南京邮电大学
人工智能（作业）

题　　目　质谱预测

专　　业　计算机科学与技术

学生姓名　胡国裕

班级学号　B20031624

指导老师　张洁

指导单位　计算机学院

日期：　2024 年 6 月 1 日 至 2024 年 6 月 14 日

摘　　要

本文研究了质谱预测中的多种深度学习模型。首先，NEIMS 模型通过扩展圆形指纹和多层感知器实现了快速准确的小分子质谱预测。其次，CFM-ID 4.0 模型利用环裂解建模和连接矩阵特征提高了质谱预测的准确性。GRAFF-MS 模型结合图神经网络和手写规则，显著提升了质谱预测的精度。MassFormer 模型通过图变压器架构，解决了复杂分子关系的捕捉问题，表现出色。MoMS-Net 模型利用异质基序图神经网络对质谱进行高效预测。3DMolMS 模型通过 3D 分子卷积操作，实现了高精度的质谱谱图预测。SCARF 模型则通过前缀树生成和 Set Transformer 架构，在质谱预测中表现优越。ICEBERG 模型通过递归生成和评分模块，实现了高精度的分子碎片图生成。FraGNNet 模型结合组合碎裂方法和概率建模，能够高效、准确地预测高分辨率质谱图。本文对上述模型进行了详细分析，验证了其在实际应用中的潜力，并提出了未来的研究方向。

关键词： 质谱预测; 深度学习模型

ABSTRACT

This paper investigates various deep learning models for mass spectrum prediction. Firstly, the NEIMS model achieves rapid and accurate prediction of small molecule mass spectra using extended circular fingerprints and multilayer perceptrons. Secondly, the CFM-ID 4.0 model improves the accuracy of mass spectrum prediction through ring-cleavage modeling and connectivity matrix features. The GRAFF-MS model combines graph neural networks and handwritten rules to significantly enhance prediction accuracy. The MassFormer model, with its graph transformer architecture, excels in capturing complex molecular relationships. The MoMS-Net model efficiently predicts mass spectra using heterogeneous motif graph neural networks. The 3DMolMS model achieves high-precision spectrum prediction through 3D molecular convolution operations. The SCARF model, leveraging prefix tree generation and Set Transformer architecture, demonstrates superior performance in spectrum prediction. The ICEBERG model generates high-precision molecular fragmentation graphs through recursive generation and scoring modules. The FraGNNet model, combining combinatorial fragmentation methods and probabilistic modeling, effectively and accurately predicts high-resolution mass spectra. This paper provides a detailed analysis of these models, verifies their potential in practical applications, and proposes future research directions.

Keywords: Mass spectrum prediction;Deep learning models

目　　录

第一章　NEIMS 模型

本章介绍一种轻量级神经网络模型，称为 NEIMS[1]（Neural Electron-Ionization Mass Spectrometry），能够快速预测小分子的质谱。该模型的预测速度非常快，平均每个分子只需 5 毫秒，并且在 10 次召回准确率达到 91.8%。NEIMS 使用了一种新颖的神经网络架构，专门设计用于捕捉电子电离过程中典型的分子碎裂模式。作者通过使用 NIST 2017 质谱库的数据对模型进行了测试。实验结果表明，NEIMS 模型的预测能力与之前的机器学习模型相当，但预测速度更快。通过将 NEIMS 预测的质谱与现有的实验质谱结合，能够显著提高质谱库的覆盖范围，从而提高未知分子样品的识别准确率。

1.1 模型介绍

NEIMS 模型旨在通过神经网络快速预测小分子的电子电离质谱。

1.1.1 扩展圆形指纹

扩展圆形指纹（Extended Circular Fingerprints，ECFP）是一种用于分子描述的方法，能够捕捉分子结构中的局部子图。ECFP 通过记录分子中原子及其周围的环境来生成固定长度的二进制向量，表示分子的结构特征。这种表示方法在化学信息学和计算机辅助药物设计中广泛使用。

ECFP 生成过程如下：

(1) 初始化：每个原子都被分配一个初始标识符（ID），通常是基于原子的类型、价态和环境等化学特性。

(2) 迭代生成：每个原子开始，在其周围一定半径范围内扩展，形成局部子图。这个半径通常被称为迭代深度或圆形指纹的“半径”，在每个迭代步骤中，将子图中的原子和键连接起来，生成新的标识符，这些标识符表示更大的结构单元。

(3) 哈希和计数：将生成的标识符通过哈希函数映射到固定长度的二进制向量中，每个子图的出现次数被记录在向量中，相应位置的值增加。

1.1.2 分子表示

首先，分子被表示为扩展圆形指纹，结构如下：

$$\mathrm{ECFP}(molecule) = [f_1, f_2, \ldots, f_n] \tag{1-1}$$

其中，f_i 表示第 i 个特征。

1.1.3 多层感知器（MLP）

这些分子特征被输入到一个多层感知器（MLP）中，该感知器由多个隐藏层组成，每层都有一定数量的节点，并使用 ReLU 激活函数和 dropout 正则化。

$$h_1 = \mathrm{ReLU}(W_1 \cdot \mathrm{ECFP} + b_1) \tag{1-2}$$

$$h_2 = \mathrm{ReLU}(W_2 \cdot h_1 + b_2) \tag{1-3}$$

$$\vdots \tag{1-4}$$

$$h_L = \mathrm{ReLU}(W_L \cdot h_{L-1} + b_L) \tag{1-5}$$

其中，W_i 和 b_i 是每层的权重矩阵和偏置向量，L 是隐藏层的总数。

1.1.4 质谱预测

在 NEIMS 模型中，为了提高预测精度，采用了双向预测（bidirectional prediction）方法，即从分子的两端同时进行预测。这个方法通过前向预测（forward prediction）和反向预测（reverse prediction）来捕捉分子的碎裂模式，确保在预测质谱图时能够更准确地反映实际的碎裂现象。

具体来说，双向预测的步骤如下：

(1) 前向预测：从分子的起始点（通常是质量数 m/z 为 0 的位置）开始，模型根据分子的特征向量逐步预测质谱图中每个位置的离子强度。

(2) 反向预测：从分子的终点（通常是分子的最大质量数）开始，模型从高质量数向低质量数进行预测。这种方法能够有效地捕捉大质量碎片的强度，这些碎片往往是由于分子中小团体的中性损失（neutral loss）而形成的。

(3) 双向预测的结合：将前向预测和反向预测的结果进行结合，使用一个门控机制（gating mechanism）来确定每个质量数位置上前向和反向预测的权重。具体公式如下：

$$p_i = \sigma(g_i) \cdot p_i^{(f)} + (1 - \sigma(g_i)) \cdot p_i^{(r)} \tag{1-6}$$

其中，σ 是 sigmoid 函数，g_i 是通过特征计算得到的门控值。

1.1.5 损失函数

在 NEIMS 模型中，损失函数是一个修改后的均方误差损失函数（mean-squared-error loss function），它根据质谱图中的特性进行了调整，以提高模型在实际应用中的表现。具体来说，损失函数的形式如下：

$$L(I,\hat{I}) = \sum_{k=1}^{M(x)} \left(\frac{m_k^{0.5} I_k}{\left(\sum_{k=1}^{M} \left(m_k^{0.5}\right)^2\right)^{0.5}} - \frac{m_k^{0.5} \hat{I}_k}{\left(\sum_{k=1}^{M} \left(m_k^{0.5} \hat{I}_k\right)^2\right)^{0.5}} \right)^2 \tag{1-7}$$

其中:

- I 是真实的质谱图强度向量。
- $\hat{I}$ 是预测的质谱图强度向量。
- I_k and $\hat{I}_k$ 分别是真实和预测的第 k 个质量数位置的强度。
- m_k 是第 k 个位置的质量数（m/z）。
- $M_{\max}$ i 是质谱图中非零强度值的最大索引。

通过这种加权的均方误差损失函数，模型能够更好地捕捉质谱图中重要的高质量数位置的特征，提高整体预测精度。模型使用随机梯度下降法（stochastic gradient descent）和 Adam 优化器进行优化。

1.1.6 质量过滤

为了提高匹配准确性和效率，论文中使用了质量过滤（mass filtering）。通过设定一个质量公差窗口（例如 ±5 Da），仅包括与查询分子质量相差不大的分子候选者。这减少了待匹配的候选光谱数量，从而提高了匹配效率。论文中提到，应用质量过滤后，NEIMS 模型的召回率（recall@10）提高到 91.7%。这意味着在应用质量过滤后，每个查询分子平均只需要匹配大约 6696 个候选光谱。

1.1.7 库匹配评估（Library Matching Evaluation）

库匹配评估的目的是测试 NEIMS 模型通过增强参考库进行光谱匹配的性能。具体步骤如下：

(1) 将预测的质谱图与 NIST 2017 主库中的实验质谱图结合，形成一个增强的参考库。

(2) 对于每个查询光谱，计算它与增强参考库中所有光谱的相似性。具体公式如

下：

$$\text{Similarity}(I_q, I_l) = \frac{\sum_{k=1}^{M_{\max}} (m_k^{0.5} I_{q,k} \cdot m_k^{0.5} I_{l,k})}{\sqrt{\sum_{k=1}^{M_{\max}} (m_k^{0.5} I_{q,k})^2} \cdot \sqrt{\sum_{k=1}^{M_{\max}} (m_k^{0.5} I_{l,k})^2}} \tag{1-8}$$

(3) 记录正确光谱的排名。例如，如果查询光谱与增强参考库中相似度最高的光谱匹配，它的排名为 1。

1.2 增强库

在 NEIMS 模型中，加入预测质谱图形成增强参考库的原因有几个，主要是为了弥补现有质谱库的覆盖问题，并提高对未包含在现有质谱库中的分子的识别能力。具体来说，有以下几个原因：

(1) 覆盖问题：现有的质谱参考库，如 NIST 2017 主库，虽然包含大量的质谱图，但仍然存在覆盖不足的问题。例如，NIST 主库中仅包含几百到几百万个分子的质谱图，而现实中可能存在的分子种类远远超过这个数量。

(2) 识别未包含的分子：当一个查询分子在现有的质谱库中没有对应的实验质谱图时，传统的方法无法准确识别该分子。而通过 NEIMS 模型预测出的质谱图，可以补充到参考库中，增加对这些未包含分子的识别能力。

(3) 提高匹配准确性：通过增加预测质谱图，增强参考库可以提供更多的候选匹配，从而提高查询光谱与库中光谱匹配的准确性。即使查询分子在实验库中没有完全匹配的光谱，通过预测的质谱图也可以提高匹配的可能性。

(4) 增强模型的泛化能力：通过训练 NEIMS 模型来预测质谱图，模型可以学习到分子碎裂模式的更广泛的特征，从而在面对新分子时具有更好的泛化能力。

1.3 论文结论

NEIMS 模型在增强的光谱库中实现了高匹配性能，并且在查询集的分子预测中表现优异。其性能略优于现有的机器学习模型，同时显著提升了预测速度。

(1) 双向预测模式的贡献：NEIMS 模型的高性能部分归功于双向预测模式，尤其是反向预测模式。反向预测模式能够更准确地预测由于小中性基团损失而产生的大碎片的强度。

(2) 光谱相似度的提高：观测到在库匹配任务中的性能提升也对应于预测光谱与真实光谱相似度的提高。高质量数区域的预测准确性在改进后显著提高，这对于分子鉴定至关重要。

1.3.1 模型的改进建议

NEIMS 目前无法建模离子碎片中对应于同位素的强度峰。如果对峰位置精度更高的光谱数据进行训练，应该能够根据 m/z 峰位置的小数值精确确定原子身份。

质量过滤提高了 NEIMS 的性能约 6%。在实验设置中，如果能够大致知道样品的分子质量，则可以通过质量过滤提高匹配的准确性。

1.3.2 未来工作方向

(1) 使用图卷积分子表示，尤其是基于键的表示，有望以稍高的计算成本提高预测准确性。

(2) 将 NEIMS 与迁移学习方法结合，可以针对特定质谱仪预测光谱，从而提高匹配性能。

(3) NEIMS 的轻量级框架使其能够快速生成大量分子的光谱预测，这些预测光谱可以直接用于质谱软件中，扩展可通过质谱鉴定的分子覆盖范围。

第二章　CFM-ID 4.0

在代谢组学中，质谱（MS）是最常用的方法，用于鉴定和标注代谢物。传统方法依赖于与实验获得的参考光谱库进行匹配，但由于这些光谱库的覆盖范围有限，这种方法存在一定局限性。为了克服这一问题，研究人员开发了 CFM-ID（Competitive Fragmentation Modeling for Metabolite Identification），这是一种能够准确预测给定化合物结构的 ESI-MS/MS 光谱的计算机程序。最新版本 CFM-ID 4.0，通过从分子拓扑结构中学习参数、引入新的环裂解模型、扩展手写规则库以涵盖更多化学类等改进，提高了光谱预测的准确性和化合物鉴定的效果[2]。

2.1 环裂解建模

在 CFM-ID 4.0 中，环裂解建模被改进为两步序列过程，具体如下：

1. **初步裂解**：环中的一个化学键断裂，生成一个中间碎片。记为：

$$f_0 \rightarrow f_{\text{inter}} \tag{2-1}$$

2. **进一步裂解**：中间碎片中的另一个化学键断裂，生成最终的离子碎片和中性丢失碎片。记为：

$$f_{\text{inter}} \rightarrow f_1 \tag{2-2}$$

3. **中间碎片处理**：中间碎片由于化学不稳定性，持续时间被固定为零，确保它们不会在预测的质谱中生成峰值。

2.2 连接矩阵特征

CFM-ID 4.0 通过引入基于连接矩阵的化学结构表示方法，提取与质谱相关的拓扑特征，具体如下：

1. **图的表示**：将化学结构表示为一个根图 $G = (V, E, VL, EL, R)$，其中：

- V 是顶点集
- E 是边集
- VL 和 EL 分别为顶点和边的标签
- R 为根顶点

2. **特征提取**：

- **启发式字符串生成**：通过广度优先遍历生成启发式字符串 $s(v)$。
- **顶点索引**：基于启发式字符串对顶点进行索引，选择与根顶点最相关的子图

SG。

3. **邻接矩阵生成**：从子图 SG 中生成邻接矩阵 M_{adj}，并将其转化为张量 T_{adj} 和 T_{vertex}。

4. **张量表示**：张量 T_{adj} 的第一和第二轴代表邻接矩阵，第三轴存储每个化学键的特征。

$$T_{adj}(i,j) = \begin{cases} 0 & \text{if } (i,j) \notin E \\ \text{feature vector} & \text{if } (i,j) \in E \end{cases} \tag{2-3}$$

将张量展平为一维向量，并与表示每个顶点特征的向量连接在一起，生成最终的特征向量。

2.3 手写规则扩展

CFM-ID 4.0 扩展了手写规则库，涵盖了更多化学类，具体如下：

1. **规则库扩展**：增加了针对酰基肉碱、酰胆碱、黄酮醇、黄酮、黄烷酮和黄酮苷等化学类的手写规则。

2. **SMIRKS 规则**：使用 SMIRKS 字符串描述化学反应，这些反应生成特定的碎片。规则还包括从实验光谱中获得的相对强度信息。

3. **模块化化合物**：规则库能够识别模块化结构的化合物，如脂质，通过对头基和尾基的组合进行预测。这种方法不仅提高了预测的质量，还加快了预测速度。

2.4 结论

在本研究中，我们介绍了 CFM-ID 的改进版本 CFM-ID 4.0。CFM-ID 4.0 包含以下几个关键改进：

1. 引入了一种新的张量表示法来描述化学结构的拓扑，从而更好地提取与质谱相关的拓扑特征。2. 采用了新的环裂解建模方法，将环裂解过程建模为两步序列过程，以提高预测的准确性。3. 扩展了基于规则的方法，涵盖了更多的化学类，如酰基肉碱、酰胆碱、黄酮醇、黄酮、黄烷酮和黄酮类糖苷等，以增强这些特定化学类化合物的 MS/MS 光谱预测能力。

我们通过对多个 ESI-MS/MS 数据集进行评估，发现 CFM-ID 4.0 在光谱预测和化合物鉴定任务中均显著优于早期版本的 CFM-ID。具体而言，CFM-ID 4.0 在以下几个方面表现出显著的改进：

- 平均 Dice 系数在所有能量水平下均提高了 31.8%（[M + H]□ 光谱）和 8%（[M - H]□ 光谱）。

- 平均点积分数在所有能量水平下均提高了 26.7%（[M + H]□ 光谱）和 20.6%（[M - H]□ 光谱）。
- 在化合物识别任务（MS2C）中，CFM-ID 4.0 能够正确识别 208 个输入光谱中的 147 个，相比之前的版本提高了 22.5%。

尽管 CFM-ID 4.0 在性能上超越了其前任，但在准确性和运行时间方面仍有进一步改进的空间。未来的研究方向包括：

- 使用更深的碎片图，以更好地处理复杂的化学结构。这可能会引入额外的计算复杂性，但对于复杂化合物的预测会更为准确。
- 采用更高效的方法来探索碎片图，从而提高计算效率。
- 使用更先进的分子建模方法，如图神经网络，以进一步提高预测性能。
- 使 CFM-ID 能够处理除 QTOF 仪器外的其他质谱仪器收集的 MS 光谱数据，如 Orbitrap 质谱仪。

总之，CFM-ID 4.0 通过结合环裂解建模、连接矩阵特征提取和手写规则扩展，显著提高了 ESI-MS/MS 光谱预测和化合物鉴定的准确性。新版本不仅提高了预测的准确性，还为特定化学类提供了更快的预测方法，展示了其在代谢组学领域的重要应用前景。我们相信，通过进一步的改进和优化，CFM-ID 4.0 将成为代谢组学研究中的一个重要工具。

第三章　GRAFF-MS 模型

本章提出了一种利用图神经网络（GNN）来高效预测小分子的高分辨率质谱的方法[3]。传统的质谱预测方法在捕捉高分辨率质量信息和可处理性之间存在权衡。本文通过将质谱预测表述为从分子图到化学式概率分布的映射，解决了这一问题。研究发现，大部分质谱可以通过仅占所有观测化学式 2% 的固定词汇近似，从而实现高效的质谱预测。提出的 GNN 架构 GRAFF-MS 在预测误差和检索准确性方面优于现有方法。

3.1 模型架构介绍

质谱预测的一个主要挑战是输出空间的建模：质谱是一个变长的实值元组集合，预测的 m/z 坐标需要非常高的精度。之前的方法在捕捉高分辨率的 m/z 信息和学习问题的可处理性之间存在权衡。将质谱建模为前体的化学子公式的概率分布，可以有效避免之前方法中的问题。

3.1.1 图神经网络编码器

GRAFF-MS 使用图同构网络（Graph Isomorphism Network, GIN）作为基础架构，对分子图进行编码。具体步骤如下：

(1) **输入表示：**

- 分子图 $G = (V, E, a, b)$ 包含节点集 V（代表原子）、边集 E（代表键）、节点特征 a（化学元素）和边特征 b（键阶）。
- 节点特征 a_i 和边特征 b_{ij} 使用 DGL-LifeSci 工具中的标准原子和键特征器生成。

(2) **特征嵌入：**

- 使用多层感知器（MLP）对节点、边和实验协变量进行嵌入。实验协变量包括归一化碰撞能量、前体离子类型、仪器型号和同位素峰的存在。
- 将拉普拉斯特征转化为节点位置编码，使用 SignNet 进行编码。

(3) **消息传递和更新：**

- 使用 GIN 卷积操作进行消息传递，并对节点和边特征进行更新。更新后的节点特征和边特征通过残差连接加速训练。
- 最终生成分子的密集向量表示 x_{mol}。

(4) **池化和解码：**

- 使用注意力池化操作对节点特征进行池化，生成分子的全局表示，并结合嵌入的协变量特征。

- 通过一个 MLP 将分子的全局表示解码为光谱表示 x_{spec}。

3.1.2 固定词汇表近似

在高分辨率质谱预测中，质谱由许多峰组成，每个峰表示一种化学子公式的质量（m/z）和强度（intensity）。由于质谱数据的复杂性和多样性，直接预测每个可能的子公式会带来很高的计算复杂度。因此，本文提出使用固定词汇表来近似质谱数据。这些固定词汇表是通过分析大量训练数据，选出那些最常见和最重要的子公式，减少计算负担并提高预测效率。具体步骤如下：

(1) **收集训练数据：**
- 从训练数据中提取所有可能的产品离子公式和中性损失公式。每个质谱中的峰都对应一个或多个子公式。

(2) **统计公式频率：**
- 对每个质谱中的每个峰，使用质量分解（mass decomposition）方法计算出该峰可能对应的所有化学子公式。
- 统计这些子公式在训练数据中出现的频率。这个过程可以理解为一个加权统计过程，峰的强度作为权重。

(3) **选择高频公式：**
- 根据统计的频率，对所有子公式进行排序。
- 择出现频率最高的前 K 个产品离子公式和中性损失公式，组成固定词汇表。
- 将选出的高频产品离子公式和中性损失公式合并，形成固定词汇表 $\hat{F}(P)$。

3.1.3 模型训练

在质谱数据中，每个质谱峰（m/z，intensity）表示某个化学子公式的质量-电荷比和相对强度。由于质谱仪器的有限分辨率，一个观测峰可能对应多个子公式。因此，单一子公式的预测并不能完全代表观测峰，需要将这些可能的子公式进行边际化处理。模型使用交叉熵损失函数来训练模型。损失函数的具体形式为：

$$\text{Loss} = -\sum_{n=1}^{N}\sum_{i\in S_n} y_{ni}\log\left(\sum_{f\in F_n^i}\hat{y}_f\right) \tag{3-1}$$

其中：
- N 是训练集中质谱的数量。
- S_n 是第 n 个质谱中的峰值集合。

- y_{ni} 是第 n 个质谱的第 i 个峰值的强度。
- F_n^i 是与第 n 个质谱的第 i 个峰值匹配的公式集合。
- $\hat{y}_f$ 是模型预测的公式 f 的概率。

在进行模型训练时，主要包括以下几个步骤：

(1) 构建图神经网络模型：使用图同构网络（GIN）作为基础架构，对分子图进行编码，生成分子的嵌入表示 x_{mol}。将嵌入表示 x_{mol} 输入到一个全连接层，预测每个子公式的概率 y_f。

(2) 计算损失函数：：使用峰边际交叉熵损失函数计算观测峰的强度 y_i 和预测峰的强度 y_f 之间的交叉熵损失。

(3) 模型优化：使用梯度下降算法（如 Adam 优化器）对模型参数进行优化，最小化损失函数。

3.2 结论

本文提出了一种新型的基于图神经网络（Graph Neural Networks, GNN）的架构 GRAFF-MS，用于高效预测高分辨率质谱图。该架构在不牺牲质量-电荷比（m/z）分辨率的前提下，实现了高效的计算，并在多个数据集上表现出色。以下是论文的主要结论：

- **创新性方法**：提出了一种将质谱预测表述为从分子图到化学公式概率分布的映射的新方法。通过这种方法，能够捕捉高分辨率的 m/z 信息，同时保持学习问题的可处理性。
- **固定词汇表的有效性**：研究发现，大多数小分子质谱的信号可以通过一小部分固定的高频化学子公式来解释。使用固定词汇表近似大大简化了光谱预测问题，并且对预测性能影响较小。
- **优异的实验结果**：GRAFF-MS 在多个数据集上的实验结果显示，该方法在预测误差和运行时间方面均优于现有的最先进方法。具体来说，GRAFF-MS 在 NIST-20、CASMI-16 和 GNPS 数据集上的表现均优于 CFM-ID 和 NEIMS 等方法。
- **计算效率高**：GRAFF-MS 利用图神经网络和固定词汇表的结合，大幅度提高了计算效率。相比传统方法，GRAFF-MS 在处理大规模数据集时更加高效。
- **领域特定的修正**：通过引入领域特定的修正（如同位素状态和双重计数修正），进一步提高了模型的预测准确性和实际应用的可行性。

3.2.1 未来研究方向

作者还讨论了未来可能的研究方向，包括：

- **动态词汇表生成**：探索动态生成词汇表的方法，进一步提高模型的灵活性和泛化能力。
- **公式表示学习**：研究通用公式表示的学习方法，使得模型能够更好地捕捉化学子公式的特性。
- **扩展应用**：将 GRAFF-MS 方法扩展到更多类型的分子和质谱数据，以验证其通用性和鲁棒性。
- **改进模型架构**：在现有模型架构的基础上，进一步优化图神经网络的设计，以提高预测精度和计算效率。

第四章　MassFormer 模型

本章提出了一种新的模型——MassFormer，用于准确预测小分子的串联质谱。MassFormer 利用图变压器架构来建模分子中原子之间的远距离关系，通过化学预训练任务初始化变压器模块的参数，然后在光谱数据上进行微调[4]。实验结果表明，MassFormer 在多个数据集上均优于现有方法。

4.1 输入特征化

输入特征化是将分子转化为模型能够处理的数值表示的过程。在本文中，分子表示为分子图，图的节点表示分子中的原子，边表示原子之间的化学键，可以使用 RDKit 等工具将分子表示为图结构。接下来，我们具体讲解如何从分子图中提取节点和边的嵌入。

4.1.1 节点特征

我们从分子图中提取节点（原子）的特征。以乙醇（C_2H_6O）为例，其节点特征包括：

- 元素种类：C, H, O
- 原子度：例如，C 的度为 4，H 的度为 1，O 的度为 2
- 形式电荷：所有原子的形式电荷均为 0
- 芳香性：所有原子的芳香性均为 False

4.1.2 边特征

我们从分子图中提取边（化学键）的特征。其边特征包括：

- 键类型：如单键、双键、三键等
- 最短路径长度：用于表示两个原子之间的最短路径
- 路径上的边：用于表示路径上的具体键信息

4.2 图变压器架构

4.2.1 输入嵌入

将分子图中的每个节点和边编码为嵌入向量，这些向量包含原子的化学属性和原子之间的拓扑关系。

4.2.2 多头自注意力（MHA）

对每个节点，计算其特征的查询（Query）、键（Key）和值（Value）向量。计算节点 i 和节点 j 之间的注意力权重 a_{ij}，公式如下：

$$a_{ij} = \text{softmax}\left(\frac{(W_Q h_i)^\top (W_K h_j)}{\sqrt{d}} + b(n_{ij}) + c_{ij}\right) \tag{4-1}$$

其中，W_Q 和 W_K 是可学习的投影矩阵，$b(n_{ij})$ 是节点 i 和节点 j 之间的最短路径长度嵌入，c_{ij} 是路径上的边嵌入。

4.2.3 多层感知器（MLP）

对经过注意力加权的节点特征进行非线性变换，公式如下：

$$h_i^{(l+1)} = \text{MLP}(h_i^{(l)} + \sum_{m=1}^{M} h_i^{(m,l)}) \tag{4-2}$$

其中，M 是注意力头的数量，$h_i^{(m,l)}$ 是第 m 个注意力头在第 l 层对节点 i 的输出。

4.2.4 全局嵌入生成

使用一个特殊的“读出”节点，将所有节点的最终嵌入汇总，生成整个分子的嵌入。这个嵌入将与光谱元数据一起输入到 MLP 中，生成最终的光谱预测。

$$h_{\text{全局}} = \frac{1}{n}\sum_{i=1}^{n} h_i \tag{4-3}$$

4.3 光谱预测的实现

在 MassFormer 模型中，光谱预测是将分子嵌入与光谱元数据结合，通过一个大型的多层感知器（MLP）进行的。光谱预测的输出以稀疏正向量的形式表示，每个维度代表一个峰的位置，幅度对应峰的强度。

4.3.1 结合分子嵌入与光谱元数据

- 分子嵌入：通过图变压器处理后的分子嵌入表示整体分子的特征。例如，乙醇分子的全局嵌入向量为 [0.34, 0.52, 0.2]。
- 光谱元数据：包括前体信息、仪器参数和实验条件。例如，光谱元数据的向量表示为 [0.1, 0.2]。
- 结合嵌入：将分子嵌入向量与光谱元数据向量连接，形成新的输入向量 [0.34, 0.52, 0.2, 0.1, 0.2]。

4.3.2 输入到 MLP 进行预测

- MLP 架构：通过多层感知器（MLP）对输入向量进行处理。假设 MLP 有两层，第一层有 5 个神经元，第二层有 3 个神经元。
- 计算过程：

$$\text{第一层输出} = \sigma(W_1 \cdot [0.34, 0.52, 0.2, 0.1, 0.2] + b_1)$$
$$\text{第二层输出} = \sigma(W_2 \cdot \text{第一层输出} + b_2)$$

其中，W_1 和 W_2 是权重矩阵，b_1 和 b_2 是偏置向量，σ 是激活函数。

4.3.3 光谱预测输出

- 稀疏正向量表示：输出的光谱预测以稀疏正向量的形式表示，每个维度代表一个峰的位置，幅度对应峰的强度。例如，输出向量为 [0.0, 1.5, 0.0, 0.0, 0.8]，表示在第二个位置和第五个位置有峰，强度分别为 1.5 和 0.8。
- 解释输出：光谱中的每个峰位置对应于质谱图中的一个特定质量/电荷比（m/z）。通过模型预测的稀疏向量，可以生成对应的质谱图。

4.4 论文结论

1. **新颖的模型**：提出了一种新颖的图变压器模型 MassFormer，用于小分子的串联质谱预测。该模型通过将分子表示为图结构，并结合多头自注意力机制（MHA）和多层感知器（MLP）来捕捉分子内部复杂的关系。
2. **化学预训练任务**：MassFormer 通过化学预训练任务来初始化变压器模块的参数，然后在光谱数据上进行微调。这种预训练策略使得模型能够更好地泛化到新的数据集上。
3. **实验结果**：实验结果表明，MassFormer 在多个公开数据集（如 NIST 2020 和 MoNA）上均优于现有的最先进方法。具体来说，MassFormer 在光谱预测任务中展示了更高的平均余弦相似度，这表明其预测结果与实际质谱图更加一致。
4. **碰撞能量的影响**：验证了模型对碰撞能量变化的预测能力。实验结果显示，MassFormer 能够有效地模拟不同碰撞能量下的光谱变化，这进一步证明了其在真实应用中的潜力。
5. **解释性分析**：通过基于梯度的归因方法，展示了 MassFormer 模型在识别谱图中峰的组成关系方面的能力。这种解释性分析有助于理解模型的预测过程，并增强其可解释性。

6. **未来工作**：未来的研究方向包括进一步优化模型结构和预训练任务，以及扩展模型应用到更多类型的质谱数据上。此外，结合更多实验条件和元数据的信息，有望进一步提升质谱预测的准确性。

第五章　MoMS-Net 模型

本研究旨在通过提出一种新的质谱预测网络（MoMS-Net），利用图神经网络（GNN）和结构基序（motifs）的信息来预测质谱，从而扩展现有的质谱数据库[5]。MoMS-Net 模型通过结合基序信息，能够有效地预测复杂分子的质谱，并且在内存使用方面比图变换模型更加高效。

5.1 MoMS-Net 模型架构

MoMS-Net 模型由两个主要部分组成：分子图神经网络和异质基序图神经网络。两个部分的嵌入结果会被连接起来，然后通过多层感知机（MLP）层进行质谱预测。

5.1.1 分子图的表示

在分子图中，每个节点表示一个原子，每条边表示一个化学键。分子图可以用图结构表示为 $G = (V, E)$，其中 V 是节点集合（原子），E 是边集合（键）。

5.1.2 图卷积网络（GCN）

图卷积网络是一种用于处理图结构数据的神经网络。它通过卷积操作将每个节点的特征与其邻居节点的特征进行组合，生成新的节点表示。

5.1.2.1 节点特征初始化

每个节点的初始特征表示为一个向量，包含该原子的属性，如原子类型、度数、价态、形式电荷等。例如，碳原子和氢原子的初始特征可以分别表示为：

$$h_C = [1, 0, 0, \ldots, 0] \quad \text{（碳原子的 one-hot 编码）}$$

$$h_H = [0, 1, 0, \ldots, 0] \quad \text{（氢原子的 one-hot 编码）}$$

5.1.2.2 图卷积层

在每一层卷积操作中，每个节点的特征向量会与其邻居节点的特征向量进行组合，生成新的节点特征。图卷积的公式为：

$$h_v^{(k+1)} = \sigma\left(\sum_{u \in \mathcal{N}(v)} \frac{1}{\sqrt{d_v d_u}} W^{(k)} h_u^{(k)} + W^{(k)} h_v^{(k)}\right) \tag{5-1}$$

其中，$h_v^{(k)}$ 表示节点 v 在第 k 层的特征向量，$\mathcal{N}(v)$ 表示节点 v 的邻居节点集

合，d_v 和 d_u 分别是节点 v 和 u 的度数，$W^{(k)}$ 是第 k 层的权重矩阵，σ 是激活函数（如 ReLU）。

5.1.2.3 聚合操作

通过多层图卷积操作，GCN 能够逐步聚合每个节点的局部邻域信息，形成更高层次的特征表示。最终，得到的节点特征表示将包含更多全局结构信息。

5.2 异质基序图神经网络

异质基序图神经网络使用图同构网络（GIN）处理异质基序图，节点表示基序和分子，边表示它们之间的关系。

5.2.1 异质基序图的表示

异质基序图由两类节点和两类边构成：

- 节点：基序（Motifs）和分子（Molecules）。
- 边：分子-基序边和基序-基序边。分子-基序边表示分子包含特定基序，基序-基序边表示两个基序之间共享至少一个原子。

5.2.2 节点特征初始化

每个节点（基序或分子）的初始特征表示为一个向量，包含节点的属性信息。例如，对于一个基序节点，其初始特征可以表示为该基序在分子中的出现频率和分子量等信息。

5.2.3 边权重计算

- 分子-基序边权重使用 TF-IDF（词频-逆文档频率）计算，反映基序在分子中的重要性。

$$\text{TF-IDF}_{ij} = C(i)j\left(\log\frac{1+M}{1+N(i)}+1\right) \tag{5-2}$$

- 基序-基序边权重使用 PMI（点互信息）计算，反映基序之间的共现关系。

$$\text{PMI}_{ij} = \log\frac{p(i,j)}{p(i)p(j)} \tag{5-3}$$

5.2.4 图同构层

GIN 通过聚合每个节点的特征与其邻居节点的特征来更新节点表示。GIN 的公式为：

$$h_v^{(k+1)} = \mathrm{MLP}^{(k)}\left((1+\varepsilon)h_v^{(k)} + \sum_{u\in\mathcal{N}(v)} h_u^{(k)}\right) \tag{5-4}$$

其中，$h_v^{(k)}$ 是第 k 层节点 v 的特征向量，$\mathcal{N}(v)$ 是节点 v 的邻居节点集合，ε 是一个可学习的参数，$\mathrm{MLP}^{(k)}$ 是多层感知机。

5.2.5 结合基序质谱信息

在得到图的嵌入表示后，将基序的质谱信息结合到模型中，因为基序的碎片化模式与质谱有直接关联。这样可以进一步优化节点表示，提升质谱预测的准确性。

5.3 嵌入连接与预测

MoMS-Net 模型的嵌入连接与预测部分将两个图神经网络的嵌入结果连接起来，通过多层感知机（MLP）层进行质谱预测。

5.3.1 两个图神经网络的嵌入结果

MoMS-Net 模型包含两个图神经网络：

- 分子图神经网络（使用 GCN 处理）：生成分子图的嵌入表示。
- 异质基序图神经网络（使用 GIN 处理）：生成异质基序图的嵌入表示。

这两个图神经网络的输出分别是每个节点（原子或基序）的高维特征表示。

5.3.2 嵌入结果的连接

两个图神经网络的嵌入结果需要进行连接，以整合分子结构和基序信息：

- 分子图的嵌入：H_{molecule}
- 异质基序图的嵌入：H_{motif}

将这两个嵌入连接起来，形成一个综合的特征表示：

$$H_{\text{combined}} = [H_{\text{molecule}}, H_{\text{motif}}]$$

5.3.3 多层感知机（MLP）层

连接后的嵌入特征 H_{combined} 需要通过多层感知机（MLP）进行处理，以最终生成质谱预测结果。MLP 层通常包含多个全连接层，每层后接一个非线性激活函数

（如 ReLU）。

MLP 层的公式为：

$$H_{\text{output}} = \sigma\left(W^{(2)}\sigma\left(W^{(1)}H_{\text{combined}} + b^{(1)}\right) + b^{(2)}\right) \tag{5-5}$$

其中，$W^{(1)}, W^{(2)}$ 是权重矩阵，$b^{(1)}, b^{(2)}$ 是偏置项，σ 是激活函数。

5.3.4 预测质谱

通过 MLP 层处理后，最终的输出 H_{output} 表示质谱预测结果。这个结果通常是一个向量，其中每个元素代表质谱中某一 m/z 值的相对强度。

5.4 结论

在本论文中，我们提出了基于基序的质谱预测网络（MoMS-Net），该模型通过分子结构中的基序信息来预测质谱。基序在预测分子性质的任务中起着重要作用，因为它们直接与分子中的功能基团相关，并提供了关于分子之间关系的宝贵信息。我们应用合并和更新方法从数据集中生成基序词汇表，以表示各种基序大小和功能基团。我们对不同大小的基序词汇表和不同的模型架构进行了测试。MoMS-Net 在从分子结构预测质谱方面优于其他深度学习模型。尽管 GNN 在考虑长距离依赖性方面存在局限性，但我们的模型通过在图级别引入基序来有效地考虑长距离依赖性。此外，我们的模型与图转换器相比需要更少的内存。我们发现，基序的真实质谱在预测分子的质谱时非常有用，尽管预测的质谱可能包含更多的小峰和错误峰。在未来的工作中，我们将努力改进基序质谱的初始化方法，并引入正则化技术以防止错误峰。此外，我们计划将 MoMS-Net 应用于更大的分子和蛋白质。

第六章　3DMolMS 模型

质谱技术，特别是与气相色谱（GC）或液相色谱（LC）联用的质谱（MS），已经被广泛应用于化学化合物的表征和结构解析。串联质谱（LC-MS/MS）能够检测化合物在碰撞池中发生高能碰撞后生成的碎片离子，是一种用于复杂样品中化合物鉴定和定量的关键技术。然而，当前的质谱数据参考库有限，许多新化合物的 MS/MS 谱图无法通过现有方法识别。因此，研究人员提出了一些计算方法来预测化合物的 MS/MS 谱图，以扩展参考光谱库，从而改进化合物鉴定[6]。

6.1 3DMolMS 模型

3DMolMS 模型是一种基于 3D 分子卷积操作的深度神经网络，专为从化合物的 3D 构象预测其 MS/MS 谱图而设计。模型的结构和方法如下：

6.1.1 输入编码

输入的点集包括每个原子的 x, y, z 坐标以及一系列原子属性，如原子类型、邻接原子数、原子质量等。

6.1.2 3D 分子卷积（3DMolConv）

这是模型的核心操作。通过对点集应用对称函数来提取 3D 结构的全局特征。这种操作不仅考虑了点之间的距离，还考虑了点之间的方向信息，从而捕捉到化学键和非键角度的信息。

公式如下：

$$f(\{x_1, x_2, \ldots, x_n\}) = g(h(x_1), h(x_2), \ldots, h(x_n)) \tag{6-1}$$

其中，f 是输入点集的表示函数，h 是对每个元素 x_i 的基本操作，g 是聚合函数。我们采用最大池化函数作为 g，并采用序列卷积层作为 h。

具体来说，每一层的基本操作（称为 3DMolConv）如下：

$$x_i^{l+1} = x_i^l + \sum_{j \in N(x_i^l)} d(x_i^l, x_j^l) W_1^l \cdot \varphi(x_i^l, x_j^l) W_2^l \cdot x_j^l \tag{6-2}$$

其中，W_1^l 和 W_2^l 分别是距离和方向的滤波器，x_j^l 是 x_i 的最近邻居之一。距离 $d(x_i, x_j)$ 计算为：

$$d(x_i, x_j) = \|x_i - x_j\| \tag{6-3}$$

角度 $\varphi(x_i, x_j)$ 计算为：

$$\varphi(x_i, x_j) = \sum_{k \in N(x_i)} e_{ij}^T e_{ik} \tag{6-4}$$

其中，边 e_{ij} 的向量表示为：

$$e_{ij} = x_i^T x_j \tag{6-5}$$

6.1.3 编码器

编码器由六个 3DMolConv 层组成，每层的输出维度依次增大，从 64 维到 1024 维。编码器提取的特征向量与元数据（如前体类型和碰撞能量）结合，形成最终的潜在表示向量。

6.1.4 解码器

解码器由五个全连接层组成，输入潜在表示向量并输出预测的 MS/MS 谱图。

6.2 数据处理

- 过滤掉峰数少于 5 的质谱。
- 限制质荷比（m/z）范围在 0 到 1500 之间。
- 仅保留由常见原子（如 C, H, O, N, F 等）组成的分子。
- 只保留常见前体类型的谱图。

6.3 模型训练

3DMolMS 模型在 PyTorch 框架下实现，训练过程中采用余弦相似度作为损失函数。为了增强模型的泛化能力，模型在不同的实验条件下（如不同的碰撞能量）进行训练和微调。

6.4 预测性能

在实验数据上的评估结果显示，3DMolMS 模型能够高精度预测 MS/MS 谱图，且其性能优于现有的质谱预测算法（如 NEIMS、MassFormer 和 CFM-ID 4.0）。具体表现为：

- 正离子模式下的余弦相似度为 0.691，负离子模式下为 0.478。
- 通过少量谱图的微调，模型能够适应不同实验室和不同仪器获取的 MS/MS 谱图。
- 学习到的分子表示可以用于增强其他化学性质的预测，如液相色谱的洗脱时

间和离子迁移谱的碰撞截面。

6.5 结论

在本文中，我们设计了一种基于 3D 分子卷积操作的深度神经网络模型——3DMolMS，用于从化合物的 3D 构象预测其 MS/MS 谱图。通过对多个实验数据集的评估，3DMolMS 模型表现出色，主要结论如下：

- **准确的 MS/MS 谱图预测**：3DMolMS 模型能够高精度预测质谱谱图，其预测性能优于现有的质谱预测算法，如 NEIMS、MassFormer 和 CFM-ID 4.0。正离子模式下的余弦相似度为 0.691，负离子模式下为 0.478。
- **泛化能力**：3DMolMS 模型通过对少量谱图进行微调，能够适应不同实验室和不同仪器获取的 MS/MS 谱图，显示出较强的泛化能力。
- **化学性质的预测**：3DMolMS 模型学习到的分子表示不仅能够用于谱图预测，还能用于其他化学性质的预测，如液相色谱的洗脱时间和离子迁移谱的碰撞截面。通过迁移学习，模型的预测精度得到了增强。

6.6 研究展望

- **扩展参考光谱库**：准确预测的谱图可以扩展现有的参考光谱库，与预测的碰撞截面和洗脱时间结合使用，有望进一步提高化合物鉴定的准确性。
- **大规模化合物鉴定应用**：未来的研究将致力于将 3DMolMS 模型应用于大规模化合物鉴定，特别是非靶向代谢组学中的应用。
- **改进模型性能**：通过增加训练数据集的覆盖范围和多样性，3DMolMS 模型性能仍有进一步提升的空间。

第七章　SCARF 模型

质谱分析在生物样本中发现临床相关的代谢产物中起着关键作用，但现有的预测工具存在一些局限性。这篇论文提出了一种新的中间策略，通过将质谱视为分子公式的集合来进行预测，每个公式由一组原子组成[7]。

7.1 前缀树生成过程

SCARF-Thread 通过以下步骤生成产品公式：

- **节点表示**：每个节点表示当前前缀公式。我们使用一个嵌入向量来表示这个前缀，记为 c'。
- **多标签二分类问题**：在每个节点，我们将预测可能的下一个原子的计数，将其视为多标签二分类问题。即，对于每个原子类型，预测其计数是 0 还是一个正整数。
- **递归生成**：从根节点开始，逐步扩展每个节点，直到生成完整的产品公式。

7.2 具体实现

7.2.1 输入表示

对于一个原始分子 M，我们将其嵌入为一个上下文向量 c。这个上下文向量包含了分子图的嵌入表示，以及当前前缀公式的信息。

7.2.2 前缀嵌入

在每一步，我们将当前前缀的嵌入向量 c' 与上一步预测的原子计数进行拼接，形成新的输入向量。

7.2.3 神经网络预测

使用一个多层感知器（MLP）预测每个原子的可能计数，形成下一层节点。

具体公式如下：

$$c' = [\text{gnn}(M), \text{counts}(f'_{<j}), \text{counts}(F - f'_{<j}), \text{one-hot}(j)] \tag{7-1}$$

其中，

- $\text{gnn}(M)$ 是原始分子图的嵌入表示，
- $\text{counts}(f'_{<j})$ 表示当前前缀公式的计数，
- $\text{counts}(F - f'_{<j})$ 表示原始公式减去当前前缀的计数，

- one-hot(j) 是当前预测的原子类型的 one-hot 编码。

预测公式为：

$$p(f'_j = a|f'_{<j}, M) = \alpha\sigma(\text{MLP}_F(c'))_a + (1-\alpha)\sigma(\text{MLP}_D(c'))_{F_j - a} \tag{7-2}$$

其中，

- MLP_F 和 MLP_D 分别用于预测原子计数和差分计数的多层感知器，
- α 是一个加权系数，表示这两种预测的权重，
- σ 是 sigmoid 函数。

7.3 输入表示

对于每个生成的产品公式，我们使用一个上下文向量 c' 来表示。这个上下文向量包含以下信息：

- 原始分子的嵌入表示（通过图神经网络得到）。
- 产品公式的计数表示。
- 产品公式与前缀公式的差异表示。

具体公式如下：

$$c' = [\text{gnn}(M), \text{counts}(f_i), \text{counts}(F - f_i)] \tag{7-3}$$

其中，

- $\text{gnn}(M)$ 是原始分子图的嵌入表示，
- $\text{counts}(f_i)$ 表示产品公式的计数，
- $\text{counts}(F - f_i)$ 表示原始公式减去产品公式的计数。

7.4 Set Transformer

我们使用 Set Transformer 结构进行强度预测。这是一种适用于集合到集合映射的问题的神经网络结构，能够考虑到集合中元素之间的相互作用。

7.5 强度预测

7.5.1 输入嵌入

我们将每个产品公式表示为一个向量，包含原始分子图的嵌入表示和产品公式的计数表示。

7.5.2 Transformer 编码

使用 Set Transformer 对这些输入嵌入进行编码，捕捉公式之间的相互关系。

7.5.3 强度预测

通过编码后的表示，预测每个公式对应的强度值。

预测公式为：

$$y_i = \text{SetTransformer}(c') \tag{7-4}$$

这里，y_i 是产品公式 f_i 对应的强度值。

7.6 结论概述

这篇论文提出了一种新的方法，称为 SCARF（Subformulae Classification for Autoregressively Reconstructing Fragmentations），用于从分子预测质谱。通过使用前缀树结构解码分子子公式，再通过 Set Transformer 预测这些公式的质谱峰强度，SCARF 成功地在准确性和效率上超越了现有的方法。

7.7 主要贡献

- **两步预测过程：**
  - 首先生成分子子公式集合。
  - 然后预测每个子公式的质谱峰强度。
- **前缀树结构的使用：**
  - 通过前缀树逐步解码分子子公式，有效地减少了组合爆炸问题。
  - 这种方法能够保持物理合理性，避免生成无效的质谱峰。
- **Set Transformer 的应用：**
  - Set Transformer 能够捕捉分子子公式之间的相互关系，提高了强度预测的准确性。
  - 这种结构适用于集合到集合的映射，能够有效处理质谱峰的预测任务。

7.8 实验结果

论文在两个独立的数据集（NIST20 和 NPLIB1）上进行了实验证明，结果表明 SCARF 在质谱预测任务中表现优越：

- **准确性：**
  - SCARF 在实验中表现出较高的预测准确性，尤其是在复杂分子结构的

预测中。

- 与现有的基于片段化和离散化的方法相比，SCARF 在精度上有显著提升。

- **物理合理性：**
  - SCARF 的预测结果能够提供合理的物理解释，避免生成不符合化学规则的质谱峰。
  - 这种物理合理性有助于研究人员更好地理解和解释质谱数据。
- **计算效率：**
  - SCARF 在预测速度上也有明显优势，能够高效地处理大量的分子数据。
  - 实验结果显示，SCARF 比传统方法快了几个数量级，使其在大规模数据处理和实时应用中更具优势。

7.9 未来工作

论文指出，未来的研究方向包括：

- **进一步优化模型结构：**
  - 探索更多种类的神经网络结构和注意力机制，以进一步提高预测精度。
  - 研究如何在更大规模和更复杂的数据集上应用 SCARF。
- **扩展应用领域：**
  - 将 SCARF 应用于其他类型的质谱数据，如蛋白质质谱和代谢物质谱。
  - 探索 SCARF 在其他科学领域中的潜在应用，如材料科学和药物发现。
- **结合其他数据源：**
  - 结合其他类型的数据（如核磁共振数据和红外光谱数据），提高分子结构解析的准确性。
  - 开发多模态模型，将不同类型的数据融合在一起，提供更全面的分子解析工具。

第八章　ICEBERG 模型

在化学和生物学研究中，质谱分析（特别是串联质谱 MS/MS）是一种重要的技术，用来识别和分析分子。通过将分子打碎成碎片，我们可以得到一张碎片光谱图。这张图上的每一个峰值都代表了分子的一个碎片。这些碎片的信息可以帮助科学家识别分子的结构[8]。

8.1 生成模块（Generate）

生成模块（Generate）通过递归地预测分子中的打碎事件来生成分子的可能碎片。每个打碎事件都是指从分子中移除某个原子或断开某个键，从而产生新的分子碎片。这个过程被重复进行，直到生成一系列的分子碎片。

8.1.1 有向无环图（DAG）

生成模块的核心是构建一个有向无环图（DAG），该图代表了分子的所有可能碎片。在这个图中，根节点是原始分子，每个子节点是通过断开某个键或移除某个原子生成的碎片。

8.1.2 使用图神经网络（GNN）

为了预测打碎事件，生成模块使用了图神经网络（GNN）对分子进行编码。GNN 能够捕捉分子的结构信息，并为每个原子生成一个嵌入向量，这些向量表示该原子的特征。

8.1.3 预测打碎事件的公式

生成模块预测每个原子周围的断键概率。具体步骤如下：

8.1.3.1 初始分子编码

将输入分子编码为一个图结构，图中的节点表示原子，边表示化学键。

8.1.3.2 图神经网络嵌入

使用 GNN 对分子图进行编码，生成每个原子的嵌入向量。令 $GNN(M)$ 表示分子的图嵌入，$GNN(S(i))$ 表示第 i 个碎片的图嵌入。对每个原子的嵌入向量计算如下：

$$GNN(S(i))_j$$

这里，$GNN(S(i))_j$ 表示在第 i 个碎片中，第 j 个原子的图嵌入。

8.1.3.3 计算断键概率

为了预测每个原子的断键概率，我们使用以下公式：

$$p\left(F[S(i)_j] \mid S(i), M, C\right) = g_{\text{Generate},\theta}(M, S(i), C)_j \tag{8-1}$$

在这个公式中：

- $F[S(i)_j]$ 表示在第 i 个碎片中，第 j 个原子周围的断键事件。
- $g_{\text{Generate},\theta}$ 表示生成模型的神经网络函数。
- M 表示原始分子，$S(i)$ 表示当前的分子碎片，C 表示上下文信息（如离子化附加物类型）。

为了计算断键概率，我们将以下信息进行拼接，并使用多层感知器（MLP）进行预测：

$$\begin{aligned} p\left(F[S(i)_j] \mid S(i), M, C\right) = \text{MLP}\,([\,& GNN(M), \\ & GNN(M) - GNN(S(i)), \\ & GNN(S(i))_j, \\ & \text{Onehot}(b), \\ & \text{Enc}(f_i), \\ & \text{Enc}(f_0 - f_i)]) \end{aligned} \tag{8-2}$$

在这个公式中：

- $GNN(M)$ 是原始分子的图嵌入。
- $GNN(M) - GNN(S(i))$ 表示原始分子和当前碎片的嵌入差异。
- $GNN(S(i))_j$ 表示当前碎片中第 j 个原子的嵌入。
- Onehot(b) 表示断裂的键的数量的独热编码。
- $\text{Enc}(f_i)$ 和 $\text{Enc}(f_0 - f_i)$ 分别表示当前碎片和原始分子的化学式的编码。

8.1.4 递归生成

生成模块递归地应用上述步骤，对每个生成的分子碎片继续进行打碎预测，直到生成一个包含所有可能碎片的 DAG。通过这种方式，模型能够高效地生成分子的所有可能碎片，而不需要穷举所有可能的打碎组合。

8.1.5 示例

假设输入分子是苯甲酸乙酯（benzocaine），生成模块的工作流程如下：

1. 初始编码：将苯甲酸乙酯编码为图结构。

2. 图神经网络嵌入：使用 GNN 生成每个原子的嵌入向量。
3. 预测断键：计算每个原子的断键概率，例如预测 C-O 键最有可能断裂。
4. 生成碎片：断开 C-O 键，生成新的分子碎片。
5. 递归处理：对生成的每个碎片重复上述步骤，继续预测和断裂新的化学键，生成更多碎片。

8.1.6 物理约束

生成模块结合了物理约束，以确保生成的碎片符合实际的化学反应规律。例如，模块只允许生成化学上稳定的碎片，并考虑可能的氢转移和同位素效应。

8.2 评分模块（Score）

评分模块（Score）的目标是对每个生成的分子碎片进行评分，并预测其在质谱中的强度。这个过程需要考虑到每个碎片的质量以及可能的氢重排和同位素效应。

8.2.1 使用 Set Transformer 架构

为了处理碎片的评分和强度预测任务，Score 模块使用了 Set Transformer 架构。Set Transformer 是一种能够处理集合数据的 Transformer 模型，它具有不变性和等变性，适合处理分子碎片这样的集合数据。

8.2.2 碎片的特征表示

对每个生成的分子碎片，我们使用多层感知器（MLP）来生成每个碎片的隐藏表示：

$$\begin{aligned} h_i = \text{MLP}\,([\,&GNN(M),\\ &GNN(M) - GNN(S(i)),\\ &GNN(S(i)),\\ &\text{Onehot}(b),\ \text{Enc}(f_i),\ \text{Enc}(f_0 - f_i)) \end{aligned} \tag{8-3}$$

在这个公式中：

- h_i 是第 i 个碎片的隐藏表示。
- $GNN(M)$ 是原始分子的图嵌入。
- $GNN(M) - GNN(S(i))$ 表示原始分子和当前碎片的嵌入差异。
- $GNN(S(i))$ 表示当前碎片的图嵌入。
- Onehot(b) 表示断裂的键的数量的独热编码。

- $\text{Enc}(f_i)$ 和 $\text{Enc}(f_0 - f_i)$ 分别表示当前碎片和原始分子的化学式的编码。

8.2.3 预测碎片强度

使用 Set Transformer 对所有生成的碎片进行编码，并预测每个碎片在不同氢转移情况下的强度：

$$y_{i\delta} = g_{\text{Score},\theta}(M, S(i), T, C)_\delta y_{i\delta} = g_{\text{Score},\theta}(M, S(i), T, C)_\delta \tag{8-4}$$

在这个公式中：

- $y_{i\delta}$ 是第 i 个碎片在氢转移 δ 情况下的预测强度。
- $g_{\text{Score},\theta}$ 是 Score 模块的神经网络函数。
- M 是原始分子，$S(i)$ 是第 i 个分子碎片，T 是生成的有向无环图（DAG），C 是上下文信息。

8.2.4 使用注意力机制加权预测

为了综合不同氢转移情况下的预测强度，Score 模块使用了注意力机制：

$$\alpha_{i\delta} = \text{Softmax}_{k \in M(i,\delta)} \left(\text{MLP}_{\text{attn}} \left(\text{Transformer}(h_0, h_1, \ldots, h_{|T|})_k \right) \right)_{i,\delta} \tag{8-5}$$

在这个公式中：

- $\alpha_{i\delta}$ 是第 i 个碎片在氢转移 δ 情况下的注意力权重。
- MLP_{attn} 是用于计算注意力权重的多层感知器。
- Transformer 是 Set Transformer 模块，用于对所有碎片进行编码。

最终的强度预测是不同氢转移情况下的预测强度的加权和：

$$y_m = \sigma \left(\sum_i \sum_\delta \alpha_{i\delta} y_{i\delta} I[M(i,\delta) = m] \right) \tag{8-6}$$

在这个公式中：

- y_m 是质量为 m 的最终强度预测。
- σ 是 Sigmoid 激活函数。
- $I[M(i,\delta) = m]$ 是一个指示函数，当第 i 个碎片在氢转移 δ 情况下的质量等于 m 时，该函数值为 1。

8.3 结论

这篇论文提出了一种新的分子碎片图生成方法，称为 ICEBERG（Inferring Collision-induced-dissociation by Estimating Breakage Events and Reconstructing their

Graphs），其主要特点如下：

1. **方法创新**：ICEBERG 利用神经网络模拟分子碎片生成和评分的过程。该方法结合了化学信息学中的经典方法和现代神经网络技术，通过预测可能的断裂事件和对结果片段进行评分，来实现对质谱数据的高精度预测。
2. **模型性能**：在 NPLIB1 和 NIST20 两个数据集上的评估结果表明，ICEBERG 在预测光谱的余弦相似度方面优于现有的其他方法。例如，在 NPLIB1 数据集上，ICEBERG 的平均余弦相似度达到 0.627，相比于 MassFormer 和 FixedVocab 的 0.568 提高了 10%。
3. **光谱检索**：ICEBERG 在光谱检索任务中表现出色，特别是在复杂天然产物分子上。相比于次优模型，ICEBERG 在 NPLIB1 数据集上的 Top 1 检索准确率提升了 9%（相对提升 46%），在 NIST20 数据集上的 Top 10 检索准确率也有超过 5% 的提升（相对提升 7.5%）。
4. **结果解释性**：ICEBERG 能够解释每个预测峰值对应的分子碎片，这在处理复杂分子时尤为重要。通过观察某些模式和示例，可以发现预期的断键情况，例如碳氧和碳氮键更容易断裂。
5. **未来工作**：尽管 ICEBERG 在准确性方面表现出色，但其计算成本较高。未来的工作将考虑如何更有效地结合片段预测和公式预测方法，以实现更高的准确性和速度。此外，模型准确性还可以通过建模其他变量（如碰撞能量、附加类型等）进一步提升。

总之，ICEBERG 方法在分子质谱预测和检索任务中表现出色，展示了其在复杂分子结构解析中的巨大潜力。

第九章　FraGNNet 模型

论文主要研究了一种新的深度概率模型，用于预测质谱图。论文介绍了从复杂混合物中识别小分子的重要性，尤其是在液体样本的化学组成分析中。质谱（MS/MS）是广泛应用于分子识别的工具，但现有的方法通常依赖于与已知光谱库的匹配，而这些光谱库的覆盖面有限[9]。

9.1 递归碎裂

步骤概述：

- FraGNNet 首先将分子表示为一个无向图，其中节点代表分子的原子，边代表原子之间的共价键。这个图称为分子的重原子骨架图（Heavy Atom Skeleton）。
- 然后，使用递归边移除算法生成可能的分子碎片。

具体细节：

- **分子图表示：**分子表示为无向图 $G = (V, E)$，其中 V 是节点集合，代表分子的原子，E 是边集合，代表原子之间的共价键。
- **重原子骨架图：**定义为不含氢原子的最大连通子图 $H(G)$，通过移除氢原子和相关边得到。
- **碎片生成：**通过递归地移除边，生成所有可能的连通子图。每个子图代表一个可能的分子碎片。
- **近似碎裂图：**由于完全枚举所有可能的子图计算量非常大，FraGNNet 采用了近似方法，只考虑距根节点 r d 跳以内的节点，生成一个近似碎裂图 $F_d(H(G))$。

公式和定义：

- 近似碎裂图定义为 $F_d(H(G)) = (V_{Fd}, E_{Fd})$，其中每个节点 $n \in V_{Fd}$ 对应一个重原子子图 $G_n \in S(G)$。
- 通过对每个子图 G_n 的氢原子进行建模，计算可能的化学式集合 $\{f^n_{-j}, \ldots, f^n_j\}$ 和相应的质量集合 $\{m^n_{-j}, \ldots, m^n_j\}$，其中 j 是氢原子数的容差参数。

9.2 概率模型

步骤概述：

- 使用图神经网络（GNN）参数化的概率模型预测这些碎片的分布。
- 通过预测每个碎片的概率分布，生成整体的质谱分布。

具体细节：

- **概率分布定义：**
  - $P_\theta(n)$：对碎裂图节点 n 的离散概率分布。
  - $P_\theta(f|n)$：条件概率分布，给定节点 n 的化学式 f 的概率。
- **联合分布**：通过联合分布 $P_\theta(n,f) = P_\theta(n)P_\theta(f|n)$ 表示不同碎片的生成过程。
- **图神经网络（GNN）参数化：**
  - **分子 GNN**（Molecule GNN）：对输入分子图 G 进行操作，生成原子和键的嵌入向量。
  - **碎片 GNN**（Fragment GNN）：结合碎裂图和分子嵌入信息，预测 $P_\theta(n)$ 和 $P_\theta(f|n)$。
  - GNN 使用多层感知器（MLP）处理节点和边信息，通过迭代更新节点状态，生成最终的嵌入表示。

公式和定义：

- Molecule GNN 的更新规则：

$$h_a^{(l+1)} = \text{MLP}\left(h_a^{(l)} + \sum_{u \in N(a)} \text{ReLU}(h_u^{(l)} + h_b^{(l)})\right) \tag{9-1}$$

- 其中，$h_a^{(l)}$ 是第 l 层中节点 a 的嵌入向量，$N(a)$ 是节点 a 的邻居节点集合，$h_b^{(l)}$ 是边的嵌入向量。

9.3 高分辨率质谱预测

步骤概述：

- 将化学式的分布转换为质谱分布，实现高分辨率峰值预测。
- 通过预测的质谱分布，生成最终的质谱图。

具体细节：

- **质谱表示**：质谱 Y 表示为一组元组 $\{(m_j, P(m_j))\}_j$，每个质量 m_j 具有相应的概率 $P(m_j)$。
- **分布转换**：通过计算每个化学式 f 对应的质量 m，将化学式分布 $P_\theta(f)$ 转换为质量分布 $P_\theta(m)$。
- **高斯混合模型**：质谱分布 $P_\theta(m)$ 表示为一组一维高斯分布的混合：

$$P_\theta(m) = \sum_f P_\theta(f) N(\mu(f), \sigma(f)) \tag{9-2}$$

- 其中 $\mu(f) = \text{mass}(f)$ 是化学式 f 的质量，$\sigma(f)$ 是标准差。

公式和定义：

- 质谱分布的计算：

$$P_\theta(m) = \sum_f P_\theta(f) N(\mu(f), \sigma(f)) \tag{9-3}$$

- 使用狄拉克 δ 函数简化计算：

$$P(m|f) = \delta(\text{mass}(f)) \tag{9-4}$$

通过以上步骤，FraGNNet 实现了高效且精确的质谱预测，能够生成高分辨率的质谱图，并且具有良好的可解释性和扩展性。

9.4 结论

9.4.1 主要贡献

- 论文介绍了一种新的深度概率模型 FraGNNet，用于质谱预测。该模型结合了组合碎裂方法和概率建模，能够高效、准确地预测高分辨率质谱图。
- FraGNNet 使用结构化的潜在空间，提供了对定义质谱的底层过程的洞察。通过这种方式，FraGNNet 不仅在预测误差方面达到了最新性能，还在基于检索的质谱到化合物（MS2C）任务中超过了现有的 C2MS 模型。

9.4.2 实验结果

- 通过与强基线模型的比较，论文展示了 FraGNNet 在质谱预测和化合物检索任务中的优越性能。
- 实验结果表明，FraGNNet 能够有效提高质谱预测的准确性，并在化合物检索任务中表现出色，验证了其在实际应用中的潜力。

9.4.3 模型优势

- FraGNNet 在分辨率、可扩展性和可解释性方面满足了关键要求。
- 该模型能够预测高分辨率的质谱图，并通过结构化潜在空间提供对质谱生成过程的深入理解。
- 此外，FraGNNet 具备可扩展性，能够应用于大规模光谱库的生成和补充，显著提高了检索基于质谱的化合物识别的覆盖率。

9.4.4 未来工作方向

- 虽然 FraGNNet 在许多方面表现出色，但仍有改进的空间。论文指出，计算碎裂图是一个计算瓶颈，需要进行递归的边移除操作。这可能会限制模型在

更大规模数据集上的应用。

- 另一个需要解决的问题是复杂反应（如环化反应）的建模，这些反应引入了碎片空间的组合爆炸，增加了模型的复杂性。
- 未来的研究可以探索使用化学反应采样的方法来识别无法通过简单键断裂生成的碎片。或者，FraGNNet 可以与更灵活的 C2MS 模型（如离散化预测模型）进行集成，以捕获更多的质谱峰值。
- 论文还建议增加对未合并光谱预测的支持和扩大前体加合物覆盖范围，以提高 FraGNNet 在实际质谱到化合物任务中的适用性。

9.5 总结

FraGNNet 通过将组合碎裂方法与神经网络相结合，提出了一种新颖且高效的质谱预测模型。该模型不仅在质谱预测和化合物检索任务中取得了最新的性能，还通过结构化潜在空间提供了对质谱生成过程的深入理解。论文的实验结果验证了 FraGNNet 在实际应用中的潜力，并为未来的研究指出了改进方向和新可能性。

通过这篇论文，研究团队展示了深度概率模型在质谱预测中的应用前景，并为进一步的研究提供了坚实的基础。

第十章　全文总结与展望

10.1 全文总结

本文研究了多种用于质谱预测的深度学习模型，分别是 NEIMS、CFM-ID 4.0、GRAFF-MS、MassFormer、MoMS-Net、3DMolMS、SCARF、ICEBERG 和 FraGNNet。这些模型利用不同的神经网络架构和特征表示方法，实现了高效、准确的质谱预测。具体而言，NEIMS 模型通过扩展圆形指纹和多层感知器实现快速准确的小分子质谱预测；CFM-ID 4.0 模型利用环裂解建模和连接矩阵特征提高质谱预测的准确性；GRAFF-MS 模型结合图神经网络和手写规则，显著提升质谱预测的精度；MassFormer 模型通过图变压器架构解决复杂分子关系的捕捉问题；MoMS-Net 模型利用异质基序图神经网络进行高效预测；3DMolMS 模型通过 3D 分子卷积操作实现高精度质谱谱图预测；SCARF 模型通过前缀树生成和 Set Transformer 架构在质谱预测中表现优越；ICEBERG 模型通过递归生成和评分模块实现高精度的分子碎片图生成；FraGNNet 模型结合组合碎裂方法和概率建模，高效、准确地预测高分辨率质谱图。本文验证了这些模型在实际应用中的潜力，并提出了未来的研究方向。

10.2 后续工作展望

尽管本文研究的模型在质谱预测任务中取得了显著进展，但仍有进一步改进的空间。未来的研究可以从以下几个方面展开：

1. **模型优化**：进一步优化现有模型的结构，提高计算效率，解决计算瓶颈问题，如递归的边移除操作。
2. **复杂反应建模**：研究如何更好地建模复杂反应（如环化反应），这些反应引入的碎片空间的组合爆炸增加了模型的复杂性。
3. **化学反应采样**：探索使用化学反应采样的方法来识别无法通过简单键断裂生成的碎片，或将 FraGNNet 与更灵活的 C2MS 模型（如离散化预测模型）集成，以捕获更多的质谱峰值。
4. **光谱预测支持**：增加对未合并光谱预测的支持和扩大前体加合物的覆盖范围，以提高在实际质谱到化合物任务中的适用性。
5. **多模态数据融合**：结合其他类型的数据（如核磁共振数据和红外光谱数据），开发多模态模型，将不同类型的数据融合在一起，提供更全面的分子解析工具。

通过解决这些领域的问题，未来的研究可以进一步推进质谱预测领域的发展，

提高这些模型的实际应用性和准确性。

致　　谢

感谢 ChatGPT 在格式问题，中英文摘要以及公式理解方面给予的极大帮助。

参考文献

[1] J. N. Wei, D. Belanger, R. P. Adams, et al. Rapid prediction of electron–ionization mass spectrometry using neural networks[J]. ACS central science, 2019, 5(4): 700-708

[2] F. Wang, J. Liigand, S. Tian, et al. Cfm-id 4.0: more accurate esi-ms/ms spectral prediction and compound identification[J]. Analytical chemistry, 2021, 93(34): 11692-11700

[3] M. Murphy, S. Jegelka, E. Fraenkel, et al. Efficiently predicting high resolution mass spectra with graph neural networks[C]. International Conference on Machine Learning, 2023, 25549-25562

[4] A. Young, B. Wang, H. Röst. Massformer: Tandem mass spectrum prediction for small molecules using graph transformers[J]. arXiv preprint arXiv:2111.04824, 2021,

[5] J. Park, J. Jo, S. Yoon. Mass spectra prediction with structural motif-based graph neural networks[J]. Scientific Reports, 2024, 14(1): 1400

[6] Y. Hong, S. Li, C. J. Welch, et al. 3dmolms: prediction of tandem mass spectra from 3d molecular conformations[J]. Bioinformatics, 2023, 39(6): btad354

[7] S. Goldman, J. Bradshaw, J. Xin, et al. Prefix-tree decoding for predicting mass spectra from molecules[J]. Advances in Neural Information Processing Systems, 2023, 36: 48548-48572

[8] S. Goldman, J. Li, C. W. Coley. Generating molecular fragmentation graphs with autoregressive neural networks[J]. Analytical Chemistry, 2024,

[9] A. Young, F. Wang, D. Wishart, et al. Fragnnet: A deep probabilistic model for mass spectrum prediction[J]. arXiv preprint arXiv:2404.02360, 2024,