通过降维识别代谢组样品中的化学物质

A PREPRINT

埃米勒 · 阿南德 *
Computing and Mathematical Science
California Institute of Technology
Pasadena, CA, 91125
eanand@caltech.edu

查尔斯 · 斯坦哈特
Cosmic Dawn Center, Niels Bohr Institute
University of Copenhagen
Copenhagen, Denmark
steinhardt@nbi.ku.dk

马丁 · 汉森
Department of Environmental and Resource Engineering
The Technical University of Denmark
Lyngby, Denmark
marthan@dtu.dk

ABSTRACT

确定水源中的污染物的过程随着农药和重金属等污染物的复杂性而不断发展。常规程序目标分析用于判断水样是否安全饮用，该程序会从某个已知清单中查找特定物质；然而，由于任何此类污染物清单在结构上的复杂度增加，这样的清单不幸地是非详尽的：我们不明确知道哪些物质应该被列入这个清单。

这引发了以下基本问题：在实验确定哪些物质是污染物之前，如何解决识别水中的所有物质的采样问题？

在这里，我们提出一种方法，该方法建立在 Liigand et al. [2020] 的工作基础上，通过非目标分析开发随机森林回归模型来预测样本中每个分子的化学式及其各自的浓度。

这项工作利用降维和线性分解技术，使用欧洲质量银行代谢组库 [Oberacher et al., 2020] 的数据呈现更准确的无监督模型以识别新样品中的化学式。

Keywords 降维 · 化学识别 · 非负矩阵分解 · 质谱 · 主成分分析 · 卷积滤波器 · 光谱鉴定

*This work was done while the author was visiting the Niels Bohr Institute at the University of Copenhagen in Denmark, and the research was funded by a Caltech Summer Undergraduate Research Fellowship (SURF) in 2020.

1 介绍

提供安全饮用水的需求是不容置疑的。科学家目前使用目标分析来寻找水中的污染物。目标分析是在样本上进行的一种狭窄搜索，以识别特定物质。然而，这引出了一个统计抽样问题：虽然科学家知道如何测试和过滤特定物质，但他们如何先验的确定这些测试应涵盖哪些物质？一种新兴的水样分析方法是非目标分析 [Liigand et al., 2020, Alder et al., 2006, Cajka and Fiehn, 2016]，即对水样本进行广泛搜索，利用质谱数据检测已知和未知化学物质。该领域的非目标分析需要一个基于数据的方法，利用机器学习来学习将质谱仪的数据映射到化学公式的函数。因此，确保安全饮用水供应并识别代谢组（生物样品中的化学物质集合）组成的这一问题的关键在于采用一种基于数据的方法来学习这个函数，并使用它来识别和预测代谢组中每种化学物质的浓度。2020 年，Liigand et al. [2020] 尝试通过开发一个超参数正则化随机森林回归器来解决这个问题。随机森林回归模型是一种离散机器学习模型，它在一个数据的随机子集上训练一组随机化的回归树。该模型通过对每个随机树回归器的预测进行平均来估计值。Liigand 等人的模型通过在每棵树 [Liigand et al., 2020] 上强制执行最大叶节点大小为 120 个节点实现了强大的正则化条件（一种防止模型过度拟合数据的方法）。因此，在输入质谱仪的数据时，随机森林回归模型 [Chen and Guestrin, 2016] 对输出空间进行离散化以估计决策边界；在这样做时，它学习了一些化学物质与其质量光谱之间的底层关系片段 [Gupta and Anand, 2020]。Liigand 等人的超参数化模型产生了一个平均误差为 8.8%[Liigand et al., 2020]，这个误差可以通过改进来降低，因为离散模型忽略了质量光谱和相应代谢物轮廓之间可能存在的连续关系。这可以通过 t-随机近邻嵌入（t-SNE）论点来证明：

t-SNE[van der Maaten and Hinton, 2008] 是一种流行的数据可视化工具。它为每个点创建一个概率分布（使用以该点为中心的归一化高斯分布），并通过随机梯度下降算法最小化这个分布与一组点在 $\mathbb{R}^2$ 中的分布之间的 Kullback-Leibler 散度。在 $\mathbb{R}^2$ 中，T-SNE 倾向于形成可以用单一变量定性描述的点群；因此，集群的数量为描述数据所需的大约最小变量数提供了线索 [van der Maaten and Hinton, 2008]。图 1 显示了一个用氯原子存在与否进行颜色编码的 t-SNE 图（在这里，含氯化合物是黑色的而非含氯化合物是黄色的）。图 1b 表明，由于存在三个集群，因此在降维形式下存在可靠的三变量聚类。观察到彩色标签中的方差是合理的，因为这些聚类无法区分相似和不相似的数据点（t-SNE 算法根本不会检查这一点）。因此，我们解决这个问题的方法可以潜在地产生大约为 0 的平均真实误差 [Waggoner, 2021]，因为我们不对分布做任何假设。

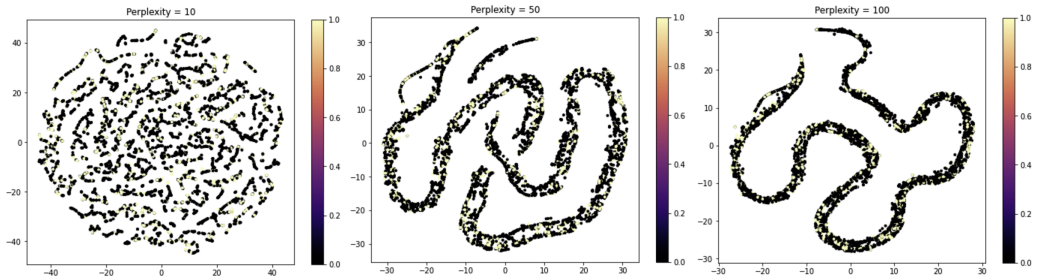

图 1： t-SNE 应用于困惑度分别为 10、50 和 100 的对象，训练了 500 次迭代。低困惑度值形成局部和复杂的结构。高困惑度值形成具有更显著聚类的全局结构。

2 预备知识

我们提供以下基本定义。

定义 2.1 (KL 散度)**.** 对于定义在相同样本空间 $\mathcal{X}$ 上的离散概率分布 P 和 Q，Kullback-Lieblerg（KL）散度由以下给出

$$D_{\mathrm{KL}}(P\|Q) = -\sum_{x\in\mathcal{X}} P(x)\log\frac{Q(x)}{P(x)} \tag{1}$$

t-SNE 模型通过 KL 散度分数 [van der Maaten and Hinton, 2008] 定义。

t-S 随机 N 邻居 E 嵌入 (t-SNE). 给定一组 N 高维对象 $x_1, \ldots, x_N$，t-SNE 首先计算概率 p_{ij}，其中 $p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N}$。这里，

$$p_{i|j} = \begin{cases} \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}, & i \neq j \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

接下来，t-SNE 学习一个 d 维映射 $y_1, \ldots, y_N$（具有 $y_i \in \mathbb{R}^d$）其中点 y_i 和 y_j 之间的相似度 q_{ij} 由 $\mathrm{KL}(P\|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$ 给出。这里，

$$q_{ij} = \begin{cases} \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|y_k - y_l\|^2)^{-1}}, & i \neq j \\ 0, & \text{otherwise} \end{cases}, \tag{3}$$

其中使用梯度下降 [Lin et al., 2023a,b, 2024] 来最小化 KL 散度。

3 方法

本节描述了数据处理技术和将化学信号编码为 SMILES 字符串的方法。

3.1 数据采集与预处理

逆向工程质谱过程的理想机器学习模型接受一个质量光谱作为输入，并预测对应于该光谱的物质名称。在构建这样的模型之前，重要的是我们只使用定性纯净的数据对其进行训练。这需要对数据进行优化。人类代谢组数据库和欧洲质量银行数据库利用质谱仪生成了 1755357 个峰对应的 85582 种不同物质的色谱图和质量光谱。每个光谱包含许多峰，而每个峰代表在质谱仪中被偏转的离子化样本片段（如图 2 所示）。峰值的高度是样品中该片段的相对强度，而峰值的位置则是该片段的质量与电荷比。为了确保我们选择的数据点具有高质量，我们仅提取其对应光谱中有超过 3 个峰且每个峰都超过了预设信噪比（SNR）阈值的化学物质。为了比较不同的光谱，我们将质量电荷值进行了离散化处理。虽然分箱会导致一些信息损失，但这对于这个问题至关重要，因为否则将不可能比较连续光谱，并且它标准化了整个数据库中的不同子数据集，从而考虑到了数据发布方式的不同。

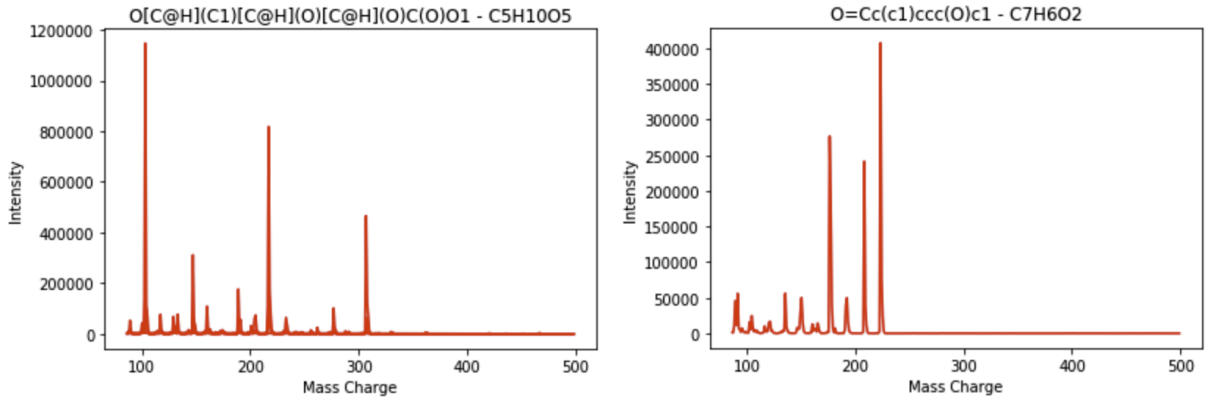

图 2: 化合物 $C_5H_{10}O_5$ 和 $C_7H_6O_2$ 的光谱可视化。数据来自欧洲质谱库和人类代谢组数据库，图形通过 Python 的 Matplotlib 图形引擎生成。

3.2 化学编码是开发连续模型的障碍

在包含数百个变量的数据集中寻找全面关系的标准方法是通过将数据拟合到神经网络上。我们的研究表明，由于其不可处理性：具体而言，相似的化学物质（结构相似的化学物质）具有不同的光谱；因此，不存在任

何可能的连续损失函数供神经网络尝试最小化开始，因此不能使用神经网络来预测质谱仪与其相应化学物质之间的关系。深入分析表明，为了使神经网络能够预测任何现有关系，它需要理解输出，并且这要求输出空间（化学品名称）遵循数值编码。可以使用多种形式的编码，所有这些都存在问题。首先，分子的名字通过 SHA-256 哈希函数被编码成一个名为 InChIKey 的 16 字符 ID。一种编码策略是使用独热编码，在这种编码中，由 0 组成的布尔列向量仅在特定位置有一个 1。然而，这会产生 $O(2^{2^{16n}}) \in O(2^{2^n})$ 的空间复杂度，对于神经网络来说太大了。另一种编码策略是通过为每个独特的化学物质分配一个自然数来使用整数编码。但是，整数编码假设相邻的数字对应于相似的化学物质，这是不正确的。最后，对化学品进行编码的唯一其他可能方法是直接通过图神经网络对化学品的结构进行编码。这不是有意义的解决方案，因为图神经网络层的输出会被作为输入提供给人工神经网络，这两个网络具有不兼容的架构层。

4 从零开始：追踪并摧毁假设

4.1 主成分分析（PCA）

我们首先尝试使用一个简单的模型来预测光谱关系，该模型采用流行的 PCA 方法 [Pearson, 1901] 进行已知假设。我们假设所有的光谱都是由一组正交基向量的线性组合形成的，每个基向量的实际解释是一个未知的原子或功能基团。这一假设是合理的，因为化合物可以被视为各种原子或功能基团的组合。因此，我们使用主成分分析（PCA）[Pearson, 1901]，这是一种降维算法，用于重新推导高维数据的基础向量。具体来说，我们将光谱传递给 PCA 算法，希望能够找到哪些基础向量能够充分表达化学光谱。

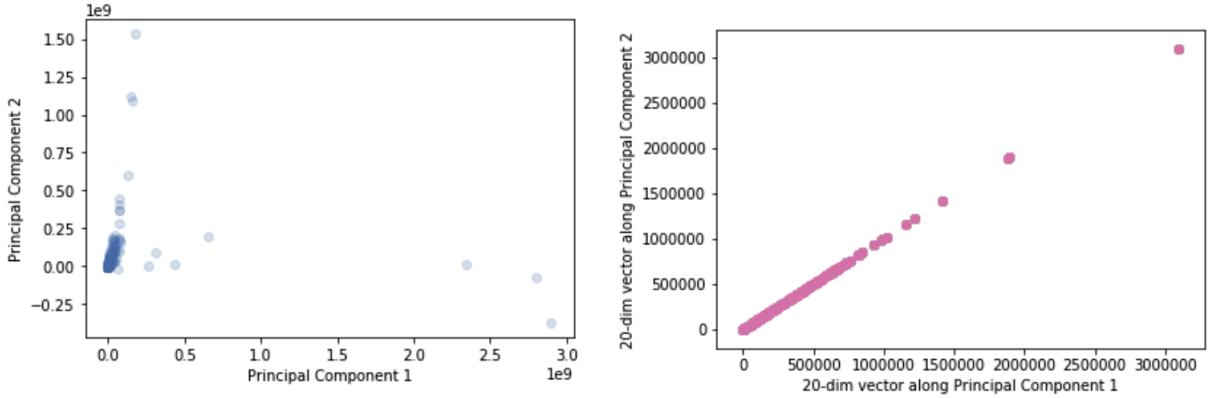

图 3： 20 维谱矩阵及其由 PCA 算法生成的基础向量的前两个维度分别表示。这些点是由 Python Scikit-learn Decomposition-PCA 库生成，并使用 Python Matplotlib 图形引擎进行散点图绘制。

然而，这种方法出现了两个问题：

1. 任何维度大于 1 的基向量不是唯一的。一个具体的例子是在 $\mathbb{R}^2$ 中，任何一组向量的基向量只是任意一对正交向量。因此，当我们试图将 2660 维分箱数据减少到合理的规模（小于 100）时，我们提取了正交归一化的基向量，其中没有一个基向量对应于任何光谱。如果我们知道将提取的基向量叠加成物理可实现解的旋转矩阵，并且我们不知道这一点，这仍然不能解决问题，因为 PCA 生成的每个基向量的系数非常相似；因此，重建的化合物与原始化合物相比太不相同而没有任何意义。因此，PCA 并不能为数据提供足够的分解。

2. 作为 (1) 的结果，算法生成的一些基向量的系数可以是负数。由于基向量对应于特定物质的存在，因此负系数对应于非物理上有效的负强度解。

4.2 非负矩阵分解（NMFA）

由于 PCA 无法分解数据，我们将假设降低一个等级，并不再假定光谱是由任何一组正交基向量的线性组合形成的，而是由稀疏正交基向量的线性组合形成。此外，为了使解决方案具有物理意义，我们假设所有系数都不是负数，因为所有的峰都必须是正值。因此，我们开发了一个非负矩阵分解（NMFA）模型 [Lee and Seung, 2000]，该模型将光谱矩阵 X 近似分解为一个包含 N 列权重的矩阵 W 和一个包含 N 行基向量的矩阵 H。我们将 NMFA 模型应用于光谱以识别权重和基向量，并通过实验优化 N，以便在保持数据足够强度的同时减少 N。为了优化 N，我们需要确定重构的光谱是否足够接近原始光谱。一般来说，验证学习到的分解能否重建原始光谱需要使用朴素矩阵乘积花费 $O(N^3)$ 时间，直到比特复杂性问题 Anand et al. [2024a]。这可以通过使用时间复杂度为 $O(N^\omega)$ 的快速矩阵乘法来改进，其中 $\omega < 2.371$ [Alman and Williams, 2024]。

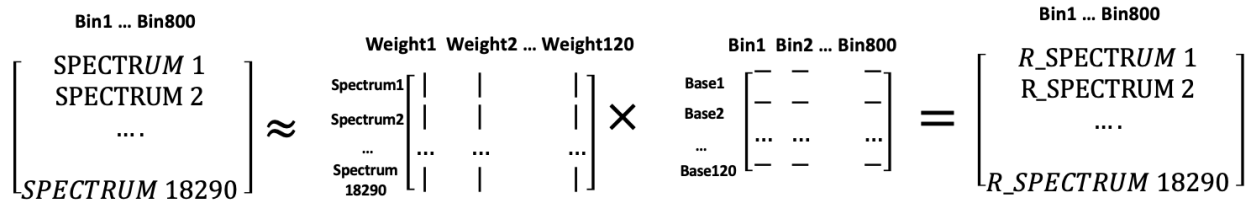

图 4：重建光谱的流程图

为了使用较少的主要成分重建原始矩阵，我们采用上述方法。需要注意的是，W 的第 i 列与 H 的第 i 行的乘积形成了第 i 谱的重建。利用这一过程，我们绘制了原始光谱和重构光谱之间的光谱矩阵差异绝对值之和（图 5）。

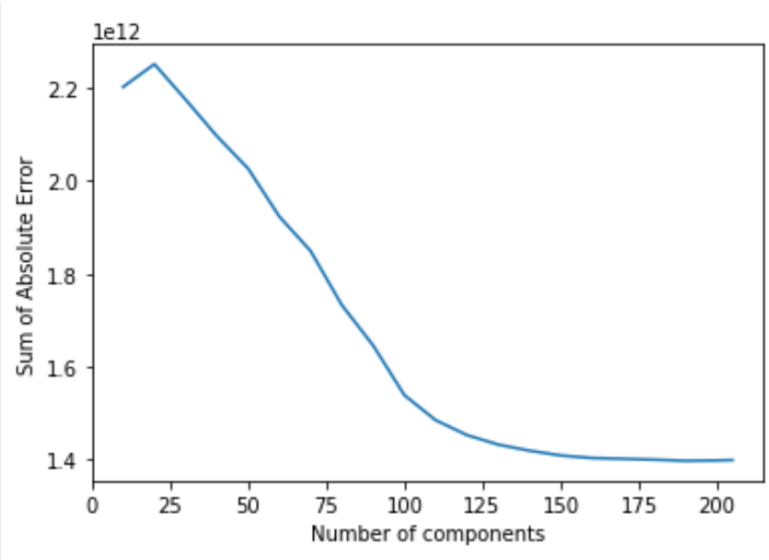

图 5： 原始光谱与重构光谱之间光谱矩阵差异的绝对值之和，绘制于变化的组分数下。

图 5 表明，在大约 125 个组件时，更多的组件不会显著降低 NMFA 误差。因此，我们选择了 120 个组件并运行了非负矩阵分解算法来生成 H 和 W 矩阵。然后我们绘制了各种物质的实际光谱和重构光谱，并观察到重构的准确性水平有所不同。图 6 展示了一个样本的相对高质量重构。在整个重构过程中，我们发现与实际光谱不一致的重构光谱在水平位置上略有不同，但在垂直尺度上有很大差异。

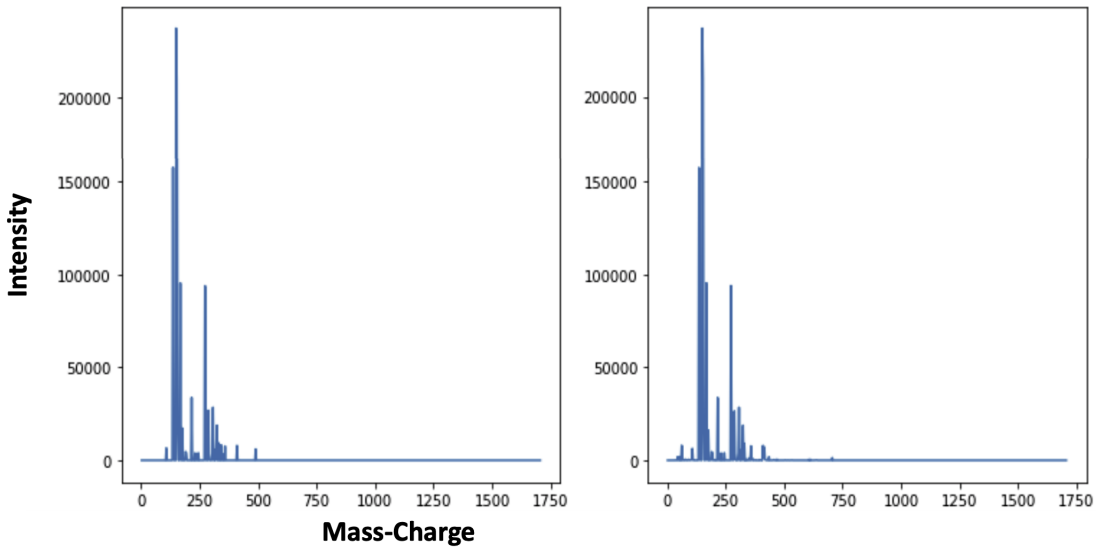

图 6: 实际光谱（左）和重构光谱（右）的 $C_{10}H_{16}N_5O_{13}P_3$。重构光谱使用了 Python 的 Scikit-learn NMFA 库，并用 Python 的 Matplotlib 图形引擎绘制。

我们然后可视化由非负矩阵分解生成的实际和重构的光谱矩阵（如图 7 所示），并注意到在高层次的观点下，重构中的差异可以忽略不计，这是所期望的结果。

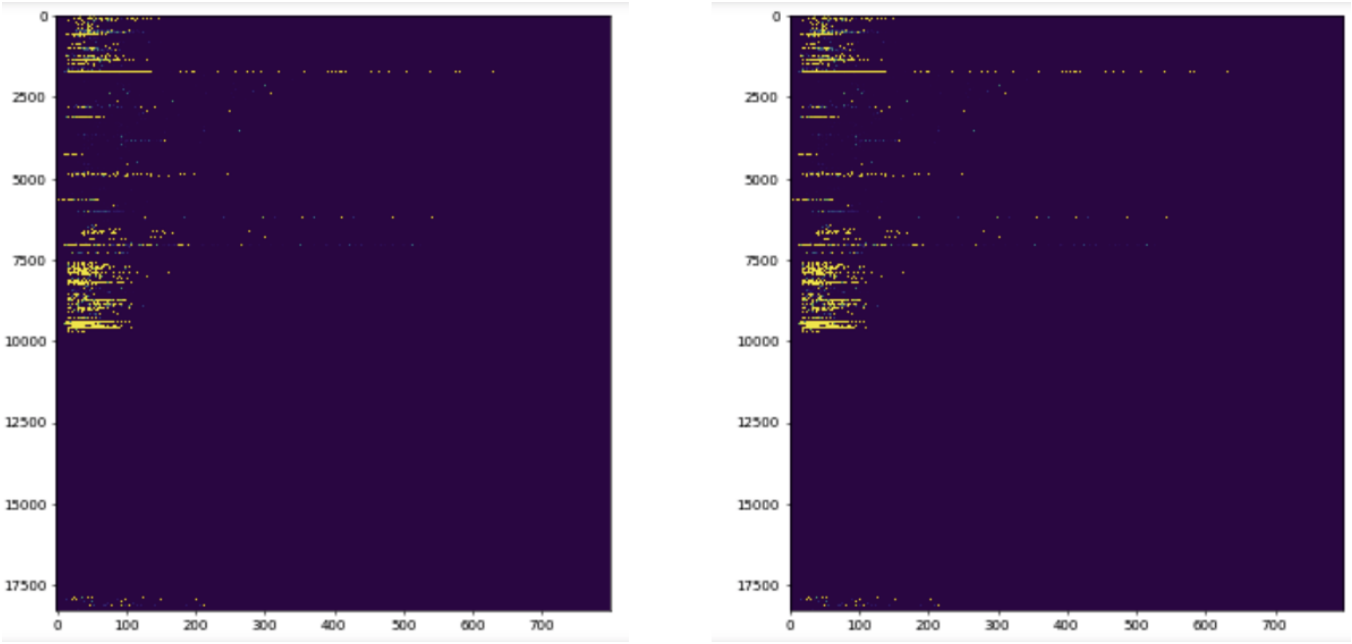

图 7: 实际的谱矩阵（左）和重构的谱矩阵（右）。可视化是使用 Python 的 Matplotlib 图形引擎和 Python 的 Imshow 功能生成的。

此外，我们在对数图上绘制了实际和重构的光谱矩阵（将每个图从相应的光谱矩阵的最低值缩放到最高值）。这样做后，重构中的差异变得更加明显。

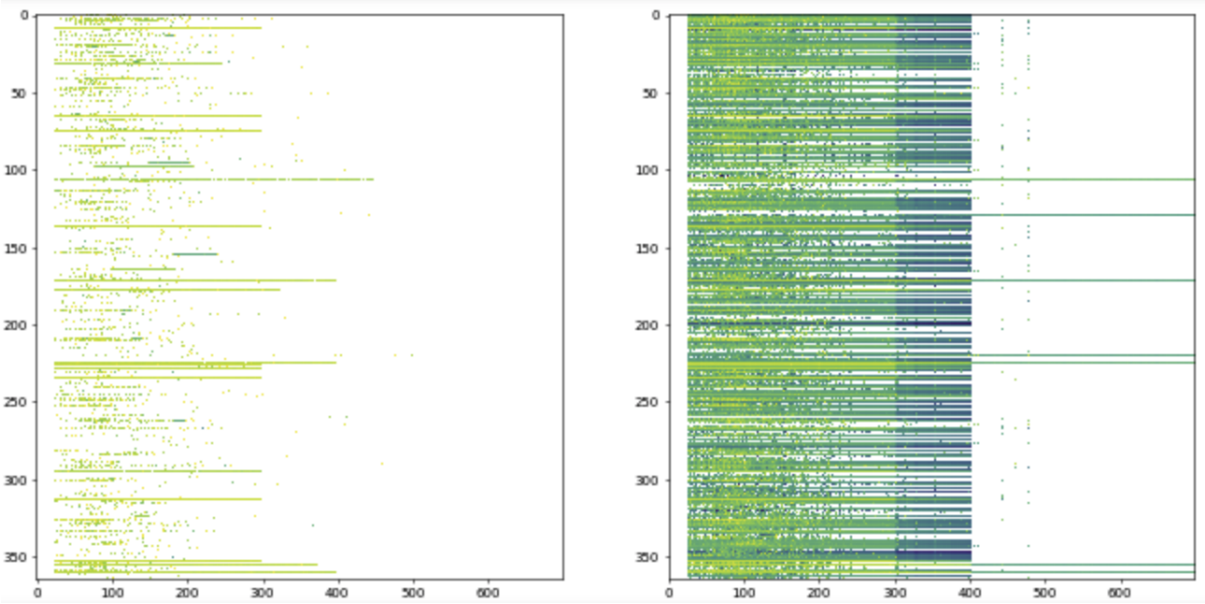

图 8: 实际的频谱矩阵（左）和重构的频谱矩阵（右）以对数尺度绘制。

5 提取 NMF 基并开发识别算法

每个基向量的理想物理解释是它代表某个“令牌”，其中每个令牌是一个单原子元素或一个作为所有化学化合物构建块的功能组。因此，如果 NMFA 分解正确地产生了一组令牌，它们将由基向量表示。在这种情况下，基础的光谱将对应于一个显著峰，该峰值代表这个较小的令牌。由于这是可证伪的，我们随机绘制了 2 个基向量（见图 9），并确认这些基向量确实包含了一个显著峰，这证实了与小型构建块结构相对应的关系。此外，我们还确认这种模式贯穿于所有基向量中。

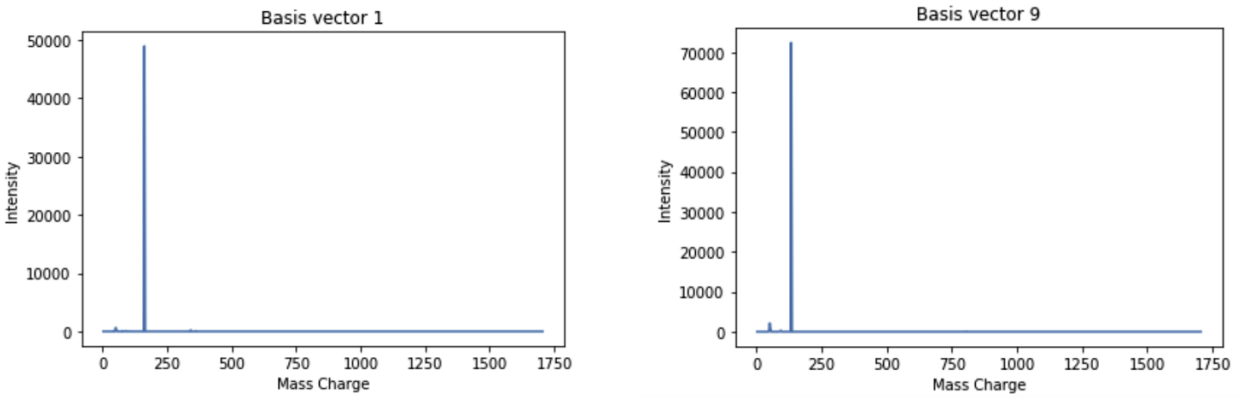

图 9: 从使用 120 个成分的 NMFA 分解产生的权重矩阵中提取的第一个（左侧）和第九个（右侧）基向量。

6 提取算法

之前我们展示了 NMFA 提取的基础向量与可能的标记一致，其中我们将标记定义为单原子元素或可能是更大化学化合物构建模块的小功能组。然而，如果不进行深入分析，我们不能天真地认为这是真实的。

因此，我们开发了以下算法（SEARCHER），该算法在化合物中搜索各种标记物以确定它们的质量电荷比。由于数据集中的大多数化合物是有机的且仅一次离子化，质量电荷比实际上与原子质量相同。因此，通过将 SEARCHER 算法输出的原子质量与已知标记物的原子质量进行比较，我们可以判断基本向量是否具有物理意义。反过来，这为检查基本向量是否确实代表真实的可解释标记物提供了更好的统计测试。

Algorithm 1 搜索者

Require: Token

对数据集中的所有化合物进行分词，并查找包含该标记的所有化合物。
查找每个包含该标记的化合物的重构光谱。
在每个包含该标记的化合物中，对每个 bin 中的强度值取平均。
找到包含最高强度的区间（质量-电荷）。
返回相应区间的质量-电荷。

7 SEARCHER 算法的结果

7.1 在氧气上运行 SEARCHER

当搜索算法运行在对应氧元素的'O' 标记上时，我们得到了令人惊讶的结果（图 10）。首先，算法收敛到了一个特定的基础向量。当检查这个基础向量时，它包含了一个主要位于质荷比 31.14505 处的单一峰值，这正是氧的原子质量。请注意，这一结果是在从未告知算法任何物质的原子质量的情况下获得的，因此可以得出结论，基础向量在表示标记时具有物理意义。

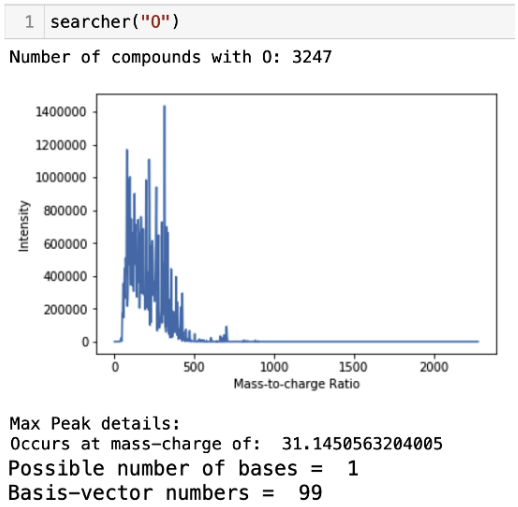

图 10： 搜索算法在对应氧元素的'O' 标记上的输出。这里最高的峰值出现在质荷比为 31.145 处，这是氧元素的原子质量，保留两位有效数字。

7.2 在氯上运行 SEARCHER

我们同样在氯代币上运行了该算法，并获得了 22.87 的原子质量（图 11），而平均氯离子的原子质量为 35.5 amu。为了说明这一点，我们注意到 22.87（精确到两位有效数字）与钠的原子质量相符，而且在这个数据集中，许多数据收集者将他们的化合物溶解在盐水溶液（高浓度的氯化钠溶液）中，在那里钠离子很可能被质谱仪捕捉到了。因此，我们认为该算法仍然正常工作，并且这是一个公平的结果。

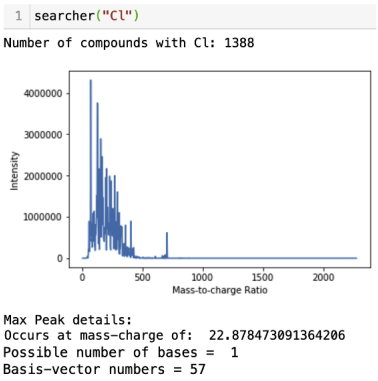

图 11：搜索算法在对应氯元素的'Cl' 标记上的输出。这里最高的峰值出现在 22.87，精确到两位有效数字。钠的原子质量。我们怀疑这是因为数据集中的所有溶液都溶解在盐水（高浓度的氯化钠）中，这干扰了读数。

7.3 在碳和氢上运行 SEARCHER

当在 Carbon 令牌（图 12-左）和 Hydrogen 令牌（图 12-右）上运行算法时，我们获得了完全相同的结果。特别是，算法将 Carbon 和 Hydrogen 令牌收敛到一个基向量，并且对应的质荷值为 31.42（精确到 4 位有效数字）。我们知道这是一个错误结果，因为碳的原子质量是 12.01（精确到 2 位有效数字）而氢的原子质量是 1.01（精确到 2 位有效数字），这实际上与在 Oxygen 令牌上运行 SEARCHER 算法时产生的原子质量相同。

我们再次怀疑该算法工作正常，我们将这些结果归因于我们正在处理的数据类型：数据集中约 85%的化合物是有机物（其中约 70%含有氧）。具体来说，这些含氧化合物主导了平均化合物的质量谱，而氧的质量谱又超过了碳和氢的质量谱，这些都是合理的选择作为标记。

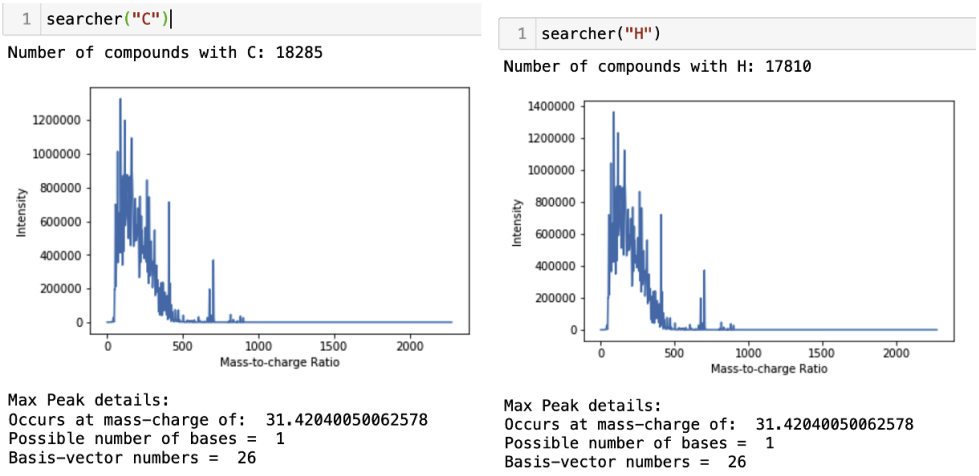

图 12：搜索算法对'C' 标记（左）对应碳和'H' 标记（右）对应氢的输出。请注意，这些光谱完全相同，并且也与氧的光谱一致。这是因为数据集中大多数化合物是有机物，包含碳、氢和氧原子。然而，真正的碳和氢基向量并未显现出来，因为氧的质量电荷比盖过了碳和氢的质量电荷比。

纠正这个问题（以及它的其他变体）并不简单，因为在朴素的方法中，在找到并确认每个物理标记正确之后，我们必须通过减去 c（该标记的光谱）、$c \in \mathbb{R}$ 来调整频谱矩阵，以最小化对应于该标记的频谱峰值。然而，找到 c 是一个难题，因为这些标记中的大多数的光谱包含随着 c 显著变化的小峰，并且由于不同的光谱

强度差异很大，这在确定正确精度（编码字节所需的位数）时造成了统计问题，即 c。

此外，一个存在的清晰计算界限是，在每次调整光谱矩阵并确定一个标记后，必须在修改后的矩阵上再次运行 NMF 算法，尽管执行 NMFA 的时间成本很高。然而，可能存在不依赖上述简单方法及其不经济时间复杂性的最优解法。

8 讨论

8.1 误差和噪声的潜在来源

该方法的主要噪声来源是数据。欧洲质量银行是由多个组织提供的大量小数据集的集合。这些组织的数据收集精度各不相同，而且它们发布数据的方式也不同。例如，图 13 展示了两种不同化学物质的光谱，每种均由不同的组织发布。第二种化学物质（图 13 - 右上）每个峰传达的信息明显多于第一种化学物质（图 13 - 左上）。然而，通过分箱（将数据的质量电荷区域离散化，并让每个箱值成为该区域内找到的最大强度），我们可以消除标准化的缺乏问题（如图 13 所示的更可比较的图表中下层所见）。尽管这确实使数据足够标准化并允许我们执行比较不同光谱的基本任务，但分箱确实引入了信息损失，并且没有简单的方法来克服这个问题。

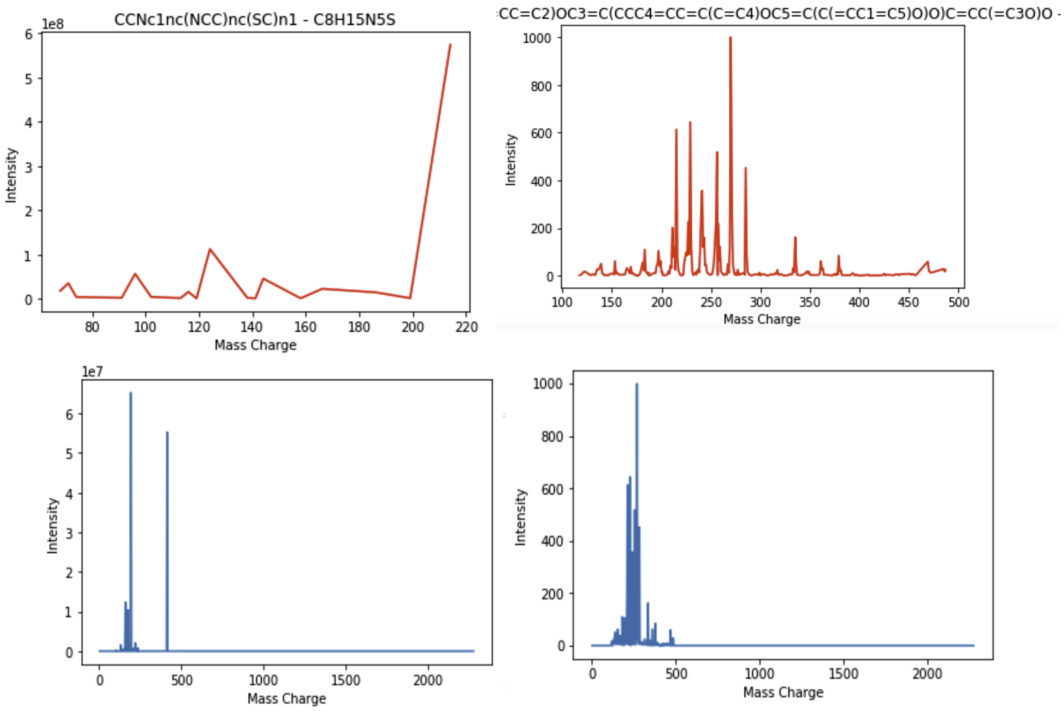

图 13: 非标准化数据（顶部）。分箱数据（底部）。显然，分箱允许光谱之间的比较。质量-电荷单位是 kg/C，强度单位是任意的（这是一个衡量不同碎片与检测器碰撞力度的标准，并且这个衡量标准依赖于检测器的灵敏度），因此只有相对数量才是重要的。

数据中的一个潜在误差来源在于问题的化学性质。当质谱仪使用样本时，它会将样本汽化并用电子轰击以使其以各种方式碎裂。这些碎片代表了质谱图中的峰。质量-电荷比通过测量片段在质谱仪中偏转的程度来对应于该片段的质量。它还测量片段撞击检测器的力量，从而为该片段分配一个强度值。质谱仪内部的碎裂并不一定是唯一的，因为大化合物可以以多种方式碎裂。因此，数据收集组重复此过程多次以获得所有可能的碎片/峰，这引入了由于随机误差/不确定性 Djoumbou-Feunang et al. [2019], Xu et al. [2007] 而丢失碎片和在不同质量-电荷值下重复碎片的可能性。

8.2 非负矩阵分解精度

非负矩阵分解（NMF）算法在数学上等同于一个聚类算法，其目标是最小化 $||X - WH||, H \geq 0, W \geq 0$。然而，当该算法应用于我们的谱矩阵 X 时，在使用了适当数量的主成分（基向量）的情况下，无法将 $||X - WH||$ 最小化为 0。请注意，“合适”这个词在这里带有误导性含义。当主成分的数量等于 85582（数据集中化合物的数量，也是谱矩阵中的行数），最小化可以完美地进行，因为每个光谱都被分配到一个基向量上。然而，这样做我们并没有学到任何新的知识，并且无法提取关于谱矩阵及其构成部分的底层信息。因此，降维的目标是在尽可能减少主成分数量的同时保留尽可能多的准确性。因此，当我们选择 125 个主成分时，我们故意在模型中引入了一些噪声。但是，我们也通过建议手动识别这 125 个基向量的标记过程来对抗这个问题，以确保结果仍然具有物理意义。这种模型有可能比任何离散模型表现得更好，而我们获得的结果就是这一潜力的证明。下一节详细介绍了该项目未来的工作内容，以及需要解决的问题以便识别基向量。

9 未来工作

为进一步解决此问题，未来的工作可以集中在理解如何找到残差基向量。这样做可以让研究人员识别出非主导的基向量作为物理标记，从而避免氧气光谱占据碳和氢的光谱的问题（以及其他类似情况）。此外，众所周知，各种原子或功能团的存在对整个分子的特性有很大影响。这种效应是非线性的，因为这样一个标记的存在会改变整个分子的光谱景观。这不能通过诸如 PCA 或 NMFA 这样的线性分解模型准确建模，可能需要新的分解方法，可能是通过对 Tucker 分解过程进行适当的修改。一旦得到基向量，未来的工作可以集中在从已知的“标记”中重建所有可能对应化合物上。最后，另一种方法是考虑多智能体 RL 方法来模拟各种分子 [Anand and Qu, 2024, Anand et al., 2024b] 之间的相互作用。

10 致谢

我们感谢以下机构的支持：

1. 加州理工学院暑期研究计划和宇宙黎明中心——尼尔斯 · 玻尔研究所
2. 马丁 · 汉森教授 - 丹麦奥胡斯大学
3. 朱晓敏 - 丹麦奥胡斯大学
4. 瓦迪姆 · 鲁萨科夫 - 宇宙黎明中心，尼尔斯 · 玻尔研究所，哥本哈根大学，丹麦
5. 罗伯特 · 菲利普斯教授 - 美国加州理工学院
6. 人类代谢组数据库和欧洲质谱库光谱库

参考文献

Lutz Alder, Kerstin Greulich, Günther Kempe, and Bärbel Vieth. Residue analysis of 500 high priority pesticides: Better by gc–ms or lc–ms/ms? *Mass Spectrometry Reviews*, 25(6):838–865, 2006. doi:10.1002/mas.20091. URL https://doi.org/10.1002/mas.20091.

Josh Alman and Virginia Vassilevska Williams. A refined laser method and faster matrix multiplication. *TheoretiCS*, 3, 2024.

Emile Anand and Guannan Qu. Efficient reinforcement learning for global decision making in the presence of local agents at scale. *arXiv preprint arXiv:2403.00222*, 2024. URL https://arxiv.org/abs/2403.00222.

Emile Anand, Jan van den Brand, Mehrdad Ghadiri, and Daniel Zhang. The bit complexity of dynamic algebraic formulas and their determinants. *arXiv preprint arXiv:2401.11127*, 2024a.

Emile Anand, Ishani Karmarkar, and Guannan Qu. Mean-field sampling for cooperative multi-agent reinforcement learning. *ArXiv*, abs/2412.00661, 2024b. URL https://arxiv.org/abs/2412.00661.

Tomas Cajka and Oliver Fiehn. Toward merging untargeted and targeted methods in mass spectrometry-based metabolomics and lipidomics. *Analytical Chemistry*, 88(1):524–545, 2016. doi:10.1021/acs.analchem.5b04491. URL https://doi.org/10.1021/acs.analchem.5b04491.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016. doi:10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

Yannick Djoumbou-Feunang, Jarlei Fiamoncini, Alberto Gil de-la Fuente, Russell Greiner, Claudine Manach, and David S. Wishart. Biotransformer: A comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *Journal of Cheminformatics*, 11(1):2, 2019. doi:10.1186/s13321-018-0324-5. URL https://jcheminf.biomedcentral.com/articles/10.1186/s13321-018-0324-5.

Srijan Gupta and Emile Timothy Anand. Cryogenic high-q mechanical resonator. *LIGO Voyager Document Control Center*, 2020.

Daniel Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2000.

Jaanus Liigand, Tingting Wang, Joshua Kellogg, Jørn Smedsgaard, Nadja Cech, and Anneli Kruve. Quantification for non-targeted lc/ms screening without standard substances. *Nature Research Scientific Reports*, 2020.

Yiheng Lin, James A Preiss, Emile Timothy Anand, Yingying Li, Yisong Yue, and Adam Wierman. Online adaptive policy selection in time-varying systems: No-regret via contractive perturbations. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=hDajsofjRM.

Yiheng Lin, James A Preiss, Emile Timothy Anand, Yingying Li, Yisong Yue, and Adam Wierman. Learning-augmented control via online adaptive policy selection: No regret via contractive perturbations. In *ACM SIGMETRICS, Workshop on Learning-augmented Algorithms: Theory and Applications 2023*, 2023b.

Yiheng Lin, James A. Preiss, Fengze Xie, Emile Anand, Soon-Jo Chung, Yisong Yue, and Adam Wierman. Online policy optimization in unknown nonlinear systems. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 3475–3522. PMLR, 30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/lin24a.html.

Herbert Oberacher, Michael Sasse, Jean-Philippe Antignac, Yann Guitton, Laurent Debrauwer, Emilien L. Jamin, Tobias Schulze, Martin Krauss, Adrian Covaci, Noelia Caballero-Casero, Kathleen Rousseau, Annelaure Damont, François Fenaille, Marja Lamoree, and Emma L Schymanski. A European proposal for quality control and quality assurance

of tandem mass spectral libraries. *Environmental sciences : an international journal of environmental physiology and toxicology*, 32(1):43, 2020. doi:10.1186/s12302-020-00314-9. URL https://hal.science/hal-01410982.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9: 2579–2605, 2008. URL http://www.jmlr.org/papers/v9/vandermaaten08a.html.

Philip D Waggoner. *Modern dimension reduction*. Cambridge University Press, 2021.

Raymond Naxing Xu, Leimin Fan, Matthew J. Rieser, and Tawakol A. El-Shourbagy. Recent advances in high-throughput quantitative bioanalysis by lc – ms/ms. *Journal of Pharmaceutical and Biomedical Analysis*, 44(2): 342–355, 2007. doi:10.1016/j.jpba.2007.02.006. URL https://doi.org/10.1016/j.jpba.2007.02.006.