arXiv:2601.00941v1 [q-bio.QM] 2 Jan 2026

串联质谱分子式与结构预测的比较分析

Xujun Che、Xiuxia Du* 和 Depeng Xu*

北卡罗来纳大学夏洛特分校
{xche,xdu4,dxu7}@charlotte.edu

摘要。 基于液相色谱-质谱（LC-MS）的代谢组学和暴露组学旨在测量生物样本中可检测到的小分子。研究结果有助于提出假设，发现代谢变化和疾病机制，并提供有关环境暴露及其对人类健康影响的信息。代谢组学和暴露组学的发展得益于液相色谱的高分辨能力和质谱的高质量测量准确度。然而，由于现有谱库远未覆盖液相色谱-串联质谱（LC-MS/MS）所捕获的广阔化学空间，这类研究中的大多数信号仍无法通过传统谱库检索予以鉴定或注释。为应对这一挑战并充分发挥代谢组学和暴露组学的潜力，研究人员开发了多种计算方法，根据串联质谱预测化合物。已发表的这些方法评估采用了不同的数据集和评价方式。为筛选适用于实际应用的预测工作流程，并确定有待进一步改进的领域，我们对最先进的预测算法进行了系统评估。具体而言，我们评估了不同类型加合物的分子式预测和结构预测准确率。所得结果确立了切合实际的性能基准，指出了关键瓶颈，并为进一步改进基于质谱的化合物预测提供了指导。

关键词： 非靶向代谢组学 · 质谱法 · 化合物预测 · 深度学习。

1 引言

代谢组学和暴露组学旨在大规模研究细胞、组织或生物体中的小分子，对理解复杂的生物系统和发现新的生物标志物至关重要。液相色谱与质谱法联用是代谢组学和暴露组学研究的主要分析平台之一，因为它能够广泛覆盖化学空间。基于质谱法的代谢组学和暴露组学有一项基础任务：将化合物或化合物类别归属到串联质谱，这对于将原始分析数据转化为可用于指导行动的生物学见解至关重要。然而，这仍然是一项挑战，所得串联质谱中大多数仍无法识别，

由此形成了庞大而沉重的暗物质；要充分发挥基于质谱法的代谢组学和暴露组学的真正潜力，就必须阐明这些暗物质。

液相色谱根据化合物的化学性质分离复杂混合物中的化合物。这种分离降低了化学复杂性，防止不同分析物相互干扰，并能够更准确地检测化合物。随后，串联质谱（MS/MS）通过使相应的前体离子碎裂，产生具有特征性的碎裂模式，从而提供化合物的结构信息，而这些碎裂模式是分子鉴定的基础。传统上，未知的 MS/MS谱会与已知 MS/MS谱库进行匹配。谱库中与未知 MS/MS谱最相似的化合物会被指认为该未知 MS/MS谱对应的化合物。这个过程称为化合物鉴定（有强有力证据支持的可靠指认）或注释（有一定证据支持的指认）。基于谱库搜索的方法已被证明非常有效。然而，与液相色谱-质谱平台能够检测的庞大化合物空间相比，现有 MS/MS谱库仍然十分有限。因此，在典型的代谢组学或暴露组学研究中，大多数 MS/MS谱在谱库搜索后仍无法鉴定。

为应对这一挑战，人们开发了基于串联质谱预测化合物的计算工作流程。化合物预测克服了串联质谱谱库的局限性，能够为可能产生未知串联质谱的化合物提供极有价值的线索，尽管其置信度无法达到谱库匹配的水平。化合物预测通常包含两个连续步骤：分子式预测 和结构解析。

在分子式预测中，首先生成并排序一系列可能的候选分子式。其中一种算法是 Sirius [5]。它结合了同位素模式建模与基于 MS/MS质谱的概率碎裂树构建，对所有候选项进行评分，从而选出与一级质谱和串联质谱证据最一致的分子式。该算法生成的结果具有可解释性，但在速度和复杂性方面表现不佳。为克服这些局限，MIST-CF [8]和 FIDDLE [9] 等基于深度学习的方法应运而生。这些数据驱动模型通过捕捉复杂的质谱模式提升了性能，并取得了较为理想的准确率，但通常需要大型训练数据集，而且其预测结果的透明度可能较低。

对于每个分子式候选项，结构预测可以通过两种策略进行：数据库搜索或从头生成。CSI:FingerID 使用前一种方法：基于 SIRIUS 生成的碎裂树预测分子结构指纹，然后计算预测指纹与结构数据库中已知化合物指纹之间的相似度 [6]。可以理解的是，这种方法仅适用于已知分子，因此会遗漏新结构；考虑到典型代谢组学实验中的大量质谱被认为源自现有结构谱库中不存在的未知化合物，这是一项关键局限 [13, 11, 1]。从头生成结构的方法包括 MADGEN [16], 、MSNovelist [14], 和 DiffMS [2], 等。这些从头预测方法难以应对

这一任务的高度复杂性，尤其难以区分结构异构体。即使采用高分辨率质谱法，多个分子式候选项仍可能无法区分；在尝试鉴定同量异位或异构化合物时，这一问题会更加棘手 [2]。

除了这些固有的技术挑战之外，当前用于评估计算模型的方法往往也存在局限，使其实际效用受到质疑。首先，许多基准数据集在规模和覆盖范围方面存在问题。例如，广泛使用的全球天然产物社交分子网络（GNPS）数据库 [15] 仅包含已知天然产物中估计约 2.5% 的质谱 [10, 4], 这表明，在此类数据集上训练和评估的模型可能无法代表更广阔的化学空间。其次，一种常见的评估范式是采用严格的结构差异划分 [3], ，即有意选择与训练集中的分子在结构上明显不同的分子作为测试集。尽管这种方法能严格检验模型的泛化能力，但它过分强调这一方面，却牺牲了实际适用性。在许多真实的代谢组学研究中，研究人员关注的是识别特定已知家族中的代谢物，而这些家族中的分子通常具有相似的结构。当前对广泛泛化能力的关注，并未反映这些常见的实际应用场景。

为弥合学术基准评测与实际应用之间的差距，我们在大规模、全面的数据集上评估了最先进的计算方法的性能。具体而言，我们使用美国国家标准与技术研究院（NIST）23 串联质谱谱库 [12]和北美质谱数据库（MoNA） [7],中的串联质谱开展了系统评估，这些数据分别代表高质量的人工整理的质谱图和多样化的社区贡献数据。我们实施了一个两阶段流程：首先预测分子式，然后为每个候选分子式预测分子结构。为了反映实际的代谢组学场景，我们采用了随机数据划分，而不是结构差异性划分。此外，我们还进行了加合物类型分层分析，以揭示不同加合物之间的性能异质性，从而提供汇总统计数据所掩盖的可操作洞见。

我们的评估表明，对于主要加合物类型，计算化合物预测已达到具有实际应用价值的准确率水平。对于 $[M+H]^+$，训练后的模型取得了超过 86% 的分子式预测准确率和超过 67% 的结构预测准确率。然而， $[M+H\text{-}2H_2O]^+$和 $[M\text{-}H+2Na]^+$ 等不太常见的加合物类型则表现出严重的性能下降，这凸显了根本性挑战：要解决这些挑战，仅靠简单的数据扩展远远不够，还需要架构创新。通过分别在提供真实分子式的预言机条件下，以及在预测分子式的现实流程中评估结构生成，我们得以将总预测误差分解为各个阶段的误差贡献。我们发现，分子式预测中的误差传递到结构生成时，造成的影响意外地有限。这些发现为确定未来研究方向的优先级奠定了基础，并为在现实世界中的代谢组学应用中部署这些方法提供了最佳实践指导。

2 实验部分

2.1 数据集

NIST 23 谱库：一个商业级高质量质谱库，在该领域被广泛视为黄金标准。它包含经过仔细整理并接受严格质量控制的质谱，是在受控条件下评估预测准确率的理想基准数据集。我们使用了其中的 MS/MS质谱，并依据以下六项标准对其进行筛选：

(1) 电离极性：正离子。(2) 碎裂类型：高能碰撞解离。(3) 电荷态：仅包含单电荷前体离子（电荷 $= 1$）。(4) 元素：仅限 H、C、N、O、P、S、F（涵盖大多数具有生物学意义的代谢物）。(5) 加合物类型： $[M+H]^+$、 $[M+Na]^+$、 $[M+K]^+$、 $[M+H\text{-}H_2O]^+$、 $[M+H\text{-}2H_2O]^+$、$[M\text{-}H+2Na]^+$。(6) 碰撞能量：每个分子-加合物对最多选取三张质谱，优先选择碰撞能量中位数对应的谱图，以减少冗余。

MoNA（北美 MassBank：）这是一个由代谢组学社区贡献内容的开放获取质谱数据库。与 NIST 不同，MoNA 展现了真实世界质谱数据的多样性和异质性，其中包括来自采用不同实验方案和仪器设备的各个实验室的贡献。我们下载了液相色谱串联质谱正离子模式谱图，并依据以下标准对其进行筛选：

(1) 元素：仅限 H、C、N、O、P、S、
F。（2）加合物类型：与 NIST 23 相同。

数据集统计： 表1总结了最终数据集的统计信息，表2展示了各数据集中不同加合物类型的谱图分布。后者揭示了显著的不平衡现象，反映了真实世界的情况，即某些加合物占据主导地位。值得注意的是， $[M+H]^+$ 在两个数据集中都占据了大多数谱图，在 NIST 中约占 56.1%，在 MoNA 中约占 78.6%。相比之下， $[M+H\text{-}2H_2O]^+$ 和 $[M\text{-}H+2Na]^+$ 仅占数据的一小部分，某些情况下谱图数量不足 30 条。

表1。 筛选后两个数据集的总体统计信息。（分子量 $=$ 分子量）

指标	NIST	MoNA
质谱数	86,435	76,196
# 独特分子	19,511	11,894
# 独特分子式	9,630	6,244
分子量范围（道尔顿）	82.11–1,704.01	45.08–2,680.17

2.2 化合物预测的计算方法

化学式预测

SIRIUS：当前最先进的基于碎裂树的分子式预测方法 [5]。该方法首先枚举与观测到的前体质量相符、且处于指定质量容差范围内的候选化学式。对于每个候选化学式，SIRIUS 会为 MS/MS 峰分配可能的子化学式，并构建碎裂树，以表示从前体到观测碎片的可能中性丢失路径。该碎裂树采用最大后验（MAP）估计框架进行评分，该框架同时考虑基于化学知识的碎片合理性，以及对 MS/MS 谱的解释质量。SIRIUS 使用由领域知识参数化的人工设计评分函数，其中包括预期的碎裂模式、常见的中性丢失和同位素分布。该方法还可选择纳入 MS1 同位素模式信息，以提高分子式判别能力；不过，为了与 MIST-CF（下文将介绍）进行公平比较，我们仅关注基于 MS/MS 的评分部分。值得注意的是，SIRIUS 是一种基于规则的方法，其参数由人工设计，因此无需训练。

MIST-CF：一种基于神经网络的方法，用于根据串联质谱谱图对化学式候选项进行排序 [8]。该方法采用基于能量的建模框架，并使用化学式Transformer架构。给定候选化学式及其对应的加合物类型，MIST-CF 首先在指定的质量容差内，为串联质谱的碎片峰分配合理的子化学式。随后，使用正弦嵌入对这些带有子化学式注释的峰进行编码，并通过Transformer网络处理，从而学习对候选分子式与观测到的碎裂模式之间的兼容性进行评分。与基于碎裂树的方法不同，MIST-CF 直接从数据中学习评分函数，无需显式构建树，因此能够高效评估大量候选项。在我们的实验中，我们使用了一个按照 MIST-CF 流程在 NPLIB1 数据集上预训练的模型，并进行了调整，使其侧重于我们关注的六种加合物类型。此外，我们分别在 NIST 和 MoNA 数据集上从头训练了两个独立模型。

分子结构生成

MSNovelist：一种基于长短期记忆网络的序列到序列模型，根据分子指纹预测结果生成以 SMILES 字符串表示的分子结构 [14]。该模型分两个阶段运行：首先，指纹预测模型根据串联质谱谱图推断概率分子指纹；其次，LSTM解码器在预测指纹的引导下自回归地生成 SMILES 字符。我们的实验使用了 SIRIUS 软件中的 MSNovelist 实现。其底层循环神经网络模型基于来自 HMDB、COCONUT 和 DSSTox 数据库的 1,232,184 个化学结构进行训练。

DiffMS：一种基于离散图扩散的条件分子生成模型 [2]。该方法由两个主要组件构成：谱编码器和图解码器。编码器采用与 MIST-CF 相同的 Formula Transformer 架构，从串联质谱中提取结构信息，并生成能够表征碎裂模式的固定维度嵌入。解码器在分子图上实现离散扩散过程，从随机初始化开始，对键类型进行迭代去噪，同时以谱图嵌入和已知化学式为条件。

在实验中，我们使用了作者发布的预训练 DiffMS 模型，该模型基于 NPLIB1 数据进行训练。为进行领域自适应，我们分别在 NIST 和 MoNA 数据集上训练了两个独立的模型，这两个模型均使用预训练解码器的权重进行初始化。我们在两种实验设置下评估了 DiffMS，以区分完整预测流程中不同的误差来源：

(1) **DiffMS (预言机)**：在此设置中，我们向 DiffMS 解码器提供真实的化学式。该设置代表结构预测性能的上限，旨在隔离解码器在给定正确化学式的情况下生成正确分子图的能力。在此设置下产生的任何错误，都反映了结构生成模型本身的局限性，或 MS/MS 数据固有的歧义（例如，无法区分立体异构体，或具有相似碎裂模式的结构异构体）。

(2) **DiffMS（MIST-CF）：**在此设置中，我们构建了一个完整的端到端流程：首先由 MIST-CF 根据 MS/MS 谱预测候选化学式，然后将排名前 5 的分子式预测分别输入 DiffMS 解码器进行结构生成。对于这五个化学式中的每一个，DiffMS 都会生成分子结构，从而得到扩展的结构候选项集合，以反映分子式预测步骤中的不确定性。我们根据结构生成频率对最终的结构候选项进行排序。该设置下的性能反映了完整的注释流程，并量化了分子式预测不确定性如何传递至结构生成过程。

对于预言机设置，我们为每张光谱生成了 100 个分子结构候选项。对于 MIST-CF 配置，我们针对排名前 5 的预测分子式中的每一个分子式，分别生成 20 个结构候选项，因此每张光谱最多可得到 100 个候选项。

2.3 评估设置

指标： 我们采用了质谱文献中的标准指标。对于分子式预测，我们计算了前 K 准确率。对于结构预测，我们使用了两个指标：(1) 基于 InChIKey-14 匹配的前 K 准确率（允许立体异构体存在差异）；(2) 最大 Tanimoto相似度。所有指标均报告前 K 个预测结果中的最佳匹配。

数据划分：我们通过随机划分，将 NIST 和 MoNA 数据集分别划分为训练集、验证集和测试集。这种策略是有意

选择的，而非结构差异性划分，因为它能保留加合物类型和化学性质的自然分布，从而反映真实的应用场景，在这些场景中，目标化合物往往具有相似的结构特征。两个数据集的加合物类型分布都严重不平衡，其中 $[M+H]^+$占主导地位，而若干罕见加合物类型的代表性则严重不足。表2列出了采用这种平衡策略后各数据集的划分统计信息。

表2. 按加合物类型划分的数据集统计。

	NIST				MoNA			
加合物类型	训练集总数		验证集测试		训练集总数		验证集测试集	
$[M+H]^+$	50,182	40,073	1,143	100	59,854	59,654	100	100
$[M+Na]^+$	13,376	10,790	298	100	10,997	10,797	100	100
$[M+K]^+$	354	292	6	30	5,197	4,997	100	100
$[M+H-H_2O]^+$	17,433	13,912	441	100	102	81	10	11
$[M+H-2H_2O]^+$	4,133	3,295	91	100	21	16	2	3
$[M-H+2Na]^+$	957	786	21	91	25	20	2	3
总计	86,435	69,148	2,000	521	76,196	75,565	314	317

3 结果与讨论

3.1 分子式预测：MIST-CF 与 SIRIUS

总体性能比较：表3总结了 SIRIUS 和 MIST-CF 在 NIST 和 MoNA 数据集上的分子式预测准确率。我们报告了 Top-1、Top-5 和 Top-10 准确率指标，这些指标反映了领域专家通常会检查多个排名靠前候选项的实际情形。MIST-CF 在两种配置下进行评估：在外部数据集（NPLIB1）上预训练，以及在目标数据集的训练集划分上重新训练（域内）。SIRIUS 在 NIST 上表现强劲（Top-5：0.758），但在 MoNA 上表现有限（0.435）。预训练的 MIST-CF 表现中等（Top-5 为 0.655–0.685）。域内重新训练显著提升了 MIST-CF 的性能：Top-1 准确率在 NIST 上提高了 8.2 个百分点，在 MoNA 上提高了 34.2 个百分点，MoNA 的 Top-5 准确率达到 0.892。然而，NIST 再训练呈现出意外的趋势：虽然 Top-1 有所提升（0.378 对比 0.296），但 Top-5 和 Top-10 均有所下降（0.503 对比 0.685，0.622 对比 0.727）。这可能由三个因素导致：（1）MIST-CF 的对比训练优先优化 Top-1 区分能力，而非排序校准；（2）NIST 受控且高质量的质谱有助于模型学习数据集特定伪影，从而提升 Top-1 表现，却降低排序鲁棒性；（3）NIST 的提升幅度小于 MoNA（8.2% 对比 34.2%），这表明 NPLIB1 已包含类似数据，限制了再训练的收益，同时引入了过拟合。

按加合物类型分层的分子式预测性能：虽然汇总指标能够提供有用的概览，但会掩盖不同加合物类型之间至关重要的性能差异。表4列出了所有评估配置下按加合物类型细分的前5名分子式预测准确率，揭示了显著差异，并表明这些差异具有重要的实际意义。

表3. SIRIUS 和 MIST-CF 的总体分子式预测准确率。SIRIUS 无需训练，而 MIST-CF 则分别评估其基础性能和经领域自适应再训练后的性能。粗体数值表示各测试配置下的最佳性能。

	NIST			MoNA		
方法	前1	前5	前10	前1	前5	前10
SIRIUS	0.284	**0.758**	**0.825**	0.193	0.435	0.500
MIST-CF（预训练）	0.296	0.685	0.727	0.402	0.655	0.744
MIST-CF（重新训练）	**0.378**	0.503	0.622	**0.744**	**0.892**	**0.915**

表4。按加合物类型分层的前5名分子式预测准确率。

	NIST			MoNA		
加合物类型	SIRIUS	MIST-CF（预训练）	MIST-CF（重新训练）	SIRIUS	MIST-CF（预训练）	MIST-CF（重新训练）
$[M+H]^+$	0.940	0.970	**0.990**	0.730	**0.869**	**0.869**
$[M+Na]^+$	0.750	0.840	**0.950**	0.520	0.780	**0.930**
$[M+K]^+$	0.667	0.233	**0.800**	0.160	0.310	**0.990**
$[M+H-H_2O]^+$	**0.950**	0.850	0.160	0.727	**0.909**	0.364
$[M+H-2H_2O]^+$	**0.950**	0.840	0.150	**1.000**	0.667	0.000
$[M-H+2Na]^+$	**0.176**	0.000	0.143	0.000	0.000	0.000

主要加合离子表现优异：对于 $[M+H]^+$ 和 $[M+Na]^+$，重新训练后的MIST-CF在两个数据集上的准确率均达到 $>$0.86，其中NIST上的准确率分别为0.990和0.950。SIRIUS在这些常见类型上也表现良好（0.730–0.940），这证实分子式注释已在大多数代谢组学应用中达到实用水平。

失水加合离子呈现出意想不到的模式：值得注意的是，失水加合物（$[M+H-H_2O]^+$ 和 $[M+H-2H_2O]^+$）表现出相反的性能趋势：SIRIUS 的准确率最高（0.727–1.000），而重新训练后的 MIST-CF 则表现灾难性失效（0.000–0.364）。在 NIST 上，SIRIUS 对两种类型的准确率均达到 0.950，而重新训练后的 MIST-CF 准确率则骤降至 0.160 和 0.150，较预训练模型的表现（0.850 和 0.840）大幅下降了 80 个百分点。这表明：（1）基于规则的碎裂树方法更适合处理失水歧义，因为多个分子式可能产生相似的质量变化；（2）NIST 再训练可能对占主导地位的加合物模式过拟合，因而失去了对化学性质模糊的情况进行泛化的能力；（3）这些类型的训练样本量较小（表2），加剧了过拟合。

极度稀有：$[M-H+2Na]^+$ 仍未解决：$[M-H+2Na]^+$ 加合物在所有方法中的准确率都接近于零（0.000–0.176），这反映了数据极度稀缺，同时也引发了一个问题：该加合物类型是否已被正确确定。

3.2 结构预测：DiffMS 与 MSNovelist

整体性能：表5展示了不同配置下的分子结构预测性能。

表5. 整体结构预测性能（前五项指标）。DiffMS（Oracle）使用真实分子式，而 DiffMS（MIST-CF）使用 MIST-CF 的前五项分子式预测。MSNovelist 已预训练，且独立于我们的训练配置。

	NIST		MoNA	
方法	准确率	Tanimoto系数	准确率	Tanimoto系数
MSNovelist	0.447	0.619	0.249	0.405
DiffMS（预训练，预言机）	0.077	0.310	0.088	0.298
DiffMS（预训练，MIST-CF）	0.052	0.282	0.066	0.272
DiffMS（重新训练，预言机）	**0.664**	**0.818**	**0.634**	0.812
DiffMS（重新训练，MIST-CF）	0.662	0.817	0.630	**0.818**

误差传播分析：一个引人注目的发现是，重新训练后的 DiffMS（MIST-CF）流程表现与 DiffMS（Oracle）设置非常接近。Oracle 与 MIST-CF 配置之间的准确率差距极小：在 NIST 上相差 0.2 个百分点（0.664 对 0.662），在 MoNA 上相差 0.4 个百分点（0.634 对 0.630）。鉴于分子式预测并不完美（例如，重新训练后的 MIST-CF 在 NIST 上的 Top-5 准确率为 0.503，见表3），这种稳健性令人惊讶。这种韧性可能源于两个因素：（1）多分子式策略（为 Top-5 个分子式生成结构）能够减轻错误的影响；（2）谱图嵌入提供了结构约束，可以部分弥补分子式错误。这表明，结构生成（而非 Top-5 范围内的分子式预测）才是主要瓶颈。Tanimoto分数几乎没有变化，这证实即使流程未能成功，生成的结构在化学上仍然合理。

为进一步了解流程的错误特征，我们进行了详细的失败分析（图1）。图1a表明，结构生成对分子式错误具有较强的鲁棒性，即使分子式不正确，仍能保持较高的Tanimoto相似度。图1b将失败模式分解为结构生成瓶颈（b1）、分子式传播错误（b2）和罕见加合物坍缩（b3）。图1c显示，异构体区分仍然具有挑战性，对于$C_{20}H_{32}O_3$ 异构体的准确率仅为66.7%。

领域适应的影响：领域自适应对于结构预测至关重要。域内（重新训练的）模型的表现远远优于通用（预训练的）模型。比较预言机配置可知，在NIST上，重新训练的DiffMS与预训练的DiffMS之间的准确率差距为58.7个百分点（0.664对0.077）；在MoNA上则为54.6个百分点（0.634对0.088）。这种对领域自适应的敏感性显著高于分子式预测中观察到的情况，反映出结构生成的复杂性更高。

按加合物类型分层的结构预测性能： 表6
展示了按加合物类型分层的Top-5准确率。
主要加合物类型达到实用准确率：对于 $[M+H]^+$这一主要加合物类型，DiffMS（MIST-CF）在两个数据集的域内设置下均实现了超过 0.67 的 Top-5准确率。结合 $>$0.86 的分子式预测准确率，这表明其在最常见场景中具有实际应用价值。

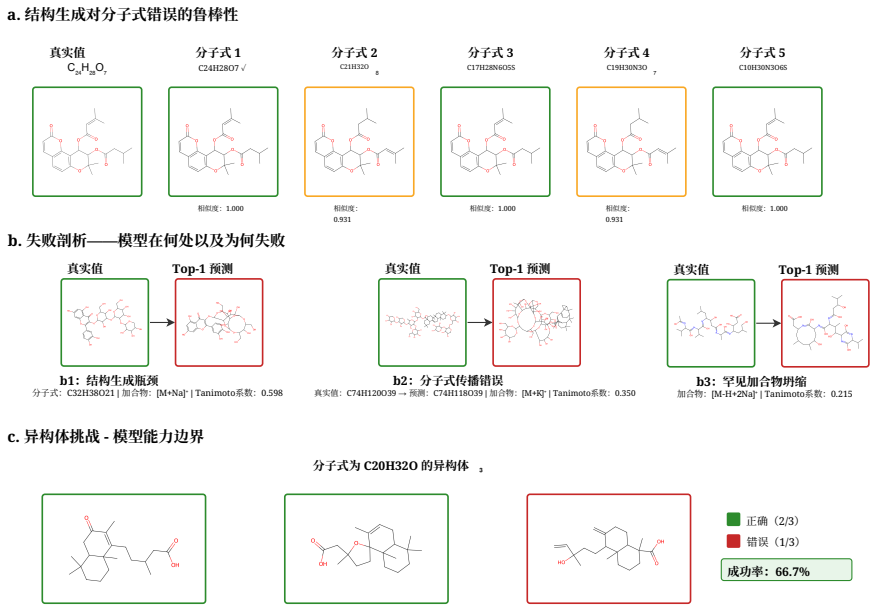

图 1. 流程性能的诊断分析。(a) 结构生成对分子式错误的鲁棒性。(b) 故障模式分解：结构生成瓶颈 (b1)、分子式传播错误 (b2) 和罕见加合物坍缩 (b3)。(c) 异构体区分挑战：对于 $C_{20}H_{32}O_3$，成功率为 66.7%。

罕见加合物类型的性能严重下降： 罕见加合物类型 $[M+H\text{-}2H_2O]^+$ 和 $[M\text{-}H+2Na]^+$ 在 MoNA 上的 Top-5 准确率仅分别为 0.333 和 0.000。

3.3 罕见加合物类型的数据增强

鉴于较不常见的加合物类型出现了严重的性能下降，我们通过使用 NIST 质谱补充 MoNA 有限的训练样本，测试数据稀缺是否是限制因素；其中包括 2,000 $[M+H\text{-}H_2O]^+$、4,032$[M+H\text{-}2H_2O]^+$ 和 951 $[M\text{-}H+2Na]^+$。尽管训练数据大幅增加，数据增强仍未能提升任何较不常见加合物类型的分子式预测和结构预测性能。这些结果表明，数据质量和相关性比原始数据量更为重要。

3.4 实际意义与建议

重新思考评估范式：当前的基准测试工作强调结构不相似性划分，将泛化置于实际相关性之上。在真实的代谢组学研究中，研究人员通常会研究相关的化合物家族，因此预期它们具有结构相似性。我们的随机划分策略更能反映此类情境，并表明，只要域内数据充足，最先进的模型就能达到实用的性能（> 0.86 前 5 个分子式， > 0.67 前 5 个

表6. 按加合物类型分层的前5名结构预测准确率。

数据集	加合物类型	MSNovelist	DiffMS			
			预训练†	预训练‡	重新训练†	重新训练‡
NIST	$[M+H]^+$	**0.800**	0.170	0.170	0.730	0.760
	$[M+Na]^+$	0.410	0.050	0.040	0.530	**0.540**
	$[M+K]^+$	0.300	0.067	0.033	**0.667**	**0.667**
	$[M+H-H_2O]^+$	0.550	0.040	0.010	**0.750**	0.720
	$[M+H-2H_2O]^+$	0.320	0.010	0.000	0.560	**0.570**
	$[M-H+2钠]^+$	0.176	0.132	0.055	**0.758**	0.725
MoNA	$[M+H]^+$	0.470	0.270	0.202	**0.680**	0.677
	$[M+Na]^+$	0.150	0.010	0.020	**0.620**	**0.620**
	$[M+K]^+$	0.090	0.000	0.000	**0.610**	0.600
	$[M+H-H_2O]^+$	0.455	0.000	0.000	**0.818**	**0.818**
	$[M+H-2H_2O]^+$	**1.000**	0.000	0.000	0.333	0.333
	$[M-H+2钠]^+$	0.000	0.000	0.000	0.000	0.000

†预言机：使用真实分子式进行结构预测；‡流程：使用 MIST-CF 预测的分子式。

$[M+H]^+$的结构准确率）。我们建议同时报告域内和域外结果，以指导其在现实世界中的应用。

数据整理建议： 数据增强实验揭示了三项挑战：(1) 高质量 NIST 光谱与异质 MoNA 数据之间存在分布不匹配，限制了迁移能力；(2) 失水加合离子的光谱存在歧义，使鉴定问题难以适定；(3) 模型容量有限，难以同时学习数据集特定规则和加合物特定规则。因此，我们建议：(1) 聚焦于目标化学空间的域内数据；(2) 混合参考质谱与社区质谱图，兼顾质量和多样性；(3) 通过针对稀有电离模式进行定向采集，缓解加合物不平衡；(4) 采用先进的数据增强方法（例如生成式方法或领域自适应方法）。

未来研究方向 我们的研究发现了若干有前景的研究方向：(1) 对较少见加合物的持续性差距表明，需要采用专用架构或元学习；(2) 整合保留时间或碰撞截面（CCS）或可消除歧义；(3) 领域自适应对于实现跨来源泛化至关重要。

4 结论

本研究表明，化合物预测现已具备常规应用的可行性，但仍存在关键瓶颈。按加合物分层的结果揭示，限制准确率的是结构生成，而非分子式预测。对较少见加合物而言，简单的数据扩展效果不佳，这凸显了专门模型、多模态整合和领域自适应的必要性。我们的分析既验证了现有工具，也为实现更全面、更贴近实际的化合物预测指明了方向。

参考文献

1. Bittremieux, W., Wang, M., Dorrestein, P.C.：谱库在记录代谢组学社区知识方面发挥的关键作用。Metabolomics**18**(12), 94 (2022) 2. Bohde, M., Manjrekar, M., Wang, R., Ji, S., Coley, C.W.：DiffMS：以质谱为条件生成分子的扩散生成方法。arXiv 预印本 arXiv:2502.09571 (2025) 3. Bushuiev, R., Bushuiev, A., de Jonge, N., Young, A., Kretschmer, F., Samusevich, R., Heirman, J., Wang, F., Zhang, L., Dührkop, K., 等：MassSpecGym：用于分子发现与鉴定的基准测试。Advances in Neural Information Processing Systems **37**, 110010–110027 (2024) 4. van Der Hooft, J.J., Mohimani, H., Bauermeister, A., Dorrestein, P.C., Duncan, K.R., Medema, M.H.：连接基因组学与代谢组学，描绘特化代谢多样性。Chemical Society Reviews **49**(11), 3297–3314 (2020) 5. Dührkop, K., Fleischauer, M., Ludwig, M., Aksenov, A.A., Melnik, A.V., Meusel, M., Dorrestein, P.C., Rousu, J., Böcker, S.：SIRIUS 4：将串联质谱快速转化为代谢物结构信息的工具。Nature Methods **16**(4), 299–302 (2019) 6. Dührkop, K., Shen, H., Meusel, M., Rousu, J., Böcker, S.：利用 CSI:FingerID，根据串联质谱搜索分子结构数据库。Proceedings of the National Academy of Sciences **112**(41), 12580–12585 (2015) 7. Fiehn Laboratory：北美 MassBank（MoNA）（2025），https://mona.fiehnlab.ucdavis.edu/ 8. Goldman, S., Xin, J., Provenzano, J., Coley, C.W.：MIST-CF：从串联质谱推断化学式。Journal of Chemical Information and Modeling**64**(7), 2421–2431 (2023) 9. Hong, Y., Li, S., Ye, Y., Tang, H.：FIDDLE：一种根据串联质谱预测化学式的深度学习方法。bioRxiv 第 2024–11 页（2024） 10. de Jonge, N.F., Louwen, J.J., Chekmeneva, E., Ca muzeaux, S., Vermeir, F.J., Jansen, R.S., Huber, F., van der Hooft, J.J.：MS2Query：可靠且可扩展的基于 MS2 质谱的类似物搜索。Nature Communications **14**(1), 1752 (2023) 11. Monge, M.E., Dodds, J.N., Baker, E.S., Edison, A.S., Fernández, F.M.：鉴定生命暗分子的挑战。Annual Review of Analytical Chemistry**12**(1), 177–199 (2019) 12. 美国国家标准与技术研究院：串联（MS/MS）质谱谱库（2023），https://www.nist.gov/srd/nist-standard-reference-database-1a 13. da Silva, R.R., Dorrestein, P.C., Quinn, R.A.：揭示代谢组学中的暗物质。Proceedings of the National Academy of Sciences **112**(41), 12549–12550 (2015) 14. Stravs, M.A., Dührkop, K., Böcker, S., Zamboni, N.：MSNovelist：从质谱从头生成结构。Nature Methods **19**(7), 865–870 (2022) 15. Wang, M., Carver, J.J., Phelan, V.V., Sanchez, L.M., Garg, N., Peng, Y., Nguyen, D.D., Watrous, J., Kapono, C.A., Luzzatto-Knaan, T., 等：通过全球天然产物社会分子网络共享质谱数据并进行社区策展。Nature Biotechnology**34**(8), 828–837 (2016) 16. Wang, Y., Chen, X., Liu, L., Hassoun, S.：MADGEN：借助质谱实现从头分子生成。arXiv 预印本 arXiv:2501.01950 (2025)