# 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 Negtive），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 transfrom）[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，这方面的技术涉及了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](https://doi.org/10.1016/j.chroma.2022.463086)[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](https://doi.org/gbrjtm)[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](https://doi.org/b58xrd)[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](https://doi.org/10.1021/acs.analchem.5b01521)[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](https://doi.org/bxbwnj)[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](https://doi.org/f82r32)[J]. Nature Methods, 2016, 13(9): 741–748.

[7] Tsugawa H, Ikeda K, Takahashi M, 等. [A Lipidome Atlas in MS-DIAL 4](https://doi.org/10.1038/s41587-020-0531-2)[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](https://doi.org/10.1021/acs.analchem.9b04811)[J]. Analytical Chemistry, 2020, 92(1): 588–592.

[9] Watrous J, Roach P, Alexandrov T, 等. [Mass Spectral Molecular Networking of Living Microbial Colonies](https://doi.org/10.1073/pnas.1203689109)[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](https://doi.org/10.1038/s41592-019-0344-8)[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](https://doi.org/10.1038/nbt.3597)[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](https://doi.org/10.1007/s11306-017-1258-z)[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](https://doi.org/10.1007/s00216-019-02351-7)[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](https://doi.org/10.1371/journal.pcbi.1006089)[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](https://doi.org/10.1186/s13321-017-0219-x)[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](https://doi.org/10.1186/1471-2105-11-148)[J]. BMC Bioinformatics, 2010, 11(1): 148.

[17] Böcker S. [Searching Molecular Structure Databases Using Tandem MS Data: Are We There Yet?](https://doi.org/10.1016/j.cbpa.2016.12.010)[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](https://doi.org/10.1038/s41467-021-23986-0)[J]. Nature Communications, 2021, 12(1): 3718.

[19] Tsugawa H. [Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software](https://doi.org/10.1016/B978-0-12-409547-2.14645-1)[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](https://doi.org/10.1186/s13321-016-0115-9)[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](https://doi.org/10.1073/pnas.1509788112)[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](https://doi.org/10.1038/s41467-019-09550-x)[J]. Nature Communications, 2019, 10(1): 1516.

[23] Böcker S, Lipták Z. [Efficient Mass Decomposition](https://doi.org/10.1145/1066677.1066715)[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](https://doi.org/10.1007/11851561_2)[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](https://doi.org/10.1093/bioinformatics/btn603)[J]. Bioinformatics, 2009, 25(2): 218–224.

[26] Rasche F, Svatoš A, Maddula R K, 等. [Computing Fragmentation Trees from Tandem Mass Spectrometry Data](https://doi.org/10.1021/ac101825k)[J]. Analytical Chemistry, 2011, 83(4): 1243–1251.

[27] Rasche F, Scheubert K, Hufsky F, 等. [Identifying the Unknowns by Aligning Fragmentation Trees](https://doi.org/gntcfj)[J]. Analytical Chemistry, 2012, 84(7): 3417–3426.

[28] Dührkop K, Ludwig M, Meusel M, 等. [Faster Mass Decomposition](https://doi.org/10.1007/978-3-642-40453-5_5)[A]. 见: D. Hutchison, T. Kanade, J. Kittler, 等. Algorithms in Bioinformatics[M]. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, 8126: 45–58.

[29] Dührkop K, Böcker S. [Fragmentation Trees Reloaded](https://doi.org/10.1007/978-3-319-16706-0_10)[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](https://doi.org/10.1038/s41467-017-01318-5)[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](https://doi.org/10.1101/2021.03.18.435634)[R]. Bioinformatics, 2021.

[32] Ludwig M, Fleischauer M, Dührkop K, 等. [De Novo Molecular Formula Annotation and Structure Elucidation Using SIRIUS 4](https://doi.org/10.1007/978-1-0716-0239-3_11)[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, Dührkop K, 等. [Database-Independent Molecular Formula Annotation Using Gibbs Sampling through ZODIAC](https://doi.org/10.1038/s42256-020-00234-6)[J]. Nature Machine Intelligence, 2020, 2(10): 629–641.

[34] Dührkop K, Nothias L-F, Fleischauer M, 等. [Systematic Classification of Unknown Metabolites Using High-Resolution Fragmentation Mass Spectra](https://doi.org/10.1038/s41587-020-0740-8)[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](https://doi.org/10.1038/s41587-021-01045-9)[J]. Nature Biotechnology, 2021.

[36] Chong J, Soufan O, Li C, 等. [MetaboAnalyst 4.0: Towards More Transparent and Integrative Metabolomics Analysis](https://doi.org/10.1093/nar/gky310)[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](https://doi.org/10.1016/j.csbj.2018.02.005)[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](https://doi.org/10.1093/bib/bbab297)[J]. Briefings in Bioinformatics, 2021, 22(6): bbab297.

[39] Fu J, Zhang Y, Liu J, 等. [Pharmacometabonomics: Data Processing and Statistical Analysis](https://doi.org/10.1093/bib/bbab138)[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](https://doi.org/10.1186/s13040-017-0142-8)[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](https://doi.org/10.1038/s41592-020-0933-6)[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](https://doi.org/gdc9cj)[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](https://doi.org/10.1101/654459)[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](https://doi.org/bxbwnj)[J]. BMC Bioinformatics, 2010, 11(1): 395.