

基于 ChemVista 转化高度规范化的 Spectraverse 谱库，结合 Explorer 显著提升 MS/MS 谱库鉴定覆盖度

作者

于擎
安捷伦科技（中国）有限公司

摘要

本研究首先利用基于 Bioconductor 生态的生物信息学工具，对高度规范化的开源谱库 Spectraverse 进行信息补充与格式转换，并借助 Agilent ChemVista 将其无缝转化为安捷伦原生的 PCDL 格式谱库。随后，将该谱库应用于植物提取物中天然活性成分以及化工材料成分的非靶向分析。MassHunter Explorer 的分析结果表明，与仅使用官方标准谱库相比，引入 Spectraverse PCDL 后，天然产物的未知物鉴定覆盖度得到显著提升（从 104 个增加至 365 个）。另外，在分析电镀液成分时，有 62 个未知特征获得了二级质谱级别的验证。这表明安捷伦软件生态具有高度开放性，能深度融合高价值第三方开源资源，使未知物的高通量鉴定更加简捷、高效。

前言

尽管当前涌现出众多性能优异的二级质谱解析软件，但在筛查和非靶向分析中，将二级谱图与谱库进行匹配仍然是研究人员首选的化合物鉴定方式。除仪器供应商提供的官方谱库以外，数量庞大、覆盖范围更广的开源第三方谱库，正越来越多地被高分辨质谱数据分析人员所采用。然而，在使用这些开源谱库时，研究者往往面临两个主要问题。其一，谱库格式不统一。种类繁多的格式不仅无法被仪器供应商的软件直接读取，也难以被大多数第三方开源软件兼容。其二，谱库内容不规范。公共谱库中的谱图通常由用户上传，并由平台归纳整理。但由于缺乏统一的审核机制，谱图中普遍存在数据分散、质量参差不齐、元数据混乱、标注不一致等问题，给研究人员使用公共谱库造成较大困扰。

安捷伦于 2023 年发布了 ChemVista 谱库管理软件，用于整合来自多种来源的包括化合物结构、保留时间和质谱峰在内的质谱图数据。该软件的设计遵循以化合物为中心的数据结构，可避免以谱图为中心的传统文本格式（例如，.mgf 或 .msp）中普遍存在的元数据缺失、化学结构标识符（例如，InChIKey/SMILES）碎片化和命名混乱等问题，从而帮助用户以更高效、可靠的方式构建自己的谱库。目前，ChemVista 1.1 支持以下五种数据格式的导入：.cdb/.cdbx（安捷伦 PCDL）、.csv、.txt (MassBank) 和 .sdf (MassBank of North America 与 EPA CompTox Chemicals Dashboard)。用户将上述格式的谱图数据导入后，即可对谱库进行管理和编辑，并将其导出为安捷伦标准 PCDL 格式，用于 MassHunter 定性分析软件或 Explorer 中的未知化合物鉴定。

尽管如此，开源谱库的不规范问题无法仅依靠谱库格式的转换来解决。由普林斯顿大学团队开发并开源的 SpectraVerse^[1] 项目，正致力于从根源上应对这些长期存在的问题。该项目的初衷在于，系统性地解决公共 MS/MS 数据在可及性、质量和一致性方面的核心痛点，进而构建一个全面、洁净且标准化的谱库资源，为代谢物注释和计算质谱学领域的机器学习研究奠定坚实的数据基础。

本研究的目的在于充分利用 SpectraVerse 所创建的高度规范化的谱图数据，结合基于 Bioconductor 生态的生物信息学工具，对其部分缺失且影响用户体验的元数据信息进行补充，进而构建可供用户直接使用的原生 PCDL 谱库。这一工作将复杂的开源数据转换为安捷伦软件直接支持的格式，使普通用户能够省去自行编写脚本或进行复杂格式转换的麻烦。最终，该谱库可无缝加载至 MassHunter 定性分析软件或 Explorer 平台中，在未知物鉴定分析中帮助用户筛查更多潜在可能的成分，有效提升未知化合物的鉴定覆盖度。

实验部分

试剂和样品

甲醇和乙腈均为液质级，购于 Merck；甲酸为液质级，购于 Sigma；所用实验用水为 Millipore Milli-Q 超纯水系统现制备的高纯去离子水；样品为市售中药丸剂以及由用户提供的电镀液。

仪器和设备

采用 Agilent 1290 Infinity III UHPLC 与 Agilent Revident Q-TOF 联用系统，配备安捷伦双喷射流离子源 (AJS)。

样品前处理

中药丸剂样品：将中药丸剂切碎，称取 1 g，使用 10 mL 甲醇进行超声提取；取 1 mL 上清液，在 10 °C 下以 15000 r/min 离心后，取 200 µL 上清液直接上机分析。

电镀液样品：量取适量电镀液，用水/乙腈 (1/1) 稀释十倍后，直接上机分析。

色谱条件（用于中药丸剂提取液样品）

色谱柱：Agilent ZORBAX RRHD Eclipse Plus C18, 2.1 × 100 mm, 1.8 µm
流速：0.3 mL/min
柱温：40 °C
流动相 A：0.1% 甲酸水溶液
流动相 B：甲醇
进样体积：5 µL
后运行时间：3 min
梯度洗脱程序：

时间 (min)	B (%)
0	10
12	100
15	100
15.1	10
17	10

色谱条件（用于电镀液样品）

色谱柱：	Waters ACQUITY BEH Amide 色谱柱，2.1 × 100 mm, 1.7 μm
流速：	0.3 mL/min
柱温：	40 °C
流动相 A：	水（0.1% 甲酸 + 10 mM 乙酸铵）
流动相 B：	乙腈/水（90/10，0.1% 甲酸 + 10 mM 乙酸铵）
进样体积：	2 μL
后运行时间：	5 min
梯度洗脱程序：	

时间 (min)	B (%)
0	100
2	100
10	75
12	50
15	50

质谱条件

极性：	正离子模式
干燥气温度：	250 °C
干燥气流速：	7 L/min
雾化器压力：	35 psi
鞘流气温度：	350 °C
鞘流气流速：	11 L/min
喷嘴电压：	0 V
毛细管电压：	3500 V
碎裂电压：	125 V
MS 采集速率：	4 幅谱图/秒
MS2 采集速率：	10 幅谱图/秒
质量范围：	<i>m/z</i> 100–1200 (MS1)； <i>m/z</i> 40–1200 (MS2)
分离峰窗口：	窄，约 1.3 <i>m/z</i>
碰撞能量：	随 <i>m/z</i> 增加而提升（斜率：3.5； 截距：15）
每个循环最大母离子数：	10
MS/MS 阈值：	10000 响应值和 0.001%
基于母离子丰度的扫描速度：	40000 响应值/质谱图
剔除未达目标 TIC 的母离子：	否
启用主动排除：	是，重复一次，然后排除 0.1 min
纯度严格性：	70%，截留 0%
同位素模型：	常见有机分子
母离子排序：	按强度值
迭代模式：	开启，迭代次数 5

结果与讨论

Spectraverse 谱库的预处理

Spectraverse 收集了来自 GNPS、MoNA、MS-DIAL、HMDB 和 MSnLib 等公共资源的谱图，同时还特别补充了已发表但尚未收录于上述数据库的独立质谱数据，包括出版物支持信息中的谱图、MetaboLights 等原始数据仓库中的实验数据、GitHub 仓库和 Zenodo 存储库中的原始数据以及从出版物附件网站获取的谱图等。最终，总共汇集了 1,672,217 张原始 MS/MS 谱图。经严格筛选后，构建了包含 488,630 张高质量谱图的 Spectraverse 谱库（以 .mgf 格式存储）。然而，.mgf 格式的 Spectraverse 谱库无法直接被 ChemVista 使用。此外，研究期间还发现，该谱库中约有 13 万张谱图缺少化合物名称，其名称字段均被标注为“nan”。虽然这一问题对模型训练影响不大，但会显著降低化合物鉴定的便利性。因此，在将 Spectraverse 谱库真正应用于化合物鉴定之前，必须完成化合物名称补全和格式转换工作。

在本研究中，首先利用 Spectra^[2] 和 MsBackendMgf^[3] 包，将原始 .mgf 格式的 Spectraverse 谱库读入，并统一补充仪器类型和 MS 级别信息。随后，借助 webchem^[4] 工具，根据每张谱图的 SMILES 或 InChIKey，向 PubChem 查询其对应的化合物名称，并将这些名称补充至原始数据中。最后，利用 MsBackendMassbank^[5] 包，将数据逐条输出为独立的 MassBank txt 格式文档，并导入 ChemVista 中。

将数据导入至 ChemVista

完成 Spectraverse 的预处理后，即可将已转换为 MassBank txt 格式的谱库导入 ChemVista 中。具体设置如图 1 所示。

Import Files

Select a source type and files to import

Source type

MassBank (*.txt)

Select file(s)

Open file(s)

10000 files selected

Create list on import

☒

List name

Spectraverse 431-440

Description

Apply method label to imported data

☐

Import selected files

图 1. ChemVista 谱图导入选项

值得指出的是，ChemVista 本身被设计用于管理由不同厂商仪器产生的大量谱图数据，因此需要在预处理步骤中补全仪器类型和 MS 级别信息，以使用户在导出谱图时能够根据仪器类型（例如，LC-ESI-QTOF，即分离技术-电离技术-仪器类型）和 MS 级别等条件对谱图进行筛选。另外，由于 SpectraVerse 的谱图数量庞大，一次性导入全部文件将对计算机内存造成较大压力，因此建议在实践中将谱图分批导入，形成多个包含部分谱图的列表。待所有谱图导入完毕后，再使用 Merge 功能对列表进行合并，将 SMILES 和 InChIKey 等信息相同的谱图归属于同一化合物条目下。

将谱图导出为 PCDL

完成上述工作后，即可将谱图导出为 PCDL 格式。从图 2 中可以看出，大多数化合物对应于不止一张二级谱图。由于数十万种化合物包含高密度的二级碎片、同位素簇和多种加合离子元数据，单个化合物节点在底层涉及庞杂的关系型数据库写入操作。为确保大容量谱库在转化过程中的数据完整性，并避免因大规模数据写入引发系统资源过载，建议用户在实践中导出化合物谱图时采用“分批导出、整体合并”的稳健策略（例如，每批次控制在 1000 种化合物以内）；随后利用 PCDL Manager 软件的“Append to PCDL”功能进行统一整合。这一推荐工作流程能够最大程度确保生成的 PCDL 谱库在底层索引上的严谨性和长期稳定性。

The screenshot displays the ChemVista Library Manager interface. The main window shows a list of substances imported from SpectraVerse 471-489. The list includes columns for Substance Name, Formula, CAS, Mass, InChIKey, Agilent ID, and Spectra Count. The first substance listed is Sulfamethoxazole, with a Spectra Count of 29. Other substances listed include Amiloride, Phenazone, Metoclopramide, Losartan, Mefruside, Dixyrazine, Chloramphenicol, Valsartan, Isoquercitrin, Sulfamoxole, CORTISOL, GLUTAMINE, TRYPTOPHAN, THREONINE, Thiabendazole, Cortisone, Dothiepin, HISTIDINE, Nylidrin, Corticosterone, Lincomycin, CAFFEINE, Fipronil, GLUCOSAMINE, Mycophenolic Acid, Lovastatin, and 1H-Benzotriazol-1-ol.

On the right side of the interface, the chemical structure of Sulfamethoxazole is shown. Below the structure, the following information is displayed:

- CAS: C10H11N3O5
- Formula: C10H11N3O5
- InChI: InChI=1S/C10H11N3O5/c1-7-6-10(12-16-7)13-17(14,15)9-4-2-8(11)3-5-9/h2-6H,11
- InChIKey: JLKIGFTWXXRPMT-UHFFFAOYSA-N
- Mass: 253.05211
- SMILES: CC1=CC(NS(=O)(=O)C2C=CC(N)=CC=2)=NO1
- Synonyms: (empty field)

Below the chemical information, there are buttons for "View full compound data", "Edit", "View spectra", "Add to list", and "Export substance". The "View spectra" button is highlighted. The "Tags" section shows "LCMS Amenable". The "Number of spectra" is 29. The "Latest Retention Times" section shows "LC | ESI | QTOF" and "POSITIVE" with "29 spectra" listed below each.

图 2. ChemVista 导入的谱库列表，其中对许多化合物都记录了不少于一张谱图

Spectraverse PCDL 性能测试

中药丸剂提取液成分鉴定

为了评估 Spectraverse PCDL 能否显著提升二级谱图的鉴定覆盖度，本实验采用迭代 MS/MS 方法采集了市售中药丸剂提取液样品（见图 3）。随后，使用 MassHunter Explorer 对数据进行提取和对齐，并分别采用安捷伦 TCM PCDL 谱库以及 TCM PCDL 与 Spectraverse PCDL 联用的方式，对具有二级质谱信息的特征进行解析，以评估 Spectraverse PCDL 在提升化合物鉴定覆盖度方面的效果。

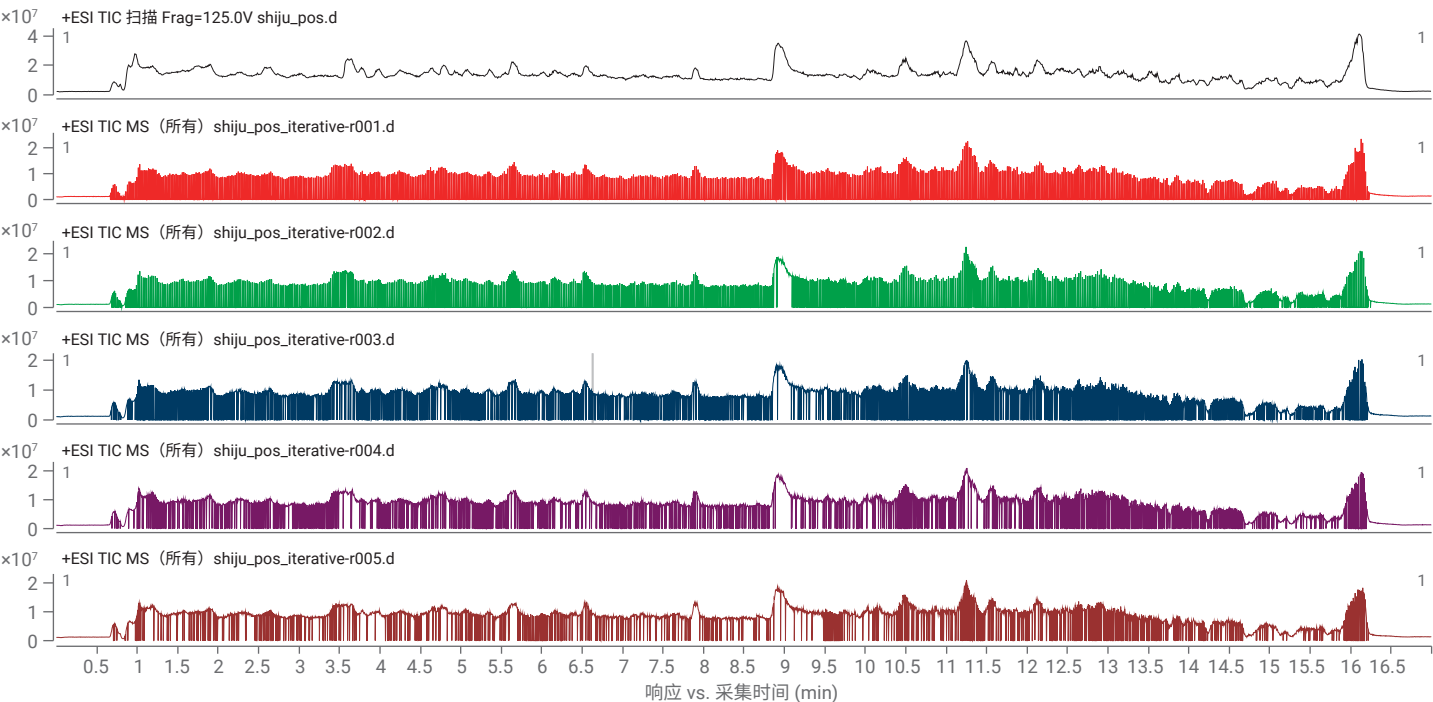


图 3. 通过迭代 MS/MS 采集的中药丸剂提取液 TIC 色谱图，其中显示，随着迭代次数的增加，能够采集到二级碎片的母离子数量显著减少

首先，仅使用安捷伦 TCM PCDL 谱库对特征进行鉴定，谱图匹配参数设置如图 4 所示。由于谱库中不包含保留时间信息，因此在“Contribution to overall score”中将 Score(RT) 的权重设为 0。另外，由于后续需要调用多个谱库对特征进行匹配，因此在“Tolerances and compound source”设置中，将谱库匹配方式选为“Search all databases/libraries”，以避免检索终止于非最优得分的匹配结果。

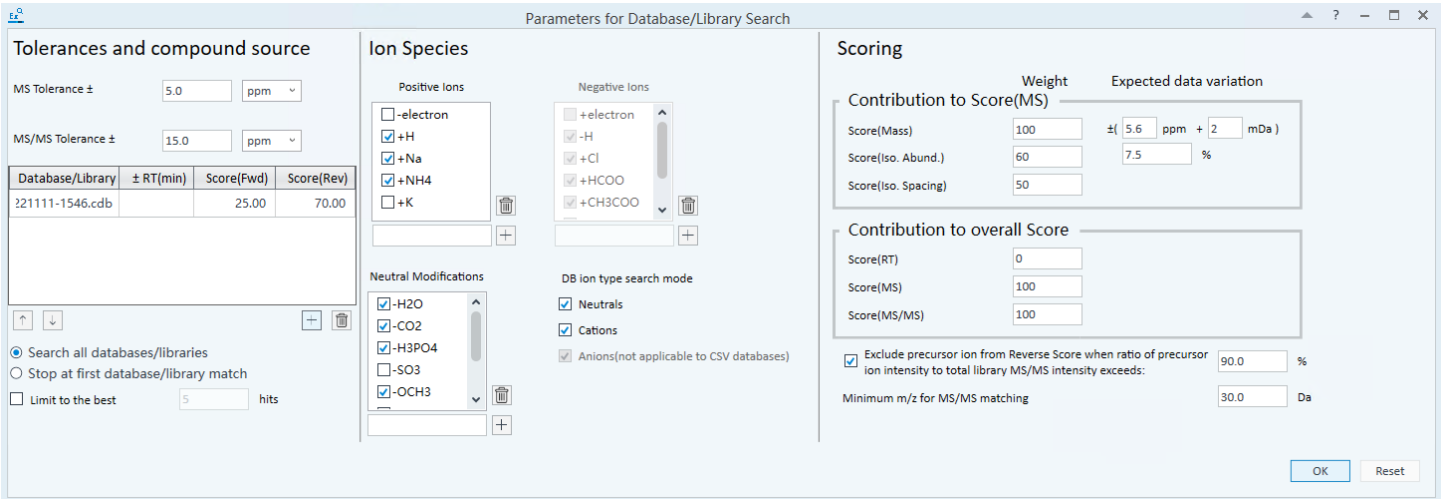


图 4. Explorer 的数据库/谱库检索设置

作为安捷伦官方针对中药及天然产物开发的专用谱库，TCM PCDL 的核心优势在于其极高的鉴定准确度。由于其谱图完全来源于药典相关高纯度标准品的实测数据，虽然其收录的化合物总数相对有限，但几乎排除了所有背景噪声和假阳性干扰。如图 5 所示，仅使用安捷伦 TCM PCDL 谱库进行匹配时，共有 104 个特征在二级质谱层面得到匹配。其余未匹配的特征可通过发送至 NIST 或 SIRIUS 进行匹配或解析。尽管 SIRIUS 具备优异的谱图解析准确度，但大多数用户仍倾向于通过谱库直接匹配未知特征，而非依赖计算或预测来确认化合物结构。

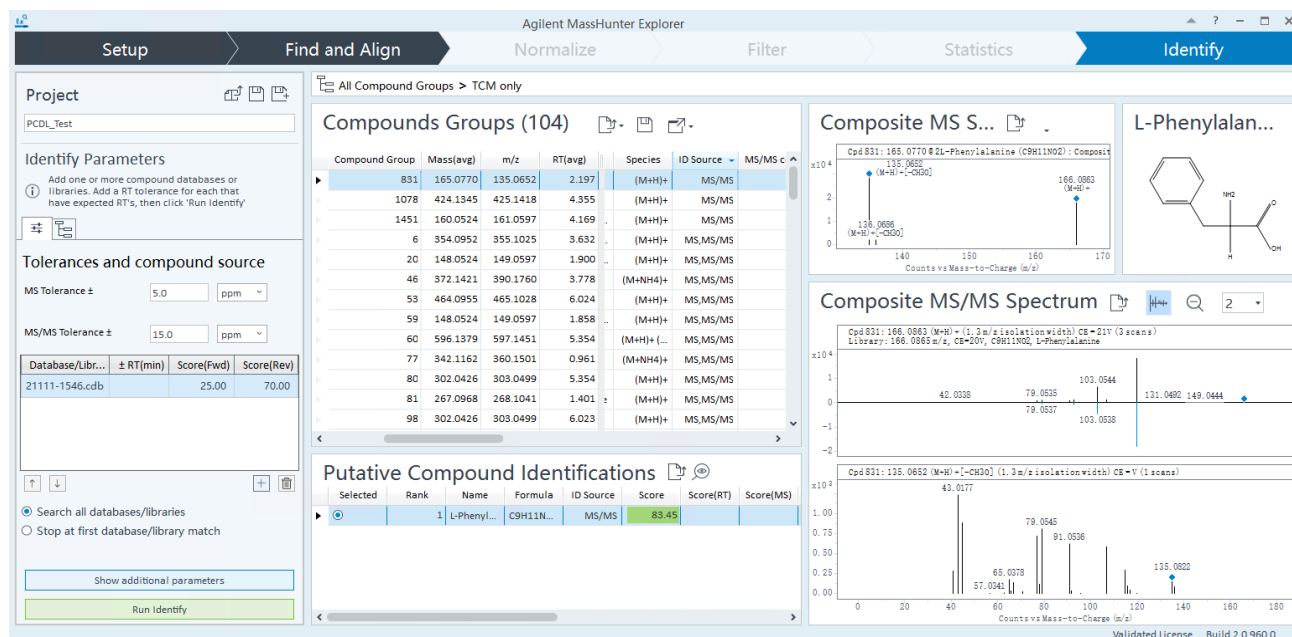


图 5. 使用安捷伦 TCM PCDL 对中药丸剂提取液的特征进行鉴定

为了充分利用 Explorer 支持同时检索多个谱库的特性，本研究进一步将安捷伦 TCM PCDL 与 Spectraverse PCDL 联合使用，对中药丸剂提取液中的特征进行匹配（见图 6）。结果显示，在相同的匹配参数设置下，共有 365 个特征获得了匹配，谱库覆盖度得到显著提升。与仅使用官方标准谱库相比，引入 Spectraverse PCDL 后，天然产物中未知物的鉴定覆盖度实现了大幅提升。这些结果验证了安捷伦软件生态的高度开放性，表明通过深度融合高价值的第三方开源资源，能够使未知化合物的高通量鉴定变得更加简捷、高效。

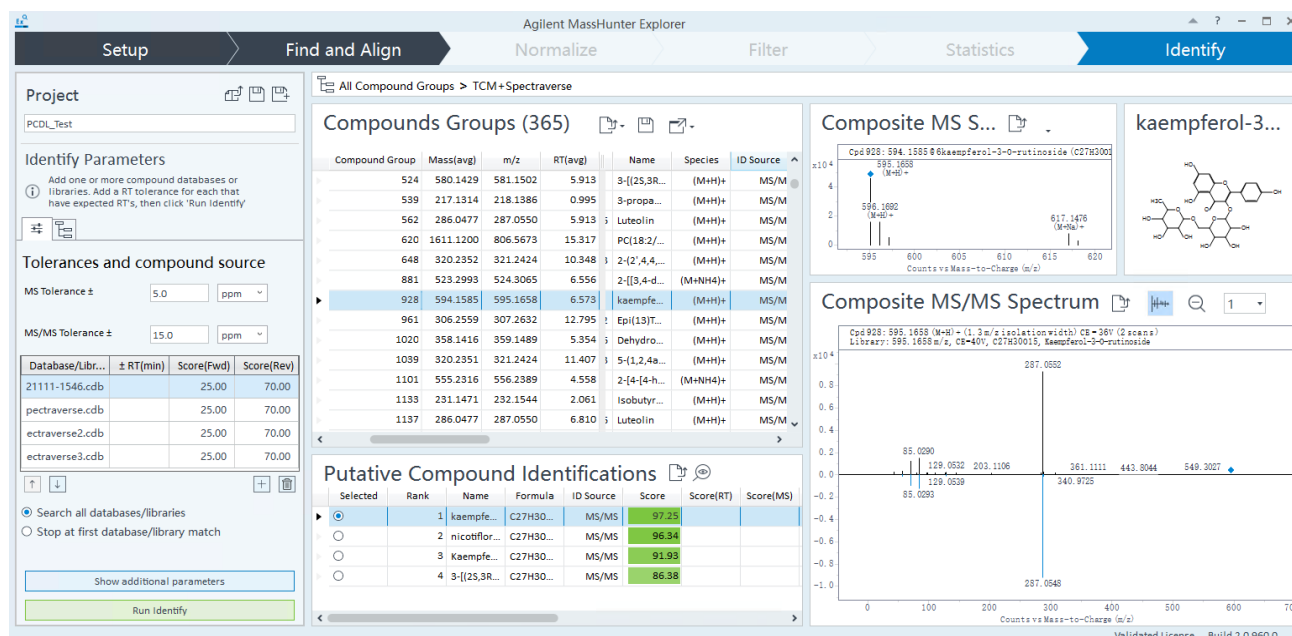


图 6. 使用安捷伦 TCM PCDL 和 Spectraverse PCDL 对中药丸剂提取液的特征进行鉴定

电镀液中的化学成分鉴定

为了进一步验证 Spectraverse PCDL 谱库在其他领域的适用性, 本研究还将其应用于某工业电镀液样品中未知成分的非靶向分析中。

在实际分析中, 由于缺乏针对化工和工业添加剂的专用质谱谱库, 此类样品的特征识别往往面临极大挑战。本研究采用迭代模式对电镀液样品进行数据采集, 并使用 Explorer 对特征进行提取和鉴定。结果如图 7 和图 8 所示。由于缺少专门对应的商品化 PCDL 可供检索, 因此直接使用 Spectraverse PCDL 对特征进行鉴定。Explorer 的匹配结果显示, 共有 62 个特征得到匹配。这一结果表明, Spectraverse PCDL 不仅在常规的天然产物研究中表现优异, 而且能够有效拓展至工业品配方解析和精细化工样品检测等领域, 具有良好的实用价值。

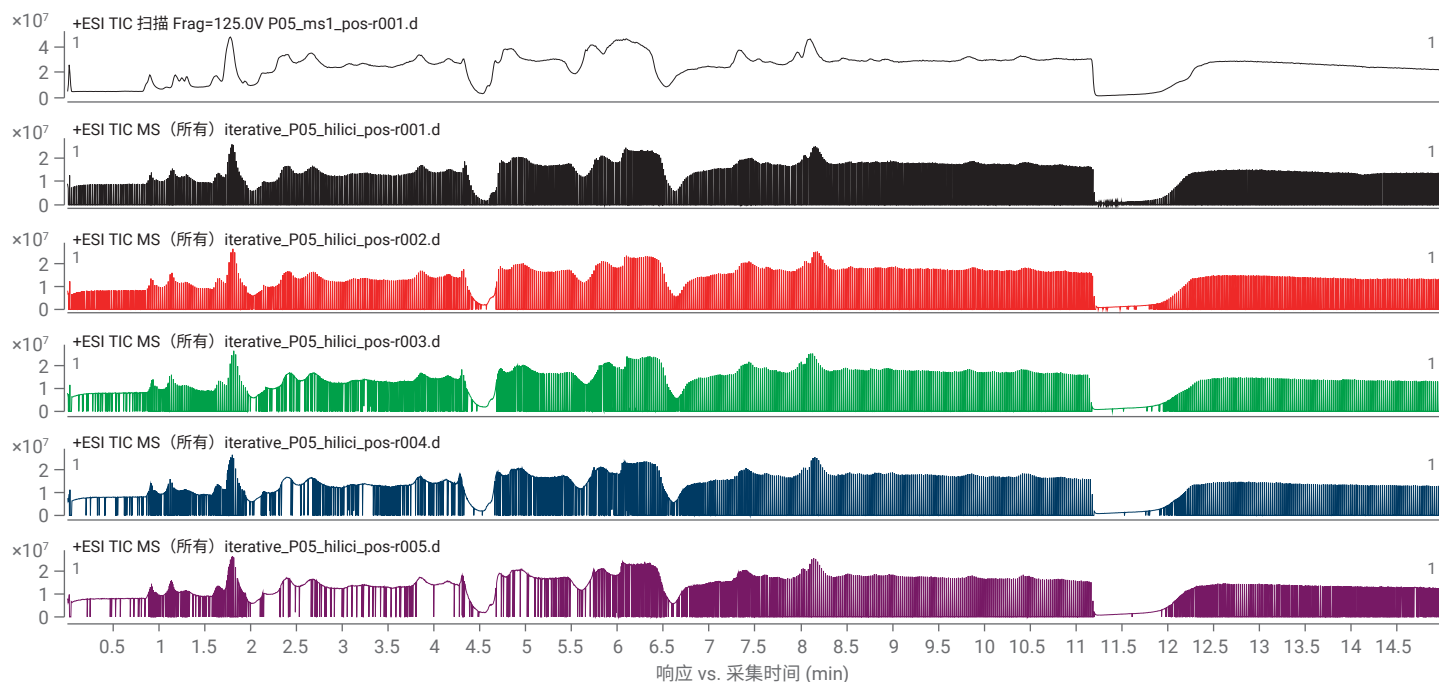


图 7. 通过迭代 MS/MS 采集的电镀液 TIC 色谱图

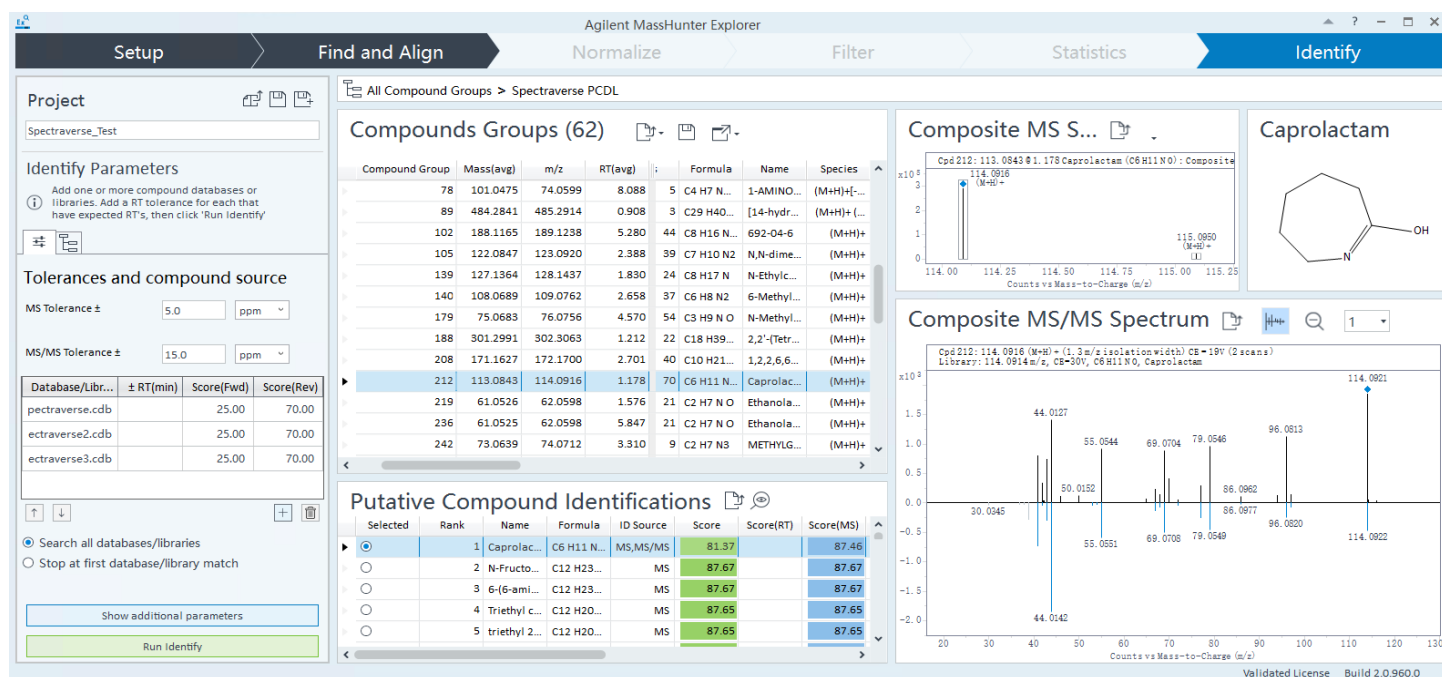


图 8. 使用 Spectraverse PCDL 对电镀液的特征进行鉴定

结论

本研究表明，将 Spectraverse 开源谱库与 Agilent ChemVista 谱库管理系统相结合，使用户能够充分利用开源、高度规范且数量庞大的第三方优质资源，实现对样品中未知化合物的高效鉴定。在叠加迭代 MS/MS 技术对二级质谱信息进行深度挖掘后，具备丰富信息的大型谱库在化合物鉴定方面的优势愈发显著。适用性测试结果进一步证明，Spectraverse PCDL 不仅适用于代谢组学和天然产物分析，还可扩展至化工材料及助剂中未知化合物的分析等领域，展现出广泛的应用前景。

参考文献

1. Vishu Gupta, Hantao Qiang, Hsin-Hsiang Chung, Ehud Herbst, and Michael A. Skinnider. Comprehensive Curation and Harmonization of Small-Molecule MS/MS Libraries in Spectraverse. *Analytical Chemistry*, 2026, 98(5): 3934-3943
2. Rainer J, Vicini A, Salzer L, Stanstrup J, Badia JM, Neumann S, Stravs MA, Verri Hernandez V, Gatto L, Gibb S and Wittin M. A Modular and Expandable Ecosystem for Metabolomics Data Annotation in R. *Metabolites*, 2022, 12(2):173
3. Gatto L, Rainer J, Gibb S (2025). MsBackendMgf: Mass Spectrometry Data Backend for Mascot Generic Format (mgf) Files. R package version 1.18.0
4. Szöcs E, Stirling T, Scott ER, Scharmüller A, Schäfer RB (2020). webchem: An R Package to Retrieve Chemical Information from the Web. *Journal of Statistical Software*, 2020, 93(13): 1-17
5. Witting M, Rainer J (2026). MsBackendMassbank: Mass Spectrometry Data Backend for MassBank record Files. R package version 1.18.2

查找当地的安捷伦客户中心：

www.agilent.com/chem/contactus-cn

免费专线：

800-820-3278, 400-820-3278 (手机用户)

联系我们：

LSCA-China_800@agilent.com

在线询价：

www.agilent.com/chem/erfq-cn



微信搜一搜

Q 安捷伦视界

www.agilent.com

DE-017147

安捷伦对本资料可能存在的错误或由于提供、展示或使用本资料所造成的间接损失不承担任何责任。

本文中的信息、说明和指标如有变更，恕不另行通知。

© 安捷伦科技（中国）有限公司，2026

2026 年 8 月 24 日，中国出版

5994-9420ZHCN