

浙江中醫藥大學

碩士學位論文

**MCnebula：基于非靶向 LC- MS/MS 技术的化学聚类可视化分
析策略快速解析复杂（中药）化学成分**

**MCnebula: Chemical clustering visualization strategy based on
non-targeted LC- MS/MS technology for rapid analysis of
complex (Traditional Chinese Medicine) chemical components
analysis**

学科专业 中藥學 中藥炮制學

学位类别 学術型 专业学位型

目 录

中文摘要	6
ABSTRACT	7
前言	9
第一部分 MCnebula的方法构建.....	12
一、材料与方法.....	12
(一) 实验材料.....	12
(二) 实验方法.....	12
1. R 的配置	12
二、结果	13
(一) MCnebula R包概览	13
1. 设计理念	13
2. 数据流	15
(二) MCnebula的算法	18
1. 整体考虑	18
2. 化学结构式和分子式	18
3. 根据最佳候选项确立Reference	18
4. 化学分类学	19
5. ABC选择算法	19
6. Cross filter stardust Classes的细节	22
(三) 数据结构	23
1. 首要Class: ‘mcnebula’ 的结构	23
2. 数据相关Class的结构	24
3. 可视化相关Class的结构	24
4. 其他Class	25
(四) 方法 (Method) 和函数 (Function)	26
1. 数据方法	26

2. 可视化方法.....	26
3. MCnebula辅助科学绘图的函数.....	27
4. 其他方法和函数.....	28
(五) MCnebula的基本使用.....	29
三、小结	29
第二部分 MCnebula的方法评估与拓展.....	30
一、材料与方法.....	30
(一) 实验材料.....	30
(二) 实验方法.....	31
1. R 的配置	31
2. 建立评估数据集	31
3. 评估的方法	32
4. 用于建立评估数据集和用于评估的R的函数	33
5. MCnebula的拓展涉及的算法.....	34
二、结果	36
(一) MCnebula的评估.....	36
1. 功能评估.....	36
2. 归类准确度评估	38
3. 鉴定准确度评估.....	41
4. 评估的报告和R代码	41
(二) MCnebula的拓展.....	42
1. 用于化学发现.....	43
2. 用于代谢组数据分析	44
三、小结	45
第三部分 基于MCnebula策略分析杜仲炮制前后的成分变化.....	45
一、材料与方法.....	45
(一) 实验材料.....	45

(二) 实验方法.....	45
1. 制备炮制前后的杜仲	45
2. 制备LC-MS的杜仲样品.....	46
3. LC-MS/MS实验条件	46
二、结果	46
(一) MCnebula对中药数据集的基础分析.....	46
(二) MCnebula对中药数据集的聚焦分析.....	53
(三) 分析的报告和R代码.....	69
三、小结	70
第四部分 基于MCnebula策略的血清代谢组学.....	70
一、材料与方法	70
(一) 实验材料.....	70
(二) 实验方法.....	71
二、结果	71
(一) MCnebula对血清数据集的整体分析.....	71
(二) MCnebula对血清数据集的聚焦分析.....	75
(三) 分析的报告和R代码.....	96
三、小结	97
结论	97
创新点	99
参考文献	100
文献综述	106

附：图目录

图1 摘要图示.....	8
图2 MCnebula数据流.....	16
图3 MCnebula过滤化学类的机制.....	21
图4 设计矩阵和对比矩阵的示例：Expected gene expression is modelled by a treatment factor.....	35
图5 设计矩阵和对比矩阵的示例：Expected gene expression is modelled by a group factor.....	36
图6 MCnebula归类的准确度评估.....	39
图7 MCnebula归类准确度的单独评估.....	41
图8 GNPS归类准确度的单独评估	42
图9 评估报告的概览.....	43
图10 在中药数据集的Child-Nebulae中追踪Top ‘Features’	48
图11 中药数据集Top ‘Features’ 的MS/MS图	49
图12 中药数据集Top ‘Features’ 的EIC图.....	51
图13 在中药数据集聚焦于Child-Nebulae中的Top ’Features’的方位	55
图14 杜仲数据集分析报告的概览.....	69
图15 血清数据集的Parent-Nebula.....	73
图16 血清数据集的Child-Nebulae	74
图17 在血清数据集的Child-Nebulae中追踪Top ‘Features’	75
图18 在血清数据集的Child-Nebulae中可视化组间Log2(Fold Change)	76
图19 血清数据集的代表ACs的Child-Nebula的深度可视化.....	78
图20 血清数据集中代表LPCs和BAs的Child-Nebulae的深度可视化.....	80
图21 血清数据集ACs、 LPCs和BAs的热图分析.....	81
图22 血清数据集ACs、 LPCs和BAs的通路富集分析.....	82
图23 血清代谢组分析报告的概览.....	97

附：表目录

表 1 MCnebula涉及的R包及其作用	12
表 2 Class: 'mcnebula'的结构	23
表 3 数据相关Class的结构	24
表 4 可视化相关Class的结构	24
表 5 MCnebula2其他Class的结构	25
表 6 MCnebula主要的数据方法	26
表 7 MCnebula主要的可视化方法	26
表 8 辅助科学绘图的函数	27
表 9 MCnebula的其他方法或函数	28
表 10 测试和拓展MCnebula用到的R包	31
表 11 评估MCnebula涉及的R函数	33
表 12 MCnebula和其他工具的功能比较	37
表 13 拓展MCnebula用于化学发现的函数	44
表 14 拓展MCnebula用于代谢组学分析的函数	45
表 15 MCnebula工作流鉴定的中药数据集的化合物（Q-value < 0.05）	56
表 16 MCnebula重新分析血清数据集Wozniak等人的Top Metabolites	84
表 17 MCnebula工作流程鉴定的血清数据集的化合物（Q-value < 0.05）	89

中文摘要

目的: 建立适应当前人工智能技术于质谱领域发展的便捷LC-MS/MS分析技术。

方法: 结合尖端的SIRIUS系列预测技术、分子网络可视化、统计筛选、R语言面向对象编程等技术，开发用于LC-MS/MS分析的工作流。

结果: 名为MCnebula (Multiple-Chemical Nebula) 的框架被建立，通过聚焦于关键化学类别和多维度的可视化，以促进质谱数据分析过程。它由三个重要步骤组成：(1) 基于丰度的化学类 (ABC, abundance-based classes) 选择算法；(2) 根据关键化学类别对 ‘Feature’ (化合物) 进行归类；(3) 可视化为多个Child-Nebulae (网络图) 并注释以化学分类和结构等。MCnebula可以应用于探索超出参考光谱库限制的未知化合物的分类和结构特征。此外，由于它具有ABC选择和可视化的功能，它对于通路分析和生物标志物的探索是直观和便捷的。MCnebula是以R语言实现的。在R语言包中提供了一系列工具，以促进MCnebula功能的下游分析，包括 ‘Features’ 筛选 (Feature selection) (主要为二元比较的统计分析)、Top ‘Features’ 的同类追踪、通路富集分析、热图聚类分析、光谱可视化分析、化学信息查询和输出分析报告等。为了说明MCnebula的广泛用途，我们分析了人类血清代谢组学数据集。结果表明，通过追踪潜在生物标志物于Child-Nebulae，‘Acyl carnitines’ 被筛选出来，这与此前的报道是一致的。我们还研究了一个植物来源的数据集，即杜仲 (*E. ulmoides*)，实现了快速的未知化合物注释和发现。

结论: MCnebula工作流功能广泛，适应于复杂的代谢组学数据分析和植物药数据分析。

主题词: 质谱，可视化，化学类，鉴定，MCnebula

ABSTRACT

Objective: Establishing a convenient LC-MS/MS analysis technique adapted to the current development of artificial intelligence technology in the field of mass spectrometry.

Methods Combine cutting-edge predictive technologies of softwares of SIRIUS, visualization of molecular network, statistical screening, and object-oriented programming in R language to develop workflows for LC-MS/MS analysis.

Results: We established a framework called MCnebula (Multiple-Chemical nebula) to facilitate mass spectrometry data analysis process by focusing on critical chemical classes and visualization in multiple dimensions. It consisted of three vital steps: (1) abundance-based classes (ABC) selection algorithm, (2) critical chemical classes to classify' features' (compounds), (3) visualization as multiple Child-Nebulae (network graph) with annotation, chemical classification and structure. Notably, MCnebula can be applied to explore classification and structural characteristic of unknown compounds that beyond the limit of spectral library. What's more, it is intuitive and convenient for pathway analysis and biomarker discovery due to its function of ABC selection and visualization. MCnebula was implemented in the R language. We provided a series of tools in the R packages to facilitate downstream analysis in a MCnebula-featured way, including feature selection (statistical analysis of binary comparisons), homology tracing of top features, pathway enrichment analysis, heat map clustering analysis, spectral visualization analysis, chemical information query and output analysis reports, etc. In order to illustrate the broad utility of MCnebula, we investigated a human-derived serum dataset for metabolomics analysis. The results indicated that' Acyl carnitines' were screened out by tracing structural classes of biomarkers which was consistent with the reference. We also investigated a plant-derived dataset of herbal *E. ulmoides* to achieve a rapid unknown compound annotation and discovery.

Conclusion: MCnebula workflows are broadly powerful and adaptable to complex metabolomics data analysis and phytopharmaceutical data analysis.

Keywords: Mass spectrometry, visualization, chemical classes, identification, MCnebula

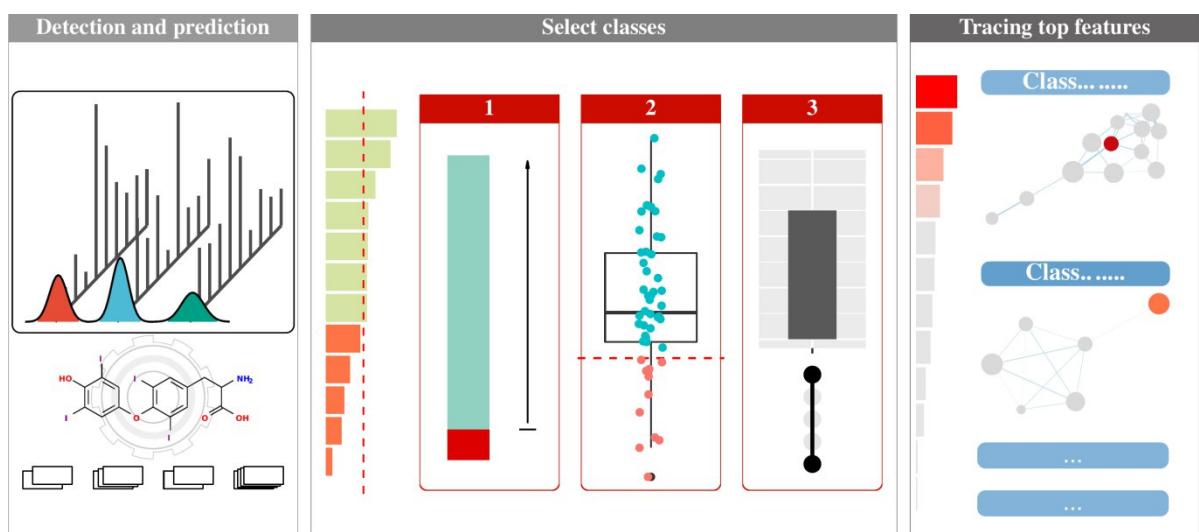


图1 摘要图示

前言

分析非靶向LC-MS/MS数据集是复杂的，由于数据量大、光谱复杂、化合物结构多样。在过去的几十年中，许多研究人员试图解决这些问题。许多技术软件或基于网络的界面被开发出来，为数据分析提供一站式的批量解决方案^[4,3,2,1]。这些解决方案应用或建议了灵活的质谱处理工具或类似的算法^[8,7,6,5]。为了减少假阳性和假阴性的结果，更多的算法已经实现了去卷积（Deconvolution）、‘Features’筛选（Feature selection）或统计过滤^[12,11,10,9]。与样品或平行样品中的化合物相对应的每一个‘Feature’都具备了用于鉴定的碎片光谱。在这种情况下，研究人员不得不面对一个问题：如何准确和快速地鉴定繁多的化合物？

直到今天，研究者们已经开发了几种策略来鉴定具有碎片光谱的化合物。**1)** 参考光谱库的匹配。一些公共可用的数据库是通过实现参考光谱的可重复使用而建立的，如MassBank、MassBank of North America（MoNA）、Global Natural Products Society（GNPS）^[4]。同时，这些碎片光谱可通过其网络服务器、第三方平台（如[CompMass](#)）或特定工具（MASST）^[13]获得。然而，与结构数据库（PubChem有超过1亿条记录）相比，光谱库的规模太小，限制了质谱的应用。为了跨越这一障碍，**2)** 通过计算机模拟碎片光谱。越来越多的研究者开发了计算机工具来模拟碎片光谱^[17,16,15,14]。一些数据库，如MoNA整理了模拟的光谱，并公开使用^[18]。**3)** 通过机器学习进行预测。这类算法从参考光谱数据集或光谱库中训练，然后“学习”如何预测化学指纹或原理，以便从结构数据库中检索出正确的结构^[21,20,19]。

计算机方法正在迅速发展。到目前为止，被称为SIRIUS^[22]的前沿技术，集成了许多先进的人工智能算法，据报道，在检索结构库时，其准确率达到了70%。这种方法有助于鉴定光谱库范围之外的代谢物。虽然计算机工具促进了化学鉴定，但仍然缺乏一个适当的框架，可以将SIRIUS纳入到用户友好的生物研究中，推进生物标志物的发现和质谱数据集的通路分析。人工注解化合物和筛选生物标志物相当耗时，而且结果受到主观因素的影响。提及用户友好的分析工具，分子网络由于其可视化和数据透明而越来越受欢迎。分子网络是一种基于光谱相关的可视化的方法，可以检测来自相似分子的光谱（所谓的光谱网络），即使这些光谱与任何已知化合物不匹配^[4]。基于分子网络的概念，我们提出了一个想法，基于化学分类的可视化的聚类可能有助于生物标志物的发现和代谢通路的分析。

化学分类的历史可以追溯到上个世纪中叶。1963年，德温特世界专利索引（Derwent World Patent Index, DWPI）首次提出了化学片段编码系统。直到近年来，像基因本体（GO）^[23]这样用分类学和本体论的化学分类被更系统地提出^[24]。ClassyFire由于其可计算性和系统性，在LC-MS数据集分析的化合物注释方面颇受欢迎^[28,27,26,25]。该分类法和本体论是强大的，对化学大有裨益。例如，有研究者提出了一种基于层次分类的方法，称为Qemistree，通过将分子关系表达为一“层级树”来分析质谱数据，这可以在样品元数据和化学本体的背景下展现同一性和特殊性^[29]。

非靶向代谢组学是组学科学的一个领域，它利用尖端的分析化学技术和先进的计算方法，全覆盖式描述复杂的生化混合物。基于LC-MS的非靶向代谢组学由于其高灵敏度、小样本量和无需分离直接进样等特点而大受欢迎^[30]。在统计学方法的帮助下，研究人员可以从数以千计的LC-MS' Features' 中筛选和确定更多信息的疾病生物标志物，以帮助设计或开发改进治疗方法，并更好地评估康复结果^[31]。这些统计方法主要涉及经典统计学和人工智能模型（如随机森林）^[32]。由于‘Features’集的复杂性或算法的稳定性，这两种方法都不可避免地导致特定的偏差^[33]。此外，在‘Features’层面的分析无法无偏差地剖析代谢物的系统效应^[34]。在这种观点下，在化学分类水平上进行分析可能是一个更好的解决方法。然而，不应该忽视同一分类层次上的代谢物的差异。例如，属于‘吲哚和衍生物’（Indoles and derivatives）的小分子对芳烃受体（AHR）有结构上的差异影响^[35]，不同的结构特征将导致不同的活动。解决这个问题的办法是将‘Feature’层面的统计和化学类层面的评估结合起来。

除了化学分类和统计分析外，聚类可视化也是一种流行的非靶向质谱数据分析工具。在过去的十年中，全球天然产物社会分子网络（GNPS MN）不断发展推进了这类方法。GNPS应用分子网络将分子的质谱基于其碎片模式的相似性连接起来^[36]。遗憾的是，GNPS的分子网络主要依赖于光谱的相似性，而不是化合物结构或分类的相似性。例如，黄酮类化合物由一个芳香环和一个与苯基相连的含氧杂环组成，由于其特殊的类别和结构的相似性，预计它们会被聚集在一起；然而，据报道，在以前的研究中，一些属于黄酮类化合物的化合物恰好不在其他黄酮类化合物的集群中^[34]。因此，在分类层面上的聚类可视化是非靶向质谱数据集的更好选择。早在2012年，首次提出了质量数据分析的可视化分子网络概念^[36]，但当时还没有通过碎片谱图预测化合物分类的计算机工具。如今，随着计算机智能分类工具的发展^[24]，有必要对可视化策略进行革新，以提高归类水平的置信度。

出于上述考虑，我们提出了一个综合框架，名为MCnebula，用于非目标LC-MS/MS数据集分析。MCnebula集成了一种新的基于丰度的类别（ABC）选择算法，用于选择化学类别。ABC选择算法的原理是(1)根据预测的概率对数千个化学类别进行初步过滤，(2)将所有的‘Features’视为一个整体，检查每个化学类别的‘Features’的数量和丰度（不同层次的分类，子结构和主导结构的分类），然后选择有代表性的类别，(3)对这些化学类别进行优度评估（关于其化合物的鉴定优度）和一致性评估（这些化学类别在MS/MS谱图中可以相互区别的程度）。最终的化学类别将用于后续的分析：可视化的Child-Nebulae，并聚焦这些化学类别/Nebulae的生物标志物或化学发现。基于统计分析的高排名‘Features’可以被设定为追踪器（Tracer），以发现更多的化学结构、光谱相似性或同化学类别的化合物。得益于SIRIUS软件的注释模块和前沿技术^[39,38,37,34,22,20]，超越了光谱库匹配的限制，MCnebula可以用于探索未知化合物。MCnebula（更新为MCnebula2，涵盖更多的工具，如ABC选择算法、Nebula可视化、统计分析和输出报告等）主要以R的面向对象编程的S4系统编写，它允许所有数据在一个对象（object）中从头到尾进行流水式分析，方便了数据处理。除了MCnebula的基本功能外，我们还提供了一个额外的‘exMCnebula2’包用于下游分析，其中包含了本研究中使用的所有分析工具，如通路富集分析、热图聚类分析、光谱可视化分析、化学信息查询等。非靶向LC/MS-MS的下

游分析很复杂，并且因数据不同而不同。`exMCnebula2`包中的额外工具可以为MCnebula的扩展应用提供一个范例。

在这项研究中，我们以MCnebula分析了两个数据集，以证明方法的广泛适用：一个是源于人类的血清数据集，与金黄色葡萄球菌菌血症（SaB）的死亡风险分析相关；另一个是源于植物的草药数据集，与中药的炮制（杜仲的炮制前后）有关。在过去，代谢组学的分析方法（非靶向LC-MS/MS）借助于集成的工具或者单功能的工具应用：‘Features’检测，统计分析（PCA, OPLS-DA等），光谱匹配，以机器预测无法以光谱匹配鉴定的化合物，或者可视化为分子网络来探索结构相似的化合物；这些方法或多或少存在上述讨论的技术发展中的局限性，或者带来不够全面的分析视角（见表12）。中药炮制的非靶向LC-MS/MS承袭于代谢组学的分析^[40]，然而面临的问题是更少可用的匹配光谱。MCnebula方法的建立将会为代谢组学分析和包括中药在内的植物药的分析鉴定带来全新的视角。

第一部分 MCnebula的方法构建

一、材料与方法

(一) 实验材料

个人笔记本电脑Surface pro7用于编程环境的搭建和R语言编程：Pop!_OS (Ubuntu) 22.04 LTS 64-bits PC (Intel Core i7-1065G7, 1.3 GHz × 8, 16 Gb of RAM)。

(二) 实验方法

所有语言和脚本编写均在配置好的VIM (version 8.2) 中进行，运行和测试在R (version 4.2.1) 中进行。必要时辅助以Bash (shell语言)。（VIM的配置：https://github.com/Cao-lab-zcmu/bash_and_vim）

1. R 的配置

以下表格为R语言编程中涉及的包以及其作用说明：

表 1 MCnebula涉及的R包及其作用

Type	Name	Description
DEPENDS	ggplot2	用于可视化的主要R包
IMPORTS	BiocStyle	为报告输出样式提供额外选择
	bookdown	为输出报告的提供样式的包
	ChemmineOB	用于化学结构可视化
	crayon	R命令行标准输出的美化
	data.table	快速读取数据
	dplyr	操纵数据框 (data.frame)
	ggimage	为子视图 (grob) 映射到主视图提供便利 (用于ggplot2)
	ggraph	ggplot2用于网络图的拓展工具
	ggsci	提供科研杂志社偏爱的调色板
	ggtext	强化ggplot2的文本可视化
	grid	用于ggplot2无能为力的复杂做图，例如可视化Child-Nebulae
	gridExtra	主要提供'arrangeGrob'函数，调整'grob'
	grImport2	提供'readPicture'和'grobify'函数，将cairosvg矢

Type	Name	Description
		量图和'grid'包绘图结合在一起
	igraph	网络数据格式的基础R包，构建基本数据格式，可以用于输出Cytoscape等软件支持的数据格式
	knitr	输出报告依赖的包
	methods	提供S4面向对象编程系统
	pbapply	可视化运行进度的包
	rlang	提供'call'、'name'、'expression'类对象和字符对象转化的工具
	rmarkdown	输出报告依赖的包
	rsvg	提供'rsvg_svg'函数将svg转化为cairosvg
	stringr	提供字符串处理的工具
	styler	将'rblock'对象中的代码格式美化
	svglite	比'svg'更轻量的svg输出工具
	tibble	提供'tibble'数据框形式，更适用于数据量大的表格的透视
	tidyverse	操纵数据框
SUGGESTS	testthat	用于测试包
...

二、结果

(一) MCnebula R包概览

1. 设计理念

MCnebula的最初R版本是简单的函数式编程（<https://github.com/Cao-lab-zcmu/MCnebula>），但它存在不少缺陷，无法满足复杂的、具有多重注释数据的代谢组学数据分析，它兼容性不够强的数据结构使得使用起来颇为困难（需要输入太多的参数）。MCnebula2 R包的设计上需要解决上个版本存在的各种问题，它需要具有以下特点：

- 简明的数据对象。它以面向对象式的编程方式实现，所有的数据存储于一个对象中（即‘mcnebula’），数据的中间处理和最终形成过程都在这个对象内部进行迭代，用户不需要知道它们是如何发生的，也不知道它们存储在哪个位置，只需要在运行结束后通过函数或者方法取得必要的数据。用户只需要掌握好这一个数据对象就能完成从头到尾的分析。

- 简单明了的参数输入。它只有最必要的参数输入需求，这些参数还需要带有默认值，降低用户的使用难度。面向对象编程具有参数化多态的特征，可以利用这一特性赋予MCnebula的方法以更强的包容性。用户可以缺省参数，缺省的参数可以以默认参数替代。默认的参数必须是易于获取的，研究者往往以它们为参考而调整进一步的分析。利用参数化多态，不带有任何参数的方法将输出该方法默认的参数的列表（例如直接输入‘cross_filter_stardust()’）。
- 安全的参数验证。R的函数式编程往往不具备参数验证的特性。R函数式编程的针对的往往是S3类对象，它们并不是严谨的对类的定义。MCnebula2新的编程形式需要对方法的输入参数的严谨验证（应用的是S4的类），防止错误的输入导致错误的输出。这是通过定义一系列的类来实现的（例如‘mcnebula’，‘nebula’，‘ggset’，‘command’等）。
- 不易出错的分析流程。MCnebula具有多个函数或方法参与的分析流程，该流程可能囊括十余个步骤，由于各个步骤的参数可定制性，它们是不能省略或相互合并的，但这将导致流程容易交叉出错，或者使用者无法理解其中先后顺序关系。因此新的编程形式需要带有数据验证环节，即对这一步需求的数据进行验证，如果缺少相应数据，将提示使用者进行上一步需要的步骤。在面向对象式的编程中，这样的验证是易于实现的，因为所有的数据都可以存储于一个数据对象中。
- 兼容性更强的结构特征。各个分析环节的分析处理要分配于不同的函数或方法中，它们内部不相交叉，只通过各个接口相互衔接。一方的错误不会导致另一方的错误，它们的错误也必须是易于排查的。MCnebula的算法依赖于SIRIUS的计算，为了获取SIRIUS的计算结果，加入API模块获取它的计算结果也是必不可少的。SIRIUS已经从版本4更迭到了版本5，它的数据存储方式发生了变动。MCnebula的数据获取模块必须具有强大的稳定性，不会因为SIRIUS的版本变化而导致全局崩溃，它是易于维护。SIRIUS对于输出数据的属性名称可能会发生改变，因此在MCnebula中，需要有自身的对于这些属性的另定义的ID名称，真正用于编程的必须是这些ID名称，而不是SIRIUS输出的名称，这样，当SIRIUS的数据结构或者属性名称发生变动时，仅仅需要更改最底层的接口，就能适应版本的变化，将维护成本降低到最小。
- 高度可定制的分析流程。MCnebula2的主要算法模块为：‘filter_structure’，‘create_reference’，‘filter_formula’，‘filter_ppcp’，‘create_stardust_classes’，‘cross_filter_stardust’，‘create_nebula_index’。‘filter_*’系列为对分子式、结构式或者化学类的候选项的筛选步骤，这些方法已经包含默认的排序和过滤方法，但它们也需要具备自定义的特征，因此，它们在设计中兼容了强大的‘dplyr’包的表达式过滤的特性，在输入参数（作用于多重属性的表达式）后对所有的Features数据集的候选项进行过滤，在最简洁优雅的方式下操作数据集。‘*stardust*’系列为可视化之前的关键步骤，它将筛选出关键的化学类用以可视化，为了缓解算法上的偏

颇，我们额外设计了‘backtrack_stardust’方法，用以对过滤掉的化学类进行追溯。这些设计使得分析流程可以灵活变幻。

- 高度可定制的数据可视化。MCnebula充分利用ggplot2包进行数据可视化。ggplot2这一明星级的可视化包具有精美、优雅、高度可定制的特性（多图层可视化是一大特点），MCnebula在设计可视化时，不应该丢失它的特性。我们设计了‘ggset’这一对象，用以将我们预设的可视化函数和参数包装（这些都是ggplot2的函数和参数）。「ggset」主要为‘layers’数据槽（Slot），实际上它存储的是预设的‘ggplot2’的各个图层。这样，MCnebula2的最终可视化是高度可定制的，通过操纵‘ggset’，经验丰富的‘ggplot2’使用者就像在编写‘ggplot2’进行绘图一样。
- 直观并且美观的数据可视化。R的ggplot2包提供了大量预设的基本的美观的图形或者主题，MCnebula2需要利用这些图形或者主题来进行可视化。此外，‘ggsci’包提供了各类顶级杂志社偏爱的色调，这些色调简洁明亮，可以丰富MCnebula2的可视化。为了充分利用这些可视化元素，MCnebula带有‘melody’对象，用于操纵可视化的调色板。
- 翔实丰富的使用说明文档。编写R包时，可以使用‘roxygen2’包进行注释，进而生成说明文档，可用R内部使用，也可形成单独的.pdf文档或其他文档（现在可获取于：
<https://github.com/Cao-lab-zcmu/MCnebula2/blob/document/reference.pdf>）。MCnebula2包为每一个用户级的方法（Method）和函数（Function）注释了使用说明，并带有示例数据和示例代码，并在必要的部分添加了细节上的算法说明，使得用户能够无障碍使用MCnebula2 R包。
- 高度可定制的输出报告。R带有一系列优秀的支持文档输出的工具包，如‘rmarkdown’，‘knitr’，‘pander’等。MCnebula在设计时，考虑了集成这些包的工具让分析流程以完整报告的形式输出。为了实现高度的自定义，我们设计存在基本工作流模块（相对固定）和自定义模块（结合其他R代码灵活部署），前者可用包装好的‘workflow’函数快速构建输出（获取代码，或直接运行直到生成报告），而后者则建立在前者的基础上增添‘section’。

2. 数据流

MCnebula工作流程致力于从头开始分析LC-MS/MS数据集，即从样品获得的原始数据开始，经过各个阶段的分析，得到一份完整的分析报告（图2）。分析过程遵循一般的MCnebula分析模板，从过滤候选化学式、结构式、化学类别，到创建可视化Nebula；它还允许自定义高级分析，在聚焦于化学类别的Child-Nebulae的帮助下，进行统计分析、‘Features’筛选（Feature selection）、聚焦关键代谢物（化合物）及其结构特征、通路富集、查询化合物同义名等（拓展功能请参考：第二部分 MCnebula的方法评估与拓展 > 二、结果 > （二）MCnebula的拓展）。

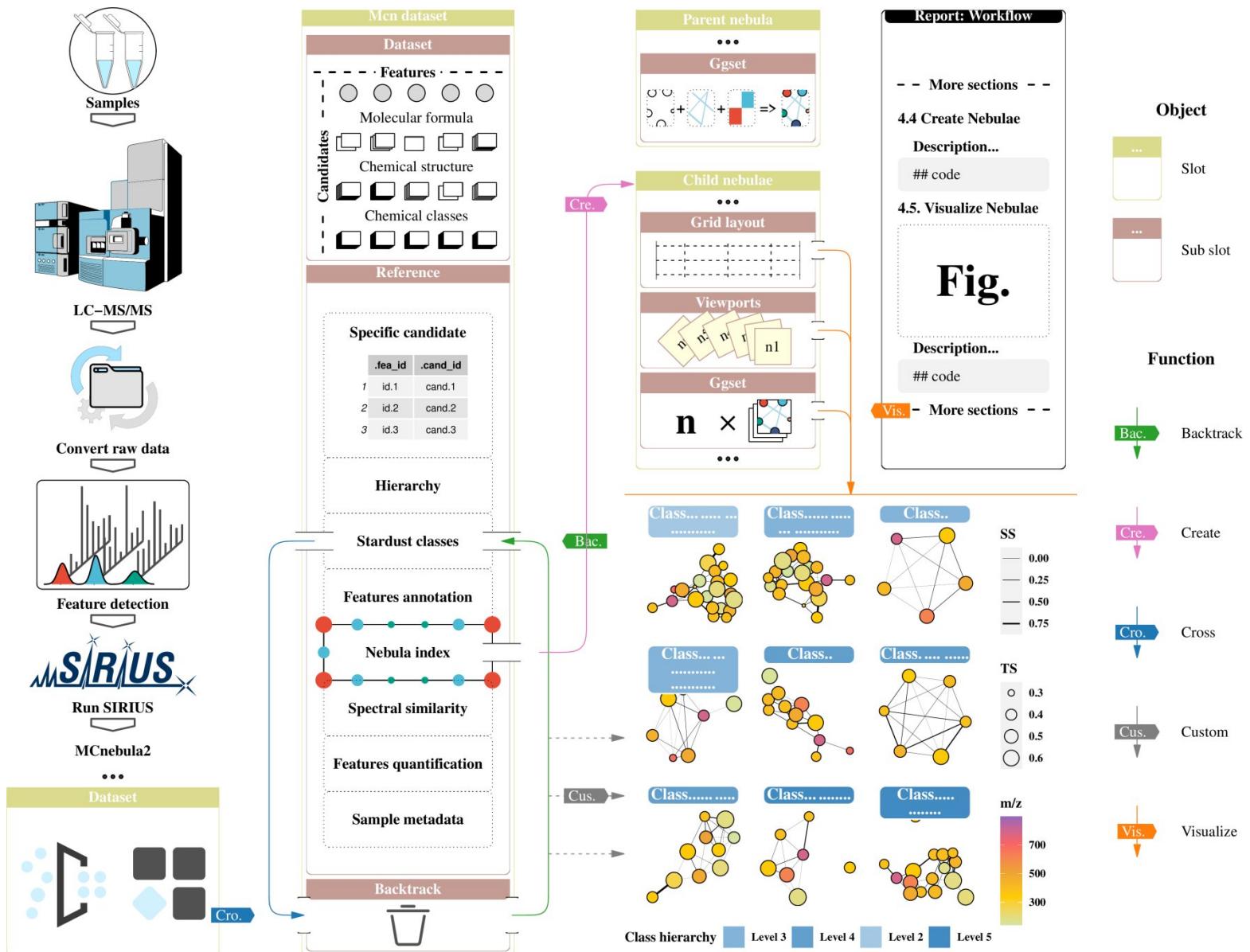


图2 MCnebula数据流

- 图2注：根据数据呈现的平台，MCnebula的工作流程可以分为两部分。第一部分是R以外的部分（在MCnebula2之前）：从Sample到LC-MS/MS，获得原始数据；Convert raw data，使用Proteowizard衍生的流行的MSconvert工具实现；对于Feature detection，用户可以用任何LC-MS处理工具实现，如MZmine、XCMS、OpenMS等；然后将.mgf或其他文件格式的MS/MS光谱导入SIRIUS进行计算。R内部的部分，MCnebula2实现了数据的整合，并在‘mcnebula’对象中创建Nebulae。

（二）MCnebula的算法

1. 整体考虑

非靶向LC-MS/MS数据集的分析一般从‘Features’检测（Feature detection）开始。‘Features’被识别为MS¹（MASS一级）数据中的‘峰’。每个‘Features’可能代表一个化合物，并以MS²（MASS级别2）光谱进行分配。然后用MS²光谱来鉴定化合物。困难主要在于对这些‘Features’进行注释以发现它们的化合物身份，并为进一步的生物研究挖掘出有益的信息。此外，非靶向的LC-MS/MS数据集通常是一个庞大的数据集，这导致整个过程的分析耗时。在此，我们采用了一种基于化学类的可视化方法，即MCnebula，来解决这些问题。

MCnebula R包本身不涉及分子式预测、结构预测和化学类的预测，因此不涉及这些部分的准确性。MCnebula通过提取SIRIUS项目的预测数据来实现下游分析。MCnebula的核心是化学类别过滤算法，也就是基于丰度的化学类（ABC）选择算法。为了详细解释ABC选择算法，我们需要从MS/MS光谱分析和鉴定化合物开始讨论。

2. 化学结构式和分子式

MS/MS谱的分析是一个推断和预测的过程。例如，我们根据分子量结合MS/MS谱的可能碎片模式来推测元素的组成，推测化合物的潜在分子式。最后，我们从化合物结构数据库中寻找确切的化合物。有时，这个过程充满了不确定性，因为有太多的因素可能影响到MS/MS数据的可靠性和推断的正确性。可以假设，在MS/MS光谱背后有复杂的潜在化学分子式、化学结构和化学类别的候选。假设我们现在有这些候选数据，MCnebula可以帮助提取这些候选数据，并根据化学结构预测的最高分获得每个MS/MS谱的唯一分子式和化学结构；在这个过程中，和大多数算法一样，我们可以根据分数对预测进行过滤。

3. 根据最佳候选项确立Reference

对以LC-MS/MS谱呈现的潜在化合物进行预测，并得到了化学分子式、结构和化学类别的候选结果（这些在SIRIUS中实现了）。这些候选化合物包括阳性和阴性结果：对于化学分子式和化学结构，阳性预测是唯一的；对于化学类别，涉及多个属于不同分类的阳性预测。无从得知确切的阳性或阴性

模式。通常情况下，我们通过得分对这些数据进行排序和过滤。有许多得分系统，例如根据同位素、质量误差、结构相似性、化学类别等等。选择哪种对候选项进行排名的分数系统取决于研究的目的。比如：

- 要找出化学结构大多是阳性的候选，通过结构得分对候选进行排名。
- 为了确定一个潜在的化合物是否属于某个化学类别，通过化学类的得分对候选化合物进行排名。

通过MCnebula中的‘filter_formula()’、‘filter_structure()’或‘filter_ppcp()’函数，可以得到得分最高的候选化合物。然而，对于三个模块（分子式、结构、化学类），有时它们的最高分候选并不一致，也就是说，他们的最高分是针对不同的化学分子式的。为了在其他模块中找到相应的数据，应该进行‘create_reference()’来建立‘Specific candidate’，作为后续数据整合的参考。我们通过得分和排名获得了唯一的化学分子式和化学结构式作为参考。但是对于化学类来说，这种方法是不够的。

4. 化学分类学

化学分类是一个复杂的系统。在这里，我们只讨论基于结构的化学分类系统^[24]，因为MS/MS谱比生物活性和其他信息更能说明化合物的结构。

根据化合物整体结构和局部结构的划分，我们可以把结构特征称为优势结构（主导结构，Dominant structure）和亚结构（Sub-structure）^[24]。相应地，在化学分类系统中，我们不仅可以根据优势结构对化合物进行分类，还可以根据亚结构进行分类。基于化合物的优势结构的化学分类很容易理解。例如，我们将把紫杉醇（Taxifolin）归入“黄酮类”（Flavones），而不是“酚类”（Phenols），尽管它的局部结构有一个“酚”的子结构。我们希望按照化合物的主要结构而不是子结构对其进行分类，因为这样的分类更简洁，包含的信息更多。但是，在MS/MS光谱分析过程中，我们有时只能根据化合物的亚结构进行化学分类，这可能是由于：结构分析过程中的不确定性；可能是未知化合物；MS/MS光谱片段信息不足。在这种情况下，我们有必要借助于子结构信息对化合物进行分类，否则我们对那些无法获得优势结构信息的化合物一无所知。

需要注意的是化学分类学的另一个方面的复杂性，即分类的层次性。例如，‘Flavones’属于其上级‘Flavonoids’；再上级‘Phynylpropanoids and polyketides’；进一步向上的分类是‘Organic compounds’。

5. ABC选择算法

在非靶向LC-MS/MS数据集中，每个‘Features’都有相应的MS/MS谱，总共可能有几千个‘Features’。ABC选择算法将所有‘Features’作为一个整体，考察属于各化学类别的‘Features’数量和丰

度（不同层次的分类、亚结构和优势结构的分类），然后选择有代表性的类（主要根据‘Features’的数量或丰度范围来筛选类），为后续分析奠定基础（图3）。

- 创建Stardust classes（Inner filter）。分类预测的后验概率（PPCP）数据归于每个‘Features’。在进行过滤时，只设置了简单的阈值条件或绝对条件来过滤化学类；不同属性之间没有交叉，‘Features’之间也没有交叉。因此，我们定义这是‘Inner filter’。
- 交叉过滤Stardust Classes（Cross filter）。化学类的数据和它们归属的‘Features’，即Stardust Classes，被结合起来，然后根据化学类进行分组。分组后，每个化学类都有一定数量的‘Features’。在过滤时，可以对组内的‘Features’数据进行统计；可以对这些数据与‘Features annotation’数据一起进行统计；还可以进行统计，将各组相互比较。由于其过滤的属性交叉，我们定义这是‘Cross filter’。

不管是全部由MCnebula函数提供的算法过滤，还是对某些化学类别进行自定义过滤，我们现在有一个叫做Nebula-Index的数据。这个数据记录了一些化学类别和归属于它们的‘Features’。随后的分析过程或可视化将以它为基础。每个化学类别被认为是一个Nebula，其分类的‘Features’是这些Nebula的组成部分。在可视化过程中，这些Nebula将被可视化为网络。从形式上来说，我们把这些在Nebula-Index数据基础上形成的‘Nebula’称为Child-Nebulae。相比之下，当我们把所有的‘Features’放在一起形成一个大的网络时，那么这个Nebula就被称为Parent-Nebula。

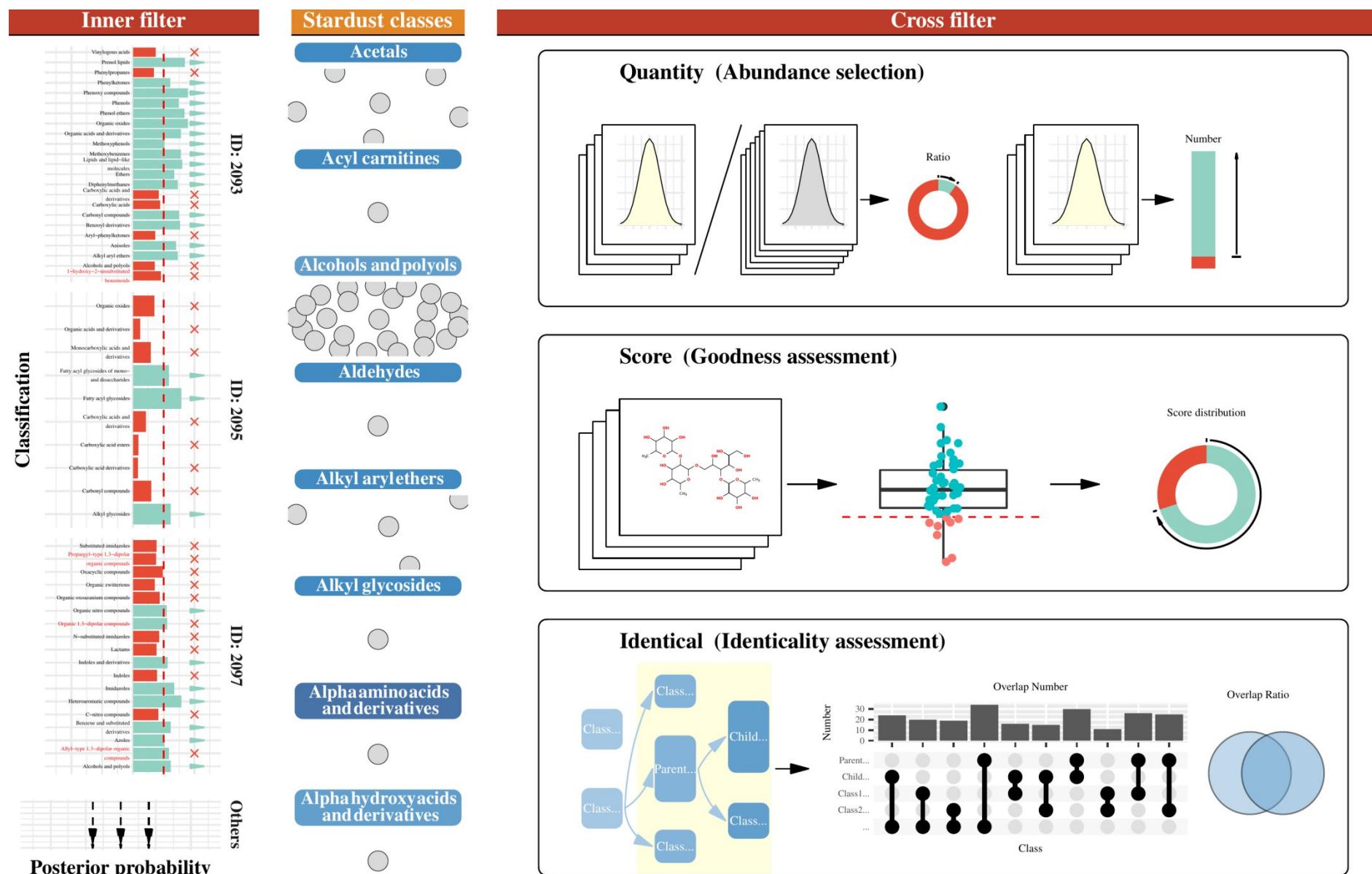


图3 MCnebula过滤化学类的机制

- 图3注：此图说明了MCnebula如何从‘Features’中过滤预测的化学类，以形成Nebulae-Index，创建Child-Nebulae。**Inner filter**通过名字的Regex匹配（名字不含阿拉伯数字）过滤化学类，并为后验概率值设定阈值。为了创建**Nebula-Index**，先前过滤的数据被按照化学类而不是‘Features’的ID重新分组。**Cross filter**通过结合Stardust Classes和‘Features’注释数据对化学类进行进一步的过滤。

6. Cross filter stardust Classes的细节

此方法是一个以下三个模块的整合（图3）：

6.1 Cross filter by ‘quantity’ (abundance selection).

为每个组设置‘Features’数量限制（每个组，即一个化学类别与其分类的‘Features’）。具有太多的‘Features’或太少的‘Features’的组将被过滤掉。这意味着化学类会被过滤掉。这些阈值为：

- 最小数量：组内的‘Features’。
- 最大比例：组内的‘Features’数量与所有组的所有‘Features’（唯一）数量相比。

这一步的目的是过滤掉那些概念范围太大或者太细微的化学类别。例如，“Organic compounds”涵盖了几乎所有可以在代谢组学数据中检测到的化合物，其范围太大，对我们的生物学研究没有任何帮助。参数的设置不是绝对的，也没有最佳的解决方案。用户可以根据研究的必要性来拟定阈值。

6.2 Cross filter by ‘score’ (Goodness assessment)

这一步将Stardust Classes的数据与‘Features’注释数据联系起来。对于每个组，对每个目标属性（连续属性，一般是化合物鉴定的评分属性，如‘Tanimoto similarity’）进行优度评估。如果该组符合所有预期的优度，该化学类将被保留；否则，该化学类将被过滤掉。优度(G)与组内的‘Features’有关。

- n : 目标属性满足阈值的‘Features’数量。
- N : 所有‘Features’的数量。

优度： $G = n/N$ 。

优度的评估与‘tolerance’和‘cutoff’参数有关。

- 预期优度，即‘tolerance’的值。
- 实际优度，与参数‘cutoff’有关。 $G = n/N$ 。

可以对多个目标属性进行优度评估。只有当化学类通过了所有目标属性的优度评估时，它才会被保留。这一步的主要目的是过滤掉那些具有太多结构鉴定度低的‘Features’的化学类。

6.3 Cross filter by ‘identical’ (identicality assessment).

为化学分类设定一个层级范围，让这个范围内的组进行比较，以确定彼此之间的一致性。对于两个组，如果分类的‘Features’几乎相同，其中一个组所代表的化学类将被排除。对两个组（A和B）的一致性评估：

- x : 属于B的A的分类‘Features’的比率
- y : 属于A的B的分类‘Features’的比率
- i : 参数‘identical_factor’的值

如果 $x > i$ 和 $y > i$ ，这两组将被认为是相同的。那么，具有较少‘Features’的组将被丢弃。这一步的目的是为了过滤掉那些可能会互相包含、范围相似的类。计算机预测方法可能无法从LC-MS/MS光谱中分辨出潜在化合物属于哪一类。

（三）数据结构

1. 首要Class: ‘mcnebula’ 的结构

表 2 Class: 'mcnebula'的结构

Inherits from	Slots	Description
mcnebula	creation_time	由'date()'创建
	ion_mode	离子模式，'pos'或者'neg'
	melody	存储调色板的类
	mcn_dataset	MCnebula的主要数据集
	statistic_set	统计数据集
project	project_version	SIRIUS版本，'sirius.v4'或者'sirius.v5'
	project_path	SIRIUS项目的文件路径
	project_conformation	项目内文件的隶属信息
	project_metadata	项目内文件的存在信息
	project_api	文件的读取和格式化函数
nebula	project_dataset	读取后存储的数据
	parent_nebula	与Parent-Nebula的可视化相关
	child_nebulae	与Child-Nebula的可视化相关
export	export_path	MCnebula的输出路径
	export_name	可视化或输出的表格中的标签名

2. 数据相关Class的结构

表 3 数据相关Class的结构

Class	Slots	Description
project_conformation	file_name	文件'subscript'对应的名称
	file_api	文件'subscript'的上级从属
	attribute_name	属性'subscript'对应的名称
project_metadata	metadata	文件的元数据表格
project_api	methods_read	读取的函数
	methods_format	格式化的函数
	methods_match	提取字符、匹配字符的函数
project_dataset	dataset	从SIRIUS项目中读取的数据集
mcn_dataset	dataset	经过初步过滤的数据集
	reference	用于可视化和后续分析的数据集
	backtrack	数据回收站
statistic_set	design_matrix	设计矩阵
	contrast_matrix	对比矩阵
	dataset	用于统计分析的数据集
	top_table	统计分析的排序结果

3. 可视化相关Class的结构

表 4 可视化相关Class的结构

Class	Slots	Description
melody	palette_set	用于追踪'feature'或者用于因子变量可视化的调色板
	palette_gradient	用于连续变量的可视化的调色板
	palette_stat	用于统计分组的调色板
	palette_col	用于化学类的区分的调色板
	palette_label	用于化学类阶层的可视化的调色板
parent_nebula	igraph	R包'igraph'生成的'igraph'类的对象，网络型数据文件
	tbl_graph	R包'tidygraph'生成的'tbl_graph'类的对象，网络型数据文件
	layout_ggraph	用于R包'ggraph'可视化的网络型数据文件
ggset		管理'ggplot2'的可视化的函数和参数的类的对

Class	Slots	Description
		象
child_nebulae	igraph	'igraph'对象的list
	tbl_graph	'tbl_graph'对象的list
	layout_ggraph	'layout_ggraph'的list
	grid_layout	R包'grid'创建的网格画板样式的对象
	viewports	R包'grid'创建的视图（viewport）对象的list
	panel_viewport	主画板的视图对象
	legend_viewport	图例的视图对象
	ggset	'ggset'的list
	structures_grob	化学结构可视化的'grob'对象（与'grid'包相关）
	nodes_ggset	绘制深度注释节点（网络图中的Nodes）的ggset的list
	nodes_grob	Nodes的grob的list
	ppcp_data	用于化学类预测的后验概率的可视化的数据集
	ration_data	同于统计数据可视化的数据集
	ggset_annotation	带有深度注释Nodes的ggset对象的list

4. 其他Class

表 5 MCnebula2其他Class的结构

Class	Slots	Description
msframe	entity	存储数据框（data.frame）
	subscript	表明来源的名称
ggset	layers	存储ggplot绘图的各个函数和响应的参数的list
command	command_name	函数的名字
	command_function	函数本体
	command_args	函数的参数
report	yaml	控制输出的yaml语言，与Rmarkdown、markdown、pandoc有关
	layers	list，存储section, character, code_block等
code_block, code_block_table, code_block_figure	codes	格式化的代码

Class	Slots	Description
heading	command_name	代码块的执行程序，一般是'r'
	command_function	将代码输出成文本的函数
	command_args	传递到执行程序的参数
.Data	heading	'heading'继承于'character'
	level	表明标题层级
section	heading	类'heading'的对象
	paragraph	'character'对象
	code_block	'code_block'对象

(四) 方法 (Method) 和函数 (Function)

1. 数据方法

表6为MCnebula应用中主要用于数据分析处理的方法 (Method)。

表 6 MCnebula主要的数据方法

Method	Description
initialize_mcnebula	初始化分析的方法
collate_data	从SIRIUS项目中获取数据的方法
filter_formula	过滤化学分子式候选选项的方法
filter_ppcp	初步过滤化学类后验概率数据集的方法
filter_structure	过滤化学结构式候选选项的方法
create_reference	明确唯一化学分子式候选选项从而承前启后的办法
create_hierarchy	创建化学类的阶层的方法
create_features_annotation	合并注释并形成注释数据表格的方法
create_stardust_classes	创建'Stardust classes'化学类数据的方法
cross_filter_stardust	过滤'Stardust classes'数据集的方法
backtrack_stardust	回溯'cross_filter_stardust'过滤掉的化学类的方法
create_nebula_index	根据'Stardust classes'创建'Nebula-Index'数据集的方法
binary_comparison	进行统计分析的方法

2. 可视化方法

表7为MCnebula中主要用于可视化的方法。

表 7 MCnebula主要的可视化方法

Method	Description
create_parent_nebula	创建Parent-Nebula初始网络型数据的方法 ('igraph'对象)
create_child_nebulae	创建Child-Nebulae初始网络型数据的方法 ('igraph'对象的list)
create_parent_layout	创建Parent-Nebula可视化样式的方法
create_child_layouts	创建Child-Nebulae可视化样式的方法（包括网络排布和画板布局）
activate_nebulae	为Parent-Nebula和Child-Nebulae创建用于可视化的'ggset'对象
annotate_nebula	绘制深入注释Child-Nebula的方法
draw_nodes	绘制深入注释的'feature'的Nodes的方法
draw_structures	绘制化学结构式的方法
compute_spectral_similarity	计算光谱相似性的方法（脱胎于'MSnbbase::compareSpectra'）
visualize	最终输出绘图的方法（Parent-Nebula或者单个Child-Nebula）
visualize_all	最终输出绘图的方法（整体Child-Nebulae）

3. MCnebula辅助科学绘图的函数

在表7提及的‘visualize_all’ 和 ‘visualize’ 方法带有一个‘fun_modify’参数，这个参数允许表8中的函数作为参数传递，以便快速实现科学绘图。

表 8 辅助科学绘图的函数

Function	Description
modify_default_child	用于'visualize_all()'的函数，相当于'modify_rm_legend' + 'modify_set_labs' + 'modify_unify_scale_limits'。此外，如果使用了'set_nodes_color'方法，并'use_tracer'参数为'True'，'modify_tracer_node'和'modify_color_edge'会自动执行。
modify_stat_child	调整'Child-Nebulae'样式，使其适合以节点颜色映射连续性变量
rev.modify_stat_child	仅用于内部执行的逆转'modify_stat_child'的函数
modify_set_labs_and_unify_scale_limits	相当于'modify_set_labs' + 'modify_unify_scale_limits'
modify_annotation_child	用于细节调整深入注释的Child-Nebula的函数
modify_rm_legend	移除图例的函数
modify_tracer_node	将Child-Nebulae用于追踪模式的函数'
modify_color_edge	调整节点的边缘颜色
modify_set_margin	调整边距

Function	Description
modify_unify_scale_limits	为所有Child-Nebulae统一属性映射的比例尺，使其结果科学规范
modify_set_labs_xy	调整x、y轴的标签，仅用于'plot_msms_mirrors'
modify_set_labs	调整Child-Nebulae图例的标签，使其与设定的'export_name'一致
...	...

4. 其他方法和函数

表 9 MCnebula的其他方法或函数

Group	Type	Name	Description
Report	Function	rblock, ...	将代码存储为'code_block'对象
	Method	include_figure, ...	将图片展示在报告文档中
Workflow	Function
	Method	workflow, ...	执行或打印MCnebula基本工作流的代码
Plot msms mirrors	Function	plot_msms_mirrors, ...	绘制MS/MS镜像图
Project.*	Function	.valid, .ate_*, .get_*, ...	与对应版本相关的SIRIUS项目的API函数
Colors	Function	.get_color_*, ...	获取哈希颜色码，主要为包'ggsci'中的配色
Default visualize	Function	.command_*, ...	包装好的'ggplot2'的代码和参数，'command'对象，用于默认的可视化
Export	Function	.get_*, ...	用于获取正式输出名称的函数
Methods	Function	.rank_*, ...	用于排序的函数
Modify ggset	Function	modify_*, ...	用于后修改'ggset'对象的一系列函数，高度定制（规范）Child-Nebulae的可视化
MODIFIED compareSpectra	Function	compareSpectra, bin_Spectra, ...	剥离于包'MSnbase'的函数或者方法，仅计算光谱相似性（'dotproduct'），速度更快
	Method	mz, intensity, ...	'lightspectrum'的getter或者setter方法（'lightspectrum'比'MSnbase'的'spectrum'更轻巧）
Yaml	Function	.yaml_*, ...	用于设定报告输出的函数
VIRTUAL slots	Method	*_dataset, *_layers, ...	虚类带有的方法，用于便捷操作对象

Group	Type	Name	Description
Clear	Function	clear_dataset, ...	用于清除'mcnebula'对象的不再用到的数据的函数
...

Name或Group中的*表示省略一系列类似名称的方法或函数

（五）MCnebula的基本使用

以下代码块展示了MCnebula（MCnebula2 R包）不带任何自定义参数的使用方法，从数据的初始化、整合和处理到Parent-Nebula和Child-Nebulae的可视化。更详细的使用示例请参考：1) 第三部分基于MCnebula策略分析杜仲炮制前后的成分变化 > 二、结果 > （三）分析的报告和R代码；2) 第四部分基于MCnebula策略的血清代谢组学 > 二、结果 > （三）分析的报告和R代码；3) 或者MCnebula2 R包的使用文档：<https://github.com/Cao-lab-zcmu/MCnebula2/blob/document/reference.pdf>

```
## 初始化
mcn <- mcnebula()
mcn <- initialize_mcnebula(mcn, "sirius.v4", ".")
ion_mode(mcn) <- "pos"
## 数据整合和处理
mcn <- filter_structure(mcn)
mcn <- create_reference(mcn)
mcn <- filter_formula(mcn)
mcn <- create_stardust_classes(mcn)
mcn <- create_features_annotation(mcn)
mcn <- cross_filter_stardust(mcn)
mcn <- create_nebula_index(mcn)
## 可视化
mcn <- compute_spectral_similarity(mcn)
mcn <- create_parent_nebula(mcn)
mcn <- create_child_nebulae(mcn)
mcn <- create_parent_layout(mcn)
mcn <- create_child_layouts(mcn)
mcn <- activate_nebulae(mcn)
visualize(mcn, 'parent')
visualize_all(mcn)
## 自定义分析
## ...
```

三、小结

LC-MS/MS数据的分析具有挑战性，因为其数据量大，潜在的未知化合物的信息多，而且参考谱库有限。研究人员往往需要花很多时间从这个‘黑匣子’中找出有意义的化合物，然后再进行下一步的研究。MCnebula可以帮助研究人员快速关注潜在的标志物或有意义的化合物，它将全谱识别与机器预测相结合，在多维视图中对Child-Nebulae进行可视化，并通过统计分析来追踪Top ‘Features’并找到类似物。ABC选择算法可以总结出数据集中具有代表性的化学类别，并获得该类别的 ’Features’，因此研究的整体方向是无偏差的。同时，它也是统计分析的有效保证，为下一步的追踪分析产生“校正”的Top

‘Features’：基于‘Features’水平的统计分析结果可能会因为信息的丢失而产生偏差，在化学类水平的基础上进行过滤可以在一定程度上防止偏差的产生。Child-Nebula是在ABC选择算法得到的化学类别的基础上绘制的，实现了将巨大的非目标数据集可视化为一个单一图形的目标。ABC选择算法的参数是可以主观调整的，它们应该根据研究对象的化学类别的丰富程度来确定。一般来说，我们的默认参数用来获取根据数据集的种类丰富的化学类，并过滤掉那些在概念范围内过大或过小的化学类。

MCnebula的R包已上传至github (<https://github.com/Cao-lab-zcmu/MCnebula2>)，用户可以在R命令行中输入 `remotes::install_github('Cao-lab-zcmu/MCnebula2')` 轻松安装它。

第二部分 MCnebula的方法评估与拓展

一、材料与方法

(一) 实验材料

个人笔记本电脑Surface pro7用于编程环境的搭建和R语言编程：Pop!_OS (Ubuntu) 22.04 LTS 64-bits PC (Intel Core i7-1065G7, 1.3 GHz × 8, 16 Gb of RAM)。

工作站用于运行和测试数据集：Pop!_OS (Ubuntu) 22.04 LTS 64-bits workstation (Intel Core i9-10900X, 3.70GHz × 20, 125.5 Gb of RAM)

用于获取参考光谱数据集的网站：<http://prime.psc.riken.jp/compms/msdial/main.html#MSP>

在测试阶段中，数据集被上传到GNPS服务器用于比较分析，现在这些数据集是可获取的：

- 1) original dataset: FBMN:

<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=05f492249df5413ba72a1def76ca973d>.
MolnetEnhancer:

<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=9d9c7f83fa2046c2bf615a3dbe35ca62>;

- 2) medium noise dataset: FBMN:

<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=c65abe76cd9846c99f1ae47ddbd34927>;
MolnetEnhancer:

<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=7cc8b5a2476f4d4e90256ec0a0f94ca7>;

- 3) high noise dataset: FBMN:

<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=62b25cf2dcf041d3a8b5593fdbf5ac5e>;
MolnetEnhancer:

<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f6d08a335e814c5eac7c97598b26fb80>.

(二) 实验方法

1. R 的配置

除了表1涉及的R包以外，在测试和拓展中还用到了额外的R包：

表 10 测试和拓展MCnebula用到的R包

Type	Name	Description
IMPORTS	aplot	用于凭借ggplot绘图
	classyfireR	R的Classyfire的API，查询化合物的分类系统
	FELLA	用于通路富集的R包
	ggtree	用于可视化层次聚类的ggplot2的拓展包
	Hmisc	使用'capitalize'将句首单词首字母大写
	magrittr	使用管道符 '%>%' 和 '%<%'
	pdftools	用于处理.pdf的输出
	png	用于读写.png格式的图片
	rcdk	用于分子量的计算和同位素模式的模拟
	RCurl	用于下载PubChem数据库
	scales	用于可视化

2. 建立评估数据集

2.1 MS/MS噪声模拟

GNPS MS/MS库的光谱集（正离子模式，以获得更多的光谱数据）被用于评估（.msp文件）（<http://prime.psc.riken.jp/compms/msdial/main.html#MSP>）。由于参考库中的碎片光谱通常具有较高的品质，在用于评估库的匹配性时，可能会导致过拟合。为了解决这个问题，参考CANOPUS的报道^[41]，我们在这些MS/MS图谱中加入了‘噪音’。简而言之，‘噪音’包括质量偏移、峰强度偏移和插入的噪声峰；这些偏移的大小系数是从正态分布的函数中随机抽取的。总的来说，我们模拟了两种模式的“噪音”（中度噪音和高度噪音）。噪音的模拟是在自定义R脚本中实现的。这些算法和参数与文献^[41]平行。我们将这些数据集指定为原数据集（Origin dataset）、中度噪音数据集（Medium noise dataset）和高度噪音数据集（High noise dataset）。噪音模拟的细节如下：

2.1.1 全局质量偏移

通过从 $N(0, \sigma_{mb}^2)$ （正态分布）中抽取一个随机数 δ^* ，然后将每个质量峰值 m 移动 δ^*m 来模拟全局质量移动。标准差 σ_{mb} 被选为 $\sigma_{mb} = (10/3) \times 10^{-6}$ （中等噪声）或 $\sigma_{mb} = (15/3) \times 10^{-6}$ （高噪声），因此 $3\sigma_{mb}$ 区间代表中等噪音的10 ppm移动，高噪音的15 ppm移动。

2.1.2 个体质量偏移

对每个质量为 m 的峰值，通过从 $N(0, \sigma_{md}^2)$ 中抽取一个随机数 δ ，并将峰值移动 δm 来模拟个体质量偏差。选择标准偏差 σ_{md} 是为了使 $3\sigma_{md}$ 的区间代表中等噪音的10 ppm移动和高噪音的20 ppm移动。

2.1.3 峰强度偏移

强度变化是在光谱中模拟的。每个峰的强度都乘以从 $N(1, \sigma_{id}^2)$ 中抽取的单个随机数 ϵ 。中等噪音的方差选择为 $\sigma_{id}^2 = 1$ ，高噪音的方差选择为 $\sigma_{id}^2 = 2$ 。从每个峰值强度中减去光谱最大峰值强度的0.03倍。如果一个峰值强度低于光谱中最大强度的千分之一的阈值，该峰值就被丢弃。

2.1.4 额外的噪声峰

额外的“噪声峰”被添加到光谱中。在预先处理原数据集时，从碎裂光谱中收集一个“噪声峰”的库，使用所有没有已知前体分子式的分子分解的峰。对于每个光谱，这些‘噪音峰’的 αn 被添加到光谱中，其中 n 是光谱中的峰数， $\alpha = 0.2$ 为中等噪音， $\alpha = 0.4$ 为高噪音。“噪声峰”的强度被调整为贡献和接收光谱中的最大峰值强度的相应比例。“噪声峰”从“噪声峰”的库中随机抽出，并添加到光谱中。

2.2 模拟同位素模式

另一个问题是，光谱集不具备同位素模式。在真正的LC-MS处理中（‘Features’检测），同位素峰被分组和合并，这有利于SIRIUS检测一些特定的元素^[37]。为了模拟同位素模式，我们使用‘rcdk’R软件包中的‘get.isotopes.pattern’函数来获得同位素质量和它的丰度^[42]。此外，这些质量被认为是加合物类型（Adduct）增加或减少的确切质量（Exact mass）。对于这些同位素模式的‘强度’，我们模拟为相对强度，即同位素的丰度乘以100的值。这些‘同位素峰’被合并到其化合物的MS¹列表中。所有的光谱集合都被格式调整，以适应MCnebula工作流程或基准方法的输入（.mgf文件和‘Features’量化表.csv）。

3. 评估的方法

三个模拟数据集（Origin dataset, Medium noise dataset, High noise dataset）都是用MCnebula工作流程和基准方法（GNPS）运行。当这些数据被放入SIRIUS 4命令行界面（CLI）（4.9.12版）进行计算时，具有空碎片峰的MS/MS光谱被自动过滤。此外，为了减少计算时间，过滤掉了超过800m/z前体

的化合物。这些被过滤掉的化合物被排除在最终准确性评估之外。原始数据中共有8782个MS/MS谱图，经过过滤或排除后，共有7524个化合物用于最终评估。

在ClassyFire^[24]的辅助下，对分类准确性进行评估。细节上，我们遍历了原始的.msp光谱文件，以整理这些化合物的元数据，包括结构注释。这些化合物的国际化学识别码（InChIKey）可用于ClassyFire数据库的检索。然而，由于ClassyFire只支持那些之前已经在其服务器上进行过结构分类的化学标识，我们注意到所有的InChIKeys都被否决了。为了解决这个问题，我们采用了这些InChIKeys的第一个哈希块（InChIKey平面，代表分子骨架）来连接PubChem应用编程接口（API）（<https://pubchemdocs.ncbi.nlm.nih.gov/pug-rest>）^[43]。因此，我们得到了所有可能的InChIKeys的异构体（立体异构）^[44]。小分子的分类取决于它的分子骨架，因此这些拥有相同InChIKey平面的化学特征在分类上是相同的。我们将获得的InChIKey列表传递给ClassyFire以获得化学分类。在R脚本中，一旦任何同分异构体的InChIKey与取得了分类数据，这个分子骨架的获取状态结束。最后，所有这些化学注释都被整理、整合并指定为标准参考（关于获取这些数据的工具，请参考：第二部分MCnebula的方法评估与拓展 > 二、结果 > （二）MCnebula的拓展 > 1. 用于化学发现）。

MCnebula和基准方法在算法和分类结果方面的差异使它们不能在完全相同的水平上被评估。我们分别评估了这两种方法。对于MCnebula，在评估准确性之前，我们审视了从原数据集（Origin dataset）的预分析中产生的类别。超过一半的类是基于子结构类进行分类的，如‘Organic carbonic acids and derivatives’、‘Hydroxy acids and derivatives’。这些类的亚结构很小，是化合物中的化学官能团。ClassyFire的原则是选择化合物中最主要的结构类进行替代^[24]。但是，从药物发现的角度来看，结构决定药效，许多药理作用可能取决于这些子结构；另一方面，质谱的注释不容忽视任何细微的信息。为了在算法中找到更多的普遍性特征，我们在结果中保留了这些类别。基准方法中没有亚结构分类，因此我们在评估中忽略了这些类别。然而，其余的类仍然可能是亚结构类。我们为评估指定了三个级别，即‘True’、‘Latent’、‘False’。‘True’表示分类后的类别与ClassyFire的一致。‘Latent’表示分类后的类与ClassyFire不一致，但它们的‘Class’级别（Level 3）的父类与ClassyFire的一致。‘False’表示分类与ClassyFire的完全不一致。

为了评估类别或结构的识别，我们用InChIKey planar将结果与标准结果进行矩阵合并。为了评估化学结构的识别，一旦识别的化学结构与标准化学结构相一致（通过 InChIKey planar 匹配），我们就把它定为‘True’。事实上，这种评估忽略了立体化学。

4. 用于建立评估数据集和用于评估的R的函数

手动实现所有的评估几乎是不可能的。为了快速得到评估结果以及将结果可视化，R的函数被编写，随后用于评估的脚本（获取这些函数以及相关评估数据：<https://github.com/Cao-lab-zcmu/exMCnebula2/blob/master/inst/extdata/evaluation.tar.gz>）：

表 11 评估MCnebula涉及的R函数

Function	Description
msp_to_mgf	用于将.msp格式光谱文件转化为SIRIUS需求的.mgf
formula_adduct_mass	将分子式转化为各种加合离子类型，并计算Exact Mass
get_adduct_mass	计算加合离子的Exact Mass
element_extract	提取字符串分子式中的元素
formula_reshape_with_adduct	字符串分子式转变为各种加合离子
collate_as_noise_pool	提取为噪声峰的库
tol_mergeEx	根据Mass容差合并表格
mass_shift	根据正态分布随机偏移Mass
spectrum_add_noise	在光谱（data.frame）中添加噪声
mgf_add_anno.gnps	将.mgf格式化为GNPS服务器接受的.mgf格式
simulate_gnps_quant	模拟量化数据
stat_classify	统计归类结果
stat_identification	统计鉴定结果
visualize_stat	可视化统计结果
visualize_statComplex	合并可视化三重数据集
visualize_comparison	可视化比较结果
visualize_summary	多个水平上总结并可视化
visualize_idRes	可视化鉴定结果
...	...

5. MCnebula的拓展涉及的算法

5.1 统计分析的算法

MCnebula整合了‘limma’包（差异表达分析的R包，RNA-序列和微阵列的分析）中的函数（主要为：‘limma::makeContrasts’，‘limma::lmFit’，‘limma::eBayes’），并将其打包用于代谢组学数据的差异分析^[45]。LC-MS的‘Features’量化矩阵和基因表达矩阵是相似的，都有相应的解释变量（样品信息）和因变量（基因表达值或‘Features’量化值），只是一个代表基因表达水平，另一个代表代谢物水平。我们将‘Features’的峰面积水平归一化，并对其进行转化（log2），利用样品的元数据信息建立设计矩阵和对比矩阵^[45]；因此，即使数据本身与基因无关，也可以利用‘limma’包的工具进行差异分析。‘limma’是一个强大的使用线性模型进行差异分析的软件包，不仅可以处理解释变量为因子的简单实验设计（如对照组与模型组），而且可以处理解释变量为协变量的复杂实验设计（如包含时间序列的组）。然而，我们的打包方法只适合于实验设计：其中解释变量是因子变量，设计矩阵没有截点

(代码：‘model.matrix(~ 0 + group)’)^[46]。由于其简单的适用性，我们称其为‘Binary comparison’。我们的评估部分没有涉及它的评价（因为它不是本研究的主要部分），但我们在两个演示数据集中使用了它，并在一定程度上进行了验证：在血清数据集中，我们将我们获得的高排名的‘Features’与Wozniak等人^[47]的数据进行了比较（见：第四部分 基于MCnebula策略的血清代谢组学 > 二、结果）；在中药数据集中，我们将获得的高排名‘Features’追溯到EIC图上（见：第三部分 基于MCnebula策略分析杜仲炮制前后的成分变化 > 二、结果）。

附Law等人关于设计矩阵的图示（其中两种类型）：

Model

$$E(y) = 1.03x_0 + 2.12x_1 + 3.00x_2 + 4.90x_3$$

$E(y) = 1.03$	= 1.03	(for control)
$E(y) = 2.12$	= 2.12	(for treatment I)
$E(y) = 3.00$	= 3.00	(for treatment II)
$E(y) = 4.90$	= 4.90	(for treatment III)

Matrix

> model.matrix(~0 + treatment)

	treatmentCTL	treatmentI	treatmentII	treatmentIII
1	1	0	0	0
2	1	0	0	0
3	1	0	0	0
4	0	1	0	0
5	0	1	0	0
6	0	1	0	0
7	0	0	1	0
8	0	0	1	0
9	0	0	1	0
10	0	0	0	1
11	0	0	0	1
12	0	0	0	1

Plot

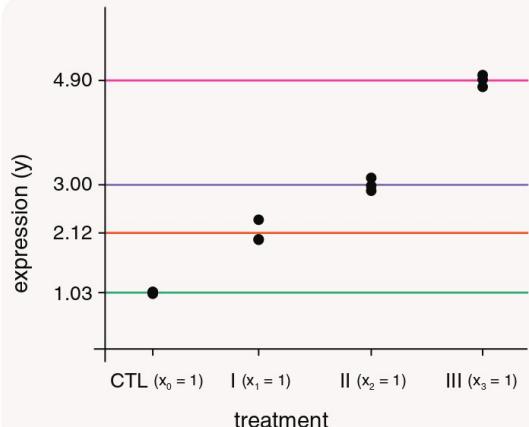


图4 设计矩阵和对比矩阵的示例：Expected gene expression is modelled by a treatment factor

Model

$$E(y) = 2.95x_1 + 4.57x_2$$

$E(y) = 2.95$	= 2.95	(for healthy group)
$E(y) = 4.57$	= 4.57	(for sick group)

Matrix

```
> model.matrix(~0 + group)
```

$$\begin{matrix} & \text{groupHEALTHY} \\ \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{pmatrix} & \left(\begin{array}{c|c} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{array} \right) \\ & \text{groupSICK} \end{matrix}$$

Plot

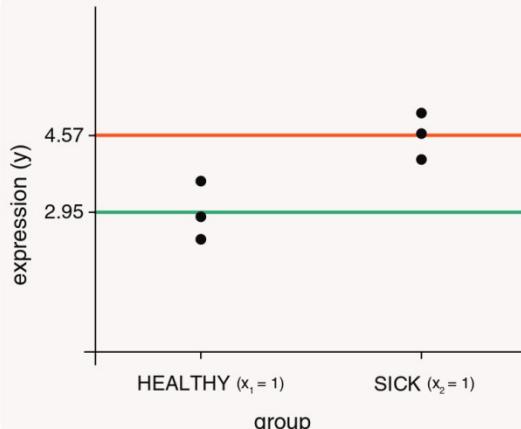


图5 设计矩阵和对比矩阵的示例：Expected gene expression is modelled by a group factor

5.2 其他的算法

请参考表13表14。

二、结果

(一) MCnebula的评估

1. 功能评估

表12为MCnebula与其他工具的功能性比较。就选择的涵盖识别、分类等指标的评估而言，MCnebula的适用范围更广。

表 12 MCnebula和其他工具的功能比较

Group	Item	MCnebula	SIRIUS	GNPS	MZmine	XCMS	MetaboAnalyst	MS-DIAL
Identificaiton	MS1	**	**	**	-	**	**	**
	Library match	***	***	***	-	**	-	**
	Machine prediction	***	***	-	-	-	-	-
Classifying	Structure based	**	-	***	-	-	-	-
	MS/MS based	***	***	-	-	-	-	-
	Select classes	***	-	-	-	-	-	-
Visualize dataset	Spectral based	**	-	***	-	-	-	*
	Classes based	***	-	-	-	-	-	-
	Indepth annotation	***	-	***	-	-	-	-
Others	Preprocessing	*	*	**	***	***	*	***
	Statistics	**	-	-	**	**	***	**
	Path enrichment	*	-	-	-	-	***	-
	Report	***	-	-	-	-	**	-
Usage	Availability	***	***	***	***	***	***	***
	Difficulty	**	*	**	***	***	*	*

以上*表示功能的齐全性或完善性，-表示不存在该功能

2. 归类准确度评估

我们使用一个公开的参考光谱库来评估MCnebula的分类准确性。直接使用这种参考光谱库可能会导致评估过程中的过度拟合。我们采取了模拟噪声的方法来消除这一后果。模拟噪声，即在参考光谱中加入无效的噪声数据或对现有数据进行数字移位，也模拟了类似于真实场景的数据采集：由于采集条件不同，真实情境下的光谱数据与参考光谱相比会有更多的噪声。通过在参考光谱库中加入噪声，我们现在有三个数据集用于评估（Origin dataset、Medium noise dataset、High noise dataset）（7524个化合物（光谱））。所有这三个数据集都使用MCnebula进行分析。由于参考光谱中化合物的丰富性，对于Origin dataset，我们通过使用ABC选择算法共获得了152个化学类（每个化学类都有一个相应的待评估化合物）。这152个化学类别包括根据优势（主导）结构提炼的化学类别和根据亚结构提炼的化学类别。为了便于与其他方法进行比较，我们只选择了可能是主导结构的化学类进行评价。有37个这样的化学类别被选中进行评估。为了更客观地评价MCnebula，我们选择了GNPS（Global Natural Products Social Molecular Networking）提供的分子网络，工作流包含基于特征的分子网络（FBMN，Feature-based Molecular Networking）和MolNetEnhancer模块，作为提供质谱数据的可视化聚类分析的基准方法。GNPS是一种典型的、流行的基于光谱库的质谱注释方法。原则上，它首先通过与公共光谱库进行镜像匹配来计算光谱相似度，识别出具有准确化学结构的化合物，然后根据注释的化学结构确定化学类别。

我们将这三个数据集上传到GNPS服务器，然后获得结果并用于评估。对于Origin dataset，GNPS总共得出了44个化学类别（与MCnebula平行，每个化学类别至少有50个化合物）。总共有19个共同的类别。这些类别被选来比较MCnebula和GNPS在分类数量、稳定性和相对错误率方面的准确度。在分类数量方面，MCnebula在三个数据集中的表现优于GNPS（MCnebula: 199, 178, 160; GNPS: 162, 95, 81）（图6a）。在加入噪声后的分类稳定性方面，MCnebula在两个数据集上的表现优于GNPS（MCnebula: 10.5%, 19.8%; GNPS: 41.7%, 50.1%）（图6a）。对于最后一个指标，为了评估分类的性能，它结合了稳定性的水平来计算相对错误率，而不是绝对错误率。相对错误率更好地模拟了实际应用于LC-MS/MS分析的情况，因为实际的光谱数据不仅包含噪声，还包含许多无法通过光谱匹配识别的未知化合物。在这种情况下，MCnebula在三个数据集的相对错误率评估中优于GNPS（MCnebula: 30.2%、32.9%、32.6%; GNPS: 51.9%、48.8%、47.6%）（图6a）。除了上述三个指标外，我们还对MCnebula和GNPS在19个化学类别的个体水平上进行了比较（图6b）。结果表明，MCnebula比GNPS对噪声更稳定。

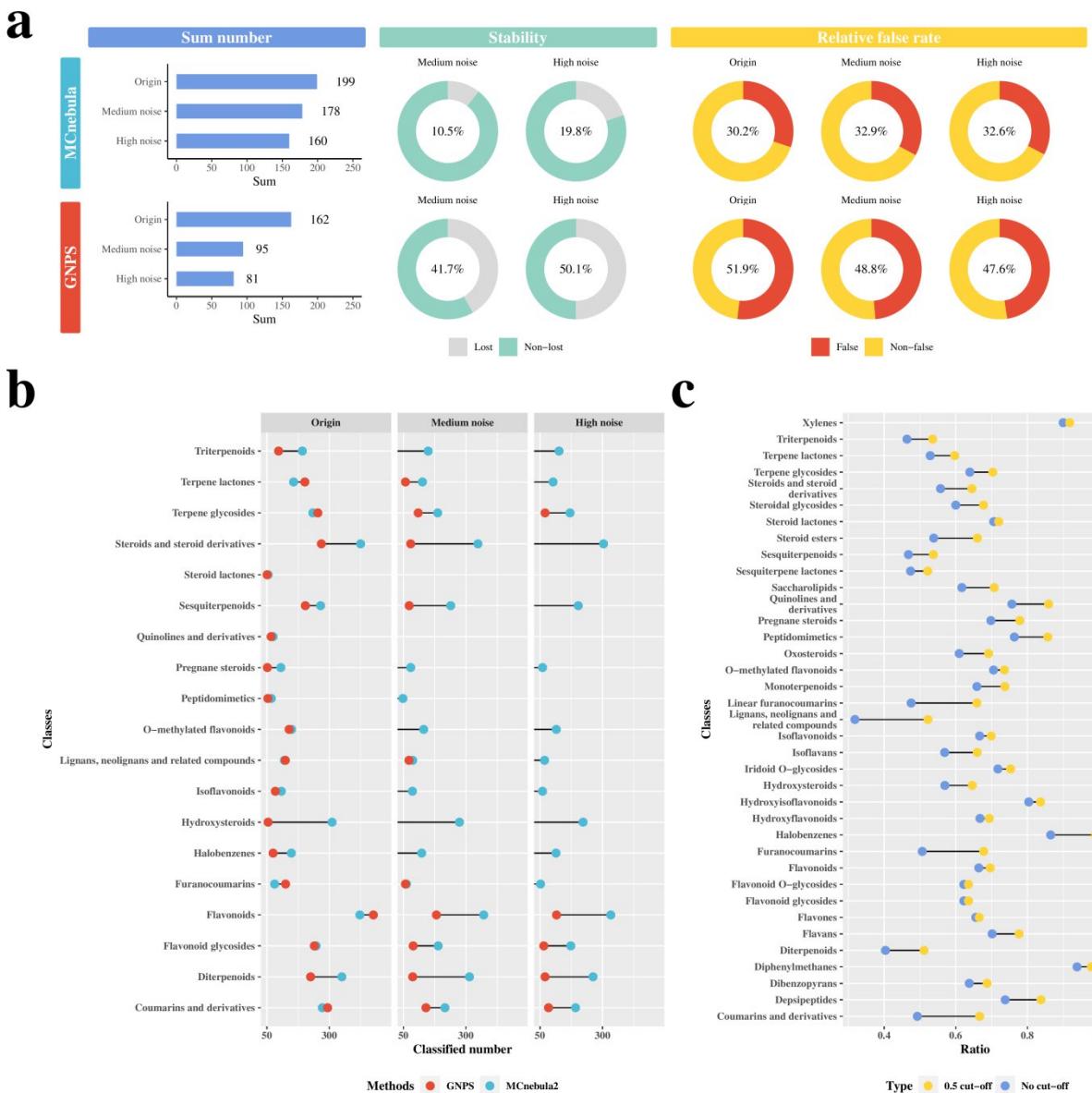


图6 MCnebula归类的准确度评估

- 图6注： **a**)为MCnebula与基准方法（GNPS）在分类数量、稳定性、相对错误率三个指标上的比较。分类数量的计算方法是选定的19个化学类别中被归类的化合物的平均总数量。稳定性的计算方法为： $S = (N_{origin} - N_x) / N_{origin}$ (N_{origin} 是产地数据集的平均总数; N_x 是中等噪声数据集或高噪声数据集的平均总数)。相对错误率的计算方法是 $R = 1 - (1 - F) \times (1 - S)$ (F 是绝对错误率; S 是稳定性，即稳定性评估中的平均损失率)。 **b**)为MCnebula和基准方法的分类数量比较。当噪声被添加到原始数据集中时，一些分类‘Features’的数量出现<50，这里设置了一个截止值(≥ 50)，将这些化学类排除在评估之外。 **c**)为MCnebula的鉴定准确度评估，图中设置了一个临界值(Tanimoto similarity ≥ 0.5)，以获得高匹配分数的化学结构进行评估。

附图7和图8，两种方法在各自水平上的评估（非平行）。



图7 MCnebula归类准确度的单独评估

- 图7注：对于**Intermediate horizontal bar plot**，为评估准确性指定了三个评估级别。‘True’ 表示归类后的类别与ClassyFire的一致。‘Latent’ 表示归类后的类与ClassyFire不一致，但其 ‘Class’ 级别的父类（由**Left tile diagram**图例说明）与ClassyFire一致。‘False’ 表示分类后的类与 ClassyFire的完全不一致。在原始数据集中加入了中度和高度的噪声，以评估MCnebula算法的稳定性。对于 ‘True’ 和 ‘False’ 的评估，箭头表示中度噪音或高度噪音会导致准确率的变化（增加或减少）。一般来说，对于每个评估的化学类别，有两对箭头，左边一对表示 ‘True’ 偏移，右边一对表示 ‘False’ 偏移。每对箭头：首先从黄色箭头指示的位置开始，在紫色箭头指示的位置结束；然后从紫色箭头指示的位置开始，在紫色条形图末端结束。准确度评估仅在分类特征数为 ≥ 50 时进行。如果噪音导致的分类数量为 < 50 ，则该类被排除在噪音评估之外。**Right horizontal bar plot**表示分类后的 ‘Features’ 数量。评估的细节见：第二部分 MCnebula的方法评估与拓展 > 一、材料与方法 > （二）实验方法 > 3. 评估的方法。

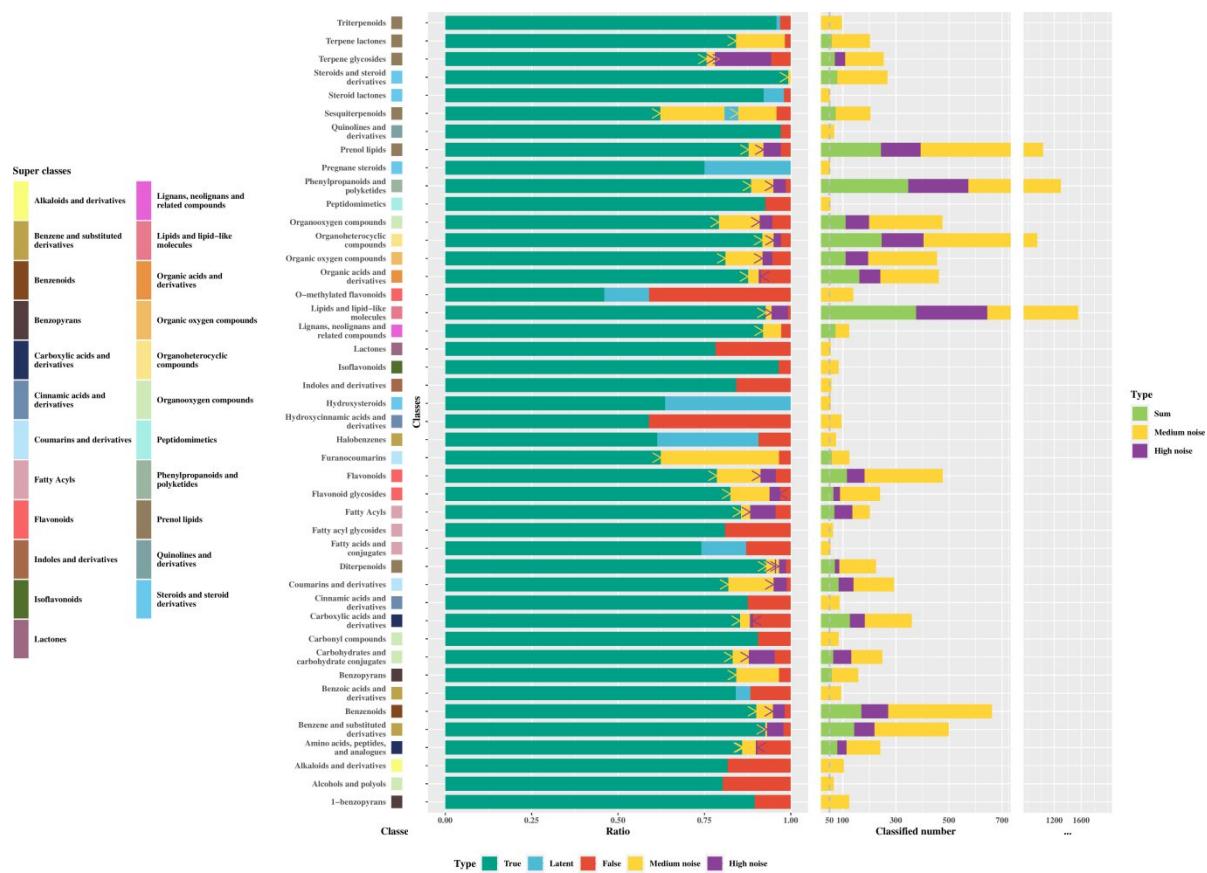


图8 GNPS归类准确度的单独评估

- 图8注：参考图7注解。

3. 鉴定准确度评估

使用MCnebula工作流程，包含8057个化合物（前体离子 $m/z < 800$ ）的源数据集，所有这些化合物都被预测为化学分子式，其中6610个化合物被预测到化学结构。这些化学结构在分类背景下被评估准确性。对于37个化学类别（图6c），平均识别的错误率为37%；平均识别的化合物数量为156个。其中，大部分的识别错误率在30%到40%之间，然而，有些类别的识别错误率相当低，如‘Long-chain fatty acids’或‘Lignans, neolignans and related compounds’。预测的化学结构的可靠性可以用一个分数来评估。Tanimoto similarity为每个预测的化学结构提供了这样一个分数（它提供了化学指纹与结构的匹配程度）。当Tanimoto similarity将临界值设定为0.5时，平均错误率为29.4%；平均鉴定的化合物数量为139个（图6c）。以上我们评估了MCnebula得到的每个化学类别的化合物的识别准确率。需要注意的是，MCnebula本身并不包含任何鉴定模块，它只利用SIRIUS预测结果中得分最高的候选化合物进行注释。关于鉴定的更多评价请参考出版物和我们以前的相关工作^[48,22]。

4. 评估的报告和R代码

以下评估的脚本和报告可见于：https://github.com/Cao-lab-zcmu/exMCnebula2/tree/master/inst/extdata/scripts_evaluation/evaluation_workflow

Evaluation of MCnebula2

Contents

1	Introduction	1
2	Set-up	2
3	Initialisation	2
4	Use simulated dataset	2
4.1	Convert .msp as .mgf	2
4.2	Query classification for compounds	2
4.3	Add noise peak	2
4.4	Output .mgf for SIRIUS	3
4.5	Output .mgf for GNPS	3
5	Evaluate MCnebula2	4
5.1	Integrate data	4
5.2	Use pre-integrate data	5
5.3	Download results of finished jobs from GNPS service	5
5.4	Evaluate accuracy of classify	5
5.4.1	Load assessment data and reference data.	5
5.4.2	Filter assessment data.	6
5.4.3	Count the results.	6
5.4.4	Post-filtering.	6
5.4.5	Distinguish the chemical class of the dominant structure.	6
5.4.6	Visualizaion	9
5.4.7	Compare with MolNetEnhancer	13
5.4.8	Summary	19
5.5	Evaluate accuracy of identification	21
6	Session infomation	24

1 Introduction

This document provides the code to evaluate MCnebula2. Most of the data used in this is already included in the package ‘exMCnebula2’. By downloading and installing the ‘exMCnebula2’ package, the following code can be run without trouble. The code in this document can be divided into two parts:

- The first part is the code used to generate the predicate dataset (for evaluation), which is time consuming to run, but we have already run it in advance and included the outcome data in the package ‘exMCnebula2’, so users do not need to rerun them, unless they are tested for feasibility.
- The second part of the code is the code for evaluating the results, including the data processing and data visualization modules, which is lightweight and can be run quickly to get the results.

图9 评估报告的概览

(二) MCnebula的拓展

在上一部分的算法阐述（见：第一部分 MCnebula的方法构建 > 二、结果 > (二) MCnebula的算法）和工作流阐述（图2）中，我们介绍了MCnebula的主体部分，它用于整合注释数据和归类可视化方面

的内容。以下我们阐述MCnebula的拓展功能，将MCnebula的算法运用于更广的方面，允许高级的自定义分析（以下内容涉及：统计分析、‘Features’筛选（Feature selection）、聚焦关键代谢物（化合物）及其结构特征、通路富集、查询化合物等）。

1. 用于化学发现

表13为拓展MCnebula用于化学信息查询的函数（应用请参考：第四部分 基于MCnebula策略的血清代谢组学 > 二、结果或者第三部分 基于MCnebula策略分析杜仲炮制前后的成分变化 > 二、结果）。

表 13 拓展MCnebula用于化学发现的函数

Group	Name	Description
Query inchikey	query_inchikey	批量多线程模式，通过PubChem API使用InChiKey2D（InChiKey的第一个哈希块）查询所有的InChiKey
	pubchem_get_inchikey	同上，查询单个InChiKey
Query classification	query_classification	批量多线程模式，通过ClassyFire在R的API使用InChiKey查询所有的化学分类系统
	classyfire_get_classification	同上，查询单个InChiKey
Query synonyms	query_synonyms	批量多线程模式，通过PubChem API使用InChiKey查询所有可得的化合物同义名
	pubchem_get_synonyms	同上，查询单个InChiKey
Query others	query_iupac	批量多线程模式，通过PubChem API使用InChiKey查询所有可得的化合物IUPAC名
Pick annotation	pick_class	从'query_classification'获得的数据选取唯一结果
	.filter_pick.class	'pick_class'预设的选取方法
	PickClass	与'pick_class'相关
	pick_synonym	从'query_synonyms'获取的数据中选取同义名（根据正则匹配的先后顺序）
Output identification	.filter_pick.general	'query_synonyms'预设的过滤方法
	rename_table	'format_table'的变形，仅变换表格的列名称
	format_table	根据预设的方法，调用'dplyr::*'系列函数对表格塑形（依次为：filter, arrange, distinct, mutate, select, rename），是用于MCnebula注释数据集输出的快速便捷的方法
Plot EIC stack	.filter_format	'format_table'预设的一系列参数
	plot_EIC_stack	用于绘制堆叠的EIC图，与'plot_msms_mirrors'相对应

2. 用于代谢组数据分析

表14为拓展MCnebula用于代谢组数据分析的函数（应用请参考：第四部分 基于MCnebula策略的血清代谢组学>二、结果）。

表 14 拓展MCnebula用于代谢组学分析的函数

Group	Name	Description
Alignment merge	align_merge	根据得分公式合并包含m/z和RT的两个矩阵： $Score = (1 - rt.difference / rt.tolerance) * rt.weight + (1 - mz.defference / mz.tolerance) * mz.weight$
	tol_merge	容差合并，'align_merge'的基础函数。本函数不使用for循环（因为R的循环速度太慢）而是用'merge'进行合并，这涉及一个'分箱'聚类的过程，聚类两次（向上约分和向下约分），将'箱'大小内的数值给定一个编号，进行两次合并，再根据容差设定滤除不匹配的数据。
Cross select	select_features	'Features'筛选算法，在'features_annotation'、'nebula_index'数据集和'top_table'数据集中交叉筛选，准则有化学类、Q-value（P-value的矫正）、Log(Fold Change)、Tanimoto similarity等
	cross_select	'select_features'的基础函数，'dplyr::*'系列函数的包装
	plot_heatmap	结合MCnebula2 R包快速绘制热图的函数
Dot heatmap	handling_na	为绘制热图前处理矩阵的算法：对于每一个数据子集（'Features'量化矩阵的各个分组），缺失的数值将被填充为平均值；如果这组数据都是缺失的数值，它们将被填充为零
	log_trans	为绘制热图前处理矩阵的算法：将宽数据（wide data.frame）转换为长数据（long data.frame）；对数值进行对数转换；如果有一个数值为0，则用数值列的最小值的1/10替换（设定极限值）
	init_fella	建立FELLA包用于富集的数据集
Pathway enrichment	load_fella	载入预建立的FELLA包的数据集
	enrich_fella	以多组'Features'为单位，使用FELLA包进行通路富集
	graph_fella	从富集完毕的数据中获取'pagerank'，'diffusion'，'hypergeom'的富集图的数据
	plotGraph_fella	使用ggplot2绘制富集图
	cid.to.kegg	使用'MetaboAnalystR'R包将PubChem的CID转化为KEGG的ID用于富集分析

三、小结

对于鉴定，光谱库匹配仍然是LC-MS/MS数据的主要方法，因为它具有很高的准确性。一般的化合物分类也是基于此，即首先通过光谱匹配来识别化学结构，然后根据化学结构来评估其化学类别。考虑到参考光谱库的局限性，像CANOPUS^[34]这样的分类技术应用于MCnebula中绕过了识别化学结构的第一步，而是预测可能的化学类别，即使确切的化学结构不为人所知。MCnebula将这一尖端技术与ABC选择算法相结合，实现了Child-Nebulae的可视化，这使得探索光谱库以外的未知化合物成为可能。我们将MCnebula的分类方法与GNPS进行了比较，GNPS的方法依赖于化学结构鉴定。当加入不同程度的噪声时，GNPS的分类化合物的数量与MCnebula的稳定表现相比明显减少。对于实际获得的MS/MS图谱，它们不如参考图谱好，而且含有一些噪声；事实上，现实中的MS/MS光谱更接近于有噪声的情况。这意味着MCnebula可以在一定程度上抵御噪声的干扰。在评估的最后，我们检查了MCnebula的鉴定精度。结果证实，鉴定的准确性在70%左右波动，与SIRIUS报道的一致^[22]。

用于MCnebula拓展的函数（表13和表14）被整理至额外的包‘exMCnebula2’
(<https://github.com/Cao-lab-zcmu/exMCnebula2>)。这些函数可配合MCnebula的R包
(MCnebula2) 使用，拓展MCnebula的应用，我们在随后的内容中示例了它们。

第三部分 基于MCnebula策略分析杜仲炮制前后的成分变化

一、材料与方法

（一）实验材料

杜仲 (*E. ulmoides*) 干树皮来自浙江佐力药业股份有限公司。

（二）实验方法

1. 制备炮制前后的杜仲

生杜仲和盐杜仲的样品制备方法如下。（1）生杜仲。取*E. ulmoides*干树皮的碎片或块状物，将其打成粉末，通过80目筛子以备进一步处理。（2）盐杜仲。将*E. ulmoides*干树皮的丝或块用盐水（盐的用量为*E. ulmoides*的2%，加10倍水溶解）均匀喷洒，并在密闭状态下闷30 min；然后，将树皮在60°C 的烤箱中烘干，接着在140°C 下烘烤60 min；最后，将烘烤过的树皮打成粉末，通过80目筛子以备进一步加工。

2. 制备LC-MS的杜仲样品

分别称取2 g生杜仲粉和盐杜仲粉，加入50 ml甲醇/水（1:1, v/v），然后进行超声波（20 kHz, 40 min）。超声后，将混合物过滤，得到滤液和残余物。残余物加入50 ml甲醇/水（1:1, v/v），再次用超声波（40 kHz, 250 W, 20 min）提取。混合物被过滤。然后，合并两种提取液的滤液，蒸发掉溶剂。加入甲醇/水（1:1, v/v）以重新溶解提取物，并将体积定容为5 ml。最后，通过离心（12,000 r.p.m., 10 min）得到上清液，用于进一步的LC-MS分析。

3. LC-MS/MS实验条件

LC-MS分析使用Dionex Ultimate 3000 UHPLC系统（Dionex, Germany），结合高分辨率傅立叶变换质谱仪（Orbitrap Elite, Thermo Fisher Scientific, Germany），使用Waters Acuity HSS T3柱（1.8 μm , 100 mm \times 2.1 mm, Waters公司, Milford, MA, USA）。溶剂A, 甲酸/水（0.1:99, v/v）和溶剂B, 甲酸/乙腈（0.1:99, v/v），作为流动相。分离的梯度曲线如下：0 min时2%的溶剂B, 2 min时5%的溶剂B, 10 min时15%的溶剂B, 15 min时25%的溶剂B, 18 min时50%的溶剂B, 23 min时100%的溶剂B, 25 min时2%的溶剂B, 30 min时2%的溶剂B。流速为0.3 ml/min。柱温设定在40°C。质谱分析使用配备ESI源的Orbitrap Elite仪器（Thermo Fisher Scientific, Germany）进行，在负电离模式下操作。ESI源在50°C下运行，毛细管温度为275°C，电离电压为3.5 kV，鞘内气体流量为35 L/min。调查扫描在Orbitrap质量分析器中进行，在120,000（半最大全宽）分辨率下操作。调查扫描的质量范围为100-1500 m/z，归一化碰撞能量为30 eV。分析方法被设定为分析调查扫描中信号最强的前10个离子，并启用了15秒的动态排除法。

二、结果

（一）MCnebula对中药数据集的基础分析

我们用MCnebula阐述了一味代表性中药——杜仲——在传统加工过程中涉及的化学结构多样性和化学转化。杜仲是*Eucommia ulmoides Oliv.* (*E. ulmoides*) 的树皮^[49]。杜仲经盐制，在历史上长期以来被普遍应用于治疗肾脏疾病，但其化学基础仍有待探索。现在利用MCnebula工作流对炮制前后的*E. ulmoides*进行分析。在ABC选择算法的帮助下，共获得了29个代表*E. ulmoides*的丰富成分的化学类别。随后对炮制前后的半定量数据进行了二元比较。使用函数‘select_features’ ($|\text{Log2}(\text{Fold change})| > 0.3$, $\text{Q-value} < 0.05$, $\text{Tanimoto similarity} > 0.5$) 选出前20个‘Features’（Top20），并在Child-Nebulae中进行追踪（图10）。

以MCnebula绘制Top20的MS/MS图谱和提取离子色谱图（EIC）（图11和图12）。根据图12，推测ID1642、1785和2321的‘Features’是新生成的化合物，因为与处理后相比，处理前的峰面积水平几乎为零。它们的化学结构见图11。其中，ID1642的‘Features’具有较高的正确识别概率（Tanimoto

similarity: 0.69）。根据图10，我们知道ID 1642属于‘Iridoids and derivatives’（IAD），其他的是‘Dialkyl ethers’（DE；ID 1785）和‘Phenylpropanoids and polyketides’（PAP；ID 2321）。

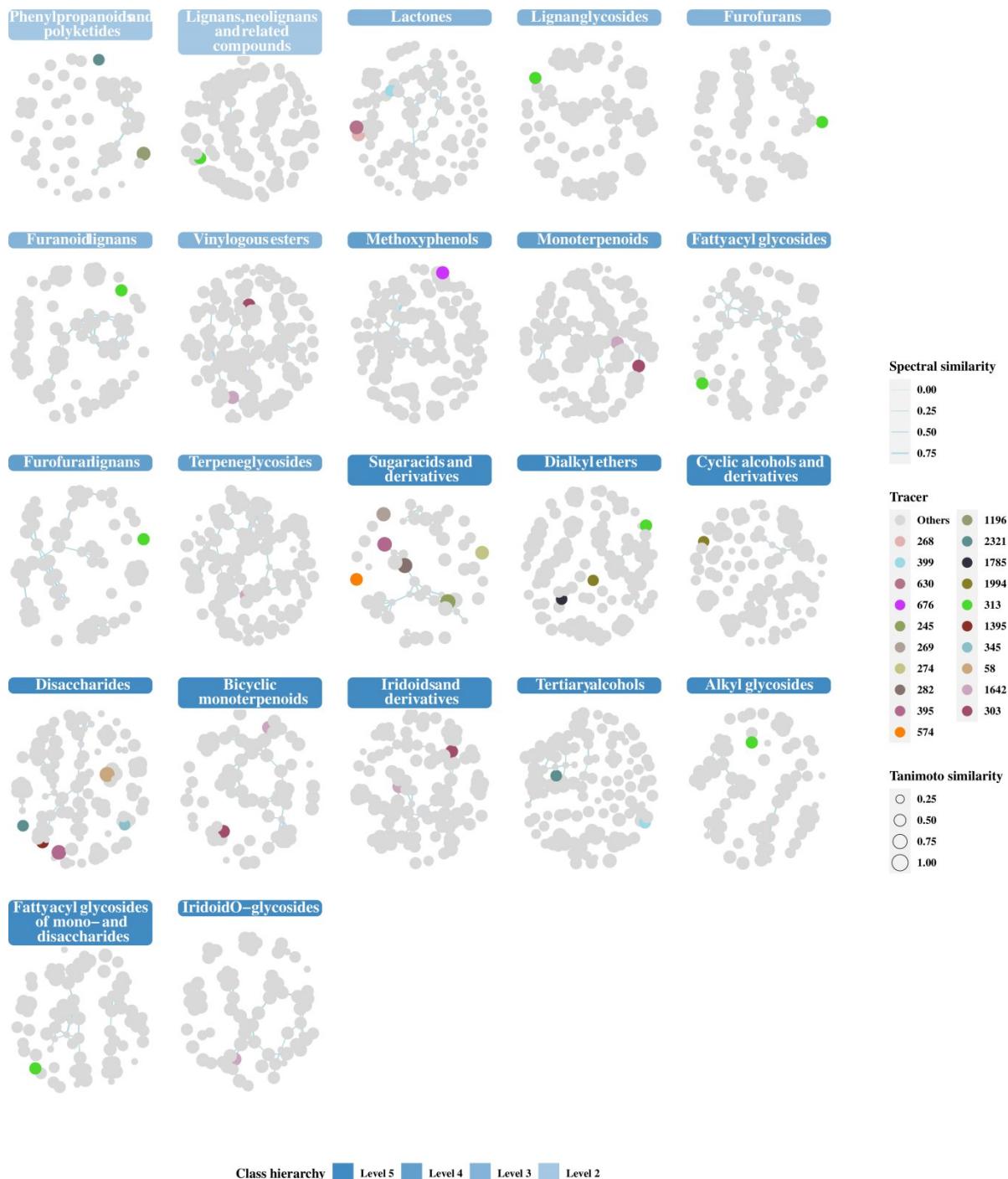


图10 在中药数据集的Child-Nebulae中追踪Top ‘Features’

- 图10注：根据统计分析的‘Features’排名，Top ‘Features’在Child-Nebulae中用不同的颜色标记。

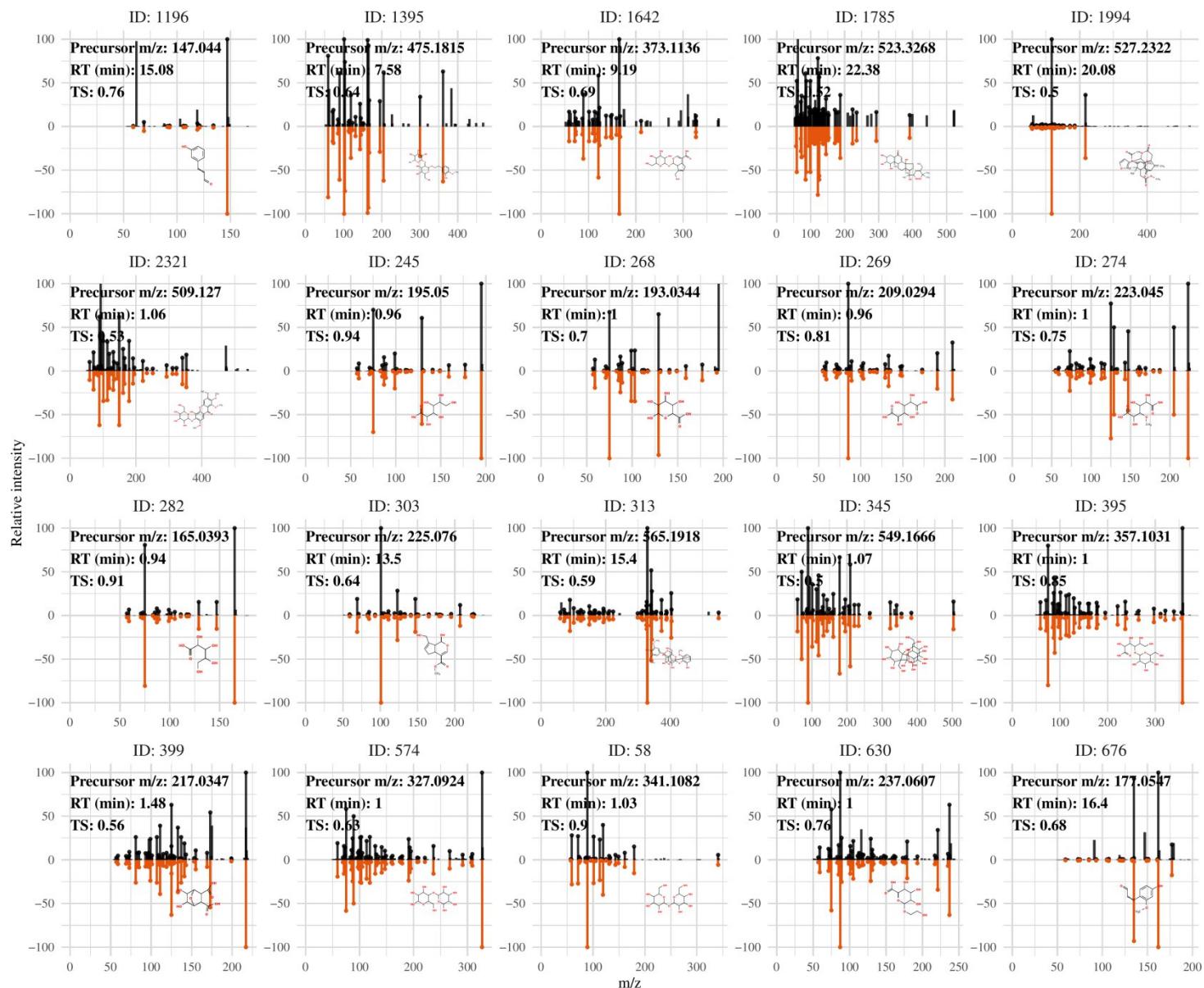


图11 中药数据集Top ‘Features’ 的MS/MS图

- 图11注：对于Top ‘Features’，镜像的MS/MS光谱图说明了原始的MS/MS光谱（黑色）和SIRIUS预测的噪声过滤的MS/MS光谱（红色）。条形图上面的点反映了相应的关系。‘Features’ 的化学结构的最佳候选项被映射到该图中。

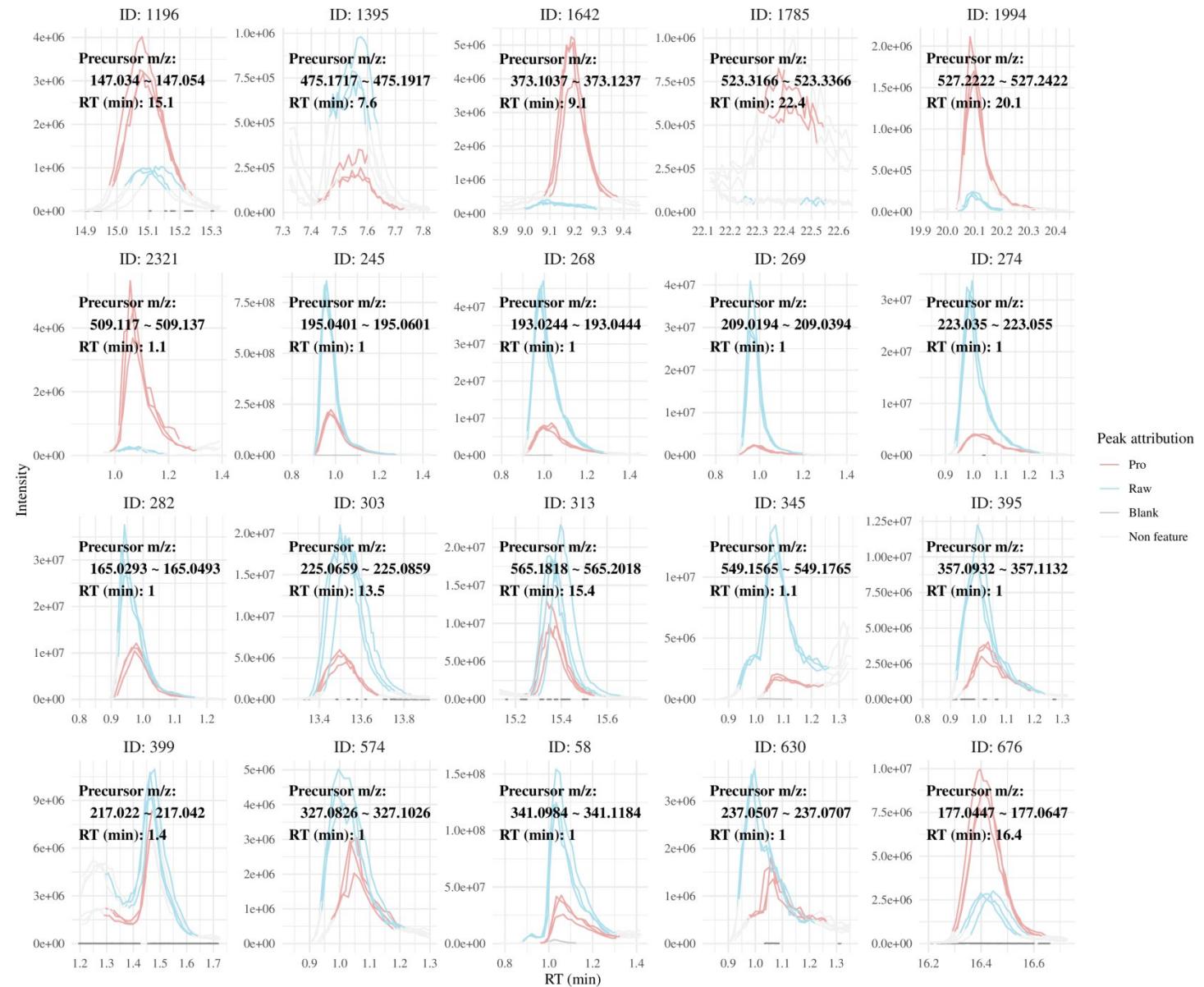


图12 中药数据集Top ‘Features’ 的EIC图

- 图12注：EIC图说明了Top 'Features'的原始峰形（通过MCnebula绘制；通过MZmine2的自动数据分析管道（ADAP）算法检测）。

（二）MCnebula对中药数据集的聚焦分析

分别对IAD、DE和PAP的Child-Nebulae进行了深度的注释。对ID 1642、1785和2321的‘Features’在Child-Nebulae中的位置进行了聚焦分析（图13a、b和c）。只有ID 1642的‘Features’有相邻的‘Features’，其识别的化学结构（ID 2110和ID 854）的母核相似。ID 2110和ID 854的‘Features’的化学结构被鉴定（Tanimoto similarity：分别为0.69和0.70）（图13d，e和f）；它们的峰面积水平在处理后有所下降和增加。根据图13d和e所示的化学结构，我们推测ID 2110的化合物在加工后部分转化为ID 854的化合物，这可能涉及化学变化如脱水和重排。这种推测解释了峰面积水平的改变。此外，化合物ID 1642（其光谱显示在图11和图12中）含量的增加也可能与化合物ID 2110的减少有关。

我们所展示的MCnebula发现重要化合物和发现化学变化的方法可以应用于探索更多的化合物（表15），但此处不再展开描述。

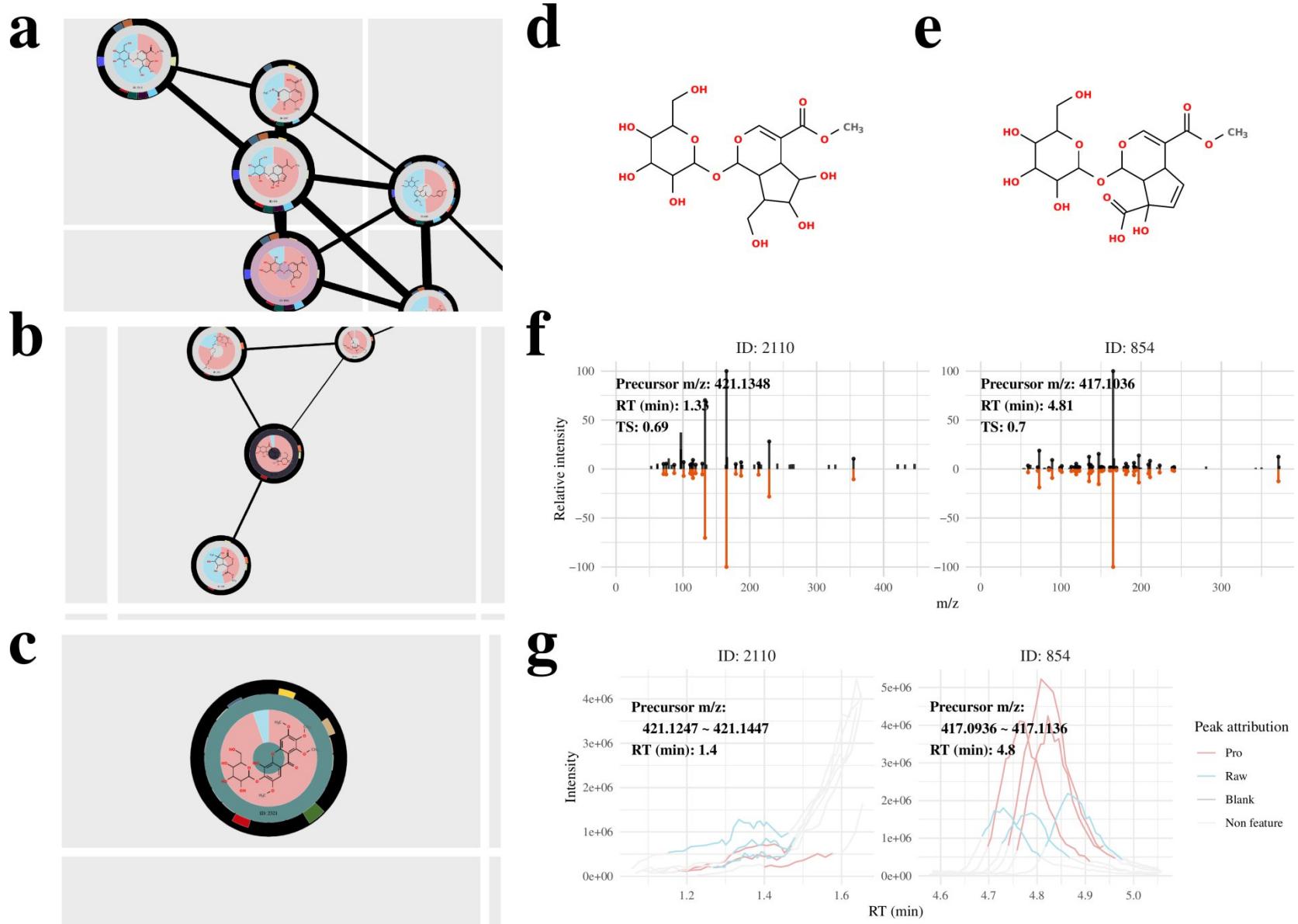


图13 在中药数据集聚焦于Child-Nebulae中的Top 'Features'的方位

- 图13注：**a**, **b**和**c**分别说明了‘Iridoids and derivatives’、‘Dialkyl ethers’或‘Phenylpropanoids and polyketides’深度注释的Child-Nebulae的局部视图。**d**和**e**分别表示ID 2110和ID 854的‘Features’（化合物）的化学结构。**f**和**g**显示了‘Features’的MS/MS图谱（参考图11的描述）和EIC图（参考图12的描述）。

表 15 MCnebula工作流鉴定的中药数据集的化合物 (Q-value < 0.05)

Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar	ClassyFire Class	Sig.	Var.
ethyl-dUTP	1644	494.9989	2.6	1.0	C11H19N2O14 P3	[M - H]-	0.67	HSZWVJYQEO ARCE	Pyrimidine deoxyribonucleoside triphosphates	***	↑↑
...	2339	555.0702	9.8	1.1	C15H22N6O13 P2	[M - H]-	0.57	YXPVCOGBO RSVTN	-	***	↑↑
D-sulfolactate	288	168.9800	-7.2	1.0	C3H6O6S	[M - H]-	0.81	CQQGIWJSIC OUON	-	***	↓↓
...	399	217.0347	-3.1	1.5	C8H10O7	[M - H]-	0.56	HZUPENFESXI QFM	Dicarboxylic acids and derivatives	***	↓↓
...	1655	327.1081	-1.3	9.2	C15H20O8	[M - H]-	0.71	CTBTYMWZD WFXTH	Phenolic glycosides	***	↑↑
...	1639	131.0338	-8.8	1.7	C5H8O4	[M - H]-	0.60	OIWJREZAJZV GQN	Hydroxy fatty acids	***	↑↑
AltA	268	193.0344	-5.1	1.0	C6H10O7	[M - H]-	0.70	AEMOLEFTQB MNLQ	Glucuronic acid derivatives	***	↓↓
dec-6-enoate	482	169.1223	-6.8	21.2	C10H18O2	[M - H]-	0.53	IZOFWCYKC WUJBY	Medium-chain fatty acids	***	↓↓
Geniposidinsaure	1642	373.1136	-1.2	9.2	C16H22O10	[M - H]-	0.69	ZJDOESGVOW AULF	Iridoid O-glycosides	***	↑↑
4-O-methyl-d-glucaric acid	274	223.0450	-4.2	1.0	C7H12O8	[M - H]-	0.75	MDGDZMIEJC CPTI	Glucuronic acid derivatives	***	↓↓

Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar	ClassyFire Class	Sig.	Var.
digalactosyl ononitol	345	549.1666	-1.2	1.1	C19H34O18	[M - H]-	0.50	HUMKCYVGI CAQIT	C-glycosyl compounds	***	↓↓
1-deoxyribitol-5-monophosphate	267	215.0320	-2.9	1.0	C5H13O7P	[M - H]-	0.68	YPXGTKHZRC DZTL	Monoalkyl phosphates	***	↑↑
galactonate	245	195.0500	-5.1	1.0	C6H12O7	[M - H]-	0.94	RGHNXZEOK UKBD	Medium-chain hydroxy acids and derivatives	***	↓↓
cellobionate	395	357.1031	-2.2	1.0	C12H22O12	[M - H]-	0.85	JYTUSYBCFIZ PBE	Fatty acyl glycosides of mono- and disaccharides	***	↓↓
7-methyl-9-oxodec-7-enoic acid	761	197.1174	-4.9	21.2	C11H18O3	[M - H]-	0.65	DEIQQSNDU HJNT	Medium-chain fatty acids	***	↑↑
Hydroxyhydroquinone	1535	125.0232	-9.5	3.1	C6H6O3	[M - H]-	0.83	GGNQRNBDZ QJCCN	Hydroxyquinols and derivatives	***	↑↑
Adenox	385	264.0719	-7.3	1.0	C10H11N5O4	[M - H]-	0.56	ILMNNSCQOSG KTNZ	Purines and purine derivatives	***	↓↓
...	2331	361.0900	-8.0	1.1	C18H18O8	[M - H]-	0.53	UKGMMLDJBE MVQLQ	Phenolic glycosides	***	↑↑
Genipin	303	225.0760	-3.9	13.5	C11H14O5	[M - H]-	0.64	AZKVVWQKM DGGDSV	Iridoids and derivatives	***	↓↓
cipatrijugin A	1994	527.2322	6.6	20.1	C29H36O9	[M - H]-	0.50	CABATAPDJF TPNH	Tricarboxylic acids and derivatives	***	↑↑
TS-70	414	453.1006	-4.9	1.0	C14H23N4O11 P	[M - H]-	0.51	VBXZSBKAJF XURR	Xanthines	***	↓↓
3-hydroxycinnamaldehyde	1196	147.0440	-8.1	15.1	C9H8O2	[M - H]-	0.76	DCHPWNOJPJ RSEA	Cinnamaldehydes	***	↑↑

Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar	ClassyFire Class	Sig.	Var.
saccharate	269	209.0294	-4.4	1.0	C6H10O8	[M - H]-	0.81	DSLZVSRJTY RBFB	Glucuronic acid derivatives	***	↓↓
Sophorotriose	410	503.1610	-1.6	1.1	C18H32O16	[M - H]-	0.80	UQBIAGWOJD EOMN	Oligosaccharides	**	↓↓
Lactotrehalose	58	341.1082	-2.1	1.0	C12H22O11	[M - H]-	0.90	HDTRYLNUV ZCQOY	O-glycosyl compounds	**	↓↓
...	451	101.0231	-13.4	0.9	C4H6O3	[M - H]-	0.56	VTERBIYJBW DXDT	Beta-hydroxy aldehydes	**	↓↓
2-Hydroxy-3-methylbenzaldehyde	1271	135.0440	-8.8	4.5	C8H8O2	[M - H]-	0.59	IPPNXSAJZO TJZ	Aryl-aldehydes	**	↑↑
Ribonate	282	165.0393	-7.1	0.9	C5H10O6	[M - H]-	0.91	QXKAIJAYHK CRRA	Sugar acids and derivatives	**	↓↓
Octose	431	239.0764	-3.6	1.0	C8H16O8	[M - H]-	0.63	ZEPAXLPHES YSJU	Octoses	**	↓↓
...	2321	509.1270	-6.1	1.1	C23H26O13	[M - H]-	0.53	FGJBUXQEYD EVAL	Xanthenes	**	↑↑
Atranol	1178	151.0390	-6.9	3.5	C8H8O3	[M - H]-	0.56	JASONGFGOL HLGB	Aryl-aldehydes	**	↓↓
...	1785	523.3268	-1.6	22.4	C29H48O8	[M - H]-	0.52	KWOLRPNEM KAFDT	Stigmastanes and derivatives	**	↑↑
...	1395	475.1815	-1.3	7.6	C21H32O12	[M - H]-	0.64	OSEBKVVIVK EDHT	O-glycosyl compounds	**	↓↓
...	630	237.0607	-3.9	1.0	C8H14O8	[M - H]-	0.76	ZPOANLXNW QACPS	Glucuronic acid derivatives	**	↓↓
Gluceptate	281	225.0608	-3.6	0.9	C7H14O8	[M - H]-	0.56	KWMLJOLKU	Sugar acids and	**	↓↓

Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar	ClassyFire Class	Sig.	Var.
								YYJFJ	derivatives		
2-deoxycellobiose	574	327.0924	-2.6	1.0	C11H20O11	[M - H]-	0.63	ZLSHPOOYAK KJSB	O-glycosyl compounds	**	↓↓
...	313	565.1918	-1.6	15.4	C27H34O13	[M - H]-	0.59	GZMNUFHJD KKCRX	Lignan glycosides	**	↓↓
...	484	391.1244	-0.5	3.4	C16H24O11	[M - H]-	0.76	BGKSXZCAO MVVOF	Iridoid O-glycosides	**	↓↓
...	447	307.1911	-1.3	20.9	C18H28O4	[M - H]-	0.50	FCCFXOHVER VHQR	Lineolic acids and derivatives	**	↓↓
...	676	177.0547	-5.9	16.4	C10H10O3	[M - H]-	0.68	MRCGVXARH KOYKU	Methoxyphenols	**	↑↑
4ciw	328	217.0709	-3.9	1.5	C9H14O6	[M - H]-	0.56	HVNZLWXJZR WNNG	Cyclitols and derivatives	**	↓↓
saccharolactone	297	191.0189	-4.3	1.0	C6H8O7	[M - H]-	0.75	XECPAIJNBX COBO	Gamma butyrolactones	**	↓↓
1b-Hydroxycholate	1939	423.2747	-1.2	23.6	C24H40O6	[M - H]-	0.55	UYVVLXVBE QAATF	Tetrahydroxy bile acids, alcohols and derivatives	**	↑↑
aldehydo-L-arabinuronate	381	163.0237	-7.1	1.1	C5H8O6	[M - H]-	0.68	VQUZNVATT CZTQO	Sugar acids and derivatives	**	↓↓
...	925	231.1231	-3.0	19.4	C11H20O5	[M - H]-	0.51	MFSMRLNNR GLWQD	Medium-chain hydroxy acids and derivatives	**	↓↓
D-xylobionate	454	297.0820	-2.5	1.0	C10H18O10	[M - H]-	0.68	DGXURXUMS DNLNT	O-glycosyl compounds	**	↓↓
Ficusesquilignan B	641	583.2178	-1.2	19.3	C31H36O11	[M - H]-	0.56	GWAGNJOOC	Furanoid lignans	**	↓↓

Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar	ClassyFire Class	Sig.	Var.
								ODQGM			
...	921	261.1340	-1.3	15.1	C12H22O6	[M - H]-	0.59	BQRISQXLSX CXSC	Medium-chain fatty acids	**	↓↓
...	276	185.0809	-5.4	6.3	C9H14O4	[M - H]-	0.56	BVHSPGBGH GXBMY	Iridoids and derivatives	**	↓↓
O-Gluur-gal	518	355.0874	-2.2	1.0	C12H20O12	[M - H]-	0.71	SYZNAVOGL CTGII	Fatty acyl glycosides of mono- and disaccharides	**	↓↓
Tanegool	592	375.1445	-1.3	15.7	C20H24O7	[M - H]-	0.60	MWQRAOGW LXTMIC	Tetrahydrofuran lignans	**	↓↓
Orcinoside A	287	583.2025	-1.3	12.1	C27H36O14	[M - H]-	0.60	SHOPWAJXD MLALR	Phenolic glycosides	**	↓
4-oxononanoate	1110	171.1016	-6.4	19.7	C9H16O3	[M - H]-	0.74	PRDIIROHTW NJDB	Medium-chain keto acids and derivatives	**	↓↓
Cytospolide F	1299	285.1703	-1.4	17.6	C15H26O5	[M - H]-	0.53	ATNGVHJGYI WYSL	Oxocins	**	↑↑
inositol	262	179.0551	-5.7	0.9	C6H12O6	[M - H]-	0.60	CDAISMWEU UEBRE	Cyclohexanols	**	↓↓
...	938	719.2399	-0.7	9.1	C31H44O19	[M - H]-	0.68	FDHKIHNRBM ODHH	Iridoid O-glycosides	**	↓↓
Methyl-D-galabioside	636	355.1239	-2.0	1.1	C13H24O11	[M - H]-	0.61	FHNIYFZSHC GBPP	O-glycosyl compounds	**	↓↓
p-Hydroxyphenylacetaldehyde	280	135.0440	-8.9	10.2	C8H8O2	[M - H]-	0.56	IPRPPFIAVHP VJH	Phenylacetaldehydes	**	↓
3-oxooctanoate	420	157.0858	-7.6	20.2	C8H14O3	[M - H]-	0.66	FWNRRWJFO	Medium-chain keto	**	↓↓

Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar	ClassyFire Class	Sig.	Var.
								ZIGQZ	acids and derivatives		
Undecanedioic acid, 5-oxo-	862	229.1075	-3.1	18.6	C11H18O5	[M - H]-	0.56	MEEMMABYF KSQGY	Medium-chain fatty acids	**	↓↓
Anhydroarabinose	598	149.0444	-7.7	0.9	C5H10O5	[M - H]-	0.74	PYMYPHUHK UWMLA	Monosaccharides	**	↓↓
bisvertinoquinol	1866	497.2227	9.3	20.3	C28H34O8	[M - H]-	0.57	ZXWVVZIMJS PORF	Cyclic ketones	**	↑↑
9,12-octadecadiynoic acid	549	275.2012	-1.8	23.3	C18H28O2	[M - H]-	0.55	KDYILQLPKV ZDGB	Long-chain fatty acids	**	↓↓
Geniposide	351	387.1292	-1.3	12.4	C17H24O10	[M - H]-	0.73	IBFYXTRXDN APMM	Iridoid O-glycosides	**	↓
L-beta-hydroxymyristate	1384	243.1960	-2.4	24.1	C14H28O3	[M - H]-	0.79	ATRNZOYKS NPPBF	Long-chain fatty acids	**	↑↑
...	448	307.1911	-1.3	21.5	C18H28O4	[M - H]-	0.53	XVRGWTKW KCRREP	Lineolic acids and derivatives	*	↓↓
...	1807	453.2853	-1.1	21.8	C25H42O7	[M - H]-	0.57	VBPODEAYFU FNNY	Leucothol and grayanotoxane diterpenoids	*	↑↑
Gulonolactone	607	177.0394	-6.0	0.9	C6H10O6	[M - H]-	0.65	SXZYCXMUP BBULW	Gamma butyrolactones	*	↓
...	1692	177.0396	-5.2	1.9	C6H10O6	[M - H]-	0.72	ZDDQAAZBPZ GPRB	Alpha hydroxy acids and derivatives	*	↑↑
...	1560	377.2092	-1.7	23.9	C18H35O6P	[M - H]-	0.66	MEIOLAIVIVF AFI	Long-chain fatty acids	*	↑↑
...	603	181.0497	-5.5	8.8	C9H10O4	[M - H]-	0.53	YRJIMTHQIPO SJR	O-methoxybenzoic acids and derivatives	*	↑↑

Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar	ClassyFire Class	Sig.	Var.
...	2379	437.2903	-1.3	21.8	C25H42O6	[M - H]-	0.56	SYHUKNSTHJ ICMX	Tetrahydroxy bile acids, alcohols and derivatives	*	↑↑
Klyxumine A	1778	455.2645	-1.1	20.2	C24H40O8	[M - H]-	0.51	MXPJWUTXZS OEAT	Eunicellane and asbestinane diterpenoids	*	↑↑
Asebotoxin I	1751	425.2541	-1.0	20.8	C23H38O7	[M - H]-	0.52	UVIOAKNWF GGRCJ	Leucothol and grayanotoxane diterpenoids	*	↑↑
...	785	499.2697	-0.9	22.7	C29H40O7	[M - H]-	0.59	SYUSCNJPCS QFSK	Dicarboxylic acids and derivatives	*	↓
...	778	205.0342	-5.6	1.1	C7H10O7	[M - H]-	0.64	FHTUHRLYAP RFEZ	Sugar acids and derivatives	*	↓
...	854	417.1036	-0.6	4.8	C17H22O12	[M - H]-	0.70	MLZKYGRNL PVMHX	Iridoid O-glycosides	*	↑
...	945	403.1605	-1.3	2.6	C18H28O10	[M - H]-	0.50	MQSFSEGMDJ UVRU	Iridoid O-glycosides	*	↓↓
Oxypaeoniflorin	771	495.1505	-0.7	13.1	C23H28O12	[M - H]-	0.57	FCHVXNVDF YXLIL	Terpene glycosides	*	↓
plantagineoside C	1060	507.1864	-1.6	18.1	C25H32O11	[M - H]-	0.56	KNGLPVKCH BEQMT	Linear diarylheptanoids	*	↓↓
7,8-diketopelargonic acid	819	185.0808	-6.0	12.2	C9H14O4	[M - H]-	0.66	LUCODFUPVS YZRC	Medium-chain keto acids and derivatives	*	↑
...	577	595.2025	-1.2	15.8	C28H36O14	[M - H]-	0.56	OCUPPIIMZPH ETJ	Lignan glycosides	*	↓
...	528	185.0446	-4.9	1.9	C8H10O5	[M - H]-	0.50	GVLJZCSVFF	Carbonic acid	*	↓

Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar	ClassyFire Class	Sig.	Var.
								YGON	diesters		
...	459	341.1964	-1.7	19.9	C18H30O6	[M - H]-	0.55	DRPDZKNSZV IOQR	Long-chain fatty acids	*	↑
Cellobiosan	2345	323.0980	-1.1	1.5	C12H20O10	[M - H]-	0.75	LTYZUJSCZCP GHH	O-glycosyl compounds	*	↑↑
Muralioside	1210	379.1244	-0.6	4.6	C15H24O11	[M - H]-	0.61	DCMBJUSPXC DZSO	Iridoid O-glycosides	*	↓
Botryosphaerinone	555	209.1175	-3.9	18.7	C12H18O3	[M - H]-	0.63	AUAYBDNKD WXWEX	Cyclic ketones	*	↓↓
...	596	363.2148	-7.9	23.9	C21H32O5	[M - H]-	0.50	NKOQQNIPH LJDL	Fatty alcohol esters	*	↓↓
...	504	165.0548	-5.8	3.0	C9H10O3	[M - H]-	0.52	QAWBTVDGX WOTRD	Aryl ketones	*	↓
...	738	311.0977	-2.2	1.1	C11H20O10	[M - H]-	0.76	AXHPKHDTO XXPGU	Fatty acyl glycosides of mono- and disaccharides	*	↓
...	1653	163.0388	-7.7	12.2	C9H8O3	[M - H]-	0.56	NFZXTKAGXS VCRD	Cinnamaldehydes	*	↑
Wikstroemol	857	405.1549	-1.6	17.4	C21H26O8	[M - H]-	0.55	JVQPMMSYMH ZSFNV	Lignans, neolignans and related compounds	*	↓
m-Hydroxybenzaldehyde	631	121.0282	-11.0	10.1	C7H6O2	[M - H]-	0.78	IAVREABSGI HHMO	Benzoyl derivatives	*	↑
...	2349	369.1038	-0.3	1.5	C13H22O12	[M - H]-	0.61	WRWIKVNJQ WUUJX	Glucuronic acid derivatives	*	↑↑
6,11-dioxododecanoic	595	227.1282	-3.0	18.7	C12H20O4	[M - H]-	0.69	OIAVTFGRDN	Medium-chain keto	*	↓↓

Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar	ClassyFire Class	Sig.	Var.
acid								UKCM	acids and derivatives		
Threose	427	119.0336	-11.2	0.9	C4H8O4	[M - H]-	0.83	YTBSYETUW UMLBZ	Monosaccharides	*	↓↓
Macedonine	599	213.0760	-3.9	6.9	C10H14O5	[M - H]-	0.56	DIIADJQOLF WUFJ	Iridoids and derivatives	*	↓
...	467	215.0554	-3.4	2.6	C9H12O6	[M - H]-	0.51	FTHDNRBKSL BLDA	Tricarboxylic acids and derivatives	*	↓
Threonate	284	135.0286	-9.8	1.0	C4H8O5	[M - H]-	0.88	JPIJQSOTBSS VTP	Sugar acids and derivatives	*	↓
...	1023	439.0854	-6.5	8.0	C19H20O12	[M - H]-	0.55	WVHDGXKUZ IDTIN	Phenolic glycosides	*	↑
YOEUNKPREOJHBW-UHFFFAOYSA-	1247	165.0183	-6.5	7.8	C8H6O4	[M - H]-	0.56	YOEUNKPRE OJHBW	Benzoin acids	*	↑
5-Vinylresorcinol	1265	135.0439	-9.2	16.2	C8H8O2	[M - H]-	0.63	LYECQNYXC RATFL	Resorcinols	*	↑↑
Cyclopassifloic acid E	1739	551.3581	-1.6	23.8	C31H52O8	[M - H]-	0.57	DXAVXXNAJI NCIJ	Cycloartanols and derivatives	*	↑↑
...	229	329.2328	-1.6	20.6	C18H34O5	[M - H]-	0.73	MDIUMSLCYI JBQC	Long-chain fatty acids	*	↓
1jlx	652	472.1815	-2.0	12.3	C21H31NO11	[M - H]-	0.54	MYDRTQFLX CNCAG	N-acyl-alpha-hexosamines	*	↓
Crescentin II	266	201.0761	-4.0	2.6	C9H14O5	[M - H]-	0.63	HYJIDONKRR KLEJ	Iridoids and derivatives	*	↓
Gallicynoic Acid E	372	325.2017	-1.2	21.2	C18H30O5	[M - H]-	0.56	IRDUBXGVW QSMKQ	Long-chain fatty acids	*	↓↓

Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar	ClassyFire Class	Sig.	Var.
...	1708	285.1704	-1.4	18.0	C15H26O5	[M - H]-	0.52	XHJITGZAWU NULF	Long-chain fatty acids	*	↑
...	526	147.0652	-7.2	3.7	C6H12O4	[M - H]-	0.50	JIXHYWCLUO GIMM	Carboxylic acids	*	↓
...	712	197.1174	-4.8	19.2	C11H18O3	[M - H]-	0.52	PGTPCJJMOJC WHV	Medium-chain fatty acids	*	↓
7-hydroxysecoisolariciresinol	928	377.1600	-1.6	15.8	C20H26O7	[M - H]-	0.55	VPDBTIFHPU YJJJ	Dibenzylbutanediol lignans	*	↓
Asaolaside	1120	613.2129	-1.5	12.5	C28H38O15	[M - H]-	0.51	PNJYEHPHHQ UERT	Terpene glycosides	*	↓
Heptose	640	209.0657	-4.9	1.1	C7H14O7	[M - H]-	0.63	YPZMPEPLW KRVLD	Heptoses	*	↓
(-)-8-hydroxyjasmonic acid	845	225.1125	-3.2	13.0	C12H18O4	[M - H]-	0.69	IQGLAWZCM QYBPA	Jasmonic acids	*	↓
nonanedioate	1683	187.0965	-5.9	16.6	C9H16O4	[M - H]-	0.54	BDJRBEYXGG NYIS	Medium-chain fatty acids	*	↓
...	721	699.2135	-1.0	14.1	C31H40O18	[M - H]-	0.51	WHSZKNNBJ ANJCB	Fatty acyl glycosides of mono- and disaccharides	*	↓
...	2209	517.1395	-2.9	0.9	C18H30O17	[M - H]-	0.85	SHNIAUOFCS LFGS	Oligosaccharides	*	↓↓
...	1300	275.1495	-1.8	18.6	C13H24O6	[M - H]-	0.57	RLLLKKWXYJ LXOU	Fatty acid esters	*	↓
9-hydroperoxylinolenic acid	428	309.2068	-1.2	23.4	C18H30O4	[M - H]-	0.68	QVUYXFFTZ MKSBG	Lineolic acids and derivatives	*	↓↓

Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar	ClassyFire Class	Sig.	Var.
Fulgidic acid	498	327.2173	-1.4	20.8	C18H32O5	[M - H]-	0.69	MKYUCBXUUSZMQB	Lineolic acids and derivatives	*	↓
...	1674	185.0446	-5.0	2.6	C8H10O5	[M - H]-	0.50	HEWAJTLQEJUAGF	Carbocyclic fatty acids	*	↑↑
...	637	505.1711	-0.9	19.1	C25H30O11	[M - H]-	0.56	YXCPQONSMLKCHP	Iridoid O-glycosides	*	↓
Scandoside	285	389.1088	-0.5	2.9	C16H22O11	[M - H]-	0.79	ZVXWFPTVHBWJOU	Iridoid O-glycosides	*	↓
1-O-methylnyasicoside	1086	491.1550	-1.8	17.3	C24H28O11	[M - H]-	0.61	PFEGVXNNPUKZKH	Fatty acyl glycosides of mono- and disaccharides	*	↓
2,3-dihydroxyoctanoic acid	797	175.0964	-6.8	6.4	C8H16O4	[M - H]-	0.65	HSOWPFQKR RJHNS	Sugar acids and derivatives	*	↓
Gallicynoic Acid B	2021	267.1598	-1.4	20.8	C15H24O4	[M - H]-	0.50	USYGFJALDR CDGW	Long-chain fatty acids	*	↓
...	317	327.2172	-1.5	21.2	C18H32O5	[M - H]-	0.77	ADHVWICOFL YDST	Lineolic acids and derivatives	*	↓
...	2111	219.0501	-4.1	1.1	C8H12O7	[M - H]-	0.53	HVHHMEMCU LPBED	C-glycosyl compounds	*	↓
15-Keto-PGE0	1657	353.2328	-1.5	21.3	C20H34O5	[M - H]-	0.56	CDUVSQMTLOYKTR	Prostaglandins and related compounds	*	↑↑
Wikstrone	724	403.1393	-1.4	18.2	C21H24O8	[M - H]-	0.51	FCUUNRWODYPBEE	Tetrahydrofuran lignans	*	↓↓
mannosylglycerate	682	267.0716	-2.3	1.1	C9H16O9	[M - H]-	0.74	DDXCFDOPXBPUJC	Fatty acyl glycosides of mono- and disaccharides	*	↓

Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar	ClassyFire Class	Sig.	Var.
alpha-hydroxy-glutarate	841	147.0288	-7.6	1.6	C5H8O5	[M - H]-	0.69	HWXBTNAVR SUOJR	Short-chain hydroxy acids and derivatives	*	↑
...	572	417.1398	-1.0	9.9	C18H26O11	[M - H]-	0.61	PTPQTMGOUI HFNR	Iridoid O-glycosides	*	↓
...	1036	181.0497	-5.3	8.1	C9H10O4	[M - H]-	0.50	ADAYLBXFC VGNKY	p-Hydroxybenzoic acid esters	*	↑
Randioside	384	371.0980	-1.1	8.0	C16H20O10	[M - H]-	0.70	IKFVJCMZZB WMML	O-glycosyl compounds	*	↑
...	1219	431.2645	-1.3	21.6	C22H40O8	[M - H]-	0.53	BFYRWXTXDF KEXDD	Saccharolipids	*	↓
Brassylate	982	243.1596	-2.3	22.6	C13H24O4	[M - H]-	0.81	DXNCZXXFR KPEPY	Long-chain fatty acids	*	↓↓
...	264	311.2224	-1.3	22.2	C18H32O4	[M - H]-	0.64	RGRKFKRAFZ JQMS	Lineolic acids and derivatives	*	↓
...	557	207.0654	-4.5	12.4	C11H12O4	[M - H]-	0.50	PWVAUDNAZ BCJOG	Hydroxycinnamic acids and derivatives	*	↓
10-Hydroxyligustroside	895	539.1762	-1.5	12.4	C25H32O13	[M - H]-	0.61	AHTRGGWSB FOEEG	Terpene glycosides	*	↓
Benz-Me-galp-Ac-glup	519	486.1975	-1.2	14.7	C22H33NO11	[M - H]-	0.50	QRWGFRUDU VXHOQ	N-acyl-alpha-hexosamines	*	↓
...	1621	567.2073	-1.8	15.4	C27H36O13	[M - H]-	0.54	LTFQGPBDLO LKNW	Lignan glycosides	*	↓

Synonym: 化合物或其立体异构的别名。

ID: MCnebula分析中'Features'的唯一ID编号。

Err.: Mass Error (ppm), 前体离子分子量和理论分子量的偏差。

RT: Retention time, 保留时间。

Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar	ClassyFire Class	Sig.	Var.
---------	----	------------------	------	-------------	---------	--------	----	--------------------	------------------	------	------

Formula: Molecular Formula。

TS: Tanimoto similarity。

InChIKey planar: InChIKey的首个哈希块代码，代表分子骨架。

ClassyFire Class: ClassyFire分类系统中的该化合物的归类，- 表示该化合物在ClassyFire Web中未查询到。

Sig.: Q-value (Pro vs Raw, P-value的FDR矫正) 代表的显著性，*** 表示Q-value < 0.001, ** 表示Q-value < 0.01, * 表示Q-value < 0.05。

Var.: Log2(Fold Change) (HM / HS) 代表的变化水平，↓↓ 或 ↑↑ 代表|log2(FC)| > 1, ↓ 或 ↑ 代表|log2(FC)| > 0.3。

(三) 分析的报告和R代码

以下用于杜仲数据分析的R代码和报告可见于：https://github.com/Cao-lab-zcmu/exMCnebula2/tree/master/inst/extdata/scripts_evaluation/eucommia_workflow

Analysis on *E. ulmoides* dataset

Contents

1 Abstract	1
2 Introduction	1
3 Set-up	2
4 Integrate data and Create Nebulae	2
4.1 Initialize analysis	2
4.2 Filter candidates	3
4.3 Filter chemical classes	3
4.4 Create Nebulae	5
4.5 Visualize Nebulae	5
5 Nebulae for Downstream analysis	6
5.1 Statistic analysis	6
5.2 Set tracer in Child-Nebulae	9
5.3 Quantification in Child-Nebulae	10
5.4 Annotate Nebulae	10
5.5 Query compounds	13
5.6 Plot MS/MS spectra of top ‘features’	16
5.7 Discover more around top ‘features’ in Child-Nebulae	18
6 Session infomation	20
Reference	22

1 Abstract

Untargeted mass spectrometry is a robust tool for biological research, but researchers universally time consumed by dataset parsing. We developed MCnebula, a novel visualization strategy proposed with multidimensional view, termed multi-chemical nebulae, involving in scope of abundant classes, classification, structures, sub-structural characteristics and fragmentation similarity. Many state-of-the-art technologies and popular methods were incorporated in MCnebula workflow to boost chemical discovery. Notably, MCnebula can be applied to explore classification and structural characteristics of unknown compounds that beyond the limitation of spectral library. MCnebula was integrated in R package and public available for custom R statistical pipeline analysis. Now, MCnebula2 (R object-oriented programming with S4 system) is further available for more friendly applications.

2 Introduction

We know that the analysis of untargeted LC-MS/MS dataset generally begin with feature detection. It detects ‘peaks’ as features in MS¹ data. Each feature may represents a compound, and assigned with MS² spectra. The MS² spectra was used to find out the compound identity. The difficulty lies in annotating these features to discover their compound identity, mining out meaningful information, so as to serve further

三、小结

在草药数据集分析中，MCnebula提供了一个快速的方法：用于化合物注释和探索化学类别范围内的Child-Nebulae化学变化。*E. ulmoides*的主要成分是木脂素、环稀醚萜类、酚类、黄酮类、甾体类和萜类^[50]。在我们的研究中，通过ABC选择算法得到的化学类别包括‘Lignans, neolignans and related compounds’（LNARC）和‘Iridoids and derivatives’（IAD），以及‘Monoterpene glycosides’。黄酮类化合物由‘Phenylpropanoids and polyketides’（PAP）涵盖^[24]，酚类化合物可在‘Methoxyphenols’中找到。黄酮类化合物与类固醇相似，在选定的结果中没有保留‘Flavonoides’和‘Steroids and steroid derivatives’，因为它们在*E. ulmoides*（树皮）中没有LNARC和IAD那么丰富。许多在LNARC和IAD化学类别中鉴定的化合物（表15）在以前关于*E. ulmoides*的LC-MS/MS分析的研究中被报道^[52,51]。我们根据对处理前后‘Features’量化水平变化的统计比较，获得了Top‘Features’。其中一个变化很大甚至是新产生的化合物（ID：1642）在Child-Nebulae中被追踪到。我们推测它与两个结构相似的化合物有转化关系。这个例子很好地说明了MCnebula在分析植物来源的化合物方面的应用，特别是在快速识别和探索化学变化方面。值得注意的是，与人源性代谢物的参考光谱库相比，植物源性化合物的参考光谱库或数据库要少得多，虽然过去也构建了一些特定的植物源性化合物数据库^[53]，但缺乏足够的碎片光谱进行全面的库匹配。在MCnebula的帮助下，可以实现对植物源性化合物复杂成分的快速解析。

第四部分 基于MCnebula策略的血清代谢组学

一、材料与方法

（一）实验材料

共245个来自MASSIVE（<https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp>）的LC-MS/MS数据（.mzML）（ID号：MSV000083593）（空白、对照和样品）用于MCnebula的应用和示例^[47]。

工作站用于下载和初步处理数据集：Pop!_OS (Ubuntu) 22.04 LTS 64-bits workstation (Intel Core i9-10900X, 3.70GHz × 20, 125.5 Gb of RAM)

个人笔记本电脑Surface pro7用于随后的MCnebula分析：Pop!_OS (Ubuntu) 22.04 LTS 64-bits PC (Intel Core i7-1065G7, 1.3 GHz × 8, 16 Gb of RAM)。

（二）实验方法

我们重新分析了来自MASSIVE（ID号：MSV000083593）的245份LC-MS/MS数据（.mzML）（空白、对照和样品）^[47]。MZmine2（2.53版）进行了‘Features’检测。检测工作流程主要涉及：**1)**自动数据分析管道（ADAP）进行峰检测和去卷积^[9]，**2)**同位素峰值查找，**3)**平行样品峰对齐，**4)**缺失峰再寻找（Gap filling）。当导出MS/MS光谱（.mgf）供SIRIUS 4软件计算时，MS/MS谱被合并到一个列表中，并有30%的峰值计数阈值过滤。‘Features’检测工作流程参照FBMN预处理和SIRIUS计算的先决条件。输出的.mgf用SIRIUS 4软件（4.9.12版）运行，与SIRIUS^[22]、ZODIAC^[39]、CSI:fingerID^[20]、CANOPUS^[34]进行计算。特别是，SIRIUS被设置为检测碘元素。MCnebula软件包被用于后续的数据分析。所有的后续分析都被组织成简明的代码，并以报告的形式输出。

京都基因和基因组百科全书（KEGG）的代谢途径富集分析分别用‘Lysophosphatidylcholines’（LPCs）和‘Bile acids, alcohols and derivatives’（BAs）进行。我们使用InChIKey2D来匹配代谢通路中的化合物。具体来说，为了避免由于立体异构而导致的鉴定结果偏差，我们使用InChIKey平面通过PubChem API获得所有可能的InChIKeys。在这个步骤中，也获得了这些化合物的PubChem CID。MetaboAnalystR的R包被用来将PubChem CID转换为KEGG ID^[54]。许多化合物与代谢途径无关，所以这些被过滤掉了。FELLA的R软件包被用于KEGG富集与‘pagerank’算法^[55]。上述方法已被整合为函数，与MCnebula工作流程对接，这些功能可在‘exMCnebula2’包中找到（见表14）。

二、结果

（一）MCnebula对血清数据集的整体分析

血清样本收集自感染金黄色葡萄球菌菌血症（SaB）或未感染的院内患者和健康志愿者。总的来说，样本分为：1) 对照组，涉及NN（non-hospital, non-infected）和HN（hospital, non-infected）；2) 感染组，涉及HS（hospital, survival），HM（hospital, mortality）。

在对血清数据集进行LC-MS预处理时，共检测到7680个‘Features’。通过MS/MS光谱（用SIRIUS软件）预测化合物后，用MCnebula进行了后续分析。其中，6501个‘Features’被注释为预测的分子式，进一步，3449个‘Features’被注释为预测的化学结构。利用ABC选择算法，我们通过应用‘Inner filter’模块（参考：第一部分 MCnebula的方法构建>二、结果>（二）MCnebula的算法>5. ABC选择算法）过滤掉了1000多个化学类别；在进行‘Cross filter’时，进一步过滤掉了508个化学类别；对于剩下的41个化学类别，我们手动过滤掉了19个化学类别，而留下最后的22个化学类别组成了Nebula-Index，进一步可视化为Child-Nebulae。值得一提的是，被过滤掉的527（508+19）个化学类可以重新加入到分析中（使用‘backtrack_stardust’方法）。在此，通过MCnebula的基本工作流程，得到

了Parent-Nebula和Child-Nebulae（图15，图16）。通过审视Child-Nebulae，可以对血清数据集中包含的化学类别有了基本的了解。为了从Child-Nebulae中挖掘更多的信息：我们对HS组和HM组进行了二元比较（Binary comparison，见：第二部分 MCnebula的方法评估与拓展>一、材料与方法>（二）实验方法>5. MCnebula的拓展涉及的算法>5.1 统计分析的算法），根据Q值（校正后的P值）对‘Features’进行排序；前50个‘Features’被设定为‘Tracer’，在Child-Nebulae中进行标记（图17）。通过结合有关Q值的‘Features’选择算法，减少了Child-Nebulae中表现出的化学类别。在Child-Nebulae中，HM组与HS组的log₂(Fold Change)（log₂(FC)）量化被可视化（图18）。在图18中，‘Bile acids, alcohols and derivatives’（BAs）类和‘Lysophosphatidylcholines’（ACs）（图19a和b）类的总体水平明显增加，而‘Lysophosphatidylcholines’（LPCs）类的总体水平明显下降。事实上，BAs、ACs和LPCs被报道与肝功能障碍、肠道微生态平衡失调和死亡风险有关^[57,56,47]。

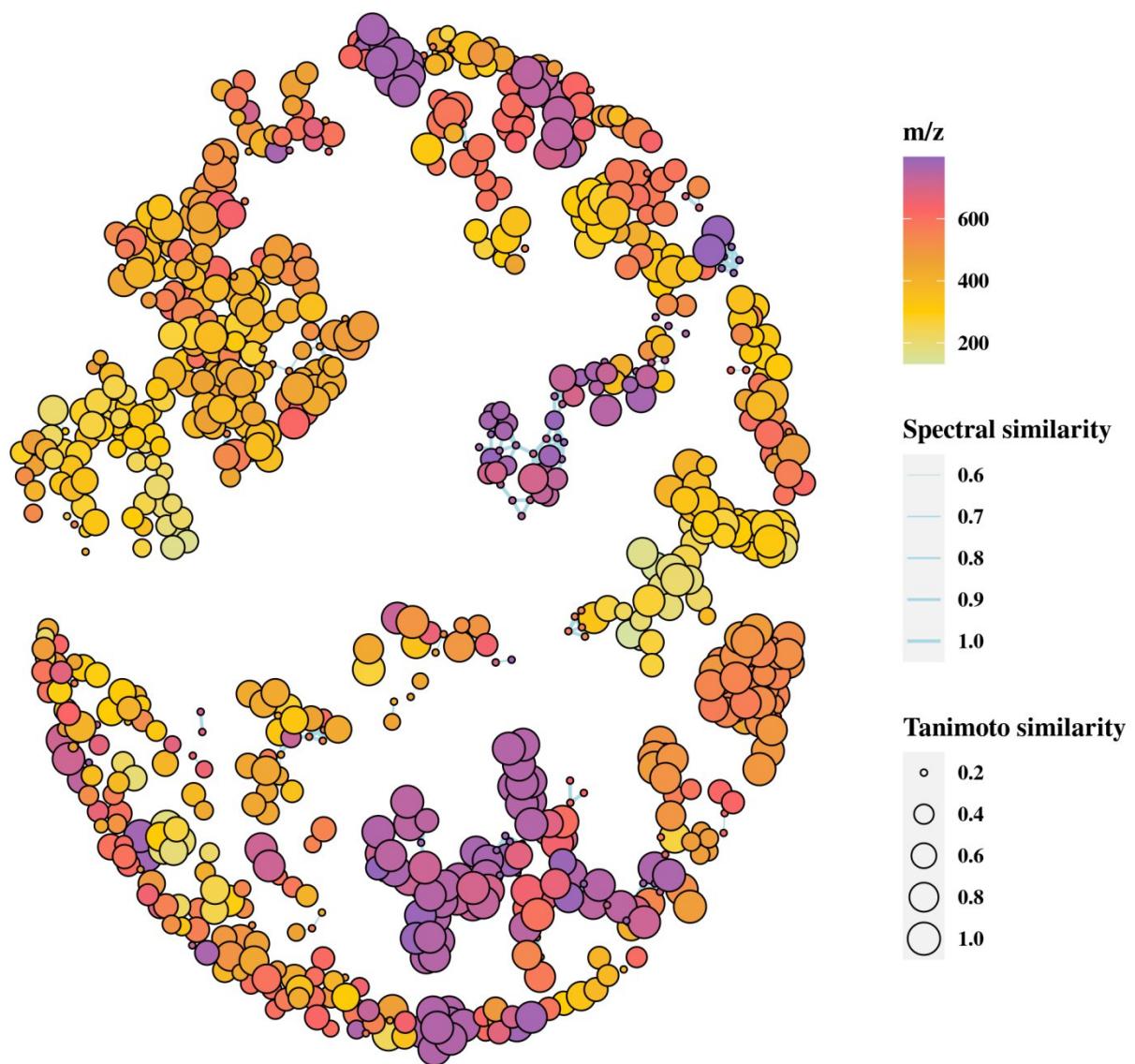


图15 血清数据集的Parent-Nebula

- 图15注：在Parent-Nebula中，‘Features’被映射为网络图中的节点（Node）。边（edge）说明了相邻‘Features’的光谱相似性。并非所有的‘Features’都显示在Parent-Nebula中，因为孤立的节点被删除。

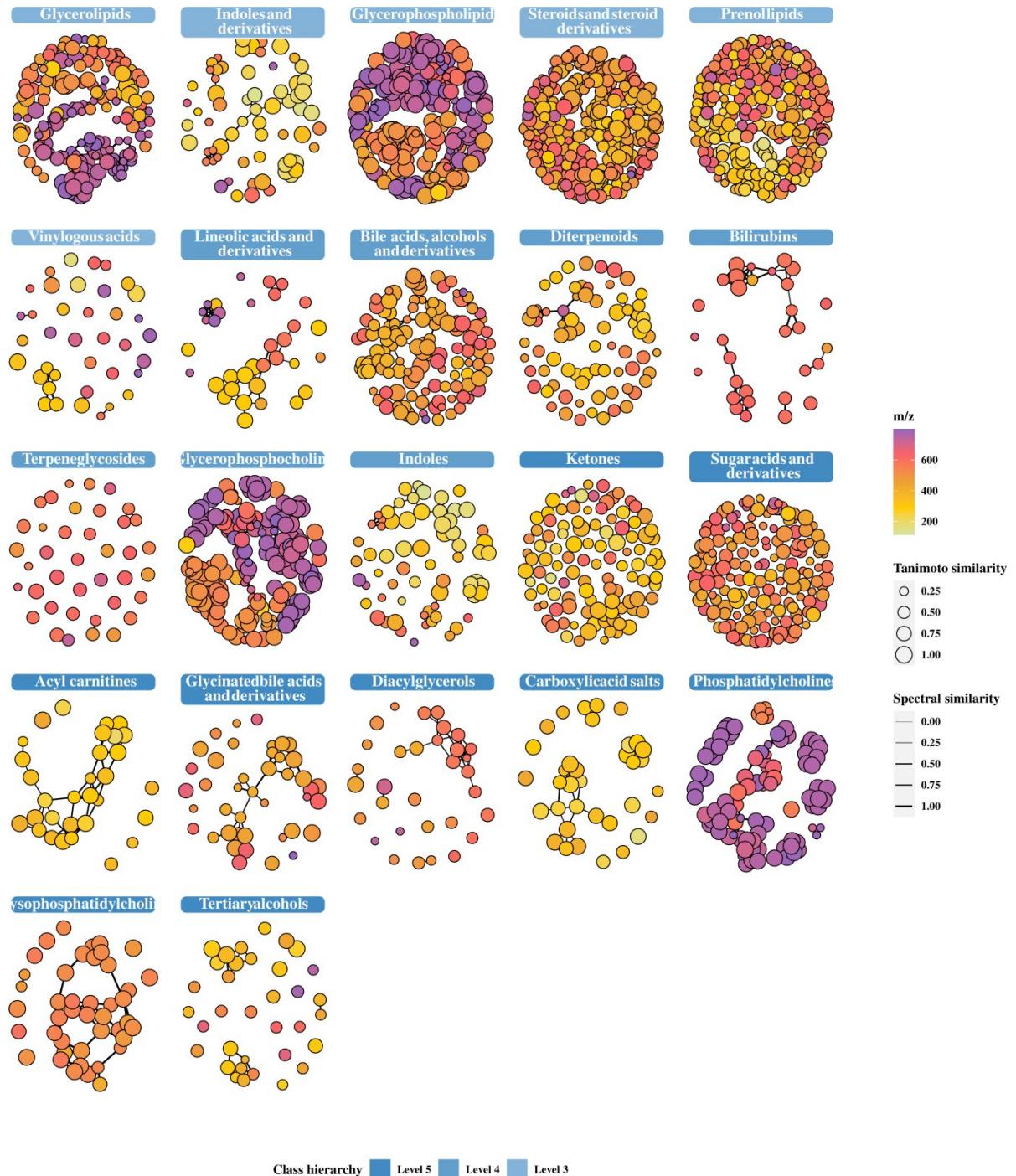


图16 血清数据集的Child-Nebulae

- 图16注：Child-Nebulae是根据Nebula-Index中的化学类别创建的。化学类别的分类‘Features’被映射到相应的Child-Nebula中。

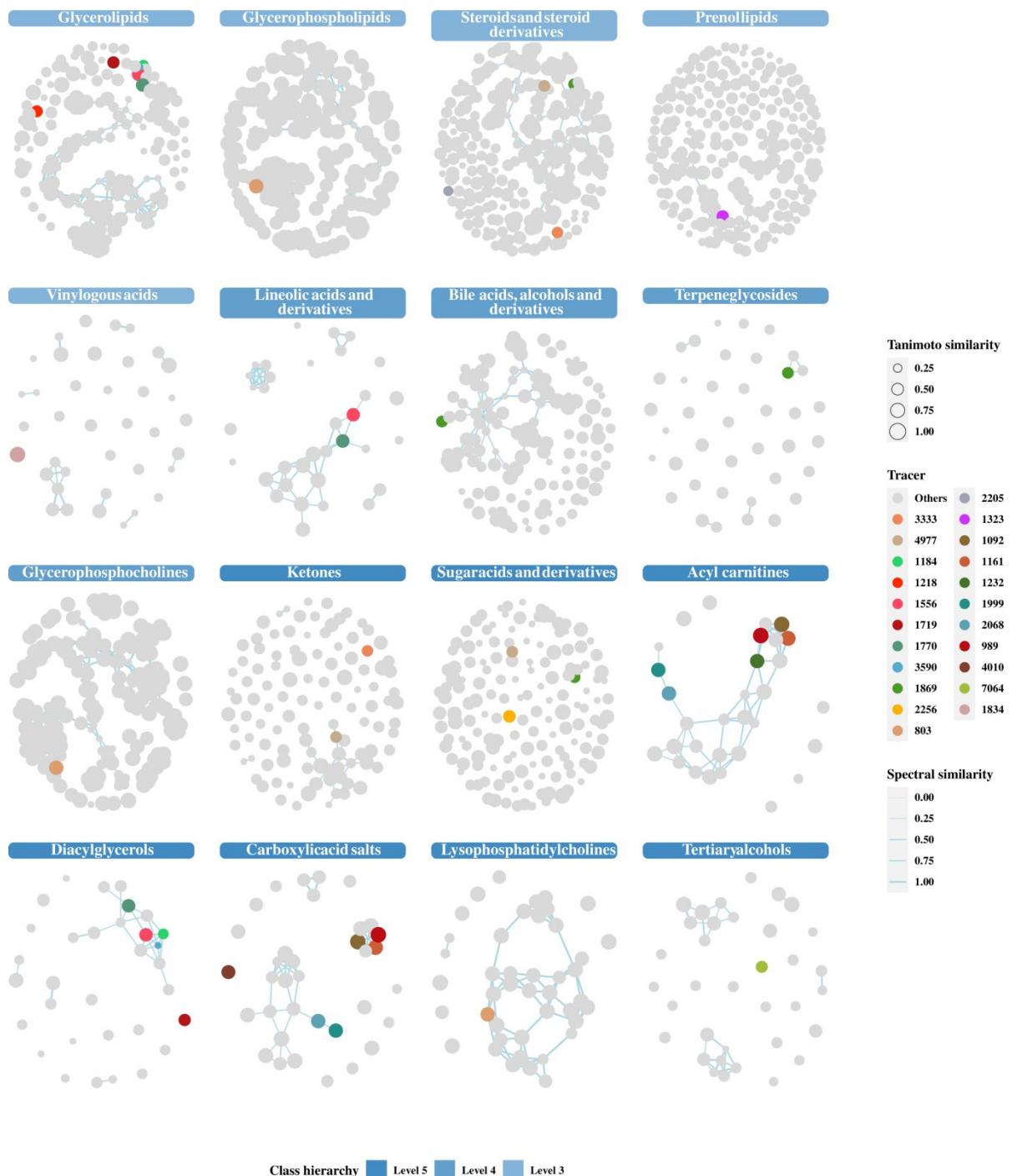


图17 在血清数据集的Child-Nebulae中追踪Top ‘Features’

- 图17注：根据统计分析的‘Features’排名，Top ‘Features’在Child-Nebulae中用不同的颜色标记。

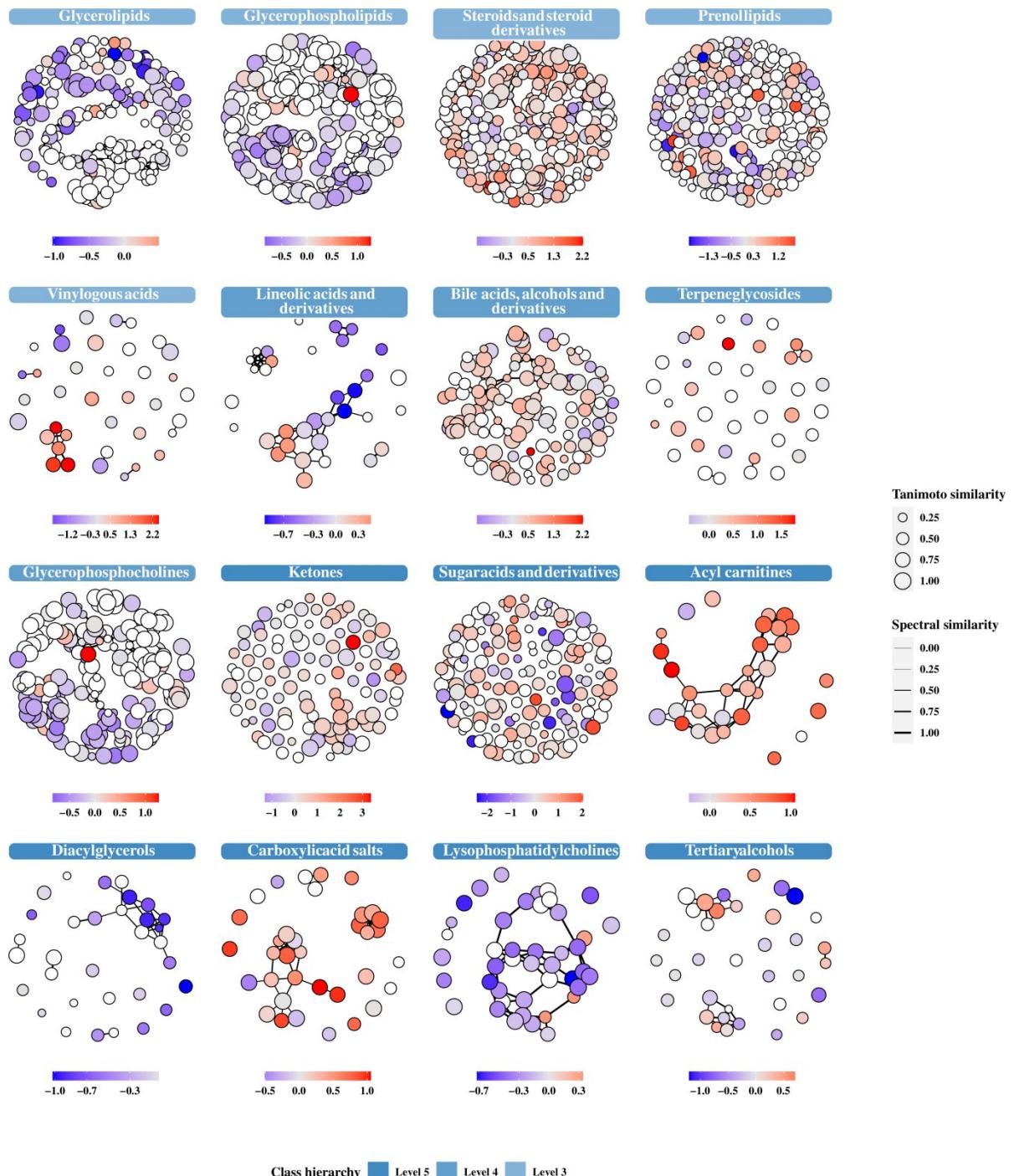


图18 在血清数据集的Child-Nebulae中可视化组间 $\text{Log}_2(\text{Fold Change})$

- 图18注：HM组与HS组的 $\text{log}_2(\text{Fold Change})$ 值在Child-Nebulae中显示为渐变颜色。白色的节点表示量化值缺失的‘Features’（这些‘Features’在我们的重新分析中被检测到，但在Wozniak等人的分析中没有^[47]）。

（二）MCnebula对血清数据集的聚焦分析

通过对Child-Nebula的深度注释可视化，这三类化合物都具有相似的结构母核，它们在NN、HN、HS和HM组中的含量也相似（图19c，图20）。随后，我们对这三类化合物进行了聚类热图分析和通路富集分析。如聚类热图所示（图21），ACs和BAs的对照组与感染组在聚类中分离，这意

味着ACs和BAs可能与SaB的感染有关。相比之下，LPCs没有显示出明显的SaB感染相关性或死亡相关性，可能是由于这类化合物对SaB疾病一般的一致性。我们对这三类化合物进行了通路富集分析（HS与HM组相比，Q值<0.05）。BAs的结果显示，四个化合物表现出与‘Bile secretion’，‘Cholesterol metabolism’，和‘Primary bile acid biosynthesis’等代谢相关（图22b）。其中， β GCS是一类具有相同母核的化合物。LPCs的结果表明，LPCs的母核结构相似的化合物意味着与一系列下游途径有关（图22c）。ACs的重要化合物在该途径中没有富集。但是，ACs在调整葡萄糖和脂肪酸代谢之间的转换中的基本作用被综述^[58]。它们的功能通过酰基的双向运输在细胞膜和线粒体之间实现（图22a）。

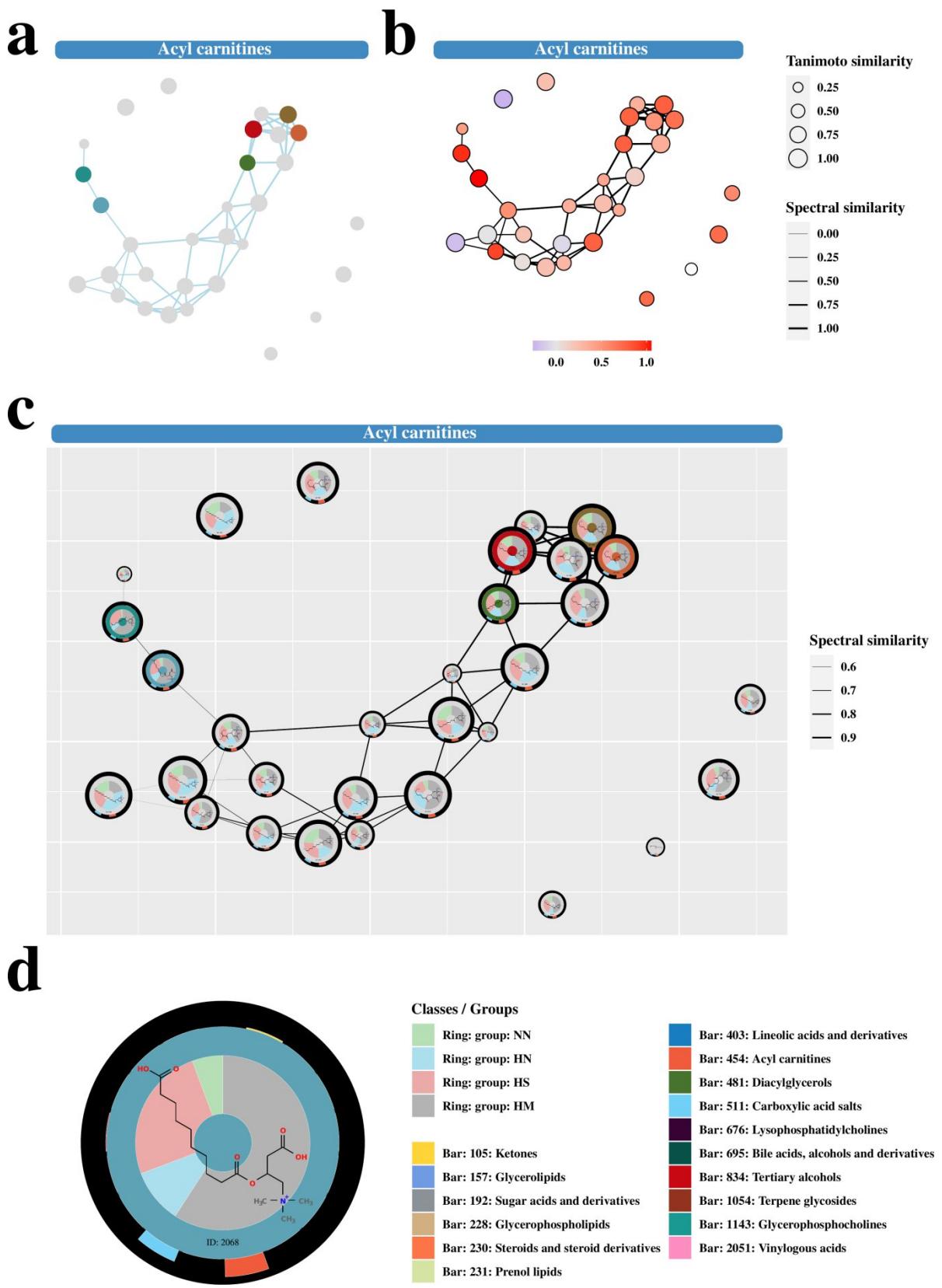


图19 血清数据集的代表ACs的Child-Nebula的深度可视化

- 图19注: **a)**, 参考图17, 审视‘Acyl carnitines’的 Child-Nebula。**b)** 参考图18。**c)** Top ‘Features’的节点用颜色标记。‘Features’ 的节点用化学结构、环形图和类预测后验概率 (PPCP) 的柱状图进行了注释。‘Features’ 的化学结构的最高候选被映射到节点中。环形图映

射出每个元数据组NN（non-hospital, non-infected），HN（hospital, non-infected），HS（hospital, survival），HM（hospital, mortality）内检测到的每个‘Features’的相对峰面
积。没有环形图的节点表示量化值缺失的‘Features’（这些‘Features’在我们的重新分析中被检
测到，但在Wozniak等人的分析中没有检测到）。柱状图为‘Features’的结构（亚结构或主导
结构）类的PPCP。**d)** ‘Features’ 2068（ID）的放大及其图例。

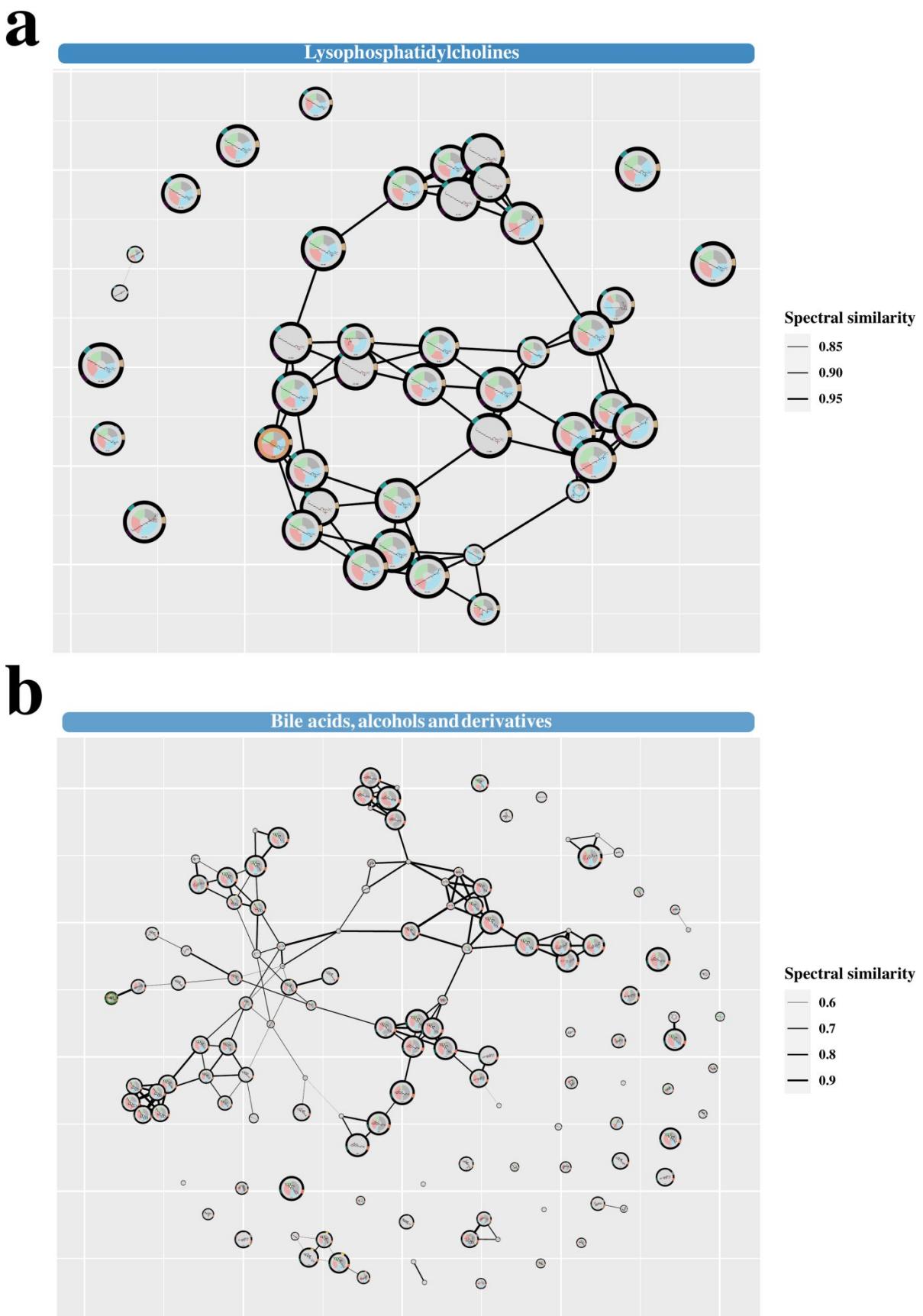


图20 血清数据集中代表LPCs和BAs的Child-Nebulae的深度可视化

- 图20注：参考图19。

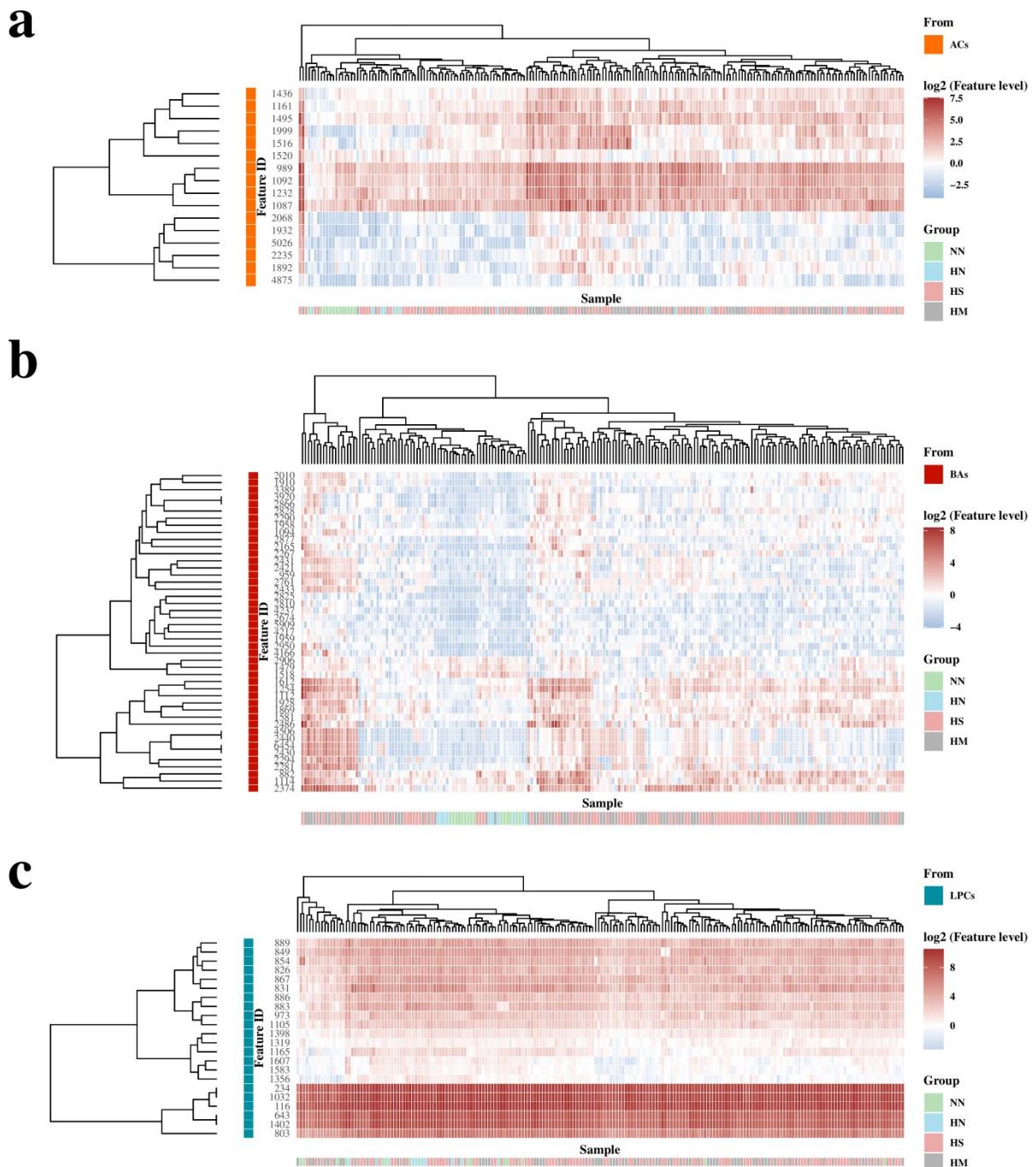


图21 血清数据集ACs、LPCs和BAs的热图分析

- 图21注: **a*、**c*和**e*显示了AC、LPC和BA的水平热图。‘Features’是通过在感染组与对照组之间或HM组与HS组之间相比选择的: Q-value < 0.05, |log₂(FC)| ≥ 0.3。

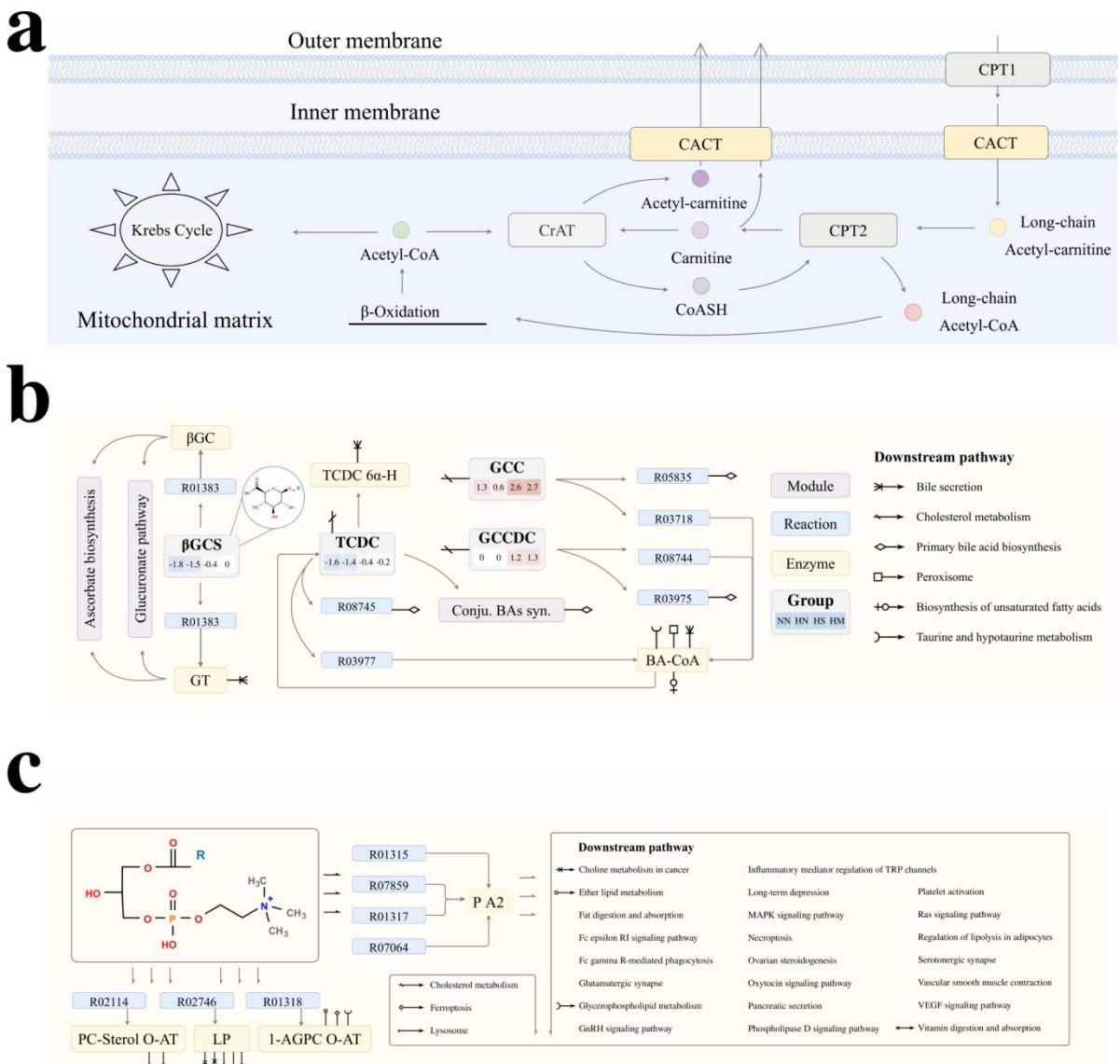


图22 血清数据集ACs、LPCs和BAs的通路富集分析

- 图22注: **a)** 线粒体中的肉碱系统。Abbreviation: CPT1, carnitine-palmitoyltransferase-1; CACT, carnitine acylcarnitine translocase; CrAT, carnitine acetyltransferase; CPT2, carnitine-palmitoyltransferase-2. **b)** 用KEGG对LPCs以pagerank算法的富集分析。Abbreviation: P A2, phospholipase A2; PC-Sterol O-AT, phosphatidylcholine-sterol O-acyltransferase; LP, lysophospholipase; 1-AGPC O-AT, 1-acylglycerophosphocholine O-acyltransferase; **c)** 用KEGG对BAs以pagerank算法富集度分析。Abbreviation: β GC, beta-glucuronidase; β GCS, beta-D-Glucuronoside; GT, glucuronosyltransferase; TCDC, taurochenodeoxycholate 6alpha-hydroxylase; GCC, Glycocholate; GCCDC, Glycochenodeoxycholate; Conju. BAs syn., ‘Conjugated bile acid biosynthesis, cholate’ ; BA-CoA, bile acid-CoA:amino acid N-acyltransferase.

在Wozniak等人的研究中^[47], 确定了五个ACs化合物。此外, 四个Top代谢物 (2-Hexadecanoylthio-1-Ethylphosphorylcholine (HEPC); sphingosine-1-phosphate (S1P); decanoyl-

carnitine; L-Thyroxine (T4)) 也被鉴定。在我们的重新分析中，除了HEPC，所有的鉴定都是一致的。在我们的重新分析中，HEPC被鉴定为1-pentadecanoyl-sn-glycero-3-phosphocholine (LPC15:0) 或其立体异构体。事实上，HEPC和LPC15:0在结构上非常相似，但在元素构成上不同（分别对应于C₂₃H₄₈NO₅PS和C₂₃H₄₈NO₇P）。在化学分类方面，它们明显不同。HEPC属于‘Organic nitrogen compounds’ (Super Class) 家族中的‘Cholines’ (Level 5)，而LPC15:0属于‘Lipids and lipid-like molecules’ 家族中的‘Lysophosphatidylcholines’ (LPCs) (Level 5)。作为MCnebula工作流程的一部分，硫元素对于SIRIUS的同位素模式是可以以高质量精度检测到的^[37]。然而，对于‘HEPC’的MS/MS光谱，没有包含硫元素的候选分子式。总的来说，我们用MCnebula工作流程鉴定了更多的化合物，许多结果与Wozniak等人^[47]的分析一致。所有鉴定的化合物都被整理（表17，用Tanimoto similarity > 0.5过滤，并用InChIKey的首个哈希块（InChIKey Planar或InChIKey2D，代表分子骨架）去掉重复的结果；共有1086个化合物，限于篇幅，仅展示Q-value < 0.05）。此外，还对Wozniak等人通过光谱库匹配没有成功鉴定、但通过MCnebula工作流鉴定出分子式或化学结构的化合物（在Wozniak等人的研究中EFS和MWU的前50名）进行了整理（表16）。

表16 MCnebula重新分析血清数据集Wozniak等人的Top Metabolites

# Original ID	# EFS Rank	# MWU Rank	# Spectral Library Match	Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar
92	9	234	-	Metronidazole-OH	1011	188.0663	-1.4	0.5	C6H9N3O4	[M + H]+	0.93	AEHPOYAOL CAMIU
132	50	286	-	1-16:0-lysoPC	867	518.3237	3.7	6.0	C24H50NO7P	[M + Na]+	0.99	ASWBNKHCZ GQVJV
225	30	31	-	Isoleucylproline	885	229.1533	-6.1	0.4	C11H20N2O3	[M + H]+	0.54	BBIXOODYW PFNDT
774	-	36	-	2-[(2-[(2-amino-3-hydroxybutanoyl)amino]-3-hydroxypropanoyl)amino]-3-(1H-imidazol-5-yl)propanoic acid	6930	344.1599	9.8	3.5	C13H21N5O6	[M + H]+	0.39	BCYUHPXBH CUYBA
1802	62	22	-	Tocris-0605	1092	288.2155	-5.2	4.1	C15H29NO4	[M + H]+	0.98	CXTATJFJDM JMIY
1425	-	50	-	3-{[(9Z)-17-carboxyheptadec-9-enoyl]oxy}-4-(trimethylammonio)butanoate	1867	456.3311	-1.9	5.4	C25H45NO6	[M + H]+	0.61	CYPBWZKNU DFJJE
822	-	41	-	-	2681	340.1744	8.6	2.8	C12H25N3O8	[M + H]+	0.41	DAKOSHWC XIPQFW
804	32	85	-	b-Uridine	1245	267.0583	-1.9	0.3	C9H12N2O6	[M + Na]+	0.42	DRTQHJPVM GBUCF
429	37	223	-	(2E,5Z,7E)-deca-2,5,7-trienoylcarnitine	957	310.2006	-2.4	3.9	C17H27NO4	[M + H]+	0.61	DUQXB EFMR JGSKH

# Original ID	# EFS Rank	# MWU Rank	# Spectral Library Match	Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar
114	17	100	Spectral Match to D-erythro-Sphingosine-1-phosphate from NIST14	Epitope ID:161059	952	380.2548	-3.3	5.8	C18H38NO5P	[M + H]+	0.99	DUYSYHSSB DVJSM
3978	-	35	-	6-Keto-decanoylcarnitine	3393	330.2279	1.1	3.6	C17H31NO5	[M + H]+	0.51	DZALQUYFN HIYDL
1189	47	281	-	Sarcine	999	137.0452	-4.6	0.3	C5H4N4O	[M + H]+	0.93	FDGQSTZJBF JUBT
835	-	25	-	sebacoylcarnitine	2068	346.2244	5.7	3.4	C17H31NO6	[M + H]+	0.84	GBFPILOKXG QZKW
247	46	32	-	9-Decenoylcarnitine	1232	314.2313	-4.1	4.6	C17H31NO4	[M + H]+	0.83	GOOCIIIXFL VRAG
777	-	43	-	bmse000410	2203	190.0498	-0.4	1.3	C10H7NO3	[M + H]+	0.75	HCZHHEIFKR OPDY
4591	-	47	-	Diacylglycerol(15:0/16:1)	1719	575.4672	4.4	9.2	C34H64O5	[M + Na]+	0.6	IHIWKNWPE CHDKE
13	11	5	-	1-18:0-lysoPC	803	546.3574	8.0	6.6	C26H54NO7P	[M + Na]+	0.84	IHNKQIMGV NPMTC
210	29	302	Massbank: Hydrocortisone	Acticort	942	363.2149	-4.7	4.1	C21H30O5	[M + H]+	0.99	JYGXADMDFJGBT
2532	7	46	-	N-{[2-(pyridin-4-ylcarbonyl)hydrazino]carbonyl}hexopyranosyl	630	365.1043	-6.8	0.3	C13H18N4O7	[M + Na]+	0.51	KPGYIOMDV KQYDK

# Original ID	# EFS Rank	# MWU Rank	# Library Match	Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar
amine												
143	26	115	TOP 8 Psoriasis feature - Unknown FeatureID=4 262	-	897	438.2993	3.2	6.6	C21H44NO6P	[M + H]+	0.52	LQICBGFMER KBCG
119	66	23	Decanoyl-L- carnitine	Tocris-0477	989	316.2469	-4.2	4.9	C17H33NO4	[M + H]+	0.99	LZOSYCMHQ XPBFU
746	2	67	-	Acisoga	981	207.1099	-2.7	0.7	C9H16N2O2	[M + Na]+	0.39	OAUYENAPB FTAQT
731	21	131	-	-	1869	615.3735	-0.7	4.7	C32H54O11	[M + H]+	0.54	ONLXJASEXI XGRM
1947	15	321	-	Carnidazol	2045	245.0691	-4.7	0.4	C8H12N4O3S	[M + H]+	0.41	OVEVHVUR WWTPFC
1231	31	106	-	L-Hypro-d3	1183	154.0459	-9.9	0.4	C5H9NO3	[M + Na]+	0.44	PMMYEEVY MWASQN
1363	10	168	-	5,11,14- Eicosatrienoicacid	1071	307.2637	1.6	8.1	C20H34O2	[M + H]+	0.44	PRHHYVQTP BEDFE
183	-	27	-	[2-hydroxy-3- [(2R,3R,4S,5R,6R)- 3,4,5-trihydroxy-6- (hydroxymethyl)oxan- 2-yl]oxypropyl] (7Z,10Z,13Z)- hexadeca-7,10,13- trienoate	2205	509.2754	6.5	4.0	C25H42O9	[M + Na]+	0.4	PSRSTOVXIJT KGG

# Original ID	# EFS Rank	MWU Rank	# Spectral Library Match	Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar
820	-	39	-	-	3286	271.1639	-5.0	3.4	C13H22N2O4	[M + H]+	0.49	RAIXJOOYPR FWSP
854	3	17	-	Opreal_213609	3333	289.2187	8.5	4.1	C19H28O2	[M + H]+	0.51	RAJWOBJTTG JROA
1947	15	321	-	7-acetyl-1,3-dimethylpurine-2,6-dione	6646	245.0626	-7.8	0.5	C9H10N4O3	[M + Na]+	0.19	RCDRLZYAK DYXMM
800	-	26	-	Dimethylguanosine	2231	312.1300	-0.8	0.7	C12H17N5O5	[M + H]+	0.99	RSPURTUNR HNVGF
670	8	95	-	-	1184	595.4932	-0.1	8.9	C36H66O6	[M + H]+	0.45	RWXWCAZL EFJOFU
289	22	185	-	7a-Hydroxy-3-oxo-4-cholestenoicacid	1104	431.3145	-2.5	6.4	C27H42O4	[M + H]+	0.85	SATGKQGFU DXGAX
835	-	25	-	L-alpha-glutamyl-L-valyl-L-valine	5569	346.2002	8.5	3.2	C15H27N3O6	[M + H]+	0.43	SOYWRINXU SUWEQ
2799	-	34	-	O-DC16:0-L-carnitine(1-)	4010	430.3178	3.4	5.2	C23H43NO6	[M + H]+	0.79	UNHCPLSWM NPZTD
814	-	24	-	N6-Threonylcarbamoyladenosine	2199	413.1410	-1.4	2.1	C15H20N6O8	[M + H]+	0.7	UNUYMBPXE FMLNW
885	103	37	-	Tocris-0526	1161	260.1839	-6.6	3.0	C13H25NO4	[M + H]+	0.91	VVPRQWTYS NDTEA
2816	39	98	-	2,2"-cyclohexane-1,2-diyl diacetic acid	3021	223.0933	-3.7	3.7	C10H16O4	[M + Na]+	0.52	VWLPAWSX KLKROQ
469	-	38	-	Diacylglycerol(18:3n6/	1770	577.4824	-0.5	8.5	C36H64O5	[M + H]+	0.73	XPGUEPBYN

# Original ID	# EFS Rank	# MWU Rank	# Spectral Library Match	Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar
				15:0)								VJPCL
3865	24	319	-	1,10-Bis(4-carboxyphenoxy)decane	1671	415.2115	-0.2	5.9	C24H30O6	[M + H]+	0.4	XRDKWFXOXXUQJS
1110	-	30	L-THYROXINE	eltroxin	1960	777.6948	1.0	4.7	C15H11I4NO4	[M + H]+	0.99	XUIIKFGFIJCVMT
755	-	48	-	O-DC8:0carnitine(1-)	1999	318.1897	-4.4	2.5	C15H27NO6	[M + H]+	0.83	YVWVEIPYMGBQPE
183	-	27	-	3-beta,19,21-Triacetoxy-14-iso-17-isopregnane-5,14-diol-20-one	4977	509.2733	-2.3	4.2	C27H40O9	[M + H]+	0.54	ZTWWQC0HUBHIGI
736	18	44	-	-	1446	168.0634	-4.7	0.4	C6H7N4O2	[M + H]+	-	-

Original ID: Wozniak等人的研究中'Features'的唯一ID编号。

EFS Rank 和 # MWU Rank: Wozniak等人的研究中对'Features'排序的两种算法。

Spectral Library Match: Wozniak等人的研究中以光谱匹配的方式得到的化合物结果。

Synonym: 化合物或其立体异构的别名。

ID: MCnebula分析中'Features'的唯一ID编号。

Err.: Mass Error (ppm), 前体离子分子量和理论分子量的偏差。

RT: Retention time, 保留时间。

Formula: Molecular Formula。

TS: Tanimoto similarity。

InChIKey planar: InChIKey的首个哈希块代码, 代表分子骨架。

表17 MCnebula工作流程鉴定的血清数据集的化合物 (Q-value < 0.05)

Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar	ClassyFire Class	Sig.	Var.
...	803	546.3574	8.0	6.6	C26H54NO7P	[M + Na]+	0.84	IHNKQIMGVN PMTC	Lysophosphatidylcholine	***	↓
...	4977	509.2733	-2.3	4.2	C27H40O9	[M + H]+	0.54	ZTWWQC0HU BHIGI	Steroid esters	***	↑↑
Etiocholanedione	3333	289.2187	8.5	4.1	C19H28O2	[M + H]+	0.51	RAJW0BJTTG JROA	Androgens and derivatives	***	↑
Isoleucylproline	885	229.1533	-6.1	0.4	C11H20N2O3	[M + H]+	0.54	BBIXOODYW PFNDT	Dipeptides	***	↑
sebacoylcarnitine	2068	346.2244	5.7	3.4	C17H31NO6	[M + H]+	0.84	GBFPILOKXG QZKW	Acyl carnitines	***	↑↑
Dimethylguanosine	2231	312.1300	-0.8	0.7	C12H17N5O5	[M + H]+	0.99	RSPURTUNRH NVGF	Purine nucleosides	***	↑
N6-Threonylcarbamoyladenosine	2199	413.1410	-1.4	2.1	C15H20N6O8	[M + H]+	0.70	UNUYMBPXE FMLNW	Purine nucleosides	***	↑
Dextrothyroxine	1960	777.6948	1.0	4.7	C15H11I4NO4	[M + H]+	0.99	XUIIKFGFIJC VMT	Alpha amino acids and derivatives	**	↓
Hexanoylcarnitine	1161	260.1839	-6.6	3.0	C13H25NO4	[M + H]+	0.91	VVPRQWTYS NDTEA	Fatty acid esters	**	↑
octanoylcarnitine	1092	288.2155	-5.2	4.1	C15H29NO4	[M + H]+	0.98	CXTATJFJDMJ MIY	Fatty acid esters	**	↑
9-Decenoylcarnitine	1232	314.2313	-4.1	4.6	C17H31NO4	[M + H]+	0.83	GOOOCHIXFL VRAG	Fatty acid esters	**	↑
...	1770	577.4824	-0.5	8.5	C36H64O5	[M + H]+	0.73	XPGUEPBYNV JPCL	Lineolic acids and derivatives	**	↓

Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar	ClassyFire Class	Sig.	Var.
6-Keto-decanoylcarnitine	3393	330.2279	1.1	3.6	C17H31NO5	[M + H]+	0.51	DZALQUYFN HIYDL	Acyl carnitines	**	↑
Transtorine	2203	190.0498	-0.4	1.3	C10H7NO3	[M + H]+	0.75	HCZHHEIFKR OPDY	Quinoline carboxylic acids	**	↑↑
Decanoyllevocarnitine	989	316.2469	-4.2	4.9	C17H33NO4	[M + H]+	0.99	LZOSYCMHQ XPBFU	Acyl carnitines	**	↑
...	1719	575.4672	4.4	9.2	C34H64O5	[M + Na]+	0.60	IHIWKNWPEC HDKE	Diacylglycerols	**	↓
O-hexadecanedioyl-L-carnitine	4010	430.3178	3.4	5.2	C23H43NO6	[M + H]+	0.79	UNHCPLSWM NPZTD	Acyl carnitines	**	↑
Octanedioylcarnitine	1999	318.1897	-4.4	2.5	C15H27NO6	[M + H]+	0.83	YVWVEIPYM GBQPE	Acyl carnitines	**	↑
...	630	365.1043	-6.8	0.3	C13H18N4O7	[M + Na]+	0.51	KPGYIOMDV KQYDK	Hexoses	**	↑
...	1867	456.3311	-1.9	5.4	C25H45NO6	[M + H]+	0.61	CYPBWZKNU DFJJE	Tricarboxylic acids and derivatives	**	↑
...	2256	305.0860	5.6	0.4	C10H18O9	[M + Na]+	0.63	FAXWMPSCK RTWTN	O-glycosyl compounds	**	↑
Triundecanoin	1218	619.4911	0.5	9.1	C36H68O6	[M + Na]+	0.56	MBXVIRZWS HICAV	Triacylglycerols	**	↓
Acisoga	1088	185.1283	-0.7	0.7	C9H16N2O2	[M + H]+	0.75	OAUYNENAPBF TAQT	N-alkylpyrrolidines	**	↑
Phenylalanylproline	1552	263.1392	0.6	2.6	C14H18N2O3	[M + H]+	0.89	WEQJQNWXC SUVMA	Dipeptides	**	↓
Stearolicacid	1323	281.2473	-0.9	9.2	C18H32O2	[M + H]+	0.60	RGTIBVZDHO	Long-chain fatty	**	↓

Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar	ClassyFire Class	Sig.	Var.
								MOKC	acids		
...	1165	564.3048	-2.3	6.0	C28H48NO7P	[M + Na]+	0.87	PDIGSOAOQO XRDU	Lysophosphatidylcholine	**	↓
...	3185	317.2683	-1.0	9.2	C18H36O4	[M + H]+	0.61	UHSVJNCEYV VOCB	Long-chain fatty acids	**	↓
4-Pyridoxate	1842	184.0596	-4.7	0.4	C8H9NO4	[M + H]+	0.80	HXACOUQIXZ GNBF	Pyridinecarboxylic acids	*	↑
Sphingosine-1-phosphate-d7	952	380.2548	-3.3	5.8	C18H38NO5P	[M + H]+	0.99	DUYSYHSSBD VJSM	Phosphosphingolipids	*	↓
3-hydroxydecanoyl carnitine	1932	332.2442	3.0	3.4	C17H33NO5	[M + H]+	0.71	VCRSQDIROU ELAR	Acyl carnitines	*	↑
...	1407	321.2784	-1.2	9.0	C21H36O2	[M + H]+	0.57	WHLKAPQFQ WYTAL	Long-chain fatty acids	*	↓
Allodihydrohydrocortisone	2407	365.2305	-4.8	3.9	C21H32O5	[M + H]+	0.82	ACSFOIGNUQ UIGE	Hydroxysteroids	*	↑
Desoxyisosteviol	5626	327.2323	8.6	4.9	C20H32O2	[M + Na]+	0.63	VHAZSZJVOS GWCB	Diterpenoids	*	↑↑
Amoxydramine	4289	272.1633	-4.5	4.2	C17H21NO2	[M + H]+	0.67	OEQNVWKW QPTBSC	Diphenylmethanes	*	↓↓
...	1928	629.3920	3.8	5.0	C33H56O11	[M + H]+	0.72	JPGSFBZIHY WZQB	Steroid glucuronide conjugates	*	↑
Hydroxymetronidazole	1011	188.0663	-1.4	0.5	C6H9N3O4	[M + H]+	0.93	AEHPOYAOL CAMIU	Nitroimidazoles	*	↑
Nudifloramide	1772	153.0653	-3.8	0.4	C7H8N2O2	[M + H]+	0.66	JLQSXXWTCJ PCBC	Nicotinamides	*	↑

Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar	ClassyFire Class	Sig.	Var.
...	897	438.2993	3.2	6.6	C21H44NO6P	[M + H]+	0.52	LQICBGFMER KBCG	Phosphoethanolamines	*	↓
...	1869	615.3735	-0.7	4.7	C32H54O11	[M + H]+	0.54	ONLXJASEXI XGRM	Diterpene glycosides	*	↑
...	939	466.3335	9.2	7.2	C23H48NO6P	[M + H]+	0.67	CDONWGCJD DHTLP	Monoalkylglycerophosphoethanolamines	*	↓
...	3021	223.0933	-3.7	3.7	C10H16O4	[M + Na]+	0.52	VWLPAWSXK LKROQ	Dicarboxylic acids and derivatives	*	↑
Diphenhydramine-N-glucuronide	3758	432.2007	-2.3	3.8	C23H29NO7	[M + H]+	0.68	OAIGZXXQYIJ BLR	Glucuronic acid derivatives	*	↓↓
...	6780	357.1298	-3.1	5.7	C18H22O6	[M + Na]+	0.70	KPDDZPBBEI ESMK	Phthalides	*	↓↓
Aminohippurate	2744	195.0765	0.3	0.7	C9H10N2O3	[M + H]+	0.57	HSMNQINEK MPTIC	Hippuric acids and derivatives	*	↑↑
...	1356	490.2896	-1.8	5.9	C22H46NO7P	[M + Na]+	0.99	VXUOFDJKYG DUJI	Lysophosphatidylcholines	*	↓
...	1892	358.2580	-2.3	4.6	C19H35NO5	[M + H]+	0.56	PKPXVGKUX UYEF	Beta hydroxy acids and derivatives	*	↑
Haematoporphyrin	1925	599.2853	-1.8	5.1	C34H38N4O6	[M + H]+	0.60	KFKRXESVM DBTNQ	Porphyrins	*	↑
...	986	592.3447	-1.2	6.5	C28H45N7O7	[M + H]+	0.56	ZDJRCWULNI PMON	Oligopeptides	*	↓
aspartylphenylalanin	2322	281.1115	-6.0	1.0	C13H16N2O5	[M + H]+	0.81	YZQCXOFQZ KCETR	Dipeptides	*	↑
2-amino-4-	3362	212.0901	3.8	0.4	C8H15NO4	[M + Na]+	0.56	ZTLDTRZYHP	Alpha amino acids	*	↑

Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar	ClassyFire Class	Sig.	Var.
methylpimelic acid								XXPS	and derivatives		
Pyroglutamylleucine	1727	243.1340	0.4	1.8	C11H18N2O4	[M + H]+	0.72	XXSAFGVAPG OYNT	Dipeptides	*	↑
19-Dfdat	3981	592.3197	-8.3	4.2	C30H51NO8	[M + K]+	0.50	HRTZUWRKG GQZSS	Aminoglycosides	*	↑↑
...	3261	257.1495	-0.3	2.9	C12H20N2O4	[M + H]+	0.83	AUBGCRXOZ PJJG	Alpha amino acids and derivatives	*	↑
Goshonoside F7	4217	631.3682	-1.1	4.4	C32H54O12	[M + H]+	0.53	HILPXFUUVM CHIZ	Diterpene glycosides	*	↑
...	7457	543.3134	-1.2	4.2	C26H48O10	[M + Na]+	0.63	KEBYFNGQBZ - SBNI		*	↑
Cortolone-3-glucuronide	2002	543.2789	-2.0	4.0	C27H42O11	[M + H]+	0.81	OKNSFVSKHQ JVRN	Steroidal glycosides	*	↑
...	1607	552.4038	2.5	8.0	C28H58NO7P	[M + H]+	1.00	UATOAILWG VYRQS	Lysophosphatidylcholine	*	↓
Triolone	5153	351.2523	-2.0	4.5	C21H34O4	[M + H]+	0.79	XBKMYUZPS UAVAK	Gluco/mineralocorticoids, progestogens and derivatives	*	↑
Adipoylcarnitine	1516	290.1590	-2.8	0.5	C13H23NO6	[M + H]+	0.84	BSVHAXJKBC WVDA	Acyl carnitines	*	↑
Glycerophosphocholine	1000	258.1095	-2.5	0.3	C8H20NO6P	[M + H]+	0.92	SUHOQUVVV LNYQR	Glycerophosphocholines	*	↓
...	1107	209.0892	-2.4	0.5	C8H14N2O3	[M + Na]+	0.52	VOLMYNNEV GLNMS	Alpha amino acids and derivatives	*	↑
...	3409	527.2858	1.3	3.8	C27H42O10	[M + H]+	0.57	VMRTWCCK WMUGSN	Triterpenoids	*	↑

Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar	ClassyFire Class	Sig.	Var.
Agamenoside J	2010	611.3789	-0.2	4.9	C33H54O10	[M + H]+	0.61	DHQWUUUFYWUJBRL	Steroidal glycosides	*	↑
...	3256	465.2475	-1.7	4.1	C25H36O8	[M + H]+	0.80	NIKZPECGCSUSBV	Steroid glucuronide conjugates	*	↑
...	886	482.3242	0.2	6.3	C23H48NO7P	[M + H]+	1.00	RJZVWDTYEWCUAR	Lysophosphatidylcholine	*	↓
9-oxooctadecanoate	1364	299.2579	-0.6	9.1	C18H34O3	[M + H]+	0.56	KNYQSFOUGYMRDE	Long-chain fatty acids	*	↓
Chamene	4294	159.1161	10.3	3.9	C10H16	[M + Na]+	0.50	MQFGIUKCEO GIGS	Branched unsaturated hydrocarbons	*	↓↓
lyso-PAF	838	482.3591	-2.9	6.8	C24H52NO6P	[M + H]+	1.00	VLBPIWYTPAXCFJ	Monoalkylglycerophosphocholines	*	↓
...	1620	597.5090	0.1	9.2	C36H68O6	[M + H]+	0.62	WFJYCOLSNJWRQK	Triacylglycerols	*	↓
Anhydroretinol	1078	269.2251	-4.8	7.8	C20H28	[M + H]+	0.56	FWNRILWHNGFAIN	Sesquiterpenoids	*	↓
indoleacrylate	805	188.0705	-0.5	0.9	C11H9NO2	[M + H]+	0.85	PLVPPCLBIEYEA	Indoles	*	↓
Skatole	1574	132.0802	-4.7	0.9	C9H9N	[M + H]+	0.76	ZFRKQXVRDFCRJG	Indoles	*	↓
...	1910	597.3631	-0.4	4.7	C32H52O10	[M + H]+	0.61	BQHIVNPPGHBFPC	Steroidal glycosides	*	↑
...	951	469.3897	1.9	8.3	C28H52O5	[M + H]+	0.60	UFLNXVJKZIUIOQ	Long-chain fatty acids	*	↓

Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar	ClassyFire Class	Sig.	Var.
S,S-Warfarin alcohol	866	311.1279	0.3	5.1	C19H18O4	[M + H]+	0.66	ZUJMMGHIYS AEOU	Hydroxycoumarins	*	↑↑
...	5282	655.2817	-2.7	3.4	C30H38N8O9	[M + H]+	0.60	SDGGGFGLMS UCLAU	-	*	↑↑
Erucylamide	30	338.3423	1.5	9.1	C22H43NO	[M + H]+	0.91	UAUDZVJPLU QNMU	Fatty amides	*	↓
1-Amhn	2666	293.1377	6.8	4.8	C12H16N6O3	[M + H]+	0.59	OFTZXSNXKY DMIL	Nitroimidazoles	*	↑↑
Racumin	5238	293.1175	0.8	4.8	C19H16O3	[M + H]+	0.61	ULSLJYXHZD TLQK	Hydroxycoumarins	*	↑↑
...	2690	255.1123	-2.2	4.5	C15H14N2O2	[M + H]+	0.50	DYAZDMVYCB GBHNX	Harmala alkaloids	*	↑
...	2437	263.2378	3.2	8.8	C18H30O	[M + H]+	0.57	TUCMDDWTB VMRTP	Fatty aldehydes	*	↓
1-dihomo-linoleoyl-GPC	854	548.3734	4.2	6.9	C28H54NO7P	[M + H]+	0.89	YYQVCMMXP IJVHY	Lysophosphatidylcholine	*	↓
beta-Indolylaldehyde	1051	146.0603	1.6	0.9	C9H7NO	[M + H]+	0.76	OLNJUISKUQ QNIM	Indoles	*	↓
...	2300	722.5133	1.9	8.9	C41H72NO7P	[M + H]+	0.90	RLLOITCRNQ RGJD	Glycerophosphoethanolamines	*	↓
...	826	572.3735	8.4	6.8	C28H56NO7P	[M + Na]+	0.81	SRZBVCCSIM MDOV	Glycerophosphocholines	*	↓
Androsta-3,5-dien-17-one	1352	271.2047	-3.6	4.5	C19H26O	[M + H]+	0.77	NINLAYUXSU KKHW	Androstane steroids	*	↓

Synonym: 化合物或其立体异构的别名。

ID: MCnebula分析中'Features'的唯一ID编号。

Synonym	ID	Precursor m/z	Err.	RT (min)	Formula	Adduct	TS	InChIKey planar	ClassyFire Class	Sig.	Var.
---------	----	------------------	------	-------------	---------	--------	----	--------------------	------------------	------	------

Err.: Mass Error (ppm), 前体离子分子量和理论分子量的偏差。

RT: Retention time, 保留时间。

Formula: Molecular Formula。

TS: Tanimoto similarity。

InChIKey planar: InChIKey的首个哈希块代码，代表分子骨架。

ClassyFire Class: ClassyFire分类系统中的该化合物的归类，- 表示该化合物在ClassyFire Web中未查询到。

Sig.: Q-value (HM vs HS, P-value的FDR矫正) 代表的显著性, *** 表示Q-value < 0.001, ** 表示Q-value < 0.01, * 表示Q-value < 0.05。

Var.: Log2(Fold Change) (HM / HS) 代表的变化水平, ↓↓ 或 ↑↑ 代表|log2(FC)| > 1, ↓ 或 ↑ 代表|log2(FC)| > 0.3。

(三) 分析的报告和R代码

以下对血清代谢组分析的脚本和报告可见于：https://github.com/Cao-lab-zcmu/exMCnebula2/tree/master/inst/extdata/scripts_evaluation/eucommia_workflow

Analysis on serum dataset

Contents

1 Abstract	1
2 Introduction	1
3 Set-up	2
4 Integrate data and Create Nebulae	2
4.1 Initialize analysis	2
4.2 Filter candidates	2
4.3 Filter chemical classes	3
4.4 Create Nebulae	4
4.5 Visualize Nebulae	5
5 Nebulae for Downstream analysis	5
5.1 Statistic analysis	8
5.2 Set tracer in Child-Nebulae	10
5.3 Quantification in Child-Nebulae	11
5.4 Annotate Nebulae	11
5.5 Query compounds	14
5.6 Pathway enrichment	17
5.7 Heatmap analysis	18
6 Verify Identification	18
7 Session infomation	23
Reference	24

1 Abstract

Untargeted mass spectrometry is a robust tool for biological research, but researchers universally time consumed by dataset parsing. We developed MCnebula, a novel visualization strategy proposed with multidimensional view, termed multi-chemical nebulae, involving in scope of abundant classes, classification, structures, sub-structural characteristics and fragmentation similarity. Many state-of-the-art technologies and popular methods were incorporated in MCnebula workflow to boost chemical discovery. Notably, MCnebula can be applied to explore classification and structural characteristics of unknown compounds that beyond the limitation of spectral library. MCnebula was integrated in R package and public available for custom R statistical pipeline analysis. Now, MCnebula2 (R object-oriented programming with S4 system) is further available for more friendly applications.

2 Introduction

We know that the analysis of untargeted LC-MS/MS dataset generally begin with feature detection. It detects ‘peaks’ as features in MS¹ data. Each feature may represents a compound, and assigned with MS²

三、小结

本部分内容应用血清代谢组学数据，说明MCnebula可用于通路分析和潜在生物标志物的发现。我们的大部分结果与报道^[47]的结果一致。此外，我们发现了更多超出光谱库匹配范围的代谢物。Wozniak等人鉴定的四个Top代谢物中，有三个与我们的重新鉴定相同，但只有一个代谢物是有争议的。Wozniak等人提到ACs化合物与SaB疾病有关联，ACs化合物在我们的研究中也被重新鉴定出来。Wozniak等人使用集合‘Features’选择（EFS）和Mann-Whitney U（MWU）测试的联合方法来筛选Top代谢物^[47]。当我们把MCnebula中集成的‘Binary comparison’方法得到的50个顶级‘Features’与W等人的联合方法得到的前50个代谢物（EFS的前50个和MWU的前50个）进行比较时，共筛选出37个重叠的代谢物，包括参考研究中的关键代谢物L-Thyroxine。根据‘Features’选择算法，Top‘Features’通常是不同的。除了一致的部分，MCnebula还揭示了与SaB疾病相关的其他化学类别的结果。我们发现了额外的化学类，即‘Lysophosphatidylcholines’（LPCs）和‘Bile acids, alcohols and derivatives’（BAs），这在Wozniak等人的研究中没有涉及。事实上，LPCs在炎症和动脉粥样硬化发展的背景下已被广泛调查^[60,59,57]。在最近的一篇综述中^[59]，充分描述了LPCs在血管炎症中的复杂作用，涉及与环境相关的促炎或抗炎作用，对先天免疫细胞和适应性免疫系统的影响等。LPCs水平的下降与一系列死亡风险增加的疾病有关^[57]。研究表明，血液中LPCs的浓度与严重的败血症或脓毒症休克有一定的相关性^[60]。据报道，LPCs与脓毒症患者的死亡率成反比^[61]。BAs的紊乱意味着肝脏功能紊乱和肠道微营养平衡的失衡^[62]。在BAs的Child-Nebulae中发现的BAs的化学多样性，由肠道微生物组决定，并允许对宿主的适应性反应进行复杂的调节。在我们的研究中，BAs的水平与SaB感染的相关性比ACs高。LPCs水平的下降表明了SaB感染的死亡风险。从LPCs到BAs，类固醇相关的类别，‘Lineolic acids and derivatives’，以及其他脂肪酸相关的类别，表明肝脏在SaB感染和死亡中起着核心作用。肝脏X受体（LXRs）在脂质代谢的转录控制中起着关键作用^[63]。LXRs通过激活溶血磷脂酰胆碱酰基转移酶3（‘lysophosphatidylcholine acyltransferase 3’，LPCAT3）来调节膜磷脂（Membrane phospholipid）组成，这与LPCs直接相关^[64]。上述化学类显示出与LXRs的相关性^[63]。

结论

LC-MS/MS数据的分析具有挑战性，因为其数据量大，未知化合物的信息多，而且参考光谱库有限。因此，我们建立了一个名为MCnebula的框架，通过关注关键的化学类别和多维度的可视化来促进质谱数据分析。MCnebula是用R语言提出的，并通过MCnebula包实现（目前为MCnebula2，<https://github.com/Cao-lab-zcmu/MCnebula2>）。作为一种综合的可视化方法，MCnebula对于没有生物信息学和计算机科学背景的研究人员来说可能更受欢迎。根据方法评估的结果，MCnebula归类的相对错误率比基准方法（GNPS）低，而其鉴定准确率高达70%。为了拓展MCnebula在生物学和化学领域的应用，我们开发了额外的工具包‘exMCnebula2’（<https://github.com/Cao-lab-zcmu/exMCnebula2>），为MCnebula工作流的应用提供了范例。为了说明MCnebula的广泛用途，我

们重新分析了一个用于代谢组学分析的人源血清数据集。结果表明，通过追踪潜在的生物标志物，‘Acyl carnitines’ 被筛选出来，这与文献一致^[47]。我们还研究了*E. ulmoides*的植物来源数据集（杜仲炮制前后），以实现快速的未知化合物注释和发现。我们的分析可以通过安装MCnebula2包和exMCnebula2包，并通过运行整理好的R代码进行重复。MCnebula在化学和生物学领域有很大的潜力。在未来，我们希望MCnebula的应用领域可以扩展到农业、食品科学、医药等领域。

创新点

第一，本研究首次设计一种基于非靶向LC-MS/MS分析技术的化学类的过滤算法，即ABC选择算法，用于代谢组学或药物分析等领域的化学类的聚焦分析。

第二，本研究提出了全新的应用于非靶向LC-MS/MS分析中数据集层面的可视化，即Parent-Nebula结合Child-Nebulae，在化学类的视角下全面、直观的审视数据集整体。

第三，本研究设计了一种全新的应用于非靶向LC-MS/MS分析的Features selection算法，通过结合统计分析和化学类，追踪高排名‘Features’（Tracing top’ features’），减少寻常Feature selection算法会有的偏倚。

第四，本研究将SIRIUS系列的尖端技术结合到了MCnebula工作流中，实现快速便捷的非靶向LC-MS/MS数据的全面分析。

第五，本研究将上述技术结合，编写R包并予以升级（MCnebula2 R包），包含严谨的数据结构（S4类存储对象）和综合简便的方法（Methods）和函数（Functions），同时包含详细的说明文档和示例数据，使用户无障碍运用R包进行分析。

第六，本研究将上述以外的其他工具结合到了‘exMCnebula2’ R包，用以示例拓展MCnebula工作流在代谢组、化学研究等领域的应用。

参考文献

- [1] Tsugawa H, Ikeda K, Takahashi M, 等. [A Lipidome Atlas in MS-DIAL 4](#)[J]. *Nature Biotechnology*, 2020, 38(10): 1159–1163.
- [2] Chong J, Soufan O, Li C, 等. [MetaboAnalyst 4.0: Towards More Transparent and Integrative Metabolomics Analysis](#)[J]. *Nucleic Acids Research*, 2018, 46(W1): W486–W494.
- [3] Tsugawa H. [Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software](#)[A]. 见: *Comprehensive Natural Products III*[M]. Elsevier, 2020: 189–210.
- [4] Wang M, Carver J J, Phelan V V, 等. [Sharing and Community Curation of Mass Spectrometry Data with Global Natural Products Social Molecular Networking](#)[J]. *Nature Biotechnology*, 2016, 34(8): 828–837.
- [5] Chambers M C, Maclean B, Burke R, 等. [A Cross-Platform Toolkit for Mass Spectrometry and Proteomics](#)[J]. *Nature Biotechnology*, 2012, 30(10): 918–920.
- [6] Röst H L, Sachsenberg T, Aiche S, 等. [OpenMS: A Flexible Open-Source Software Platform for Mass Spectrometry Data Analysis](#)[J]. *Nature Methods*, 2016, 13(9): 741–748.
- [7] Smith C A, Want E J, O’Maille G, 等. [XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification](#)[J]. *Analytical Chemistry*, 2006, 78(3): 779–787.
- [8] Pluskal T, Castillo S, Villar-Briones A, 等. [MZmine 2: Modular Framework for Processing, Visualizing, and Analyzing Mass Spectrometry-Based Molecular Profile Data](#)[J]. *BMC Bioinformatics*, 2010, 11(1): 395.
- [9] Myers O D, Sumner S J, Li S, 等. [One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks](#)[J]. *Analytical Chemistry*, 2017, 89(17): 8696–8703.
- [10] Fu J, Zhang Y, Wang Y, 等. [Optimization of Metabolomic Data Processing Using NOREVA](#)[J]. *Nature Protocols*, 2022, 17(1): 129–151.
- [11] Mahieu N G, Patti G J. [Systems-Level Annotation of a Metabolomics Data Set Reduces 25 000 Features to Fewer than 1000 Unique Metabolites](#)[J]. *Analytical Chemistry*, 2017, 89(19): 10397–10406.

- [12] Gloaguen Y, Kirwan J A, Beule D. Deep Learning-Assisted Peak Curation for Large-Scale LC-MS Metabolomics[J]. Analytical Chemistry, 2022, 94(12): 4930–4937.
- [13] Wang M, Jarmusch A K, Vargas F, 等. Mass Spectrometry Searches Using MASST[J]. Nature Biotechnology, 2020, 38(1): 23–26.
- [14] Wolf S, Schmidt S, Müller-Hannemann M, 等. In Silico Fragmentation for Computer Assisted Identification of Metabolite Mass Spectra[J]. BMC Bioinformatics, 2010, 11(1): 148.
- [15] Allen F, Greiner R, Wishart D. Competitive Fragmentation Modeling of ESI-MS/MS Spectra for Putative Metabolite Identification[J]. Metabolomics, 2015, 11(1): 98–110.
- [16] Ruttkies C, Schymanski E L, Wolf S, 等. MetFrag Relaunched: Incorporating Strategies beyond in Silico Fragmentation[J]. Journal of Cheminformatics, 2016, 8: 3.
- [17] Blaženović I, Kind T, Torbašinović H, 等. Comprehensive Comparison of in Silico MS/MS Fragmentation Tools of the CASMI Contest: Database Boosting Is Needed to Achieve 93% Accuracy[J]. Journal of Cheminformatics, 2017, 9(1): 32.
- [18] Kind T, Liu K-H, Lee D Y, 等. LipidBlast in Silico Tandem Mass Spectrometry Database for Lipid Identification[J]. Nature Methods, 2013, 10(8): 755–758.
- [19] Heinonen M, Shen H, Zamboni N, 等. Metabolite Identification and Molecular Fingerprint Prediction through Machine Learning[J]. Bioinformatics (Oxford, England), 2012, 28(18): 2333–2341.
- [20] Dührkop K, Shen H, Meusel M, 等. Searching Molecular Structure Databases with Tandem Mass Spectra Using CSI:FingerID[J]. Proceedings of the National Academy of Sciences, 2015, 112(41): 12580–12585.
- [21] Ludwig M, Dührkop K, Böcker S. Bayesian Networks for Mass Spectrometric Metabolite Identification via Molecular Fingerprints[J]. Bioinformatics (Oxford, England), 2018, 34(13): i333–i340.
- [22] Dührkop K, Fleischauer M, Ludwig M, 等. SIRIUS 4: A Rapid Tool for Turning Tandem Mass Spectra into Metabolite Structure Information[J]. Nature Methods, 2019, 16(4): 299–302.
- [23] Ashburner M, Ball C A, Blake J A, 等. Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium[J]. Nature Genetics, 2000, 25(1): 25–29.
- [24] Djoumbou Feunang Y, Eisner R, Knox C, 等. ClassyFire: Automated Chemical Classification with a Comprehensive, Computable Taxonomy[J]. Journal of Cheminformatics, 2016, 8(1): 61.

- [25] Blaženović I, Kind T, Sa M R, 等. **Structure Annotation of All Mass Spectra in Untargeted Metabolomics**.[J]. Analytical chemistry, United States: 2019, 91(3): 2155–2162.
- [26] Ernst M, Kang K B, Caraballo-Rodríguez A M, 等. **MolNetEnhancer: Enhanced Molecular Networks by Integrating Metabolome Mining and Annotation Tools**.[J]. Metabolites, 2019, 9(7).
- [27] Lee J, da Silva R R, Jang H S, 等. **In Silico Annotation of Discriminative Markers of Three Zanthoxylum Species Using Molecular Network Derived Annotation Propagation**.[J]. Food chemistry, England: 2019, 295: 368–376.
- [28] Sha B, Schymanski E L, Ruttkies C, 等. **Exploring Open Cheminformatics Approaches for Categorizing Per- and Polyfluoroalkyl Substances (PFASs)**.[J]. Environmental science. Processes & impacts, England: 2019, 21(11): 1835–1851.
- [29] Tripathi A, Vázquez-Baeza Y, Gauglitz J M, 等. **Chemically Informed Analyses of Metabolomics Mass Spectrometry Data with Qemistree**[J]. Nature Chemical Biology, 2021, 17(2): 146–151.
- [30] Wishart D S. **Emerging Applications of Metabolomics in Drug Discovery and Precision Medicine**[J]. Nature Reviews. Drug Discovery, 2016, 15(7): 473–484.
- [31] Guma M, Tiziani S, Firestein G S. **Metabolomics in Rheumatic Diseases: Desperately Seeking Biomarkers**[J]. Nature Reviews. Rheumatology, 2016, 12(5): 269–281.
- [32] Degenhardt F, Seifert S, Szymczak S. **Evaluation of Variable Selection Methods for Random Forests and Omics Data Sets**[J]. Briefings in Bioinformatics, 2019, 20(2): 492–503.
- [33] Neumann U, Genze N, Heider D. **EFS: An Ensemble Feature Selection Tool Implemented as R-package and Web-Application**[J]. BioData Mining, 2017, 10(1): 21.
- [34] Dührkop K, Nothias L-F, Fleischauer M, 等. **Systematic Classification of Unknown Metabolites Using High-Resolution Fragmentation Mass Spectra**[J]. Nature Biotechnology, 2021, 39(4): 462–471.
- [35] Platten M, Nollen E A A, Röhrlig U F, 等. **Tryptophan Metabolism as a Common Therapeutic Target in Cancer, Neurodegeneration and Beyond**[J]. Nature Reviews Drug Discovery, 2019, 18(5): 379–401.
- [36] Watrous J, Roach P, Alexandrov T, 等. **Mass Spectral Molecular Networking of Living Microbial Colonies**[J]. Proceedings of the National Academy of Sciences, 2012, 109(26): E1743–E1752.

- [37] Böcker S, Letzel M C, Lipták Z, 等. **SIRIUS: Decomposing Isotope Patterns for Metabolite Identification**[J]. Bioinformatics, 2009, 25(2): 218–224.
- [38] Dührkop K, Böcker S. **Fragmentation Trees Reloaded**[A]. 见: T.M. Przytycka. Research in Computational Molecular Biology[M]. Cham: Springer International Publishing, 2015, 9029: 65–79.
- [39] Ludwig M, Nothias L-F, Dührkop K, 等. **Database-Independent Molecular Formula Annotation Using Gibbs Sampling through ZODIAC**[J]. Nature Machine Intelligence, 2020, 2(10): 629–641.
- [40] Cai T, Guo Z-Q, Xu X-Y, 等. **Recent (2000-2015) Developments in the Analysis of Minor Unknown Natural Products Based on Characteristic Fragment Information Using LC-MS**[J]. Mass Spectrometry Reviews, 2018, 37(2): 202–216.
- [41] Hoffmann M A, Nothias L-F, Ludwig M, 等. **High-Confidence Structural Annotation of Metabolites Absent from Spectral Libraries**[J]. Nature Biotechnology, 2021.
- [42] Guha R. Chemical Informatics Functionality in R[J]. Journal of Statistical Software, 2007, 18(6).
- [43] Temple Lang D. RCurl: General Network (HTTP/FTP/...) Client Interface for R[M]. 2022.
- [44] Pletnev I, Erin A, McNaught A, 等. **InChIKey Collision Resistance: An Experimental Testing**[J]. Journal of Cheminformatics, 2012, 4(1): 39.
- [45] Smyth G K. **Limma: Linear Models for Microarray Data**[A]. 见: R. Gentleman, V.J. Carey, W. Huber, 等. Bioinformatics and Computational Biology Solutions Using R and Bioconductor[M]. New York: Springer-Verlag, 2005: 397–420.
- [46] Law C W, Zeglinski K, Dong X, 等. **A Guide to Creating Design Matrices for Gene Expression Experiments**[J]. F1000Research, 2020, 9: 1444.
- [47] Wozniak J M, Mills R H, Olson J, 等. **Mortality Risk Profiling of Staphylococcus Aureus Bacteremia by Multi-omic Serum Analysis Reveals Early Predictive and Pathogenic Signatures**[J]. Cell, 2020, 182(5): 1311–1327.e14.
- [48] Lai J, Huang L, Bao Y, 等. **A Deep Clustering-Based Mass Spectral Data Visualization Strategy for Anti-Renal Fibrotic Lead Compound Identification from Natural Products**[J]. The Analyst, 2022, 147(21): 4739–4751.
- [49] Huang L, Lyu Q, Zheng W, 等. **Traditional Application and Modern Pharmacological Research of Eucommia Ulmoides Oliv.**[J]. Chinese Medicine, 2021, 16(1): 73.

- [50] Huang L, Lyu Q, Zheng W, 等. Traditional Application and Modern Pharmacological Research of Eucommia Ulmoides Oliv.[J]. Chinese medicine, 2021, 16(1): 73.
- [51] Huang Y-X, Liu E-W, Wang L, 等. LC/MS/MS Determination and Pharmacokinetic Studies of Six Compounds in Rat Plasma Following Oral Administration of the Single and Combined Extracts of Eucommia Ulmoides and Dipsacus Asperoides.[J]. Chinese journal of natural medicines, China: 2014, 12(6): 469–476.
- [52] Hu F, An J, Li W, 等. UPLC-MS/MS Determination and Gender-Related Pharmacokinetic Study of Five Active Ingredients in Rat Plasma after Oral Administration of Eucommia Cortex Extract.[J]. Journal of ethnopharmacology, Ireland: 2015, 169: 145–155.
- [53] Sawada Y, Nakabayashi R, Yamada Y, 等. RIKEN Tandem Mass Spectral Database (ReSpect) for Phytochemicals: A Plant-Specific MS/MS-based Data Resource and Database[J]. Phytochemistry, 2012, 82: 38–45.
- [54] Pang Z, Chong J, Li S, 等. MetaboAnalystR 3.0: Toward an Optimized Workflow for Global Metabolomics[J]. Metabolites, 2020.
- [55] Picart-Armada S, Fernandez-Albert F, Vinaixa M, 等. FELLA: An R Package to Enrich Metabolomics Data[J]. BMC Bioinformatics, 2018, 19(1): 538.
- [56] Chen Y-H, Bi J-H, Xie M, 等. Classification-Based Strategies to Simplify Complex Traditional Chinese Medicine (TCM) Researches through Liquid Chromatography-Mass Spectrometry in the Last Decade (2011): Theory, Technical Route and Difficulty[J]. Journal of Chromatography A, 2021, 1651: 462307.
- [57] Krautbauer S, Eisinger K, Wiest R, 等. Systemic Saturated Lysophosphatidylcholine Is Associated with Hepatic Function in Patients with Liver Cirrhosis[J]. Prostaglandins & Other Lipid Mediators, 2016, 124: 27–33.
- [58] Melone M A B, Valentino A, Margarucci S, 等. The Carnitine System and Cancer Metabolic Plasticity[J]. Cell Death & Disease, 2018, 9(2): 228.
- [59] Knupplez E, Marsche G. An Updated Review of Pro- and Anti-Inflammatory Properties of Plasma Lysophosphatidylcholines in the Vascular System[J]. International Journal of Molecular Sciences, 2020, 21(12): E4501.

- [60] Park D W, Kwak D S, Park Y Y, 等. Impact of Serial Measurements of Lysophosphatidylcholine on 28-Day Mortality Prediction in Patients Admitted to the Intensive Care Unit with Severe Sepsis or Septic Shock[J]. Journal of Critical Care, 2014, 29(5): 882.e5–11.
- [61] Drobnik W, Liebisch G, Audebert F-X, 等. Plasma Ceramide and Lysophosphatidylcholine Inversely Correlate with Mortality in Sepsis Patients[J]. Journal of Lipid Research, 2003, 44(4): 754–761.
- [62] Perino A, Demagny H, Velazquez-Villegas L, 等. Molecular Physiology of Bile Acid Signaling in Health, Disease, and Aging[J]. Physiological Reviews, 2021, 101(2): 683–731.
- [63] Wang B, Tontonoz P. Liver X Receptors in Lipid Signalling and Membrane Homeostasis[J]. Nature Reviews. Endocrinology, 2018, 14(8): 452–463.
- [64] Zhang Q, Yao D, Rao B, 等. The Structural Basis for the Phospholipid Remodeling by Lysophosphatidylcholine Acyltransferase 3[J]. Nature Communications, 2021, 12(1): 6869.

文献综述

非靶向LC-MS/MS技术在小分子化合物分析中的应用

中文摘要

非靶向LC-MS/MS应用于小分子化合物的解析是融汇了尖端技术的学科领域。本综述就目前发展的几种重要解析技术做了概述，贯通成了一个较为完整的分析流程：1) ‘Features’ 的检测，应用了半自动或自动的计算机解析技术，分离 ‘Features’ 用于后续的定性或定量分析；2) MS/MS光谱的鉴定，应用主流的光谱匹配或者新兴的计算机技术对化合物进行鉴定或匹配，以SIRIUS为代表的新兴技术开拓了质谱用于探索未知化合物的应用；3) 统计分析，分为经典的统计理论和人工智能算法，对 Features集进行筛选从而用于下游的分析；4) 分子网络技术，在数据集的层面上可视化光谱，GNPS为这一技术提供了最全面的支持。

ABSTRACT

The application of un-targeted LC-MS/MS to the analysis of small molecule compounds is a discipline that incorporates cutting-edge technologies. This review provides an overview of several important analytical techniques that have been developed to form a more complete analytical process: 1) Detection of ‘Features’ , where semi-automatic or automatic computerized analytical techniques are applied to separate’ Features’ for subsequent qualitative or quantitative analysis. 2) Identification of MS/MS spectra, where mainstream spectral matching or emerging computerized techniques are applied to identify or match compounds. The emerging technologies represented by SIRIUS have opened up the application of mass spectrometry for the exploration of unknown compounds. 3) Statistical analysis, divided into classical statistical theory and artificial intelligence algorithms, to filter the set of Features for downstream analysis; 4) Molecular networking technology, to visualize the spectra at the level of data sets, GNPS provides the most comprehensive support for this technology.

前言

非靶向LC-MS/MS应用于小分子化合物 (< 1000 Da, 尤指代谢组学) 的解析是前沿的组学科学的一个领域，尖端的分析化学技术和先进的计算方法，全覆盖式描述复杂的生化混合物。这种技术由于其高灵敏度、小样本量、无需分离直接进样和高通量的特征而大受欢迎。伴随着它的发展和流行，越来越多的技术被提出以解决该领域的难题，比如MS/MS光谱在鉴定上的挑战。这些技术带有着21世纪人工智能技术发展的特征，机器学习被结合到LC-MS/MS分析的方方面面；这些技术的结合为质谱的应用带来了前所未有的突破。以下，本文就LC-MS/MS分析的几类主要技术的前沿进展做了概述，从 ‘Features’ 的检测开始，历经MS/MS光谱的鉴定，统计学上的筛选，还有数据集层面上的可视化。

正文

1. 非靶向LC-MS的‘Features’检测

峰检测（Peak detection, PD），又可称之为Feature detection（FD），以计算机的方式实现自动或半自动地检测LC-MS质谱数据中代表潜在化合物的连续信号峰（在一定时间内持续，构成近似正太分布曲线的峰形），并将其与其他信号峰或噪声峰分离，以便于后续定量分析或定性分析。在计算机技术还未普及之前，研究员人员偏向于手动检测并区分这些峰；然而在质谱技术愈加发展的当下，分析对象越来越复杂，分析代谢组学或者植物药（像中药）这类复杂成分，一个质谱数据中会包含代表潜在化合物的数千乃至数万个‘峰’(Features)，手动解析这些Features是不可能的，于是自动化的方法逐渐被开发。自动化伴随着真阳性（True Positive），假阳性（False Positive），真阴性（True Negative），假阴性（False Positive）的卷入和区分。真阳性是数据中代表化合物峰，是有意义于分析的峰；而假阳性则是被错误评估为代表化合物的峰，实际上是噪声干扰的峰；真阴性是计算机正确地将其排除在分析之外的噪声干扰峰；假阴性是计算机错误地将代表化合物的峰识别为噪声峰的检测。保留正向的结果（真阳性和真阴性），排除负向的结果（假阳性和假阴性），这是各类FD算法所追求的性能优化和提升。

FD根据MS/MS光谱的是否牵涉，可以分为两类（需要注意的是，在这一层面上，FD比PD所代表的含义更广）：1) 不考虑MS/MS光谱，仅从峰型上进行峰检测，这种方法更有益于区分化合物，因为某些异构类的化合物从MS²光谱上无法区分，但可以在峰形上（即保留时间不同）分离；2) 另一种更保守的做法，应用于LC-MS/MS中，检测具备任何MS²光谱的Features，严格意义上它并不区分峰形，并在随后的流程中合并相同或近似（根据光谱相似度，例如计算余弦相似度）MS²代表的Features，因此，这种做法无法区分更多的异构化合物。第一类做法在算法上有更高的要求。传统的像涉及小波变换（Wavelet transform）^[3,2,1]或者贝叶斯概率（Bayesian Probability）^[4]数学理论的解析方法难以做到完全的自动化解析，于是规则（Principle）和阈值（Threshold）被设定以纠正可能存在的条件带来的偏差（仪器条件、样品条件、操作误差等），以实现需要人工考察和设定方法和参数的半自动化分析；像建立于R的XCMS^[3]，独立的MZmine^[5]，OpenMS^[6]，MS-DIAL^[7]等是其中的代表。当下可能更流行的，涉及人工智能的算法可以实现全自动化FD，但这种做法对机器训练的数据集的量和质（涵盖各种类型的数据特征）需要足够大和足够丰富，否则无法达成正向的解析结果；深度神经网络^[8]这样的算法模型可能会是今后的热门。第二类（仅检测具备MS²光谱的Features）的FD更简单，但偏消极地回避峰的解析，而是仅保留有有解析可能的Features，理所当然的是，这种做法仅适用于定性分析，除非选择峰强度代替峰面积进行定量分析。第二类做法被许多定性分析的工具所采用，像GNPS的经典分子网络（Classical Molecular Networking）^[9]，SIRIUS 4软件的自动FD^[10]；因为更加便捷，对分析人员的技术要求更低，更易于实现，无需考虑峰检测即可将前处理流程纳入到分析管道中。

通常，非靶向质谱数据的分析从LC-MS的FD开始，获得一定数量的Features，从而进入下一阶段的化合物鉴定（或者先统计筛选）。需要注意的是，一般的LC-MS/MS的预处理流程所指的FD不仅包括单一数据的FD，还包括寻找同位素峰（Grouping Isotopic Peaks）、峰对齐（Alignment）、峰再寻找（Gap filling）等流程（见正文>4. LC-MS/MS与分子网络的Feature-based Molecular Networking）。

2. MS/MS光谱的鉴定

MS/MS光谱的解析鉴定是非靶向质谱数据分析的最大挑战，目前这一领域尚有待开拓和解决。正如本论文正文的前言部分提及的，MS/MS的解析在当下可分为（单凭化学经验分析的方法除外）：

1) 参考光谱库的匹配；2) 匹配模拟的理论碎片光谱；3) 通过机器学习进行预测。参考光谱库往往需要具备更高的品质，它需要涵盖化合物理论上能产生的绝大部分MS²信号峰，并且不包含或仅具有极少数的噪声峰，这样才能适用于以匹配的方式来鉴定化合物。光谱库的匹配仍然是主流的鉴定方法，因为它更高的准确度。然而，光谱库的匹配这一方法也是复杂的。参考光谱库往往因为其商业价值而被建立或拓展（后来可能收集于像GNPS这样的开放性网站^[11]），这些光谱还更常见为热门的代谢物，相较于PubChem所涵盖的结构库而言（超过十亿种结构），它包含的分子数量太少，对质谱技术的应用推广带来消极的影响。值得深思的是，即使理论上能够匹配的光谱，因为噪声带来的影响，也可能使得匹配无法成功，导致鉴定失败，本论文第二部分的数据已经证明了这一点。无法正确匹配的原因可能有三种，其一是数据或者参考光谱中过多的噪声带来的干扰，其二是仪器或者喷撞能量的不同带来的影响，其三是匹配的算法的性能有待优化。参考光谱库的建立需要更多的成本，为了降低成本，理论碎片光谱应运而生（有趣的是，这种光谱被称之为*In silico* Fragmentation Spectra）。理论碎片光谱拓展了光谱匹配的应用，它在原理上也能细分为两种：一种是根据化学经验设定原理，推断化合物可能产生何种碎片^[12]；第二种也与机器学习有关^[16,15,14,13]，与方法3) 所代表的机器学习预测化合物有相似之处，不同的是，它预测出光谱供于匹配，而非像方法3) 那样，从待鉴定的数据光谱中预测出分子指纹用于在结构库中搜索。接下来，我们需要谈一谈以机器预测的方式从待测数据光谱中预测出结构这一方法。

先于机器预测，我们需要先声明分子结构数据库的存在。使用分子结构数据库（Molecular Structure Database）而不是参考光谱库（Reference Spectral Library）进行匹配的方式脱离了MS/MS鉴定的‘comfort zone’^[17]，但这一方法大有裨益于发现未知化合物。像PubChem的结构库，许多化合物仅记录了它们的结构信息（以SMILES或者InChI的形式记录了——值得一提的是，SMILES和InChI都是化学结构式的线性书写方式，以便于计算机记录和计算；然而，生成唯一SMILES的算法是商业性的，因而InChI被推出；此外，InChIKey是InChI的不可逆的编码压缩形式，而InChIKey planar可以指代为它的首个哈希块代码，代表分子骨架），而不包含任何实验记录：它们存在，但还未被探索。过去的生物学研究常常局限于那些有限的已知的化合物，现在，随着人工智能技术的普及，这一藩篱似乎即将被突破。当下已经有许多优秀的用于结构数据库搜索的工具，与机器学习技术和质谱技术相结合，像MolDiscovery^[18]，MS-FINDER^[19]，MetFrag^[20]，CSI:FingerID^[21]，它们似乎只被较少一

部分的研究者用于生物学的研究（还有一种虽然运用了机器学习，但并没有搜索结构数据库来鉴定，像MetDNA^[22]，需要注意区分其准确度所代表的含义，因为结构数据库的大小是代谢物数据库或光谱数据库的几何倍以上）。结构数据库搭乘机器学习的特快列车跑得没有想像中那么快，原因在于它的关键‘限速度’：鉴定的准确度；大多数的方法都还浮动在70%以下。

SIRIUS软件系列是人工智能运用于质谱鉴定的佼佼者，它们自称为‘Cutting-edge technologies’或者‘State of art technologies’，事实似乎的确如他们所说。SIRIUS的发展由来已久，早在2005年，第一篇（似乎是）有关于它的算法的理论问世了，有关于质量分解（Mass Decomposition）^[23]。后续的报道涉及了同位素模式的运用^[24]。在2009年，第一个版本的SIRIUS问世了，它还仅包含分子式预测的模块^[25]。后来碎片树（Fragmentation Tree）的理论被结合在它的方法中^[27,26]，并对质量分解的计算的速度做了优化^[28]。随着碎片树理论的再优化^[29]，SIRIUS中用于结构数据库搜索的模块也问世了，即CSI:FingerID^[21]。SIRIUS对MS/MS鉴定领域的问题的解决的广度在拓展。相关研究人员对光谱匹配的显著性评估做了优化^[30]，这成为后来的SIRIUS的COSMIC模块的基础^[31]。在2019年，SIRIUS 4软件，一个集成了先前版本的SIRIUS和CSI:FingerID模块的工具被公布了^[10]，有趣的是，在2020年，它又一次被报道了^[32]。ZODIAC^[33]，一个基于网络算法结合来强化分子式预测的工具被设计并集成于SIRIUS 4中，通过数千个Features的MS²光谱的同时预测，它表现出比此前的SIRIUS更高的对分子式注释的准确率；ZODIAC本身不做预测，但它的算法对SIRIUS的分子式预测的候选项做了重新的排序。到了2021年，SIRIUS研究的广度又一次拓宽了，一个称之为CANOPUS的工具问世了^[34]，它以ClassyFire的化学分类学理论为基础，结合机器学习，能够对MS²光谱进行化学类的注释，马上被集成在了SIRIUS 4中。最让人震惊的是，CANOPUS对化学类的预测绕开了化学结构的鉴定，即使不知道确切的化学结构，依然能根据已有的片段信息判断化合物的化学类；这种技术的出现凸显了MS²鉴定的“魅力”，因为有的时候即使最经验丰富的化学工作者也不一定能从碎片光谱中窥探出化合物的结构全貌，这可能是光谱的信息缺失或者噪声带来的干扰，但这并不意味着这光谱就是毫无意义的，因为某些片段暗示了该化合物的亚结构，或者说，局部结构，这种技术大大推进了非靶向LC-MS/MS技术用于探索未知化合物。CANOPUS一共能预测ClassyFire体系中的2497种化学类，几乎涵盖了所有代谢物会涉及的化学类，它所报道的在交叉验证中的准确度达到了99.7%（在坚信CANOPUS的正确率之前，可能需要先了解优势结构（Dominant structure）和亚结构（Sub-structure）的区别）。后来的2021年，称之为COSMIC的的工具^[35]（源于光谱匹配，高于高匹配）的工具被集成在了SIRIUS 4中，它能通过机器预测的方式，能将不存在于光谱库的化合物以近似于光谱库匹配的高准确度的方式来鉴定。在2022年，在SIRIUS的官网（<https://bio.informatik.uni-jena.de/software/sirius>），SIRIUS 5被发布了。在2023年的最初几个月里，CSI:FingerID的网络服务（结构数据库的检索依赖于网络功能）对SIRIUS 4的支持被关闭，只有在SIRIUS 5中注册登陆才能享受到SIRIUS 5的结构库搜索功能。

3. 非靶向LC-MS的半定量分析——Statistic Analysis与Feature selection

‘Features’ 数据集来源于Features detection，一般经过FD能得到一个Features的量化表（Features Quantification Table, FQT），大多数的开源软件或者商业软件（例如，Waters的Progenesis QI）都能做到这一点。FQT以峰面积或峰强度来代表定量信息。在后续的生物学研究上，常常通过统计检验的方式来挖掘有显著性意义的Features，筛选为潜在的生物标志物。MetaboAnalyst是代表性的应用于Features统计检验的工具^[36] (<https://www.metaboanalyst.ca>)。这种通过组间差异分析来筛选出有意义的Features的做法被称之为Features selection，更早被应用于基因表达的差异性分析（Feature selection）被用于提高准确性和降低模型的复杂性，因为它可以去除冗余和不相关的特征以降低输入维度，并帮助生物学家确定将基因表达与疾病或感兴趣的表型联系起来的基本机制^[37]），后来被推广到了代谢组学。随着人工智能技术的发展，Feature selection算法的选择性也变得更加多样了^[40,39,38,37]。

4. LC-MS/MS与分子网络

自从分子网络（Molecular Networking, MN）在2012年被介绍以来^[9]，它在非靶向代谢组学的分析中就发挥着越来越重要的作用。对MN的发展贡献最大的是Global Natural Products Society（GNPS）^[11]，它以网络图可视化MS/MS光谱这一分析技术为基础，越来越发展成为一个开放的知识库，供全社会组织和分享原始、处理或注释的碎片质谱数据（MS/MS）；GNPS在数据的整个生命周期（从最初的数据采集/分析到发表后）中协助识别和发现。GNPS对分子网络的技术提供了最大的支持，从最初的经典分子网络（Classical Molecular Networking, CMN）^[9]，到后来的基于Features的分子网络（Feature-based Molecular Networking, FBMN）^[41]，还有辅助于分子网络鉴定分析的各类工具，例如用于注释传播的NAP^[42]，用于提供化学分类学注释的MolNetEnhancer^[43]等；所有的这些工具都在GNPS的网络服务中开放获取（<https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp>），并提供了细致入微的说明文档（<https://ccms-ucsd.github.io/GNPSDocumentation>）。

可以认为，GNPS提供的分子网络技术主要为两类，一种是CMN，另一种是FBMN，而其他的工具都是对这两种技术的强化工具。CMN对用户的操作需求更小，只需要将转化为开源格式的数据（例如将.raw转化为.mzML格式的数据）上传至GNPS的服务器，在较短的时间之后，用户就会收到一封完成分析的邮件；FBMN对用户有一定的分析技术需求，它需要结合用户自主进行Feature Detection（FD）处理后得到的Features量化表（FQT）和MS/MS光谱列表文件（.mgf或.msp）上传至GNPS网络服务器。CMN提供自动化的FD，但是，正如前文（正文 > 1. 非靶向LC-MS的‘Features’检测）提到的，CMN集成的FD是基于MS/MS光谱的FD，与峰形无关，它无法区分同分异构体，也无法提供峰面积为源的量化信息；此外，CMN的FD得到的Features的ID信息（对Features的编号）是服务器自主分配的，可能更难以被用户用于后续的分析中。FBMN的前处理是用户在本地自主完成的，关于ID的分配也全权在用户手中，用户可以容易地将FBMN的分析与其他分析工具相结合（例如，纳入R的BioConductor提供的生物学分析工作流

(<https://bioconductor.org/packages/release/BiocViews.html#Workflow>)，后续关于GNPS服务器的任务主要是将处理后的数据构建成分子网络（计算光谱相似性，一般是Cosine Similarity），并在参考光谱数据库进行匹配，鉴定化合物。值得细说的是，FBMN的前处理并不仅仅是FD，包括一个相对完整的质谱数据预处理流程^[41]，以MZmine结合的FBMN为例子（因为FBMN的前处理在本地完成，因此必要的工具需要自主配备和使用，FBMN对大多数的流行工具的处理后格式都提供了支持），流程包括：1) Data import; 2) Peak detection; 3) Chromatogram building; 4) Chromatogram deconvolution; 5) Isotope grouping; 6) Feature alignment; 7) MS row filter; 8) Isotope filter; 9) Gap filling; 10) Normalization; 11) Manual validation; 12) Features set export。（<https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking-with-mzmine2>）。流程具有一定的自主性，可以减少或添加处理步骤，以适应对象数据的特性；在这方面，像MZmine或XCMS这类工具提供了更灵活的选择^[44,3]，但是相对而言，它们对使用者有着更高的技术要求。

小结

非靶向LC-MS/MS的应用是前沿学科和尖端技术融汇发展的领域，被广泛运用于代谢组学、天然产物、植物药物等的分析鉴定。本文主要就LC-MS/MS在小分子领域鉴定分析的技术做了综述，这方面技术涉及了Feature Detection, MS/MS光谱的鉴定，统计分析和Feature selection，还有用于可视化分析的分子网络等。在非靶向LC-MS/MS分析的领域，以机器学习和运用为主要特征的人工智能技术正日趋融合发展，深入到分析的方方面面；在其最大挑战——MS/MS光谱的鉴定面前，以SIRIUS为代表的鉴定技术正跨出了参考光谱库的藩篱，向着分子结构数据库探索。在未来，非靶向LC-MS/MS将越来越与人工智能技术密不可分，并成为分析鉴定未知化合物的重要工具。

参考文献

- [1] Bai C, Xu S, Tang J, 等. A 《Shape-Orientated》 Algorithm Employing an Adapted Marr Wavelet and Shape Matching Index Improves the Performance of Continuous Wavelet Transform for Chromatographic Peak Detection and Quantification[J]. Journal of Chromatography A, 2022, 1673: 463086.
- [2] Myers O D, Sumner S J, Li S, 等. One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks[J]. Analytical Chemistry, 2017, 89(17): 8696–8703.

- [3] Smith C A, Want E J, O'Maille G, 等. **XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification**[J]. *Analytical Chemistry*, 2006, 78(3): 779–787.
- [4] Woldegebriel M, Vivó-Truyols G. **Probabilistic Model for Untargeted Peak Detection in LC-MS Using Bayesian Statistics**[J]. *Analytical Chemistry*, 2015, 87(14): 7345–7355.
- [5] Pluskal T, Castillo S, Villar-Briones A, 等. **MZmine 2: Modular Framework for Processing, Visualizing, and Analyzing Mass Spectrometry-Based Molecular Profile Data**[J]. *BMC Bioinformatics*, 2010, 11(1): 395.
- [6] Röst H L, Sachsenberg T, Aiche S, 等. **OpenMS: A Flexible Open-Source Software Platform for Mass Spectrometry Data Analysis**[J]. *Nature Methods*, 2016, 13(9): 741–748.
- [7] Tsugawa H, Ikeda K, Takahashi M, 等. **A Lipidome Atlas in MS-DIAL 4**[J]. *Nature Biotechnology*, 2020, 38(10): 1159–1163.
- [8] Melnikov A D, Tsentalovich Y P, Yanshole V V. **Deep Learning for the Precise Peak Detection in High-Resolution LC**[J]. *Analytical Chemistry*, 2020, 92(1): 588–592.
- [9] Watrous J, Roach P, Alexandrov T, 等. **Mass Spectral Molecular Networking of Living Microbial Colonies**[J]. *Proceedings of the National Academy of Sciences*, 2012, 109(26): E1743–E1752.
- [10] Dührkop K, Fleischauer M, Ludwig M, 等. **SIRIUS 4: A Rapid Tool for Turning Tandem Mass Spectra into Metabolite Structure Information**[J]. *Nature Methods*, 2019, 16(4): 299–302.
- [11] Wang M, Carver J J, Phelan V V, 等. **Sharing and Community Curation of Mass Spectrometry Data with Global Natural Products Social Molecular Networking**[J]. *Nature Biotechnology*, 2016, 34(8): 828–837.
- [12] Wang Y, Wang X, Zeng X. **MIDAS-G: A Computational Platform for Investigating Fragmentation Rules of Tandem Mass Spectrometry in Metabolomics**[J]. *Metabolomics*, 2017, 13(10): 116.
- [13] Chao A, Al-Ghoul H, McEachran A D, 等. **In Silico MS/MS Spectra for Identifying Unknowns: A Critical Examination Using CFM-ID Algorithms and ENTACT Mixture Samples**[J]. *Analytical and Bioanalytical Chemistry*, 2020, 412(6): 1303–1315.
- [14] da Silva R R, Wang M, Nothias L-F, 等. **Propagating Annotations of Molecular Networks Using in Silico Fragmentation**[J]. *PLoS computational biology*, 2018, 14(4): e1006089.

- [15] Blaženović I, Kind T, Torbašinović H, 等. **Comprehensive Comparison of in Silico MS/MS Fragmentation Tools of the CASMI Contest: Database Boosting Is Needed to Achieve 93% Accuracy**[J]. Journal of Cheminformatics, 2017, 9(1): 32.
- [16] Wolf S, Schmidt S, Müller-Hannemann M, 等. **In Silico Fragmentation for Computer Assisted Identification of Metabolite Mass Spectra**[J]. BMC Bioinformatics, 2010, 11(1): 148.
- [17] Böcker S. **Searching Molecular Structure Databases Using Tandem MS Data: Are We There Yet?**[J]. Current Opinion in Chemical Biology, 2017, 36: 1–6.
- [18] Cao L, Guler M, Tagirdzhanov A, 等. **MolDiscovery: Learning Mass Spectrometry Fragmentation of Small Molecules**[J]. Nature Communications, 2021, 12(1): 3718.
- [19] Tsugawa H. **Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software**[A]. 见: Comprehensive Natural Products III[M]. Elsevier, 2020: 189–210.
- [20] Ruttkies C, Schymanski E L, Wolf S, 等. **MetFrag Relaunched: Incorporating Strategies beyond in Silico Fragmentation**[J]. Journal of Cheminformatics, 2016, 8: 3.
- [21] Dührkop K, Shen H, Meusel M, 等. **Searching Molecular Structure Databases with Tandem Mass Spectra Using CSI:FingerID**[J]. Proceedings of the National Academy of Sciences, 2015, 112(41): 12580–12585.
- [22] Shen X, Wang R, Xiong X, 等. **Metabolic Reaction Network-Based Recursive Metabolite Annotation for Untargeted Metabolomics**[J]. Nature Communications, 2019, 10(1): 1516.
- [23] Böcker S, Lipták Z. **Efficient Mass Decomposition**[A]. Proceedings of the 2005 ACM Symposium on Applied Computing - SAC '05[C]. Santa Fe, New Mexico: ACM Press, 2005: 151.
- [24] Böcker S, Letzel M C, Lipták Z, 等. **Decomposing Metabolomic Isotope Patterns**[A]. 见: D. Hutchison, T. Kanade, J. Kittler, 等. Algorithms in Bioinformatics[M]. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, 4175: 12–23.
- [25] Böcker S, Letzel M C, Lipták Z, 等. **SIRIUS: Decomposing Isotope Patterns for Metabolite Identification**[J]. Bioinformatics, 2009, 25(2): 218–224.
- [26] Rasche F, Svatoš A, Maddula R K, 等. **Computing Fragmentation Trees from Tandem Mass Spectrometry Data**[J]. Analytical Chemistry, 2011, 83(4): 1243–1251.
- [27] Rasche F, Scheubert K, Hufsky F, 等. **Identifying the Unknowns by Aligning Fragmentation Trees**[J]. Analytical Chemistry, 2012, 84(7): 3417–3426.

- [28] Dürkop K, Ludwig M, Meusel M, 等. *Faster Mass Decomposition*[A]. 见: D. Hutchison, T. Kanade, J. Kittler, 等. *Algorithms in Bioinformatics*[M]. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, 8126: 45–58.
- [29] Dürkop K, Böcker S. *Fragmentation Trees Reloaded*[A]. 见: T.M. Przytycka. *Research in Computational Molecular Biology*[M]. Cham: Springer International Publishing, 2015, 9029: 65–79.
- [30] Scheubert K, Hufsky F, Petras D, 等. *Significance Estimation for Large Scale Metabolomics Annotations by Spectral Matching*[J]. *Nature Communications*, 2017, 8(1): 1494.
- [31] Hoffmann M A, Nothias L-F, Ludwig M, 等. *Assigning Confidence to Structural Annotations from Mass Spectra with COSMIC*[R]. *Bioinformatics*, 2021.
- [32] Ludwig M, Fleischauer M, Dürkop K, 等. *De Novo Molecular Formula Annotation and Structure Elucidation Using SIRIUS 4*[A]. 见: S. Li. *Computational Methods and Data Analysis for Metabolomics*[M]. New York, NY: Springer US, 2020, 2104: 185–207.
- [33] Ludwig M, Nothias L-F, Dürkop K, 等. *Database-Independent Molecular Formula Annotation Using Gibbs Sampling through ZODIAC*[J]. *Nature Machine Intelligence*, 2020, 2(10): 629–641.
- [34] Dürkop K, Nothias L-F, Fleischauer M, 等. *Systematic Classification of Unknown Metabolites Using High-Resolution Fragmentation Mass Spectra*[J]. *Nature Biotechnology*, 2021, 39(4): 462–471.
- [35] Hoffmann M A, Nothias L-F, Ludwig M, 等. *High-Confidence Structural Annotation of Metabolites Absent from Spectral Libraries*[J]. *Nature Biotechnology*, 2021.
- [36] Chong J, Soufan O, Li C, 等. *MetaboAnalyst 4.0: Towards More Transparent and Integrative Metabolomics Analysis*[J]. *Nucleic Acids Research*, 2018, 46(W1): W486–W494.
- [37] Liang S, Ma A, Yang S, 等. *A Review of Matched-pairs Feature Selection Methods for Gene Expression Data Analysis*[J]. *Computational and Structural Biotechnology Journal*, 2018, 16: 88–97.
- [38] Sharma A, Lysenko A, Boroevich K A, 等. *DeepFeature: Feature Selection in Nonimage Data Using Convolutional Neural Network*[J]. *Briefings in Bioinformatics*, 2021, 22(6): bbab297.
- [39] Fu J, Zhang Y, Liu J, 等. *Pharmacometabonomics: Data Processing and Statistical Analysis*[J]. *Briefings in Bioinformatics*, 2021, 22(5): bbab138.

- [40] Neumann U, Genze N, Heider D. **EFS: An Ensemble Feature Selection Tool Implemented as R-package and Web-Application**[J]. BioData Mining, 2017, 10(1): 21.
- [41] Nothias L-F, Petras D, Schmid R, 等. **Feature-Based Molecular Networking in the GNPS Analysis Environment**[J]. Nature Methods, 2020, 17(9): 905–908.
- [42] da Silva R R, Wang M, Nothias L-F, 等. **Propagating Annotations of Molecular Networks Using in Silico Fragmentation**[J]. A. Schlessinger. PLOS Computational Biology, 2018, 14(4): e1006089.
- [43] Ernst M, Kang K B, Caraballo-Rodríguez A M, 等. **MolNetEnhancer: Enhanced Molecular Networks by Integrating Metabolome Mining and Annotation Tools**[R]. Biochemistry, 2019.
- [44] Pluskal T, Castillo S, Villar-Briones A, 等. **MZmine 2: Modular Framework for Processing, Visualizing, and Analyzing Mass Spectrometry-Based Molecular Profile Data**[J]. BMC Bioinformatics, 2010, 11(1): 395.