质谱模拟思路

1. 问题定义

质谱预测是计算蛋白质组学中的核心任务，旨在根据给定的肽段序列精确预测其在质谱仪中碎

输入：肽段氨基酸序列（如"PEPTIDE"），可能包含翻译后修饰信息（如氧化、磷酸化等）。

输出：对应的预测质谱图。

技术挑战：传统方法将连续的m/z空间离散化为固定数量的bins，将回归问题转化为强度预测题：1、bins数量过多，计算开销大；2、绝大多数bins是没有对应的质谱峰的，输出过于稀疏

数据示例：

BEGIN IONS

TITLE=controllerType=0 controllerNumber=1 scan=4524

PEPMASS=515.937805175781

RTINSECONDS=1135.07892

CHARGE=3+

SCANS=4524

SEQ=LVHVEEPHTETVR

110.071434 94581.6953125

111.0752029 2357.2253417969

116.016983 857.7626953125

124.176239 819.0247192383

127.0207138 776.9647827148

129.101944 2067.8432617188

130.0864105 1099.2720947266

133.0431366 1262.3026123047

136.0755768 7162.3349609375

138.0563965 794.4822387695

138.0662384 5720.1733398438

其实可以看到，质谱本身就是一个标准的token格式

2. 研究现状

当前质谱预测方法主要基于深度学习，并普遍采用分bin策略：

分bin方法：将m/z范围（如0–2000 Da）划分为固定数量的区间（如30,000个bins），每个l

代表性工作：

- **Prosit**：使用Transformer编码器处理肽段，输出预定义bins的强度分布
- **pDeep**：采用LSTM网络，类似地预测binned质谱
- **MS2PIP**：使用梯度提升树或神经网络预测binned强度

3. 论文方案

3.1 模型结构

我们提出基于序列到序列的质谱生成框架，采用Transformer encoder–only架构处理输入肽段

- 氨基酸embedding（包括修饰）
- 位置embedding
- 保留时间embedding
- 电荷态embedding

输出为
[mz_token1, intensity_token1, mz_token2, intensity_token2, ..., SEP]

应当注意的是，质谱本身是没有很强的从左到右的因果关系的，其内部更倾向于上下文关系，型推理速度。

但实际训练时，应当进行测试，因为decoder是可以采用beam search来提高模型预测效果的

3.2 损失设计

问题：预测tokens 和真实tokens都是集合（无序），直接计算L1损失会导致：

- 错误对齐（如预测峰A匹配到真实峰B）。
- 未匹配峰的损失计算混乱。

匈牙利匹配解决方案：

1. 构建成本矩阵：

匹配成本矩阵：对于预测的N个峰和真实的M个峰，构建N×M成本矩阵，其中每个元素C_ij计

C_ij = λ1 × |mz_pred_i − mz_true_j| + λ2 × |intensity_pred_i − intensity_true_j|

矩阵中第一项为质荷比惩罚，第二项为强度惩罚。应当注意的是，实际使用时，化学专家通常

但实际运行中，通常采用填充处理输入，因此M==N。其中有N−M个填充峰。对于填充峰，应

2. 计算最优匹配：

基于成本矩阵C计算最小成本二分图匹配，然后作为损失。此外，可以额外增加一个余弦相似

3.3 论文创新点

1. **提出了质谱的序列化表示方法与生成式预测模型。** 我们摒弃了将连续m/z空间离散化为固
   于此，我们将质谱预测问题构建为一个序列到序列的生成任务，使模型能够直接输出连续
2. **设计了基于匈牙利匹配的集合预测损失函数。** 针对预测质谱峰与真实质谱峰之间无序对
   只对匹配的峰对进行梯度更新，有效解决了因峰序任意性导致的错误对齐问题，从而更精
3. **使用了**Encoder−only**架构获取上下文信息。** 我们探索并采用了Encoder−only的模型结构
   而不是单独的因果关系。

1419字