。近5年，以第一作者（含并列）在《Food Chemistry》《Food Research International》《Food Chemistry X》《Journal of Chromatography B》等领域内重要期刊上发表论文7篇；获授权发明专利5项；获广东省农业技术推广奖1项。

晏石娟，博士，研究员，硕士研究生导师。现任广东省农业科学院农业生物基因研究中心副主任，作物品质控制与多组学技术创新团队的负责人。德国马普植物分子生理研究所和康奈尔大学生物技术研究所访问学者。先后获广东省特支计划科技创新青年拔尖人才、广东省农科院“青年研究员”“金颖之星”等人才称号。主要从事代谢组与蛋白质组学技术研发、作物品质性状形成机理与调控研究。主持承担国家自然科学基金、省重点研发计划等科研项目18项；研究成果发表学术论文共84篇，其中以第一作者或通讯作者（含并列）在《Plant Cell》《Trends in Plant Science》《Green Chemistry》等领域内重要期刊发表论文41篇；获授权专利12项。参编著作1部。