DOI:10.3969/j.issn.1005-202X.2025.12.017

医学生物信息

SpecEmbedding:一种面向化合物鉴定的深度学习嵌入方法

熊鹏,郑浩然
中国科学技术大学计算机科学与技术学院,安徽 合肥 230027

【摘要】为应对质谱图在化合物结构多样性和实验环境差异下所表现出的异质性问题,提出一种提升质谱图间可比性的表征方法。该方法命名为SpecEmbedding,融合正弦嵌入与监督对比学习策略,旨在将高维、复杂的质谱图转化为低维向量表示。在GNPS公共数据集上对该方法进行训练与评估,并将其与主流方法进行对比。实验结果显示,SpecEmbedding在测试集上Top-1命中率指标上达到84.38%,相较目前最优方法CLERMS提高6.3%。该方法能显著增强质谱图间的可比性,有效提升化合物鉴定任务中的准确性与鲁棒性。

【关键词】质谱;化合物鉴定;表征学习;对比学习
【中图分类号】R318 【文献标志码】A 【文章编号】1005-202X(2025)12-1660-08

SpecEmbedding: a deep learning based embedding approach for compound identification

XIONG Peng, ZHENG Haoran
School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China

Abstract: To address the heterogeneity of mass spectra caused by the structural diversity of compounds and variations in experimental conditions, a novel representation method named SpecEmbedding is proposed to enhance the comparability between mass spectra. SpecEmbedding integrates sinusoidal embedding and supervised contrastive learning strategy, aiming to transform high-dimensional and complex mass spectra into low-dimensional vector representations. This approach is trained and evaluated on the public GNPS dataset, with comparion performed against mainstream methods. Experimental results show that SpecEmbedding achieves a Top-1 hit rate of 84.38% on the test set, representing a 6.3% improvement over CLERMS, the current state-of-the-art method. These findings demonstrate that SpecEmbedding significantly improves the comparability between mass spectra while effectively enhancing accuracy and robustness of compound identification tasks.

Keywords: mass spectrum; compound identification; representation learning; contrastive learning

前 言

串联质谱技术为化合物鉴定提供重要的信息资源[1-3]。质谱分析过程通常包括以下几个基本步骤:目标分子通过电离过程生成母离子,并通过信号采集获取一级质谱图(First-stage Mass Spectrometry, MS1);随后,根据设定的策略选择母离子进行碎裂,形成一系列子离子,通过信号采集得到二级质谱图(MS2)。谱图中的质荷比(m/z)与强度共同构成质谱峰,后续基于MS2谱图进行分析。在化合物鉴定过程中,将实验谱图与已知化合物谱图库中的参考谱图进行比对,通常通过计算相似度来判定化合物的身份。相似度较高的匹配结果通常被认为是该化合物的鉴定结果[4-6]。由于实验条件的差异以及化合物本身空间结构的复杂性,同一化合物在不同实验环境中可能会产生具有显著差异的质谱图。例如,化合物的异构体或不同的离子化模式可能导致谱图峰的位置、强度分布等特征不同,这些影响因素最终给鉴定带来挑战。余弦相似度方法是常用的谱图相似度计算方法[6-7]。该方法基于两张谱图之间能够匹配成功的子离子的强度值,计算两个谱图的余弦值,从而量化它们之间的相似度。空间结构导致的峰位置变化使得同一化合物产生的部分谱图之间只有少数匹配成功的子离子,导致相似度评分较低,这使得简

【收稿日期】2025-04-12
【基金项目】中国科学院战略性先导科技专项(XDB38000000)
【作者简介】熊鹏,硕士研究生,研究方向:生物信息学,E-mail: pengx@mail.ustc.edu.cn
【通信作者】郑浩然,副教授,硕士生导师,研究方向:生物信息学,E-mail: hrzheng@ustc.edu.cn

单的余弦相似度计算难以有效匹配这些不同的谱图，进而影响最终的鉴定时的准确性[8]。尽管后续也有一些变体方法对余弦进行改进，如 Modified Cosine[9]、Neutral Loss[10-11]和 Binned Cosine[12]等。由于速度上的限制以及对高质量数据的高度依赖，难以应对由于实验条件差异或化合物异构体引起的局部差异，研究者们开始探索更高效、鲁棒性更强的替代方法。

近年来，新兴的深度学习方法在化合物鉴定和谱聚类领域展现巨大的潜力，这些方法能够有效地捕捉化合物和谱图之间的复杂关系，从而提高化合物鉴定的准确性[13]。随着深度学习技术的快速发展，谱图表征方法因其出色的表现应用广泛，它们通过将谱图转化为低维向量，从而将谱图相似度问题转化为向量相似度问题。这些方法通过学习谱图的潜在特征，能够更好应对谱图的复杂性和多样性。Spec2Vec[14]基于 Word2Vec[15]模型将质谱图的 m/z 值转换为向量，但由于对原始质荷比信息进行截断操作，导致部分信息缺失。MS2DeepScore[16]采用孪生网络（Siamese Network）来预测化合物间的结构相似度，但同样依赖峰箱法，导致谱图信息可能丢失，尤其是在高分辨率数据中。GLEAMS[17]通过卷积神经网络结合肽段元数据提升蛋白质组学中谱图的鉴定能力，但同样面临稀疏性和分辨率问题。CLERMS[18]则采用无监督对比学习方法，通过利用 InfoNCE[19]损失函数，结合化合物的元数据和谱图数据进行嵌入。CLERMS 虽然有效减少了对标注数据的依赖，但未能充分学习到同类样本之间的内在联系，从而限制了泛化性能。尽管这些方法在化合物鉴定和谱图分析中取得显著进展，但它们仍然存在一些不足之处，主要体现在以下几个方面：（1）稀疏性和分辨率问题，峰箱法无法充分利用高分辨率质谱数据；（2）多数模型仅关注单个谱图的特征，忽略了不同实验条件下的谱图之间的内在关联，导致学习的泛化能力不足；（3）当前的模型更多关注不同化合物之间的全局相似度，而忽略了同类化合物之间微小的差异，而这些细微差异对化合物鉴定的精度至关重要。

针对上述问题，本文提出一种结合正弦嵌入以及监督对比学习的深度学习嵌入方法。首先，利用正弦嵌入的方式将质荷比信息转化为序列向量，能够保证充分利用现代质谱仪高分辨率的优势。除此之外，还特别考虑不同谱图之间的内在关系。与传统的谱图匹配方法相比，本文方法通过引入对比学习的核心思想，在特征空间中实现同类样本的紧密聚集与异类样本的有效分离[20-22]。具体而言，监督对比学习通过在嵌入空间中将属于同一类别的样本拉近，将不同类别的样本推远，从而确保同一化合物的不同谱图在特征空间中彼此靠近，而不同化合物的谱图则被有效地分隔开。为了实现这一目标，在训练过程中采用监督对比损失函数（Supervised Contrastive Loss, SCL）。该损失函数能够有效地推动模型学习过程中的正负样本对关系。在标准对比学习框架中，模型通常只考虑正样本与负样本之间的相对距离，而本文方法进一步引入类别层次信息，使得同一化合物的多个谱图之间的相似性得以加强，同时增强模型对不同化合物谱图的辨识能力。这一机制显著提高化合物鉴定任务中的准确性，尤其在面对多样化和大规模的质谱数据时，能够更有效地保持谱图的区分度。

1　资料与方法

1.1　数据集

本文使用公共数据库 GNPS（Global Natural Products Social Molecular Networking）[23]中的数据集来训练模型参数并评估模型性能。GNPS 数据集包含了广泛来源的化合物质谱数据，具有较高的多样性和代表性，非常适合用于化合物鉴定和质谱谱图分析的研究。在数据预处理阶段，采用 matchms[24]库的默认筛选流程以及 MS2DeepScore 中的过滤方法，去除峰数量较少和通过负电离方式得到的低质量谱图。这一处理步骤旨在去除噪声数据，确保模型训练时使用的谱图具有较高的质量，从而提高模型的泛化能力和鉴定精度。通过这些筛选步骤，最终得到了来自 23 372 种化合物的 267 702 张高质量谱图，这些谱图覆盖了化合物的多种类别和实验条件，具有较好的代表性。每个化合物都通过其 InChIKey 的前 14 位字符进行标识，InChIKey 是一种标准化的化学物质标识符，其前 14 位字符仅表示化合物的二维平面结构，而不考虑立体异构问题。后续认定如果两个化合物的 InChIKey 前 14 位字符相同则认定是同一个化合物。为了进行模型的评估和训练，从所有化合物中随机抽取 20% 的数据作为测试集，其余的数据作为训练集。为了更贴近实际化合物鉴定场景，在数据划分中进一步将样本分为查询集（Query set）与参考集（Reference set），模拟实验谱图与库中参考谱图进行比对检索的过程。具体而言，对于每个化合物，如果其对应有多张谱图，随机选取一张作为查询谱图，其余作为参考谱图；若某化合物仅有一张谱图，则将其直接划入参考集。该划分方式不仅适用于测试集，也同样应用于训练集。最终数据被划分为 4 个部分：训练-参考集（18 686 个化合物，201 254

张谱图）、训练-查询集（12 571个化合物，12 571张谱图）、测试-参考集（4 676个化合物，50 702张谱图）以及测试-查询集（3 175个化合物，3 175张谱图）。测试-参考集和测试-查询集则用于评估模型的泛化能力，测试数据中的查询集和参考集的划分方式能模拟真实应用中的实际场景，即如何根据已知谱图信息预测未知谱图的化合物。

1.2 性能评估方法

将来自同一化合物的谱图简称为“同一类”，而来自不同化合物的谱图则简称为“不同类”。在模型将谱图嵌入为低维向量后，使用余弦相似度来计算两张谱图之间的相似度。对于包含M张谱图的查询集和N张谱图的参考集，可以计算得到$M\times N$的相似度分数矩阵，并使用Top-k的方式计算命中率和召回率。命中率（Hit@k）：对于查询集中的每一个谱图，如果在分数排名前k的谱图中，存在与该谱图属于同一类的谱图，则认为这是一个成功的命中。命中率定义为成功命中的元素数量除以查询集中总的元素数量。简单来说，命中率衡量了查询谱图在前k个匹配中是否能够找到一个来自同一化合物的谱图。召回率（Recall@k）：对于查询集中的每一个谱图，统计分数排名前k的谱图中与该查询谱图属于同一类的谱图数量，然后除以参考集中与该查询谱图属于同一类的谱图总数。召回率反映参考集中与查询谱图同类的谱图被成功找到的比例。这种评估方式能够兼顾模型在匹配正确类谱图方面的准确性和在不同谱图之间检索到目标谱图的能力，是评价谱图嵌入模型性能的重要标准。具体的计算公式如下所示：

$$\begin{aligned}&\mathrm{Hit@}k=\frac{\sum_{i=1}^{M}I\left(L_i\in C_i\right)}{M}\\&I\left(L_i\in C_i\right)=\begin{cases}0, & \text{if } L_i\notin C_i\\1, & \text{if } L_i\in C_i\end{cases}\\&\mathrm{Recall}_i=\frac{\sum_{l\in C_i}\varphi\left(l,L_i\right)}{T_i}\\&\varphi\left(l,L_i\right)=\begin{cases}0, & \text{if } l\neq L_i\\1, & \text{if } l=L_i\end{cases}\\&\mathrm{Recall@}k=\frac{1}{M}\sum_{i=1}^{M}\mathrm{Recall}_i\end{aligned}\tag{1}$$

其中，L_i表示查询集中第i个元素的真实类别标签，C_i表示查询集中第i个元素对应的前k个最高分参考集元素的类别集合，T_i表示参考集中与查询集中第i个元素同属一类的所有元素的数量。

1.3 模型

模型的结构及其嵌入流程如图1所示。该模型主要由3部分组成：质荷比和峰嵌入模块、Transformer Encoder模块以及最终的由单层感知机构成的维度转换模块。m/z的嵌入采用正弦嵌入方式，思路源于Transformer[25]中采用的位置编码，在语言翻译任务中用于标识不同的词在句子中所处的位置。这种方式已经在自然语言处理领域[26]和蛋白从头测序方向[27-29]上被广泛应用。具体来说，初始时，谱图首先被拆分为m/z和intensity两部分，峰强度使用最大值进行归一化处理，为了保持维度一致且与谱峰区别开，受DreaMS[30]方法的启发，在强度向量母离子m/z位置处填充了一个常数值2。m/z通过该模块后，采用该策略初步嵌入之后，每个质荷比都投影成一个d维的向量，其中向量的偶数部分由式（2）计算，奇数部分由式（3）计算：

$$\mathrm{SE}\left(\mathrm{m/z},2i;d\right)=\sin\left(2\pi\left[\lambda_{\min}\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{2i/(d-2)}\right]^{-1}\mathrm{m/z}\right)\tag{2}$$

$$\mathrm{SE}\left(\mathrm{m/z},2i+1;d\right)=\cos\left(2\pi\left[\lambda_{\min}\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{2i/(d-2)}\right]^{-1}\mathrm{m/z}\right)\tag{3}$$

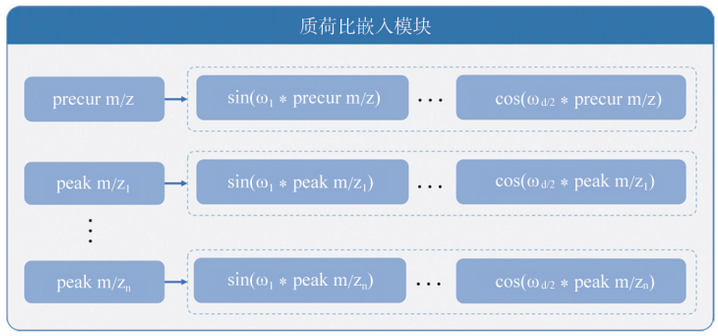
其中，SE表示质荷比的正弦嵌入，其中$\lambda_{\min}$表示质荷比的分辨率下限，$\lambda_{\max}$表示所考虑的质荷比最大范围。

通过设定$\lambda_{\min}$和$\lambda_{\max}$，模型能够适应质荷比的物理尺度和检测精度，保证后续的Transformer Encoder模块可以充分提取高分辨率的质荷比信息而不丢失。在质荷比经过正弦嵌入后，这些嵌入向量与峰强度信息一起输入到峰嵌入模块。在峰嵌入模块中，首先质荷比的嵌入向量通过一层前馈神经网络（Feed-Forward Network, FFN），接着在每个嵌入向量的末尾连接峰强度向量，再次通过一层前馈神经网络处理，最终得到谱图的嵌入表示，具体的公式如下：

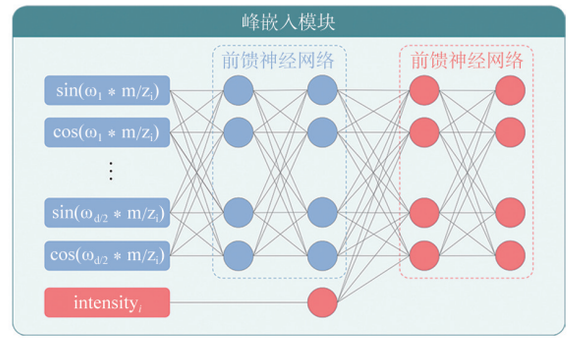
$$\begin{aligned}&\mathrm{PE}\left(\mathrm{m/z},I\right)=\mathrm{FFN}\left(\mathrm{FFN}\left(\mathrm{SE}\left(\mathrm{m/z}\right)\right)\|I\right)\\&\mathrm{FFN}\left(x\right)=\mathrm{SELU}\left(x\boldsymbol{W}_1+b_1\right)\boldsymbol{W}_2+b_2\end{aligned}\tag{4}$$

其中，||表示连接符号，SE表示质荷比的嵌入，SELU表示SELU激活函数。

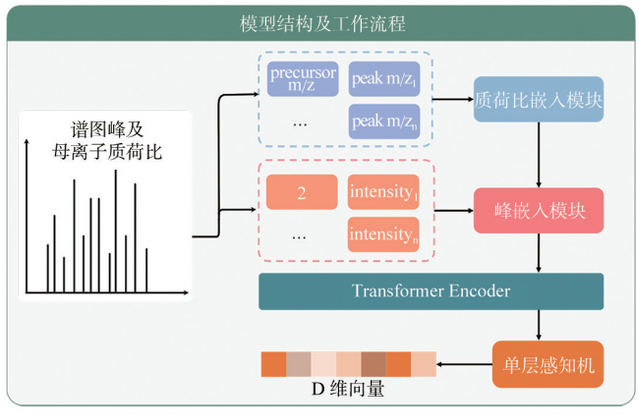
初始的谱图经过质荷比和峰嵌入后，得到处理后的序列向量，接着会被输入到Transformer Encoder模块中。Transformer Encoder模块通过自注意力机制自行捕捉谱图中峰之间的复杂关系，并生成得到新的序列向量。这些序列向量随后会经过平均池化和一个多层感知机模块进行维度转换，最终得到模型D维的嵌入向量，作为化合物鉴定的基础。最后计算该向量之间的余弦相似度得到实验谱图和多个参考谱图之间的相似度。



a：质荷比嵌入模块示意图



b：峰嵌入模块示意图



c：模型结构及工作流程

图1 SpecEmbedding 模型结构及关键模块示意图

Figure 1 Architecture of SpecEmbedding model and illustration of key modules

1.4 损失函数

SCL是一种用于学习样本表示的方法，它通过最大化同类样本之间的相似度，同时最大化异类样本之间的距离，来学习更具判别性的特征表示[22]。SCL在图像识别、自然语言处理等多个领域中得到广泛应用。具体而言，目标是将同类样本聚集在一起，而将异类样本分开。实际训练时，在训练过程中，对于每一个batch的数据，首先从原始数据中随机采样N个样本对，记为$\{x_k,y_k\}_{k=1,2,\cdots,N}$，其中$y_k$是$x_k$的标签。然后，根据$y_k$标签，随机从样本中选择一个相同标签的样本，在训练过程中，基于标签信息，随机从具有相同类别标签的样本中选取一例作为辅助样本，用于构造正样本对。该策略不同于传统的数据增强，而是通过利用类别标签实现样本的类别保持与多样化，最终得到$2N$个数据样本$\{\tilde{x}_k,\tilde{y}_k\}_{k=1,2,\cdots,2N}$，损失函数如式(5)所示：

$$L_{\text{out}}^{\text{sup}}=\sum_{i\in I}\frac{-1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp\left(z_i\cdot z_p/\tau\right)}{\sum_{a\in A(i)}\exp\left(z_i\cdot z_a/\tau\right)}\tag{5}$$

其中，$P(i)$是和i相同标签的数据集合(除i本身)，$A(i)$是除i本身外其他所有的数据集合即负样本集合。z_i表示经过模型嵌入所得的向量。τ是温度参数，用来调整分布的平滑程度。exp表示指数运算，$\cdot$表示向量的点积。

与常规的数据增强方法不同，本文采用一种基于标签的增强策略。即通过标签y_k从同类样本中抽取增强样本，这有助于在训练过程中保持样本的类别信息，并加快模型学习更具判别性的嵌入表示。通过这种策略，监督对比损失能够促进模型在化合物鉴定任务中更精确地捕捉同类谱图的特征，并有效区分不同化合物的谱图。通过利用监督对比损失函数，模型能够有效地最小化同类样本在嵌入空间中的距离，同时最大化异类样本的距离。这种方法在化合物鉴定任务中具有显著优势，因为它能够确保同一化合物产生的谱图在嵌入空间中紧密聚集，而不同化合物的谱图则被有效地分隔开，从而提高鉴定的准确性。

1.5 模型训练

本文的模型使用PyTorch框架进行搭建，在两张RTX 4090显卡上进行训练，训练总时长约为1 h。为了确保实验结果的可复现性，在训练过程中固定随机种子为42。优化器使用AdamW[31]，初始学习率为1×10^{-5}，衰减值为0.1，其余采用默认参数。总训练轮数为100，批次大小为256，训练过程中采用学习率预热策略，在前10%的训练步内，学习率从0线性增加到初始学习率，在接下来的90%的训练步中采用余弦衰减策略将学习率逐渐降低至0，这样做可以避免

训练初期存在的模型震荡问题，有利于模型收敛，还可以缓解优化器在早期阶段对动量估计不准确的问题，增强稳定性。除此之外，还采用梯度裁剪策略，裁剪值设为0.5，确保每次参数更新时梯度值不超过该值。在模型训练过程中，从训练数据中随机抽取30%数据作为验证数据，进行模型选择，保存验证集上损失值最小的模型。

为了筛选出最优的模型超参数组合，采用超参数调优框架Optuna[32]，它是基于贝叶斯优化和遗传算法采样策略，能基于指定的评估指标进行自动搜索并选取最优的超参数组合。具体而言，为每个待调节的超参数设定合理的搜索空间，Optuna利用自动剪枝算法在空间内进行高效搜索，设定搜索次数上限为100。在每次超参数搜索过程中，使用当前超参数组合对模型进行训练。训练完成后，计算训练-查询集与训练-参考集的Top-1命中率，并以此作为评估指标，选取命中率最高的超参数组合作为最终配置。各超参数搜索空间及对应的最终取值如表1所示，其中训练集上的最佳Top-1命中率为87.3%。

表1 模型超参数

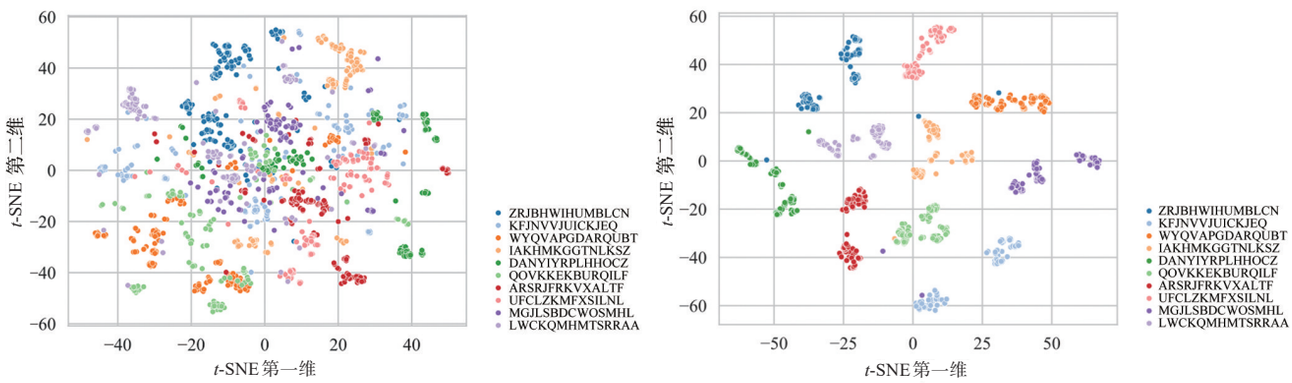
Table 1 Model hyperparameters

参数	搜索空间	最终取值
学习率lr	10^{-6}~10^{-3}	1×10^{-5}
批次大小batch size	[32, 64, 128, 256, 512]	256
嵌入维度D	[100, 200, 300, 400, 500]	300
温度系数τ	[0.005, 0.01, 0.05, 0.1]	0.05
$(\lambda_{min}, \lambda_{max})$	[(10^{-3}, 10^{3}), (10^{-4}, 10^{4})]	(10^{-3}, 10^{3})

2 结 果

2.1 嵌入向量可视化

在模型训练完成后，每个谱图被转换为一个300维的嵌入向量，而不是传统的通过峰箱法生成的高维向量。为了更直观地展示模型的效果，选择10种不同化合物生成的谱图进行可视化。为了更清晰地展示这些高维嵌入向量之间的关系，使用*t*-SNE降维方法[33]，将300维的嵌入向量降至2维，并通过散点图呈现结果。图2展示了通过*t*-SNE降维后的可视化效果。图2a是使用传统峰箱法生成的1 000维向量降维后的结果，而图2b是通过SpecEmbedding方法得到的300维嵌入向量的降维结果。从图中可以清晰看到，两者在特征空间中的分布情况有显著差异。使用峰箱法生成的高维向量经过降维后，谱图的分布较为分散，同一化合物的不同谱图样本之间没有明显的聚集趋势，且不同化合物的谱图也没有明显的区分。通过SpecEmbedding处理得到的300维嵌入向量降维后的结果显示，经过嵌入后，同类谱图的样本在特征空间中紧密聚集，而异类谱图的样本则被明显区分开。这表明SpecEmbedding能够有效地将同一化合物的不同谱图映射到相近的位置，而不同化合物的谱图则被明显分离，从而提高化合物鉴定任务中的准确性。这一可视化结果直观地验证SpecEmbedding在化合物鉴定任务中的优越性，表明它能够更好地捕捉质谱谱图之间的相似性，并有效地区分不同化合物的谱图。与传统方法相比，SpecEmbedding在特征空间中的聚集效果更强，表现出更好的分类和识别能力，进一步提升谱图间的可比性和化合物鉴定的准确性。



a：峰箱法向量降维后可视化图像　　b：本文模型方法嵌入向量降维后可视化图像

图2 *t*-SNE可视化结果

Figure 2 *t*-SNE visualization results

2.2 化合物鉴定

在化合物鉴定任务中，谱图相似度是决定鉴定结果的关键因素之一。准确的谱图匹配依赖于同一化合物生成的不同谱图之间具有较高的相似性，这对于准确鉴定未知样品至关重要。在实际操作中，通常将实验得到的查询谱图与已知谱图库中的参考谱图进行比对，通过计算它们之间的相似度分数，进而确定最有可能匹配的化合物。一般来说，鉴定结果是通过选取相似度分数最高的前几位参考谱图作为最终候选结果。

本节采用性能评估方法部分描述的指标，在测试集上全面评估所提模型的表现。为了突出本文方法的优势，还与前言部分提到的几种现有方法进行对比，测量了在不同排名阈值下的命中率和召回率，包括Top-1、Top-2、Top-3、Top-5和Top-10的表现。这些评价指标有效地反映模型在不同条件下的识别能力，以及从大量候选物中正确筛选出目标化合物的能力。具体的命中率和召回率数据见表2和表3。从表中的数据可知，本文提出的SpecEmbedding方法在命中率和召回率两个方面都优于现有的最佳方法CLERMS。在命中率方面，SpecEmbedding方法的Top-1、Top-2、Top-3、Top-5、Top-10分别比CLERMS高出6.30%、3.52%、2.67%、1.47%和0.13%；在召回率方面，SpecEmbedding方法的Top1、Top2、Top3、Top5、Top10分别比CLERMS高出2.61%、3.20%、3.98%、4.30%和3.82%。这些结果表明，SpecEmbedding方法在化合物鉴定任务中展现出了更为优异的性能，能够更准确地从大量候选化合物中识别出目标化合物，尤其在处理大规模数据和复杂样本时，展现了更强的辨识能力和更好的泛化性能，尤其是在面对多样化的化合物数据时，SpecEmbedding方法能够更有效地保持谱图间的可比性，并提高鉴定准确度，显著优于现有方法。

表2 测试集上的化合物鉴定的命中率结果(%)

Table 2 Hit rates of compound identification on the test set (%)

方法	命中率				
	Top-1	Top-2	Top-3	Top-5	Top-10
Binned Cosine	67.97	71.37	74.30	77.41	80.69
Modified Cosine	45.86	50.02	52.88	56.00	60.85
Spec2Vec	74.08	76.69	78.55	80.88	83.50
Ms2DeepScore	62.77	67.80	70.89	74.57	78.90
CLERMS	78.08	83.31	85.89	88.32	91.18
SpecEmbedding	84.38	86.83	88.56	89.79	91.31

表3 测试集上的化合物鉴定的召回率结果(%)

Table 3 Recall rates of compound identification on the test set (%)

方法	召回率				
	Top-1	Top-2	Top-3	Top-5	Top-10
Binned Cosine	20.51	25.81	29.34	33.61	38.39
Modified Cosine	12.56	15.82	17.87	20.26	23.79
Spec2Vec	22.48	28.12	32.19	36.94	42.18
Ms2DeepScore	18.16	23.40	26.87	31.16	36.48
CLERMS	24.75	33.66	39.92	47.26	57.63
SpecEmbedding	27.36	36.86	43.90	51.56	61.45

2.3 相似度分数分布

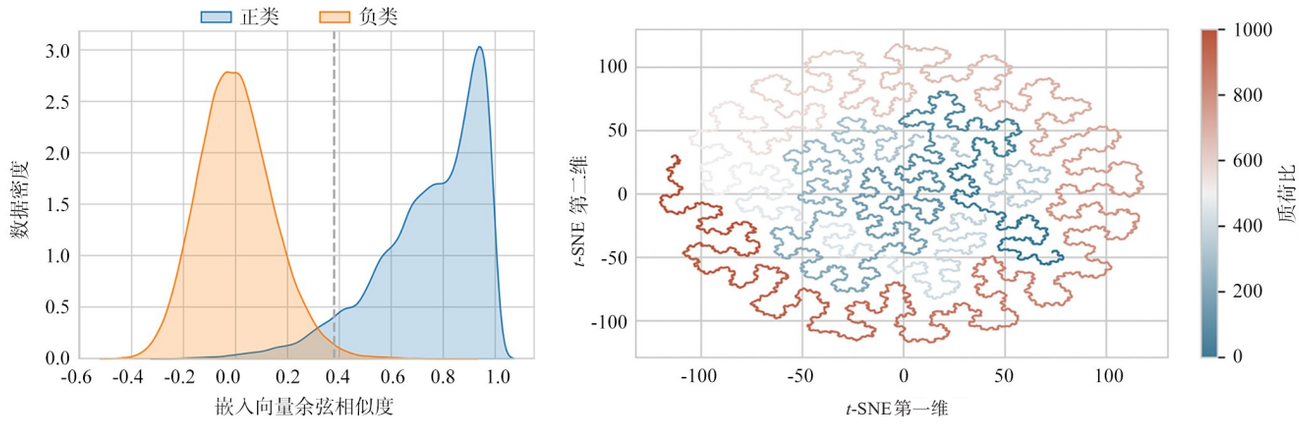
为了进一步评估其泛化能力和可靠性，从测试集的结果中随机抽取等量的正样本对(来自同一化合物的谱图)和负样本对(来自不同化合物的谱图)。随后绘制这些样本对的相似度分布图，结果如图3a所示。从图3a可知，绝大多数正样本对的相似度分数都集中在0.4以上，这表明SpecEmbedding能够有效地将来自同一化合物的不同谱图聚集在一起，从而形成紧密的类内聚类。另一方面，负样本对的相似度分数则普遍低于0.4，说明模型能够显著地区分来自不同化合物的谱图，有效地拉开类间的距离。显著的分数差异不仅证明模型在减小同类样本之间的距离和增大异类样本之间的距离方面的能力，还从侧面验证模型在识别同类谱图时的有效性及其良好的泛化能力。这一结果表明，SpecEmbedding不仅能够在训练集上表现出色，还能够在新的、未见过的样本上有效地进行谱图鉴定，进一步验证其鲁棒性和可靠性。

2.4 正弦嵌入模块有效性

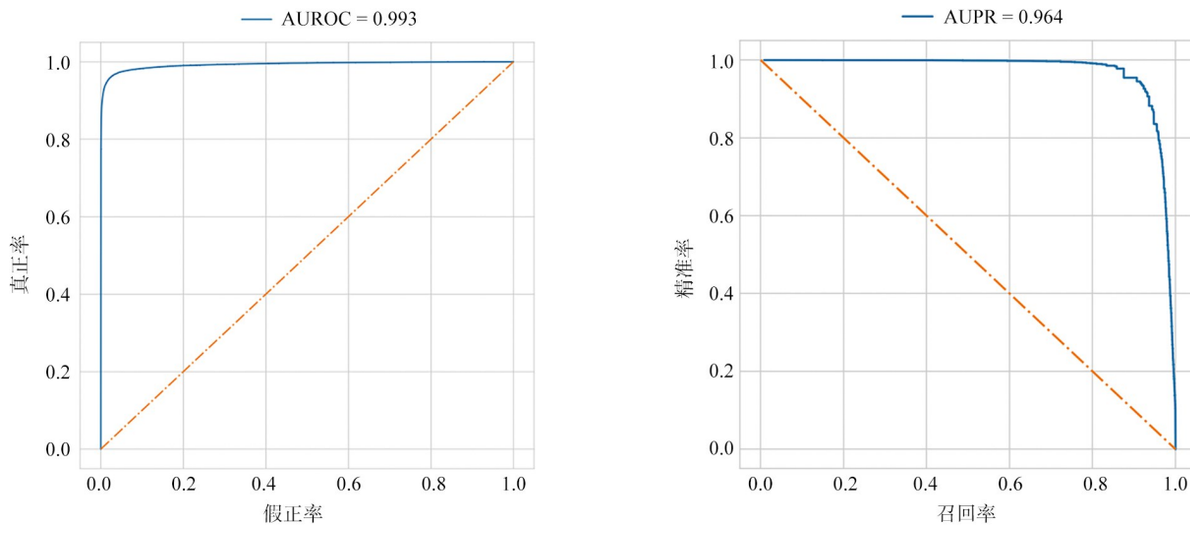
为了验证正弦嵌入模块的有效性，进行额外的实验。将0~1 000道尔顿之间的50 000个m/z值进行嵌入。随后应用*t*-SNE方法将这些嵌入向量降维并进行可视化，结果如图3b所示。从图3b可知，不同大小的m/z值在嵌入空间中得到明显的分隔。这表明SpecEmbedding模型能够有效利用质谱的高分辨率特性，精确地捕捉质谱谱图中的细节。模型在质荷比信息的嵌入过程中展现较强的分辨能力，使得不同的化合物谱图得到有效区分。这一实验结果进一步证明SpecEmbedding在质谱谱图特征表示方面的卓越能力，能够更好地捕捉质谱数据中的精细差异，从而提高化合物鉴定的准确性。

2.5 ROC (Receiver Operating Characteristic) 和 PR (Precision-Recall) 曲线测试

为了进一步验证模型区分正负样本对的能力，进行ROC和PR测试，结果分别如图3c和图3d所示。在



a：正负样本余弦相似度分布图　　b：根据质荷比值着色的质荷比嵌入的*t*–SNE投影图



c：ROC曲线　　d：PR曲线

图3 SpecEmbedding模型在测试集上的综合评估结果

Figure 3 Comprehensive evaluation results of SpecEmbedding model on the test set

ROC测试中，模型能够在保持较低假正率（False Positive Rate, FPR）的同时获得较高的真正率（True Positive Rate, TPR），这一结果表明SpecEmbedding模型能够有效区分来自同一化合物的正样本（同类谱图）和来自不同化合物的负样本（异类谱图）。高TPR和低FPR的组合表明该模型具有强大的区分能力，能够有效地识别目标化合物。为了保证评估的公平性，在PR测试中，从测试集中抽取所有正样本对，并随机选取与正样本对数量相当的负样本对。PR曲线的结果表明，SpecEmbedding模型不仅能够保持较高的精准率，还能够获得较高的召回率。这表明尽管数据集存在正负样本不平衡问题，模型依然能够高效识别正样本，并且保持较高的召回率，避免过度牺牲召回率以提高精准率的情况。

综合上述测试结果，SpecEmbedding通过正弦嵌入模块和对比学习策略的结合，能够有效利用质谱数据中的高精度特征，在化合物鉴定任务中取得优异的表现。这些结果表明SpecEmbedding模型不仅在当前的数据集上表现优异，而且具有较强的泛化能力。在ROC测试中，模型能够在低假正率的情况下取得高真正率；在PR测试中，模型不仅精准率高，而且召回率也表现出色。这些测试结果进一步证明SpecEmbedding模型在区分正负样本方面的强大能力和良好的泛化性能。

3 讨论与结论

本文提出的SpecEmbedding方法结合正弦嵌入方法和监督对比学习策略，旨在提高化合物鉴定的准确性和鲁棒性，应对化合物空间结构和实验环境为鉴定带来的挑战。实验结果表明，SpecEmbedding在处理化合物质谱图时，不仅能够有效地将同一化合物的不同谱图嵌入到低维向量空间中进行聚集，而且能够显著区分不同化合物的谱图。通过这种方式，模型在提高化合物鉴定的准确性方面取得显著的进展。现有的谱图嵌入方法往往忽视质荷比这一关键信息的有效利用，导致嵌入向量的区分度较低。为了弥补这一不足，在

模型中引入正弦嵌入模块，通过质荷比信息的嵌入进一步提升谱图特征的表达。实验结果通过*t*-SNE可视化验证了这一模块的有效性，嵌入的质荷比信息能够有效地将不同质荷比的向量分隔开，证明该嵌入模块对提升谱图区分度的贡献。在命中率评估方面，SpecEmbedding在Top-1命中率上达到84.38%，比现有方法CLERMS高出6.3%。此外，在召回率等指标上，模型也表现出了显著的优势，同时也侧面说明监督对比学习策略在谱图嵌入任务中的有效性。这一结果表明，监督对比学习不仅能够通过聚合同类样本并区分异类样本，提高化合物鉴定的精度，而且在复杂数据集上能够有效优化嵌入空间，从而提升识别能力。

本文使用InChIKey的前14位字符标识化合物，这一简化做法减少标签歧义并降低模型复杂度，但忽略立体异构信息：若两个化合物仅在手性或双键构型上不同，却被视为“同一化合物”，则可能高估模型性能。在药物研发、代谢物追踪等必须区分立体构型的场景，这种处理方式会削弱模型的实际应用价值。

综上所述，SpecEmbedding不仅为化合物鉴定提供一个高效且精准的解决方案，同时可以为质谱数据的嵌入提供一个新的可能的思路和方法。未来的研究可以进一步探索该模型在其他类型的谱图数据（如NMR或GC-MS）上的可扩展性与通用性。另外，本研究尚未对立体异构体（如手性化合物）在谱图中的细微差异进行建模和区分，这在某些特定场景下可能会限制模型的分辨能力。因此，构建能够识别或适应立体异构体差异的谱图嵌入方法，也是未来的重要研究方向之一。

【参考文献】

[1] Zang XL, Monge ME, Fernández FM. Mass spectrometry-based non-targeted metabolic profiling for disease detection: recent developments [J]. Trends Analyt Chem, 2019, 118: 158-169.

[2] Wu ZJ, Bagarolo GI, Thoröe-Boveleth S, et al. "Lipidomics": mass spectrometric and chemometric analyses of lipids[J]. Adv Drug Deliv Rev, 2020, 159: 294-307.

[3] Wolfender JL, Nuzillard JM, van der Hooft JJ, et al. Accelerating metabolite identification in natural product research: toward an ideal combination of liquid chromatography-high-resolution tandem mass spectrometry and NMR profiling, in silico databases, and chemometrics[J]. Anal Chem, 2019, 91(1): 704-742.

[4] Lam H, Deutsch EW, Eddes JS, et al. Development and validation of a spectral library searching method for peptide identification from MS/MS[J]. Proteomics, 2007, 7(5): 655-667.

[5] Lam H, Deutsch EW, Eddes JS, et al. Building consensus spectral libraries for peptide identification in proteomics[J]. Nat Methods, 2008, 5(10): 873-875.

[6] Stein SE, Scott DR. Optimization and testing of mass spectral library search algorithms for compound identification[J]. J Am Soc Mass Spectrom, 1994, 5(9): 859-866.

[7] Wan KX, Vidavsky I, Gross ML. Comparing similar spectra: from similarity index to spectral contrast angle[J]. J Am Soc Mass Spectrom, 2002, 13(1): 85-88.

[8] Bittremieux W, Schmid R, Huber F, et al. Comparison of cosine, modified cosine, and neutral loss based spectrum alignment for discovery of structurally related molecules[J]. J Am Soc Mass Spectrom, 2022, 33(9): 1733-1744.

[9] Watrous J, Roach P, Alexandrov T, et al. Mass spectral molecular networking of living microbial colonies[J]. Proc Natl Acad Sci U S A, 2012, 109(26): E1743-E1752.

[10] Guijas C, Montenegro-Burke JR, Domingo-Almenara X, et al. METLIN: a technology platform for identifying knowns and unknowns [J]. Anal Chem, 2018, 90(5): 3156-3164.

[11] Xue JC, Guijas C, Benton HP, et al. METLIN MS2 molecular standards database: a broad chemical and biological resource[J]. Nat Methods, 2020, 17(10): 953-954.

[12] Liu KY, Tao CH, Ye YZ, et al. SpecEncoder: deep metric learning for accurate peptide identification in proteomics[J]. Bioinformatics, 2024, 40(S1): i257-i265.

[13] Lu XY, Wu HP, Ma H, et al. Deep learning-assisted spectrum-structure correlation: state-of-the-art and perspectives[J]. Anal Chem, 2024, 96 (20): 7959-7975.

[14] 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.

[15] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2013: 3111-3119.

[16] Huber F, van der Burg S, van der Hooft JJ, et al. MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra[J]. J Cheminform, 2021, 13(1): 84.

[17] Bittremieux W, May DH, Bilmes J, et al. A learned embedding for efficient joint analysis of millions of mass spectra[J]. Nat Methods, 2022, 19(6): 675-678.

[18] 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.

[19] van den Oord A, Li YZ, Vinyals O. Representation learning with contrastive predictive coding[EB/OL]. (2019-01-22). https://arxiv.org/abs/1807.03748.

[20] Ye M, Zhang X, Yuen PC, et al. Unsupervised embedding learning via invariant and spreading instance feature [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2019: 6203-6212.

[21] He KM, Fan HQ, Wu YX, et al. Momentum contrast for unsupervised visual representation learning[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2020: 9726-9735.

[22] Khosla P, Teterwak P, Wang C, et al. Supervised contrastive learning [C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2020: 18661-18673.

[23] Wang MX, Carver JJ, Phelan VV, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking[J]. Nat Biotechnol, 2016, 34(8): 828-837.

[24] Huber F, Verhoeven S, Meijer C, et al. Matchms-processing and similarity evaluation of mass spectrometry data[J]. J Open Source Softw, 2020, 5(52): 2411.

[25] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2017: 6000-6010.

[26] Lauriola I, Lavelli A, Aiolli F. An introduction to deep learning in natural language processing: models, techniques, and tools [J]. Neurocomputing, 2022, 470: 443-456.

[27] Voronov G, Lightheart R, Davison J, et al. Multi-scale sinusoidal embeddings enable learning on high resolution mass spectrometry data [EB/OL]. (2023-05-05). https://arxiv.org/abs/2207.02980.

[28] Qiao R, Tran NH, Xin L, et al. Deepnovov2:better de novo peptide sequencing with deep learning[EB/OL]. (2019-05-22). https://arxiv.org/abs/1904.08514.

[29] Yilmaz M, Fondrie W, Bittremieux W, et al. De novo mass spectrometry peptide sequencing with a transformer model[C]// Proceedings of the 39th International Conference on Machine Learning. Chia Laguna Resort, Sardinia, Italy: PMLR, 2022: 25514-25522.

[30] Bushuiev R, Bushuiev A, Samusevich R, et al. Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS[J]. Nat Biotechnol, 2025: 1-11.

[31] Loshchilov I, Hutter F. Decoupled weight decay regularization[EB/OL]. (2019-01-04). https://arxiv.org/abs/1711.05101.

[32] Akiba T, Sano S, Yanase T, et al. Optuna: a next-generation hyperparameter optimization framework[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY, USA: Association for Computing Machinery, 2019: 2623-2631.

[33] van der Maaten L, Hinton G. Visualizing data using *t*-SNE[J]. J Mach Learn Res, 2008, 9(86): 2579-2605.

（编辑：陈丽霞）