

基于质谱的未知次生代谢物结构解析研究进展与展望

纪宏超¹ 李正艳^{1, 2, 3}

(1. 中国农业科学院深圳农业基因组研究所(岭南现代农业科学与技术广东省实验室深圳分中心), 深圳 518000; 2. 河南大学生命科学学院, 开封 475004; 3. 河南大学深圳研究院, 深圳 518000)

摘要: 次生代谢产物研究对于植物生长发育、环境适应以及人类健康和药物研发具有重要意义。液相色谱-质谱联用平台(LC-MS)已成为次生代谢研究的首选策略。然而, 代谢物结构解析仍受限于标准谱图库覆盖度不足的问题。由于代谢物结构数据库的覆盖度远高于标准谱图库的覆盖度, 通过人工智能方法建立代谢物结构与谱图之间的关联, 从而实现基于质谱谱图搜索结构数据库是解决这一问题的有效途径。论文综述了通过深度学习技术和生物信息学方法建立代谢物结构与谱图之间联系的三种策略, 包括通过结构预测谱图、通过谱图预测结构和通过已知推测未知, 并介绍了每种策略解决问题的思路 and 代表方法。对于每种策略, 论文讨论了其算法的优势和不足, 以及在实际应用中可能遇到的挑战。此外, 还探讨了在开发新算法和进行基准测试时应注意的因素, 以及这些因素如何影响对算法的评估。最后, 指出了融合更多正交信息是实现更加准确的代谢物注释的未来方向。

关键词: 次生代谢; 生物信息学; 质谱分析; 人工智能

DOI: 10.13560/j.cnki.biotech.bull.1985.2024-0523

Research Progress and Prospects in the Structural Annotation of Unknown Secondary Metabolites Based on Mass Spectrometry

Ji Hong-chao¹ Li Zheng-yan^{1, 2, 3}

(1. Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518000; 2. School of Life Sciences, Henan University, Kaifeng 475004; 3. Shenzhen Research Institute of Henan University, Shenzhen 518000)

Abstract: Research on secondary metabolites is of great significance for plant growth and development, environmental adaptation, as well as human health and drug development. Liquid chromatography-mass spectrometry (LC-MS) has become the preferred strategy for secondary metabolism research. However, the annotation of metabolite structures is still hindered by the insufficient coverage of standard spectral libraries. Given that the coverage of metabolite structure databases far exceeds that of standard spectral libraries, establishing the association between metabolite structures and mass spectra through artificial intelligence methods to search molecular structure databases based on mass spectrometry data is an effective approach to address this issue. This paper reviews three strategies for establishing the association between metabolite structures and mass spectra using deep learning techniques and bioinformatics methods, including structure-to-spectrum, spectrum-to-structure, and known-to-unknown strategies. It also introduces the rationale and representative methods for each strategy. For each strategy, the paper discusses the advantages and limitations of its algorithms, as well as the challenges that may be encountered in practical applications. Additionally, the paper explores that factors should be considered when developing new algorithms and conducting benchmark tests, and how these factors may affect the evaluation of algorithms. Finally, the paper points out that integrating more orthogonal information is a future direction for achieving more accurate metabolite annotation.

Key words: secondary metabolism; bioinformatics; mass spectrometry analysis; artificial intelligence

收稿日期: 2024-05-31

基金项目: 国家重点研发计划(2023YFA0915800)

作者简介: 纪宏超, 男, 博士, 研究员, 研究方向: 生物信息学; E-mail: jihongchao@caas.cn; 纪宏超同为本文通信作者

次生代谢是植物中一种重要的生命活动，其产生的次生代谢产物不仅与植物的生长发育、环境适应性密切相关，还是人类健康饮食和药物研发的重要来源。次生代谢产物的多样性和复杂性使得它们在农业生产和植物学研究中占据了举足轻重的地位。包括生物碱、黄酮类化合物、萜类、苯丙素类在内的次生代谢产物，不仅在植物中具有抗菌、抗虫、抗紫外线等防御功能，还在人类医药和食品工业中具有广泛应用。研究次生代谢物的合成调控、结构多样性与生物学功能，是代谢生物学领域的核心内容，对于深入理解植物生命活动规律、提升作物抗逆性、改良农产品质量等具有重要意义^[1-2]。

次生代谢产物的性质和功能与其独特的化学结构相关，例如，生物碱类次生代谢产物由于其碱性结构，常具有抗菌和神经活性；黄酮类化合物表现出强抗氧化和抗炎特性；萜类化合物则通常具有抗虫、抗病及芳香特性。因此，解析未知次生代谢物结构是植物代谢组学研究中的重要组成部分，有助于揭示这些化合物的生物活性和潜在应用价值。近年来，基于色谱-质谱联用平台的非靶标代谢组学研究，因其能够对生物体系中大量不同种类的代谢物进行全面、高效地分析，已成为当前次生代谢研究的首选策略。尤其是高效液相色谱-串联质谱联用技术（HPLC-MS/MS）在植物次生代谢物研究中已得到广泛应用，HPLC能够有效分离复杂的代谢物混合物，将不同化合物按其化学性质进行分离，而质谱则通过两级或多级电离和裂解，逐步将代谢物分子转变为碎片离子并检测其质荷比，从而提供关于分子结构的详细信息。尽管这项技术取得了显著进展，基于所获得的质谱谱图解析未知代谢物的结构仍然是一个亟待解决的难题。制约因素主要包括：（1）质谱数据解析复杂，难以与现有数据库中的标准谱图进行匹配；（2）缺乏完善的标准谱图库，影响了结构鉴定的准确性；（3）代谢物种类繁多、结构多样性高，增加了解析的难度。

本综述首先分析了次生代谢物结构解析中的瓶颈问题及其成因，解释了为何通过生物信息学算法来解决这些问题比等待标准谱图库扩展更为有效。其次，归纳和总结了当前的三种主要策略：即从结构到谱图、从谱图到结构以及从已知到未知，并回

顾了每种策略的代表方法，讨论了它们的特点和概念，指出了各自的优势、不足和适用范围，希望为新方法的提出提供思路 and 方向。然后，详细探讨了在算法开发过程中需要注意的问题和可能的挑战，强调了这些问题对算法开发人员的重要性。最后，从更广阔的视角展望了代谢物结构解析的未来发展方向，期望这些讨论能够为该领域的研究提供新的启示和推动。

1 代谢物结构解析的现状与瓶颈

代谢组学包括对结构各异的小分子代谢物的研究，例如脂类、核酸、氨基酸、肽类、碳水化合物、有机酸、酮类、醛类、胺类、类固醇、维生素、信号分子、激素和多酚类。高效液相色谱-串联质谱联用技术作为代谢组学研究的主要分析工具，能够分离并检测不同种类的代谢物信息。典型的代谢组学研究流程包括样本预处理、质谱检测、数据处理、模式识别等，最终得到生物学知识。代谢物结构解析是数据处理流程中的关键步骤，目的在于将质谱检测获得的代谢物信号对应到具体的代谢物结构。经预处理所获得的代谢物信号，包括色谱保留时间与峰面积、母离子质荷比与同位素分布以及二级质谱碎片等信息。目前最为常用的结构鉴定策略是将检测到的代谢物二级谱图与 GNPS、NIST 等标准质谱库进行匹配^[3]。这种方法利用了这些数据库中存储的大量已知代谢物的质谱数据，通过比较实验数据和数据库中的参考数据，计算待鉴定谱图与数据库中质谱的相似度，以鉴定出代谢物的化学结构，其相似度通常由余弦距离或欧氏距离评价。但这一策略受到两个方面的限制：第一，不同仪器类型、条件和参数的设置，导致相同代谢物可能具有不同的二级谱图^[4-5]，这会降低标准谱图库的通用性，从而减少已知代谢物在实际研究中的鉴定数量；第二，标准谱图库的覆盖范围远小于潜在的代谢物数目，标准谱图的缺失使得通过库匹配鉴定新颖代谢物几乎不可能^[6]。

尽管近些年 MASST^[7]、MS-Entropy^[8] 等更加稳健和快速的算法在一定程度上可以提高跨质谱平台的谱图匹配问题，但植物代谢组学研究的成熟度仍然相对较低，在非靶向代谢组学实验中，仅有不到

10% 的植物代谢物能够得到分析和研究^[9-10]。最新发表的植物代谢数据库 PMHub 收录了 18.9 万个代谢物^[11]。鉴于生物体内代谢网络的复杂性,尚有大量潜在代谢物尚未被发现。据估算,植物代谢物的实际数量可能高达 100 万种^[12]。相比之下,GNPS 仅收录了来自约 2.6 万个化合物的 49.6 万张谱图,NIST 20 则包含来自约 2.8 万个化合物的 93.2 万张谱图,而且并非所有都是代谢物。这意味着,拥有标准谱图的已知代谢物尚不到收录数量的 1/10。

虽然该瓶颈可以通过扩展标准谱图的数量得到部分解决,但两个限制条件使该途径难以实现。第一,标准谱图库的低通用性,使得各实验室为了得到更高质量的鉴定结果,倾向于分别自建谱图库,这需要极大的经济成本并造成资源浪费;第二,对于易于获得标准品的代谢物,其谱图往往已经得到收录,未收录的代谢物往往具有难合成和难分离的特点,这意味着收录新的标准谱图需要更高的成本和代价^[13]。

2 代谢物结构解析的策略与方法

鉴于当前已知代谢物结构种类远超过具有标准谱图的代谢物种类,一个自然的追问是:是否能够建立代谢物结构与谱图之间的联系,使得我们不再局限于搜索谱图库,而是能够通过检索结构库来直接推断未知代谢物的结构。具体实现流程:针对一个未知代谢信号的谱图,首先通过母离子质荷比、同位素分布等信息推断分子式,基于分子式从结构库中检索该未知代谢物的候选结构;然后根据已建立的代谢物结构-谱图关系,分析未知代谢信号的谱图与每个候选结构匹配的概率,并进行打分和排序;最后根据排名依次验证,得到未知代谢物的真实结构。

量子化学的知识告诉我们,化学键所处的化学位置决定了键能的强弱及断裂的趋势,这为建立化学结构与其谱图的关联提供了理论基础^[14]。但每个质谱碎片离子的形成都可能涉及多条化学键的断裂,这会导致组合爆炸的问题,再考虑到潜在代谢物的数量和复杂度,基于量子化学计算的时间成本是不可接受的。近年来,深度学习技术的进步极大推动了生物信息学领域的发展。日益提高的数据规模和

运算能力,使得通过先进架构和更高深度的神经网络,能够模拟人类专家所无法归纳的潜在关联。例如 AlphaFold2 已经可以成功从蛋白质序列准确预测蛋白质的三维结构^[15]。因此,更有希望用于生物大数据中未知代谢物结构的鉴定,是通过生物信息学方法和人工神经网络,让计算机从已有数据中去学习化学键断裂规律,从而建立代谢物结构-谱图关联模型^[16-17],常见的预测策略包括从结构到谱图,从谱图到结构和从已知到未知三种(图 1)。

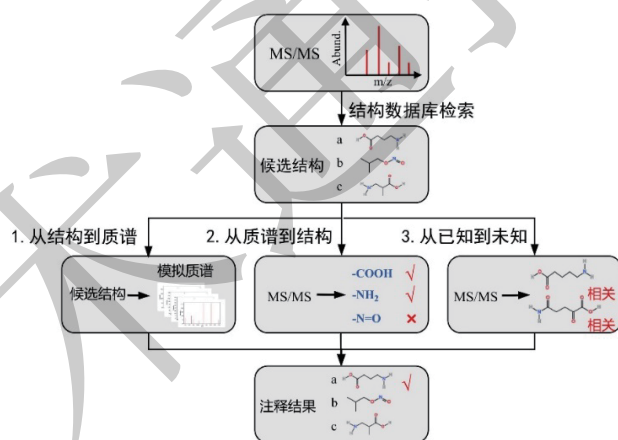


图 1 基于 MS/MS 谱图搜索分子结构数据库的不同策略

Fig. 1 Strategies for searching molecular structure databases based on MS/MS

2.1 从结构到谱图的预测策略

2.1.1 预测碎片离子及其丰度 建立代谢物结构-谱图关联模型,最直观的思路是从结构出发直接预测其质谱谱图。即通过人工归纳或利用机器学习手段,从现有标准谱图中总结裂解规律,预测代谢物中化学键断裂的概率,并由此获取代谢物最有可能产生的质谱碎片。CFM-ID 就是基于这样的思路,首先通过枚举的方式构建质谱碎片空间,然后以标准图谱的峰强度为依据,通过竞争性碎裂建模预测某个结构下通过断键得到相应质谱碎片的概率,进而得到代谢物的理论图谱^[18]。通过其网络服务器应用程序,用户可以使用该方法进行质谱预测、未知物鉴定和质谱碎片来源注释^[19]。在其迭代版本中,CFM-ID 将机器学习模型和人工归纳规律进行了结合,进一步提升了对特定种类代谢物的预测准确

性^[20-22]。质谱谱图与碰撞能等仪器参数有关,因此 CFM-ID 提供了碰撞能分别为 10、20、40 eV 下的谱图预测结果。

近年来,随着可用于训练的数据量增加,以及深度学习技术的进步,使得直接以结构为输入、以模拟质谱为输出的端到端预测方法成为可能。其流程是:首先,利用生物信息学领域已相对成熟的分子结构嵌入方法,用数值矩阵表示分子的结构信息;再通过深度神经网络提取与质谱谱图有关的信息;输出层则是质谱谱图的数值向量。这就将谱图预测问题转化为一个可以使用深度神经网络训练的多元回归问题。例如,3DmolMS 通过从分子的三维坐标建立分子网络模型,通过深度神经网络对分子进行向量化,并进一步预测质谱,其预测谱图在正离子模式和负离子模式下与实验谱图的平均余弦相似度分别为 0.691 和 0.478^[23]。MassFormer 则将分子视为拓扑结构,提取原子和化学键作为节点和边并进行嵌入学习,通过图注意力模型(graph transformer)进行谱图预测^[24]。同样地,这些模型也把质谱平台参数的各类参数进行独热编码,作为额外的信息输入到模型中,来解决参数设置对谱图的影响。

2.1.2 预测碎片离子的超集 质谱仪器参数的多样性使得上述模型无法穷尽各种参数条件,并进行相应的代谢物谱图预测。因此很多算法采取相对灵活的方式,即对于给出候选结构列表,通过已知的裂解规则,预测每种候选结构可能解离的所有碎片,共同组成了谱图中碎片离子的超集;然后针对给出的未知谱图,逆向匹配中的离子是否出现在每个候选结构的超集中;最后根据候选结构能够解释的碎片离子比例,对候选结构进行评分和排序。

例如, MetFrag 使用键离解方法产生每个离子碎片,并构建碎片树。碎片树以根节点为母离子,每个节点代表一个通过特定键位分裂分子所获得的碎片,运用启发式方法,以加速计算过程,直到达到给定的树深度,最终所有节点的并集构成了谱图碎片离子的超集^[25-26]。MAGMa 同样采用了构建碎片树的方法,区别在于其生成子结构时,需要考虑如何匹配碎片离子与观察到的质谱碎片峰。然后并对于每个候选分子的碎片层级树,计算匹配分数,并评估候选结构如何解释观察到的层级碎裂模式^[27]。

MS-Finder 在构建碎片树时则主要考虑产生相应碎片离子是否符合根据质谱实验所总结的七种氢重排规则,并根据每个候选结构所能产生的,符合规则的碎片离子所占未知谱图中碎片峰的比例进行评分和排序^[28]。

2.2 从谱图到结构的预测策略

2.2.1 预测可能的子结构 事实上,只要建立结构-谱图的关联,就可以依据一定的评分标准对候选结构进行排序来实现。因此,预测谱图并非唯一可行的方案,部分算法采用了即从谱图逆向推导至结构的预测方法,即从未知谱图出发,先利用机器学习模型预测出与之对应的代谢物结构应具备的关键特征,随后,按照这些特征,评估候选结构列表中的每一项,并基于匹配程度和预测特征进行打分与排序。FingerID 率先提出了一种由支持向量机从谱图中预测分子指纹的方法,对候选结构进行评分和排序的方案^[29]。分子指纹是一种表征化合物是否具有某个特定子结构的手段,它是一组由 0 和 1 组成的向量,每个字节代表一种子结构,1 和 0 则分别代表相应子结构的有无。此后,该方法与碎片树^[30]、多核支持向量机策略^[31]结合,用于进一步提取谱图特征,以增强指纹预测的准确度,从而得到了准确性更高的代谢物定性工具 CSI:FingerID^[32]。该工具在后续版本中又进一步通过使用贝叶斯网络算法^[33]、输入输出核回归(IOKR)算法^[34]等不断优化其性能,并整合为综合性代谢物定性工具 SIRIUS^[35]。

深度神经网络也被应用于这类指纹预测的方法中,具体而言,通过将谱图转化为向量表示,然后运用多层感知器(MLP)、卷积神经网络(CNN)等架构,进行多分类任务的预测,直接输出分子指纹向量。这种方法摆脱了对传统碎片树与核方法的依赖,实现了从谱图到分子指纹的端到端预测,从而显著提升了训练和预测的效率^[36-37]。尽管这种方法速度上有所突破,但在性能上,它仍不及结合了碎片树和核方法的 MLP 模型^[38]。因此 MIST 方法采用了更为先进的架构,它同样不需构建碎片树,但预测质谱碎片所对应的分子式,进而使用 Transformer 架构进行嵌入学习,获得与结构更为相关的谱图向

量,并在此基础上训练分子指纹分类器,并取得了更为理想的效果^[39]。

2.2.2 预测整个分子结构 随着生成式深度学习模型越来越多地被用于分子结构生成^[40-41],通过谱图数据直接预测分子结构的方法也逐渐受到更多的关注。与基于变分自编码器(VAE)的分子结构生成算法类似,它们将谱图数据通过编码器映射到隐层空间,形成隐空间的概率分布,再通过解码器输出分子结构。该方法除了能够用于候选分子的排序外,更能直接产生数据库收录以外的分子结构,为发现全新代谢物提供了可能性。然而,其缺点在于鉴定准确率相对较低。

例如,Mass2SMILES方法将编码后的谱图数据输入到由5层Transformer组成的编码器,然后作为时间卷积神经网络的一部分进行处理。最后两个输出由全连接层解码,返回SMILES字符串和编码的60个官能团的化学描述符空间编码,从而完成对化学结构和性质的预测^[42]。MS2Mol方法将质谱碎片按强度排序,分别标记为整数和小数部分,然后将其转换为片段嵌入,输入到均由Transformer组成的编码器和解码器中,通过对隐层的束搜索生成该谱图所可能对应的SMILES字符串以表示分子结构^[43]。MSNovelist方法则首先将谱图输入到SIRIUS软件预测分子指纹,再以指纹为输入进行编码和解码操作,解码器由RNN组成,并在逐字符生成SMILES的过程中,将使用的原子依次在分子式中扣除,以生成准确符合预期分子式的结构,从而实现结构预测^[44]。

2.3 从已知到未知的预测策略

2.3.1 通过代谢网络迭代预测 以上两种策略在完成模型训练后,使用时都不再需要进行标准谱图检索。这显然有利于扩展可鉴定代谢物的覆盖范围,但同样降低了对标准谱图库中已经包含的代谢物的鉴定能力。而在实际研究中,研究者事先并不知道待注释代谢物是否已被包含在标准谱图库中。通过已知推测未知的策略则避免了该问题,其特点在于其受到生物知识的驱动,首先通过检索标准谱图库鉴定代谢组中的一部分已知的种子代谢物作为起点,然后利用谱图相似性与结构相似性的关系,对这些

未知结构是否与已知代谢物具有结构和代谢途径上的关联,进而推断未知代谢物结构。

MPEA方法假定整个代谢组是通过有限数量的初始代谢产物经过各种类型和多步代谢反应构建的,因此理论上,整个代谢组可能是从初始代谢产物和代谢反应中映射出来的。该方法以肉碱被用作初始代谢产物为示例,通过常见代谢反应进行多轮递归,结合谱图信息共鉴定了97种相关代谢物^[45]。MetDNA和MetDNA2通过构建代谢网络和谱图相关网络进行多轮递归,实现从种子代谢物到整个代谢网络的结构注释^[10,46]。类似地,SGMNS通过代谢组数据库中化学结构的分子指纹相似性构建全局连接性分子网络,代替已知代谢网络进行递归,也实现了对未知代谢物的注释^[47]。

通过代谢网络的迭代预测策略往往具有特定的应用场景,其原因在于所有能够进行注释的化合物,其起点必须都是已经通过谱图检索所注释的代谢物。如果无法获得足够数量的种子代谢物,或者待鉴定代谢物与种子代谢物缺失多个中间代谢物,那么递归注释将受到影响。因此它们的输入数据必须是完整的非靶标组学数据,而不能是单张或数张独立的图谱,同时其迭代预测的准确性同样受谱图匹配精确度影响很大,种子代谢物的鉴定质量将对整个实验结论产生较大影响。

2.3.2 通过代谢物关联性预测 另一种方式是对未知代谢物与标准谱图库中已知代谢物的关联性进行预测,并以此为依据解析未知代谢物结构。这种方式不再依赖基于种子代谢物的迭代,仍然对每一张未知谱图独立处理,因此可以处理单张或者多张谱图。例如iMet方法,首先将未知谱图与标准谱图库进行相似度比较,选择可能与未知代谢物具有相似结构的已知代谢物。然后通过未知代谢物的分子质量,结合常见的代谢反应导致的质量差,推断该未知代谢物可能是已知代谢物通过哪些潜在转化来得到,进而解析未知代谢物的结构^[48]。

DeepMASS提供了另一种基于已知化合物的未知物化学空间定位的思路,该方法假设代谢物分子在化学空间内按照其骨架和官能团的差异呈规律分布,经过两个步骤解析待鉴定代谢物的结构。首先,通过待鉴定代谢物的质谱,判断其与已知代谢物的

关联性,确定其在化学空间上所处的相对位置;然后,评估候选列表中的分子所处化学空间的未知与预测位置的距离,以此为依据对候选分子进行排序^[49]。

代谢物关联性预测的理论基础在于,代谢物谱图的相似性意味着其所对应的代谢物也具有相似性。如何评价谱图的相似性,并确认这种相似性与结构的相似性高度相关,是此类算法成功的关键。传统

的相似性评价方式往往是基于谱图的余弦相似,而最近一些研究认为,基于嵌入学习^[50]、孪生网络^[51]和对比学习^[52]的方法,可能对这种相似性的评估更为有利,这些研究有望进一步推进代谢物关联性预测策略的发展。

表 1 总结了基于谱图搜索分子结构数据库的代表方法。

表 1 基于谱图搜索分子结构数据库的代表方法

Table 1 Representative methods for searching molecular structure databases based on MS/MS						
策略类型 Strategy type	方法名称 Method name	输入数据 Input data	预测依据 Prediction basis	是否有开源代码 Is there open source code available	是否有可视化软件 Is there any visualiza- tion software available	官方网站 Official website
从结构到 谱图	CFM-ID	单张 / 多张谱图	模拟谱图	是	是	https://cfmid.wishartlab.com/
	3DmolMS	单张 / 多张谱图	模拟谱图	是	否	https://github.com/JosieHong/3DMolMS
	MassFormer	单张 / 多张谱图	模拟谱图	是	否	https://github.com/Roestlab/massformer
	MetFrag	单张 / 多张谱图	碎片离子超集	是	是	https://ipb-halle.github.io/MetFrag/
	MS-Finder	单张 / 多张谱图	碎片离子超集	是	是	http://prime.psc.riken.jp/compms/msfinder/main.html
从谱图到 结构	MAGMa	单张 / 多张谱图	碎片离子超集	是	是	https://github.com/NLeSC/MAGMa
	SIRIUS	单张 / 多张谱图	分子指纹	是	是	https://bio.informatik.uni-jena.de/software/sirius
	Mass2SMILES	单张 / 多张谱图	分子结构	是	否	https://github.com/volvox292/mass2smiles
	MS2Mol	单张 / 多张谱图	分子结构	否	否	暂无
	MSNovelist	单张 / 多张谱图	分子结构	是	是	https://github.com/meowcat/MSNovelist
从已知到 未知	MPEA	非靶标组学数据	代谢反应	否	否	暂无
	MetDNA	非靶标组学数据	代谢反应	是	是	http://metdna.zhulab.cn/
	iMet	单张 / 多张谱图	代谢反应	是	是	http://imet.seeslab.net/
	SGMNS	非靶标组学数据	代谢反应	否	否	暂无
	DeepMASS	单张 / 多张谱图	化学空间定位	是	是	https://github.com/hcji/DeepMASS2_GUI

3 代谢物结构解析算法开发的挑战

自 2012 年以来,小分子鉴定大赛 (CASMI)^[53]已成功举办了 6 届。该比赛为评估分子结构数据库搜索的最新技术水平提供了一种独立客观的方式,也成为各类代谢物结构解析算法开发时常用的评估测试集。一些研究指出,在进行算法开发和评价过程中,存在一些不合理和不恰当的方式,可能会导致准确率评估上的虚假繁荣,这些情况值得研究人员关注和警惕。

3.1 警惕Metadata的使用

这里的 Metadata 包括候选化合物的引文频率、

专利、生产量或生物数据库的收录情况。它们虽然与谱图解析无关,但在候选化合物评分时,部分软件把这些信息也作为评分准则的一部分,并带来鉴定准确率上的提升。事实上,通过这些信息,即使完全忽略质谱碎片信息,也可以得到出色的评估结果, MetFrag2.2 在搜 PubChem 时,仅基于专利和 PubMed 引用次数,再加上来自 MS1 的化合物质量,就达到了 65.8% 的正确鉴定率^[13]。这些信息的引入脱离了未知代谢物注释工具的初衷,它使得解析结果更偏向于已知代谢物,不利于新颖代谢物的发现。

研究人员在创新算法时,往往需要与现有已整

合了 Metadata 的算法进行比较, 基于基准数据集的评估会让不使用 Metadata 的算法天然处于劣势, 较差的比较结果会阻碍这些算法的发表。解决这一矛盾需要两方面的努力, 一是需要为不借助 Metadata 的算法提供新的赛道, 对他们进行更为客观公正的评价和发表途径, 而不强求其一定超过现有方法; 二是需要开发新颖的基准数据集, 例如新发现或新合成化学结构, 利用这些数据进行评估时, 即使使用 Metadata 也不会带来更多的优势。

3.2 避免潜在的数据泄露

对数据的训练集和测试集划分不合理时, 机器学习模型可能会倾向于记忆而非真正学习。进行代谢物结构解析的性能评估时, 合理划分数据集尤为困难。即便训练集和测试集来自截然不同的数据库, 数据泄露的风险依然存在。一方面, 由于检测仪器和参数的不同, 同一代谢物的谱图可能在标准谱图库中有多种表现, 它们可能极为相似, 也可能完全不同, 这些谱图可能分布在不同的数据库。另一方面, 不同的标准谱图库之间也可能存在相互引用和数据重叠。以 CASMI 2016 为例, 该数据被广泛用于方法评估, 但它同时也被收录在 GNPS 和 MoNA 数据库中。因此, 如果某个方法是以 GNPS 和 MoNA 数据作为训练集, 并在 CASMI 2016 数据上进行评估, 那么这实际上已经构成了数据泄露, 严重影响评估的准确性和可靠性^[54]。

因此, 研究和审阅人员进行算法开发和评估时, 要高度警惕这种情况, 更加科学和有效地划分训练集和测试集, 例如通过对代谢物结构及其骨架特征进行划分, 确保训练集和测试集中不包含过多相似或相同谱图。并且准备发表的研究应当鼓励公开和透明的数据集划分和评估流程并开源代码。

3.3 应对公平比较的挑战

已发布且广受欢迎的未知代谢物结构解析工具, 通常以高度整合的软件包或网络服务应用的方式呈现, 这对于用户来说无疑是友好和便捷的。然而, 对于竞争性方法的开发者而言, 这种高度的集成性却构成了与之进行公平比较的困难。一方面, 这些软件用于检索候选物质的结构数据库各不相同, 对 Metadata 的利用方式也各有不同, 这会对代谢物结

构解析性能评价产生显著影响, 而这些因素与谱图解析算法本身的优劣并无直接关联。另一方面, 这些软件在模型训练时, 为了追求最佳性能, 往往涵盖了所有可用的公开数据集。因此, 竞争性算法开发者若从公开数据集中划分测试集进行性能评估, 将天然处于不利地位。反之, 从公开数据以外获得与之完全无关的测试集用于方法评估又需要高昂的成本。这些因素导致了软件之间难以进行真正意义上的公平比较。

我们希望现有的工具能够考虑提供开发者版本, 或者提供更加灵活的参数设置方法, 以便在评估和比较时, 对不同工具可以使用相同的数据源进行模型重训练, 使用相同的结构数据库进行候选化合物检索, 并且将基于 Metadata 的打分排除在排序依据以外。这些途径不仅可以用于算法评估, 也有利于研究者探索实质上更为有效的结构解析策略, 促进算法的改进和创新, 推动代谢组学研究的发展。

4 代谢物结构解析算法开发的展望

在代谢物结构注释的研究中, 基于质谱碎片建模的潜力正在被不断挖掘, 如何在此之外挖掘更多潜力, 将是未来开发新算法的重要方向。这种策略不仅有助于克服传统方法的局限性, 还能提供更加全面和准确的代谢物结构信息。我们认为以下因素未来需要被纳入考量。

4.1 结合多维测量学信息

尽管使用与代谢物结构分析无关的先验信息存在风险, 但若能更好地获取和利用具有正交性的测量学信息, 则有望为代谢物的结构注释提供更加有力的支持。其中, 典型代表包括碰撞横截面 (CCS) 和保留时间 (RT) 等。CCS 描述了代谢物在气相或液相中的物理性质, 它反映了代谢物分子与周围环境相互作用的强弱。通过测量 CCS, 我们可以获取到关于代谢物分子大小、形状以及极性等方面的信息。RT 代表了代谢物在色谱柱上的保留行为, 受到代谢物极性、分子量以及色谱柱类型等多种因素的影响。它们都可以为结构注释提供有力支持。目前已经有相当数量的 RT 和 CCS 测量数据用于机器学习算法进行建模^[55-57], 也有相当多的结果已经被报道^[51-54]。

4.2 挖掘多组学实验线索

代谢物作为中心法则的一环，在生命活动中与基因组、转录组和蛋白组之间紧密相连，相互交织成一张复杂的生命网络。因此，结合多组学信息进行综合分析，对于代谢物的结构注释具有显著的优势和潜力。例如，在生命过程中，具有相同或相近功能的酶往往会催化具有相近结构的代谢物进行生物化学反应，进而生成各自的下游代谢产物。这一特性为我们提供了一个独特的视角，即通过探索与未知代谢物相关的酶或对应基因，结合已知的代谢反应信息，有望推测出未知代谢物的结构特征。因此在多组学联合分析中，相对更成熟的蛋白组学、基因组学可以更加准确地发现差异蛋白和差异基因，而代谢物的差异与这些差异蛋白和差异基因具有一定的关联性。对于其中未知结构的差异代谢物的解析，或许通过分析那些能够鉴定结构的已知代谢物与之的关联性，并结合差异蛋白和差异基因的生物化学功能，有望能够推断它们是通过何种生物反应所得到，并解析其具体结构。

基于结构生物学的小分子-蛋白相互作用研究可能进一步为该思路提供解决方案。通过深入研究小分子与蛋白质之间的相互作用，我们可以更加精准地理解代谢物在生物体内的功能和作用机制。分子对接、动力学模拟以及深度学习模型等先进策略的应用，使得我们能够实现对小分子-蛋白结合能力的准确预测，分析差异蛋白是否可能与预期的代谢底物产生结合，并催化其转化为未知代谢物。也可以通过比较代谢物分子不同空间结构与对应催化酶的结合能力，对代谢物的空间结构做出预测。

4.3 评价注释结果可信度

在未来，我们需要解决许多问题：例如，如果注释软件的正确率可以达到50%–80%，这些信息要如何被生物学家使用？我们能否区分那些被正确识别的化合物和我们不太确定的化合物？通过注释软件而不是标准图库所获得的代谢物定性结果是否可以被期刊审稿人所认可，下游研究人员是否可以采信通过这些方式所获得的生物学知识？在未来，我们需要不断探索和创新，以进一步完善化合物识别方法，提高识别准确性和可靠性的同时，也需要给

出这些问题的答案。

参考文献

- [1] Shen SQ, Zhan CS, Yang CK, et al. Metabolomics-centered mining of plant metabolic diversity and function: past decade and future perspectives [J]. *Mol Plant*, 2023, 16 (1): 43-63.
- [2] Liu XY, Zhou LN, Shi XZ, et al. New advances in analytical methods for mass spectrometry-based large-scale metabolomics study [J]. *Trac Trends Anal Chem*, 2019, 121: 115665.
- [3] Medema MH. The year 2020 in natural product bioinformatics: an overview of the latest tools and databases [J]. *Nat Prod Rep*, 2021, 38 (2): 301-306.
- [4] Bazsó FL, Ozohanics O, Schlosser G, et al. Quantitative comparison of tandem mass spectra obtained on various instruments [J]. *J Am Soc Mass Spectrom*, 2016, 27 (8): 1357-1365.
- [5] Hoang C, Uritboonthai W, Hoang L, et al. Tandem mass spectrometry across platforms [J]. *Anal Chem*, 2024, 96 (14): 5478-5488.
- [6] Wei JN, Belanger D, Adams RP, et al. Rapid prediction of electron-ionization mass spectrometry using neural networks [J]. *ACS Cent Sci*, 2019, 5 (4): 700-708.
- [7] Wang MX, Jarmusch AK, Vargas F, et al. Mass spectrometry searches using MASST [J]. *Nat Biotechnol*, 2020, 38 (1): 23-26.
- [8] Li YY, Kind T, Folz J, et al. Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification [J]. *Nat Methods*, 2021, 18 (12): 1524-1531.
- [9] da Silva RR, Dorrestein PC, Quinn RA. Illuminating the dark matter in metabolomics [J]. *Proc Natl Acad Sci U S A*, 2015, 112 (41): 12549-12550.
- [10] Shen XT, Wang RH, Xiong X, et al. Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics [J]. *Nat Commun*, 2019, 10 (1): 1516.
- [11] Tian ZT, Hu X, Xu YY, et al. PMhub 1.0: a comprehensive plant metabolome database [J]. *Nucleic Acids Res*, 2024, 52 (D1): D1579-D1587.
- [12] Wang SC, Alseekh S, Fernie AR, et al. The structure and function of major plant metabolite modifications [J]. *Mol Plant*, 2019, 12 (7): 899-919.
- [13] Böcker S. Searching molecular structure databases using tandem MS data: are we there yet? [J]. *Curr Opin Chem Biol*, 2017, 36: 1-6.
- [14] Ásgeirsson V, Bauer CA, Grimme S. Quantum chemical calculation of electron ionization mass spectra for general organic and inorganic molecules [J]. *Chem Sci*, 2017, 8 (7): 4879-4895.

- [15] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold [J]. *Nature*, 2021, 596 (7873) : 583-589.
- [16] Cai YP, Zhou ZW, Zhu ZJ. Advanced analytical and informatic strategies for metabolite annotation in untargeted metabolomics [J]. *Trac Trends Anal Chem*, 2023, 158: 116903.
- [17] Hu GL, Qiu MH. Machine learning-assisted structure annotation of natural products based on MS and NMR data [J]. *Nat Prod Rep*, 2023, 40 (11) : 1735-1753.
- [18] 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.
- [19] Allen F, Pon A, Wilson M, et al. CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra [J]. *Nucleic Acids Res*, 2014, 42 (W1) : W94-W99.
- [20] Djoumbou-Feunang Y, Pon A, Karu N, et al. CFM-ID 3.0: significantly improved ESI-MS/MS prediction and compound identification [J]. *Metabolites*, 2019, 9 (4) : 72.
- [21] Wang F, Liigand J, Tian SY, et al. CFM-ID 4.0: more accurate ESI-MS/MS spectral prediction and compound identification [J]. *Anal Chem*, 2021, 93 (34) : 11692-11700.
- [22] Wang F, Allen D, Tian SY, et al. CFM-ID 4.0-a web server for accurate MS-based metabolite identification [J]. *Nucleic Acids Res*, 2022, 50 (W1) : W165-W174.
- [23] Hong YH, Li SJ, Welch CJ, et al. 3DMolMS: prediction of tandem mass spectra from 3D molecular conformations [J]. *Bioinformatics*, 2023, 39 (6) : btad354.
- [24] Young A, Röst H, Wang B. Tandem mass spectrum prediction for small molecules using graph transformers [J]. *Nat Mach Intell*, 2024, 6: 404-416.
- [25] Ruttkies C, Schymanski EL, Wolf S, et al. MetFrag relaunched: incorporating strategies beyond in silico fragmentation [J]. *J Cheminform*, 2016, 8: 3.
- [26] Ruttkies C, Neumann S, Posch S. Improving MetFrag with statistical learning of fragment annotations [J]. *BMC Bioinformatics*, 2019, 20 (1) : 376.
- [27] Verdegem D, Lambrechts D, Carmeliet P, et al. Improved metabolite identification with *MIDAS* and *MAGMa* through MS/MS spectral dataset-driven parameter optimization [J]. *Metabolomics*, 2016, 12 (6) : 98.
- [28] Tsugawa H, Kind T, Nakabayashi R, et al. Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER software [J]. *Anal Chem*, 2016, 88 (16) : 7946-7958.
- [29] Heinonen M, Shen HB, Zamboni N, et al. Metabolite identification and molecular fingerprint prediction through machine learning [J]. *Bioinformatics*, 2012, 28 (18) : 2333-2341.
- [30] Rasche F, Svatos A, Maddula RK, et al. Computing fragmentation trees from tandem mass spectrometry data [J]. *Anal Chem*, 2011, 83 (4) : 1243-1251.
- [31] Shen HB, Dührkop K, Böcker S, et al. Metabolite identification through multiple kernel learning on fragmentation trees [J]. *Bioinformatics*, 2014, 30 (12) : i157-i164.
- [32] Dührkop K, Shen HB, Meusel M, et al. Searching molecular structure databases with tandem mass spectra using CSI: FingerID [J]. *Proc Natl Acad Sci U S A*, 2015, 112 (41) : 12580-12585.
- [33] Ludwig M, Dührkop K, Böcker S. Bayesian networks for mass spectrometric metabolite identification via molecular fingerprints [J]. *Bioinformatics*, 2018, 34 (13) : i333-i340.
- [34] Brouard C, Shen HB, Dührkop K, et al. Fast metabolite identification with Input Output Kernel Regression [J]. *Bioinformatics*, 2016, 32 (12) : i28-i36.
- [35] Dührkop K, Fleischauer M, Ludwig M, et al. *SIRIUS* 4: a rapid tool for turning tandem mass spectra into metabolite structure information [J]. *Nat Methods*, 2019, 16 (4) : 299-302.
- [36] Fan ZL, Alley A, Ghaffari K, et al. MetFID: artificial neural network-based compound fingerprint prediction for metabolite annotation [J]. *Metabolomics*, 2020, 16 (10) : 104.
- [37] Gao S, Chau HYK, Wang KJ, et al. Convolutional neural network-based compound fingerprint prediction for metabolite annotation [J]. *Metabolites*, 2022, 12 (7) : 605.
- [38] Dührkop K. Deep kernel learning improves molecular fingerprint prediction from tandem mass spectra [J]. *Bioinformatics*, 2022, 38 (Suppl 1) : i342-i349.
- [39] Goldman S, Wohlwend J, Stražar M, et al. Annotating metabolite mass spectra with domain-inspired chemical formula transformers [J]. *Nat Mach Intell*, 2023, 5: 965-979.
- [40] Mokaya M, Imrie F, van Hoorn WP, et al. Testing the limits of SMILES-based *de novo* molecular generation with curriculum and deep reinforcement learning [J]. *Nat Mach Intell*, 2023, 5: 386-394.
- [41] Qian H, Lin C, Zhao DW, et al. AlphaDrug: protein target specific *de novo* molecular generation [J]. *PNAS Nexus*, 2022, 1 (4) : pgac227.
- [42] Elser D, Huber F, Gaquerel E. Mass2SMILES: deep learning based fast prediction of structures and functional groups directly

- from high-resolution MS/MS spectra [J]. bioRxiv, 2023. DOI: 10.1101/2023.07.06.547963.
- [43] Butler T, Frandsen A, Lighthead R, et al. MS2Mol: a transformer model for illuminating dark chemical space from mass spectra [J]. ChemRxiv, 2023. <https://chemrxiv.org/engage/chemrxiv/article-details/64f76a0279853bbd7829bf27>.
- [44] Stravs MA, Dührkop K, Böcker S, et al. MSNovelist: *de novo* structure generation from mass spectra [J]. Nat Methods, 2022, 19 (7): 865-870.
- [45] Wang L, Ye H, Sun D, et al. Metabolic pathway extension approach for metabolomic biomarker identification [J]. Anal Chem, 2017, 89 (2): 1229-1237.
- [46] Zhou ZW, Luo MD, Zhang HS, et al. Metabolite annotation from knowns to unknowns through knowledge-guided multi-layer metabolic networking [J]. Nat Commun, 2022, 13 (1): 6656.
- [47] Wang XX, Li C, Li ZF, et al. A structure-guided molecular network strategy for global untargeted metabolomics data annotation [J]. Anal Chem, 2023, 95 (31): 11603-11612.
- [48] Aguilar-Mogas A, Sales-Pardo M, Navarro M, et al. iMet: a network-based computational tool to assist in the annotation of metabolites from tandem mass spectra [J]. Anal Chem, 2017, 89 (6): 3474-3482.
- [49] Ji HC, Xu YM, Lu HM, et al. Deep MS/MS-aided structural-similarity scoring for unknown metabolite identification [J]. Anal Chem, 2019, 91 (9): 5629-5637.
- [50] Huber F, Ridder L, Verhoeven S, et al. Spec2Vec: improved mass spectral similarity scoring through learning of structural relationships [J]. PLoS Comput Biol, 2021, 17 (2): e1008724.
- [51] Huber F, van der Burg S, van der Hooft JJJ, et al. MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra [J]. J Cheminform, 2021, 13 (1): 84.
- [52] Guo H, Xue KB, Sun HM, et al. Contrastive learning-based embedder for the representation of tandem mass spectra [J]. Anal Chem, 2023, 95 (20): 7888-7896.
- [53] Blaženović I, Kind T, Torbašinović H, et al. Comprehensive comparison of in silico MS/MS fragmentation tools of the CASMI contest: database boosting is needed to achieve 93% accuracy [J]. J Cheminform, 2017, 9 (1): 32.
- [54] Hoffmann MA, Kretschmer F, Ludwig M, et al. MAD HATTER correctly annotates 98% of small molecule tandem mass spectra searching in PubChem [J]. Metabolites, 2023, 13 (3): 314.
- [55] Domingo-Almenara X, Guijas C, Billings E, et al. The METLIN small molecule dataset for machine learning-based retention time prediction [J]. Nat Commun, 2019, 10 (1): 5811.
- [56] Kretschmer F, Harrieder EM, Hoffmann MA, et al. RepoRT: a comprehensive repository for small molecule retention times [J]. Nat Methods, 2024, 21 (2): 153-155.
- [57] Zhang HS, Luo MD, Wang HM, et al. AllCCS2: curation of ion mobility collision cross-section atlas for small molecules using comprehensive molecular representations [J]. Anal Chem, 2023, 95 (37): 13913-13921.

(责任编辑 李楠)