

中国科学技术大学

University of Science and Technology of China

硕士学位论文



论文题目

面向DIA质谱数据的

无色谱分析方法研究

作者姓名

代奇

学科专业

计算机应用技术

导师姓名

郑浩然 副教授

完成时间

二〇二四年五月

中国科学技术大学

硕士学位论文



面向 DIA 质谱数据的 无色谱分析方法研究

作者姓名： 代芎

学科专业： 计算机应用技术

导师姓名： 郑浩然 副教授

完成时间： 二〇二四年五月二十三日

University of Science and Technology of China
A dissertation for master's degree



Research on Chromatography-free Analysis Method for DIA Mass Spectrometry Data

Author: Ge Dai

Speciality: Computer Application Technology

Supervisor: Assoc. Prof. Haoran Zheng

Finished time: May 23, 2024

摘要

蛋白质组学作为研究生物体内蛋白质组成、功能及其相互作用的重要学科,对于深入理解生命过程和疾病机制具有重要意义。在蛋白质组学领域中,质谱技术已成为不可或缺的分析工具,其中数据非依赖性采集 (Data-independent acquisition, DIA) 策略因其高通量、高重复性和高灵敏度等特性而备受瞩目。然而,传统的 DIA 质谱数据仍面临液相色谱洗脱时间较长和谱图复杂度较高等挑战。为解决这些问题,一些新型的 DIA 采集方法如 DI-SPA 和 4D-DIA 等应运而生。目前,针对 DIA 质谱数据的分析方法大多依赖于色谱信息,其算法流程往往较为繁琐和复杂,且容易受到不同实验平台间色谱差异的影响,难以对上述这些新型的 DIA 质谱数据进行通用且高效的分析。在此背景下,本文面向 DIA 质谱数据,研究了若干种不依赖于色谱信息的分析方法,以进一步推动 DIA 质谱数据在蛋白质组学研究中的广泛应用。

本文的主要工作包括:(1) 针对循环采集的 DI-SPA 质谱数据,提出了一种名为 RE-FIGS 的定性及定量分析方法。该方法首先通过构建线性判别分析模型实现了对肽段的准确鉴定,之后对 DI-SPA 质谱数据进行图谱分解,从而得到肽段的定量结果。与现有的分析工具相比,RE-FIGS 在肽段定性和定量方面均取得了一定的提升。(2) 针对 4D-DIA 质谱数据缺少无色谱定性分析工具这一问题,提出了一种基于深度学习的定性分析方法。该方法利用卷积神经网络分类模型对肽段图谱库中的肽段进行打分,能够在不依赖于色谱信息的前提下,对 4D-DIA 质谱数据进行准确可靠的定性分析。与已有方法相比,该方法大幅提升了肽段鉴定的数量。(3) 针对 4D-DIA 质谱数据缺少无色谱定量分析工具这一问题,提出了一种基于深度学习的定量分析方法。与传统的色谱峰面积法不同,该方法无需利用色谱信息,便可以通过卷积神经网络回归模型来预测肽段的定量值。实验结果表明,该方法不仅具备较高的定量准确性和重复性,同时在肽段定量数量方面具有显著优势。

关键词: 蛋白质组学; 数据非依赖性采集; 肽段定性; 肽段定量; 深度学习

ABSTRACT

Proteomics, as an important discipline studying the composition, function, and interaction of proteins in organisms, holds significant importance for gaining a deeper understanding of biological processes and disease mechanisms. Within the field of proteomics, mass spectrometry has become an indispensable analytical tool, and among them, the Data-independent acquisition (DIA) strategy has garnered attention due to its high throughput, reproducibility, and sensitivity. However, traditional DIA mass spectrometry still faces challenges such as long liquid chromatography elution times and high spectral complexity. To address these issues, novel DIA acquisition methods like DI-SPA and 4D-DIA have emerged. Currently, most analysis methods for DIA mass spectrometry data rely on chromatographic information, and their algorithmic processes are often cumbersome and complex. Additionally, they are susceptible to chromatographic differences among various experimental platforms, making it difficult to perform universal and efficient analysis on these novel DIA mass spectrometry data. In this context, this dissertation focuses on DIA mass spectrometry data and explores several chromatographic-independent analysis methods to further promote the widespread application of DIA mass spectrometry data in proteomics research.

The main contributions of this dissertation are as follows: (1) A qualitative and quantitative analysis method called RE-FIGS is proposed for the cyclically acquired DI-SPA mass spectrometry data. This method first achieves accurate identification of peptides by constructing a linear discriminant analysis model and then performs spectral decomposition on DI-SPA mass spectrometry data to obtain quantitative results for the peptides. Compared to existing analysis tools, RE-FIGS has achieved certain improvements in both peptide qualification and quantification. (2) Aiming at the lack of chromatographic-independent qualitative analysis tools for 4D-DIA mass spectrometry data, a qualitative analysis method based on deep learning is proposed. This method utilizes a convolutional neural network classification model to score peptides in the peptide spectral library, enabling accurate and reliable qualitative analysis of 4D-DIA mass spectrometry data without relying on chromatographic information. Compared to existing methods, this approach significantly increases the number of peptide identifications. (3) Addressing the shortage of chromatographic-independent quantitative analysis tools for 4D-DIA mass spectrometry data, a quantitative analysis method based on deep learning is presented. Unlike traditional chromatographic peak area methods, this approach

can predict the quantitative values of peptides through a convolutional neural network regression model without utilizing chromatographic information. Experimental results demonstrate that this method not only possesses high quantitative accuracy and reproducibility but also exhibits significant advantages in terms of the number of peptides quantified.

Key Words: Proteomics; Data-Independent Acquisition; Peptide Identification; Peptide Quantification; Deep Learning

目 录

第 1 章 绪论.....	1
1.1 研究背景与意义.....	1
1.2 研究现状.....	5
1.3 本文主要工作.....	8
1.4 内容安排.....	10
第 2 章 针对循环采集的 DI-SPA 质谱数据的定性及定量分析.....	13
2.1 研究背景与意义.....	13
2.2 数据预处理.....	15
2.2.1 质谱的采集与处理.....	15
2.2.2 图谱库处理.....	16
2.3 定性算法.....	17
2.4 定量算法.....	21
2.5 实验结果与分析.....	22
2.5.1 定性数量对比.....	22
2.5.2 定性参数研究.....	25
2.5.3 定量性能评估.....	28
2.6 本章小结.....	31
第 3 章 基于深度学习的 4D-DIA 质谱数据定性分析方法.....	33
3.1 研究背景与意义.....	33
3.2 数据预处理.....	35
3.2.1 质谱处理.....	35
3.2.2 图谱库处理.....	36
3.2.3 质谱的匹配与过滤算法.....	36
3.2.4 训练数据与测试数据的生成.....	38
3.3 定性模型构建.....	40
3.4 实验结果与分析.....	45
3.4.1 同源血浆数据集定性结果.....	45
3.4.2 公共血浆数据集定性结果.....	47
3.5 本章小结.....	50
第 4 章 基于深度学习的 4D-DIA 质谱数据定量分析方法.....	51
4.1 研究背景与意义.....	51
4.2 数据预处理.....	51
4.2.1 质谱的匹配与过滤算法.....	51
4.2.2 训练数据与测试数据的生成.....	53

4.3 定量模型构建.....	54
4.4 实验结果与分析.....	57
4.4.1 同源血浆数据集定量结果.....	57
4.4.2 公共血浆数据集定量结果.....	60
4.5 本章小结.....	63
第5章 总结与展望	65
5.1 工作总结.....	65
5.2 未来工作展望.....	66
参考文献.....	69

插图清单

图 1.1	自下而上的蛋白质组学分析流程	2
图 1.2	LC-MS/MS DIA 质谱数据示意图	3
图 1.3	研究内容之间的关系	10
图 2.1	DI-SPA 质谱数据采集过程	13
图 2.2	RE-FIGS 定性算法流程图	19
图 2.3	RE-FIGS 肽段打分分布图	23
图 2.4	不同分辨率的 DI-SPA 质谱数据肽段定性结果对比	24
图 2.5	100 次重复采集的 DI-SPA 质谱数据肽段定性结果对比	25
图 2.6	不同采集周期数下的肽段定性结果对比	26
图 2.7	不同注射时间下的肽段定性结果对比	27
图 2.8	最优实验参数下的肽段定性结果对比	28
图 2.9	MCF7 肽样品肽段定量结果直方图对比	29
图 2.10	多物种混合肽段样品肽段定量结果散点图对比	30
图 2.11	RE-FIGS 肽段定量结果热力图	31
图 3.1	4D-DIA 质谱数据示意图	33
图 3.2	4D-DIA 质谱数据采集流程	34
图 3.3	CNN 定性模型网络结构示意图	42
图 3.4	CNN 定性模型肽段打分分布图	46
图 3.5	LCH_Plasma 数据集肽段定性结果对比	47
图 3.6	LCH_Plasma 数据集肽段定性重复性对比	47
图 3.7	公共血浆数据集肽段定性结果对比	48
图 3.8	公共血浆数据集肽段定性结果韦恩图	49
图 3.9	PXD030327 数据集肽段定性重复性对比	49
图 4.1	CNN 定量模型网络结构示意图	56
图 4.2	LCH_MN_144-Plasma 数据集肽段定量结果韦恩图	58
图 4.3	LCH_MN_144-Plasma 数据集肽段定量结果散点图对比	59
图 4.4	LCH_MN_144-Plasma 数据集肽段定量结果热力图对比	60
图 4.5	LCH_MN_144-Plasma 数据集肽段定量重复性对比	60
图 4.6	PXD040205 数据集肽段定量结果韦恩图	61
图 4.7	PXD040205 数据集肽段定量结果散点图对比	62

图 4.8	PXD040205 数据集肽段定量结果热力图对比	62
图 4.9	PXD040205 数据集肽段定量重复性对比	63

表 格 清 单

表 2.1 A、B 样品肽段混合比例表 29

表 3.1 定性模型训练集组成 39

表 3.2 CNN 定性模型网络结构参数 43

表 3.3 实验设备及环境配置 45

表 4.1 CNN 定量模型网络结构参数 57

算 法 清 单

2.1	峰合并算法	16
2.2	诱饵库生成算法	18
2.3	RE-FIGS 定性算法	20
2.4	Fisdec 图谱分解算法	22
3.1	质谱匹配算法	37
3.2	质谱过滤算法	38
4.1	定量模型训练集标签获取算法	54

第1章 绪 论

1.1 研究背景与意义

蛋白质是组成人体一切细胞、组织的重要成分, 机体所有重要的组成部分都需要有蛋白质的参与。因此, 对蛋白质的研究可以帮助人们探索生命活动的各种机制, 并为生命科学和医学研究做出重要贡献。蛋白质组 (Proteome) 是指一个细胞或一个组织基因组所表达的全部蛋白质^[1], 而蛋白质组学 (Proteomics) 则是以蛋白质组为研究对象, 研究细胞、组织或生物体蛋白质组成及其变化规律的科学, 这一概念最早是在二十世纪九十年代初期, 由 Marc Wilkins 等研究人员提出的^[2]。蛋白质组学主要应用于寻找有效药物和药物靶点、蛋白质间的相互作用、癌症诊断和治疗以及生物信息学和基因组学研究等领域, 为人类健康事业的发展提供了重要的技术支持^[3-4]。蛋白质组学的研究内容包括蛋白质的定性和定量分析、蛋白质的结构预测、蛋白质的差异显示和比较等, 其中蛋白质的定性和定量是蛋白质组学分析流程中最基础的操作, 同时也是蛋白质组学领域的研究重点^[5]。

蛋白质组学分析主要包括两种方式。一种是自上而下的分析方式 (Top-down), 它通过直接对完整的蛋白质进行质谱分析, 从而获取蛋白质的相关信息。该方式需要使用较高分辨率和灵敏度的质谱仪器进行实验, 且对样品处理的要求较为严格, 不太适用于大规模的实验研究^[6]。另一种是自下而上的分析方式 (Bottom-up)^[7-8], 其具体步骤如下 (图1.1):

1. 对生物样品进行处理, 提取出其中的蛋白质成分
2. 使用蛋白酶对蛋白质进行酶解得到肽段混合物
3. 利用液相色谱、电泳等技术对肽段混合物进行分离, 然后将其送入质谱仪中进行离子化并生成实验质谱数据
4. 对实验质谱数据进行质谱分析, 从而鉴定出生物样品中存在的肽段并计算其含量
5. 根据肽段与蛋白质之间的对应关系, 推导出蛋白质的定性及定量结果

在上述步骤中, 前三步为生物实验, 主要由从事蛋白质组学和生物学等领域研究的科研人员负责实施; 而第四步和第五步则是数据分析, 主要由计算机科学和信息科学等领域的科研人员来完成。自下而上的分析方式具有高通量和高灵敏度等特点, 可以在较短的时间内完成生物样品的大规模蛋白质定性及定量分析, 因此也成为了当前蛋白质组学的主流分析方式^[9]。

在蛋白质组学的发展过程中, 质谱技术起到了十分关键的作用。该技术通过对实验样品进行质谱采集和分析, 从而实现对蛋白质的鉴定和定量^[10-11]。然而,

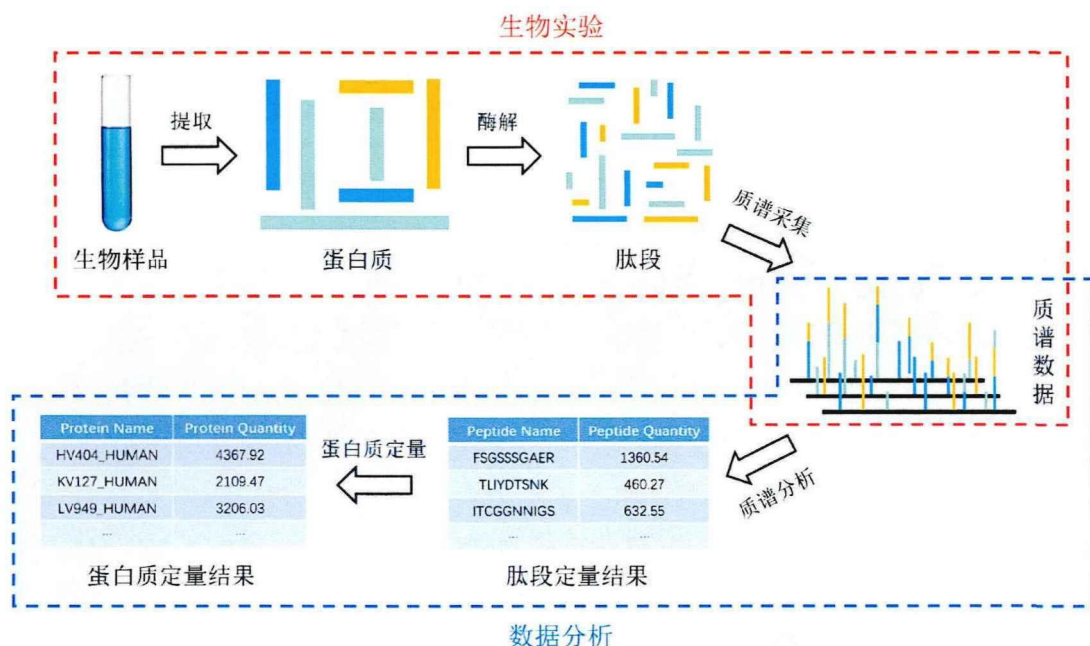


图 1.1 自下而上的蛋白质组学分析流程

单一的质谱技术对复杂蛋白质样品的分析效果并不理想。因此，在实际的蛋白质组学实验中，往往需要将质谱技术和其他分离技术联合在一起使用。在传统的蛋白质组学分析中，主要采用液相色谱-串联质谱 (Liquid chromatography-tandem mass spectrometry, LC-MS/MS) 方法^[12]。该方法涉及到液相色谱仪和串联质谱两个概念，其中液相色谱仪是一种高效分离复杂混合物中各组分的仪器，当混合样品进入液相色谱仪之后，其不同的成分会因为各自理化性质的不同在保留时间 (Retention time) 维度上以色谱峰的形式流出^[13]。而串联质谱则是蛋白质组学中的一项重要技术^[14]，这项技术可以将多个质谱分析仪耦合在一起，以提高整个系统分析实验样品的能力。在基于液相色谱-串联质谱的蛋白质组学实验中，蛋白质样品首先会在蛋白酶的作用下生成多肽混合物，多肽混合物在进入液相色谱仪后，不同的肽段会在不同的保留时间点流出，从而实现了对实验样品的色谱洗脱。色谱洗脱过程不光使实验样品得到了初步的分离，同时还可以为后续的质谱分析提供保留时间和色谱峰形状等色谱信息。在完成色谱洗脱后，需要将分离后的样品输入到质谱仪中进行串联质谱采集。在该实验中，串联质谱包括一级质谱和二级质谱。其中一级质谱会采集输入样品中的肽段母离子信号，进而形成一级质谱图 (MS1)。而二级质谱则根据特定的采集策略，从一级质谱图中选择部分或者全部肽段母离子进行窗口碎裂得到碎片离子，并生成相应的二级质谱图 (MS2)^[15]。

在对肽段母离子进行窗口碎裂并生成相应的二级质谱图的过程中，主要有两种不同的采集策略，它们分别是数据依赖性采集 (Data-dependent acquisition,

DDA)^[16-17]策略和数据非依赖性采集 (Data-independent acquisition, DIA)^[18]策略。DDA 策略主要针对高丰度肽段母离子进行小窗口的收集和碎片化, 这种采集方式可以获得单个肽段母离子相对纯净的二级质谱, 但是, 由于这种采集策略是选择性采集, 因此所获得的质谱数据的覆盖率和重复性较差。而 DIA 策略则是按质荷比预先划分多个采集窗口, 循环地碎裂每个窗口下的所有肽段母离子。这种采集方式覆盖了窗口范围内的所有肽段母离子, 所生成的二级质谱具有高度的重复性, 在高通量蛋白质组学领域有着广泛的应用前景^[19-20]。

因此, 在传统的蛋白质组学分析中, 大多采用 LC-MS/MS 技术与 DIA 策略相结合的方式来进行质谱采集, 所生成的质谱数据也被称为 LC-MS/MS DIA 质谱数据^[21]。LC-MS/MS DIA 质谱数据共包含三个维度的信息, 它们分别是质荷比、保留时间和峰强度, 其示意图如图1.2所示。在图1.2中共有四张二级质谱图(对应的保留时间分别为 t_1 、 t_2 、 t_3 和 t_4), 每张二级质谱图都包含若干个子离子峰。由于液相色谱的洗脱作用, 同一子离子的峰强度在保留时间维度可能会形成色谱峰的形状, 图1.2中的红色部分所展示的正是质荷比为 mz_1 的子离子的色谱峰曲线。

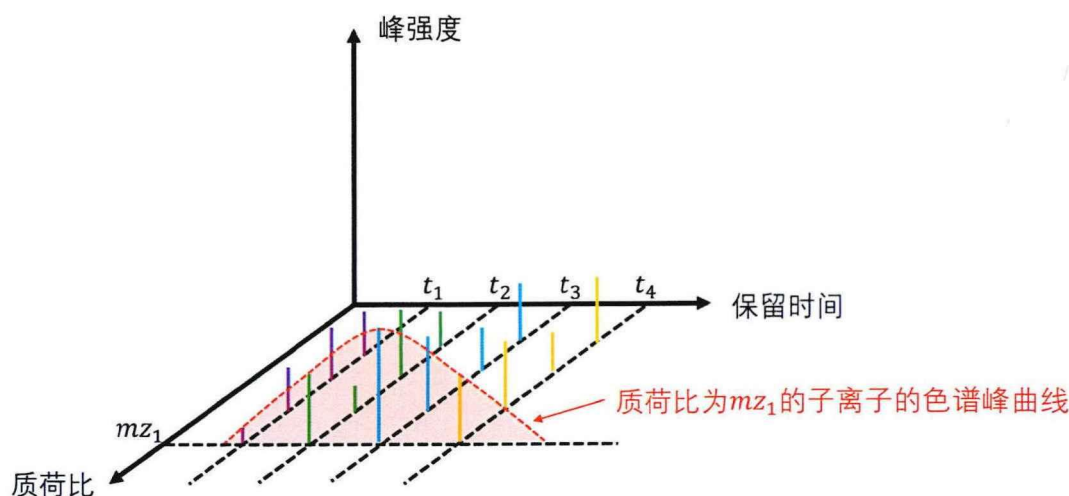


图 1.2 LC-MS/MS DIA 质谱数据示意图

目前, 针对传统的 DIA 质谱数据的定性分析主要采用图谱匹配 (Spectrum-spectrum matches, SSM)^[22]的方法。在这类方法中, 首先需要将实验质谱和肽段图谱库中的肽段图谱进行匹配, 之后通过计算 SSM 分数来评估肽段存在于实验样品中的可能性, 进而得到肽段的定性结果。在计算 SSM 分数的过程中, 需要综合考虑保留时间、子离子峰匹配情况和色谱峰的大小形状等信息。由于肽段母离子和对应的子离子往往存在着高度重合的色谱峰, 因此, 在过去的很多研究中都将色谱信息作为 SSM 打分的一个重要依据, 这一做法也显著提升了 DIA 质谱数据定性分析的准确性^[23-24]。至于 DIA 质谱数据的定量分析则主要依赖于色谱

峰面积法^[25]。这类方法需要对色谱峰进行提取、对齐、打分和积分等一系列操作,之后利用色谱峰曲线在保留时间维度下形成的面积来推算肽段的定量值。过往的实验表明,色谱峰面积法具备较高的准确性和灵敏度,可以对复杂生物样品中的肽段进行准确可靠的定量^[26]。

尽管液相色谱-串联质谱技术可以对实验样品进行准确的质谱分析,但是液相色谱的应用显著增加了实验时间,这阻碍了蛋白质组学朝着高通量的趋势发展^[27]。近年来,随着质谱技术的不断进步,越来越多的研究者开始尝试在质谱实验中缩短甚至完全省去液相色谱洗脱的时间。例如,DI-SPA (Direct infusion-shotgun proteome analysis)^[28]就是近几年提出的一种新型的 DIA 质谱数据采集方法。它使用 FAIMS (High-field asymmetric waveform ion mobility spectrometry)^[29]技术替换液相色谱,通过利用补偿电压 (Compensation voltage) 的方式对肽段母离子进行气相分离,然后再将分离后的样品送入串联质谱仪中进行分析。由于 FAIMS 对肽段母离子的分离速度明显快于液相色谱,因此这也大大缩短了整个质谱数据的获取时间,从而促进了高通量蛋白质组学的发展。但是,在删除了液相色谱洗脱环节后,所采集到的 DI-SPA 质谱数据将会缺少色谱信息,而目前的绝大多数 DIA 质谱数据定性及定量分析方法都需要使用色谱信息,DI-SPA 质谱数据将无法在现有的大多数 DIA 质谱数据分析工具上使用。因此,研究如何对 DI-SPA 质谱数据进行有效的定性及定量分析,对于提高大规模蛋白质样本的测定效率具有十分重要的意义。

此外,传统的液相色谱-串联质谱技术在处理一些复杂生物样品,例如人类血浆蛋白质样品时,其色谱分离效果并不理想。这导致最终所产生的质谱数据极其复杂,同时存在着大量的重叠峰,严重干扰了后续的质谱数据分析过程^[30]。为此,研究者开始尝试在液相色谱-串联质谱工作流程的基础上,运用新的分离技术来处理实验样品,从而进一步降低实验质谱的复杂度^[31]。这类工作的代表是 Meier 等人所提出的 4D-DIA (diaPASEF)^[32]方法,在该方法中,实验样品首先经过蛋白酶水解和液相色谱初步分离,之后进入 timsTOF 质谱仪中进行进一步的分析。与传统的 DIA 质谱采集仪器不同,timsTOF 质谱仪中包含一个双串联 TIMS (Trapped ion mobility spectrometer)^[33-34]装置,该装置可以用来捕获质谱仪中的肽段母离子,并利用不同离子的淌度差异对其进行分离。因此,4D-DIA 质谱数据不仅具备更低的谱图复杂度,同时额外增添了一维离子淌度信息。目前,针对 4D-DIA 质谱数据的定性及定量分析依然主要采用传统的 DIA 质谱数据分析方法。这类方法往往需要依赖于色谱信息,容易受到不同实验平台间色谱条件差异的影响。因此,未来仍需提出新的不依赖于色谱信息的 4D-DIA 质谱数据分析方法,以提高 4D-DIA 技术的通用性。

近年来,人工智能得到了飞速发展,在很多领域中都发挥了十分重要的作用,

同时计算能力的提升和大数据的涌现也使得深度学习逐渐崭露头角^[35-36]。目前人工智能技术已经广泛应用于蛋白质组学的相关研究中,极大地推动了这一领域的研究进展^[37]。对于蛋白质定性及定量问题而言,计算机科学领域需要关注的问题是,如何利用人工智能相关技术从复杂的质谱数据中计算出肽段的定性和定量结果,进而得到蛋白质的鉴定及定量结果。其中,根据肽段定量结果推导出蛋白质定量结果这一步骤的相关方法已经相对成熟^[38-40]。因此,在未来,计算机科学领域的研究重点是如何对质谱数据进行有效的定性及定量分析,特别是如何在不使用色谱信息的情况下实现对 DIA 质谱数据的准确定性及定量。

1.2 研究现状

近年来,基于液相色谱-串联质谱技术的蛋白质定性及定量的相关研究一直在朝着高通量的方向发展。为了实现这一目标,在对质谱数据进行采集的过程中, DIA 策略凭借其较高的重复性和较大的覆盖范围逐渐取代了传统的 DDA 策略。随着质谱技术的不断提升, DIA 策略也在不断发展,目前已经有多种 DIA 策略被应用到质谱采集的过程中。例如, SWATH^[41]通过设置多个窗口,将整个质谱范围按照质荷比划分为一系列离散的窗口,并对每个窗口中的所有离子进行碎裂和检测。Scanning SWATH^[42]通过结合四极杆的连续扫描功能,在保留了 SWATH 技术的优点的同时,提升了质谱数据的分析速度。PASS-DIA^[43]使用较小的采集窗口来碎裂肽段母离子,并利用准确的肽段前体质量来解析质谱数据,其性能明显优于原始的 DIA 策略。

由于 DIA 策略在进行质谱采集时会对窗口范围内的所有肽段母离子进行碎裂,因此产生的碎片离子覆盖范围更广,所采集到的质谱数据也会更加复杂。同时由于碎片离子的随机性,不同的肽段母离子也可能会产生同一个碎片离子,这些重复的碎片离子将会对 DIA 质谱数据的解析产生较大的干扰。目前针对 DIA 质谱数据的分析已经提出了许多方法,它们大致可以分为两大类:一类是不依赖于图谱库 (Library-free) 的方法,另一类是依赖于图谱库 (Library-based) 的方法^[44-45]。

不依赖于图谱库的 DIA 质谱数据分析方法不需要预先构建图谱库就可以对 DIA 质谱数据进行定性及定量分析,因此可以适应不同种类的蛋白质样本,具有较高的灵活性。这类方法包括 FT-ARM^[46]、DIA-Umpire^[47]和 Group-DIA^[48]等。其中, FT-ARM 将理论碎片离子的预测结果与整个采集过程中的每个二级质谱进行对比,通过计算每个图谱的点积分数来生成用于定性和定量的色谱图。DIA-Umpire 将肽段母离子和碎片离子的洗脱图谱组装成伪串联质谱,之后利用传统的数据库搜索和蛋白质推断工具就可以对这些伪图谱进行解析,从而实现了在

没有图谱库的情况下对 DIA 质谱数据进行分析。Group-DIA 则是在 DIA-Umpire 的基础上根据所有质谱数据中肽段母离子和碎片离子洗脱曲线,对母离子和子离子进行分组,之后使用分组后的结果生成伪图谱,最后再利用这些伪图谱完成后续 DIA 质谱数据分析。

而依赖于图谱库的 DIA 质谱数据分析方法则需要通过搜索图谱库来实现对 DIA 质谱数据的定性和定量分析,其中图谱库是一个包含了多个肽段图谱的数据库,可以用来匹配实验样品中的肽段^[49]。为了建立图谱库,需要对大量实验样品进行 DDA 质谱分析,从而获取高质量的肽段图谱数据。这个过程虽然需要消耗大量的时间和资源,但是可以为 DIA 质谱数据提供参考标准,这将大大提高后续蛋白质鉴定和定量的准确性^[50]。因此,在实际的蛋白质组学定量实验中,往往采用依赖于图谱库的方式对 DIA 质谱数据进行分析,接下来本文将重点介绍有关该方法的研究进展和现状。

在 DIA 质谱数据的定性方面,目前已经提出了许多相关的算法。在早期的研究中,首先会将实验质谱与图谱库中的肽段图谱进行匹配,之后提取匹配过程中的相关特征,并计算图谱匹配 (Spectrum-spectrum matches, SSM) 分数来量化实验图谱和肽段图谱的相关性。通常来说 SSM 分数越高的肽段,其存在于实验样品中的概率也会越大^[22]。因此,SSM 的评分策略对最终的蛋白质鉴定结果具有显著影响。目前,研究者们已经提出了多种 SSM 评分方式。例如, Mascot^[51] 利用子离子匹配数量和峰强度等信息来计算 SSM 分数。Comet^[52] 基于所有匹配到的碎片离子的峰强度之和来计算互相关系数,之后利用该系数生成分数直方图并计算图谱库中每个肽段的分数。SpectraST^[53] 则通过计算实验图谱和肽段图谱峰强度的点积来说明两者的匹配程度。

而在液相色谱-串联质谱技术开始流行后,由于理论上肽段母离子和它所对应的碎片离子在色谱图上应该存在着高度重合的色谱峰,因此在对 DIA 质谱数据进行定性分析时,很多研究都引入了色谱信息作为一个重要的评分标准,这一做法也显著提高了定性的准确性和可靠性。例如, MSPLIT-DIA^[23] 在计算 SSM 分数时,除了根据实验质谱和肽段图谱的碎片离子来量化匹配的相似度外,还将肽段母离子与碎片离子色谱峰的重叠程度作为打分的一个重要依据。OpenSWATH^[24] 则首先会提取目标肽段的色谱峰,并对这些色谱峰在保留时间维度上进行归一化处理,之后再对所有碎片离子的共洗脱峰进行打分,最后综合质谱数据的其他特征计算出 SSM 分数,并基于该分数得到肽段鉴定的结果。除此之外,还有 Spectronaut^[54]、Skyline^[55]、DIA-NN^[56] 等方法也都是通过综合色谱信息和图谱匹配情况来对 DIA 质谱数据进行定性分析。

在 DIA 质谱数据的定量分析方法中,根据是否采用同位素标记技术,可以将其分为有标记定量和无标记定量 (Label-free quantification, LFQ)^[57] 两种方法。相

较于有标记定量方法, 无标记定量省去了耗时较长且成本较高的同位素标记环节, 实验操作也更为简便。因此, 目前的大多数蛋白质组学实验都更倾向于使用无标记定量方法。在前期的研究中, 无标记定量分析方法主要包括谱图计数法^[58]和色谱峰面积法^[25]。谱图计数法的基本假设是, 实验样品中肽段的丰度与其被检测到的概率呈正相关, 即丰度越高, 被检测到的可能性越大, 从而对应的二级质谱的数量也会相应增加。这种方法虽然操作简便直接, 但在评估肽段相对丰度时往往不够精确^[59]。相较之下, 色谱峰面积法会构建肽段碎片离子的色谱洗脱峰, 肽段的丰度越高, 所产生的色谱峰面积就越大, 可以用该面积作为肽段的相对定量值。过往的大量实验表明, 色谱峰面积法通常比谱图计数法更加灵敏和精确^[26]。因此, 在以前的研究中, 对于 DIA 质谱数据的无标记定量分析, 主要依赖于色谱峰面积法。

在过去, 研究者们提出了多种基于色谱峰面积法的 DIA 定量分析方法, 其中一些具有代表性的研究成果至今仍广泛应用于 DIA 定量分析中。例如, Spectronaut^[54]使用机器学习算法进行色谱峰的自动识别和评分, 同时结合保留时间信息预测肽段的洗脱范围, 最后采用先进的干扰检测算法进一步提高 DIA 质谱数据的定量准确性。DIA-NN^[56]采用神经网络模型来区分色谱峰信号与噪声信号, 并通过迭代过程精准识别出最优的色谱洗脱峰。在计算肽段丰度时, DIA-NN 还会对色谱峰面积进行修正处理, 从而有效地减少定量误差。Specter^[60]和 FIGS^[61]利用混合二级质谱与肽段母离子所匹配的离子峰信息, 通过线性拟合方法来迭代计算出每个肽段母离子的定量相关系数, 之后基于这些系数构建色谱峰, 并通过计算色谱峰面积来得到肽段的定量值。以上所述的方法在处理由液相色谱-串联质谱技术产生的质谱数据时, 通常表现出良好的定量效果。然而, 由于各个实验室采用的色谱梯度存在差异, 且色谱峰面积法高度依赖于色谱信息, 因此这些方法可能会受到色谱梯度变化的干扰。此外, 液相色谱洗脱过程的耗时较长, 这也限制了高通量蛋白质组学的发展。

近年来, 随着人们对测定大规模蛋白质样本的需求日益增加, 研究者们开始尝试对 DIA 质谱采集过程中耗时较长的液相色谱洗脱步骤进行改进, 其具体措施包括缩短液相色谱洗脱的时间或采用新的色谱洗脱方法等。例如, DI-SPA^[28]使用 FAIMS^[29]技术替换液相色谱洗脱过程, 不仅具有较高的分辨率和灵敏度, 而且比传统的液相色谱-串联质谱技术更加快速、高效和简便。因此, 该方法对于推动高通量蛋白质组学的发展起到了十分重要的作用。然而, 过去的大多数 DIA 质谱数据分析方法都依赖于传统的色谱信息, 因此无法直接应用在 DI-SPA 质谱数据上。目前, 也只有极少数的 DIA 质谱数据分析工具可以对 DI-SPA 质谱数据进行定性及定量分析。其中, CsoDIAq^[62]就是一款专门针对 DI-SPA 质谱数据设计的分析工具, 它在 MSPLIT-DIA^[23]的基础上改进了图谱匹配算法和 SSM 评分

机制，从而增加了肽段的定性数量，并且可以对 DI-SPA 质谱数据进行准确的有标记定量分析。尽管 CsoDIAq 在 DI-SPA 质谱数据分析方面已经取得了进展，但整体来看，目前关于 DI-SPA 质谱数据的研究仍然相对匮乏。因此，对 DI-SPA 质谱数据进行有效分析，尤其是针对其无标记定量分析的研究，显得尤为迫切和重要。

除了致力于缩短质谱采集时间的研究外，研究者们同样开始关注如何提高 DIA 质谱数据的质量，以进行更准确的质谱分析。传统的液相色谱-串联质谱技术在处理一些复杂生物样品，如人类血浆蛋白质样品时，往往会因为其色谱分离效果不够理想，从而导致最终生成的 DIA 质谱数据复杂度过高，这为后续的数据分析带来了极大的挑战。为解决该问题，Meier 等人提出了一种名为 4D-DIA (diaPASEF)^[32] 的新型 DIA 质谱采集方法，该方法通过结合 DIA 策略和 TIMS (Trapped ion mobility spectrometer)^[33-34] 技术，不仅有效降低了质谱数据的复杂度，还额外引入了一维离子淌度信息。目前，针对 4D-DIA 质谱数据的定性及定量分析方法大多沿用了传统的 DIA 质谱数据分析方法，如 Spectronaut^[54] 和 DIA-NN^[56] 等。这些方法大多依赖于色谱信息，其算法流程往往较为繁琐和复杂。同时，由于不同实验平台的色谱条件存在差异性，因此这些依赖于色谱信息的 DIA 质谱数据分析方法在处理不同实验平台所制备的质谱数据时，容易受到色谱梯度变化的影响，从而为最终的分析结果带来一定的误差。因此，当前需要探索出不依赖于色谱信息的 4D-DIA 质谱数据分析方法，以提高 4D-DIA 方法的通用性和可靠性。

1.3 本文主要工作

在蛋白质组学领域中，如何对生物样品中的蛋白质进行准确的定性和定量分析一直是科研人员的研究重点。随着质谱技术的快速发展，人们开始采用质谱技术对实验样品进行 DIA 质谱采集，并通过分析生成的 DIA 质谱数据来确定实验样品中肽段的种类和含量，进而推导出蛋白质的定性及定量结果。在过去很长的一段时间里，液相色谱-串联质谱技术凭借其优异的灵敏度、特异性和分辨率成为了蛋白质组学的主流分析方法之一。然而，液相色谱的应用使得质谱数据的采集时间明显增加，这对高通量蛋白质组学的发展构成了阻碍。同时，液相色谱-串联质谱技术在处理一些较为复杂的生物样品时，可能会面临谱图复杂度过高的问题。为解决这些问题，研究者们提出了 DI-SPA 和 4D-DIA 等新型 DIA 质谱数据采集方法。然而，现有的大多数 DIA 质谱数据定性及定量分析方法都高度依赖于色谱信息，容易受到不同实验平台间色谱差异的影响，难以对上述这些新型的 DIA 质谱数据进行通用且准确的分析。在这样的背景下，为了满足

当前蛋白质组学对 DI-SPA 质谱数据和 4D-DIA 质谱数据定性及定量分析的需求, 本文深入研究了若干种无色谱分析方法, 主要研究内容包括以下三个方面:

1. 针对循环采集的 DI-SPA 质谱数据的定性及定量分析

DI-SPA 质谱数据是近几年提出的一种新型的 DIA 质谱数据, 其在数据采集过程中使用分离速度较快的 FAIMS 技术替换传统的液相色谱技术, 从而显著缩短了质谱采集时间。然而, 由于 DI-SPA 质谱数据缺少色谱峰信息, 因此以往那些依赖于色谱峰信息的 DIA 质谱数据分析工具无法直接应用在此类数据上。现有的 CsoDIAq 等方法虽然可以对 DI-SPA 质谱数据进行分析, 但其定性及定量效果仍存在一定的不足。为此, 本文提出了一种不依赖于色谱峰信息的 DI-SPA 质谱数据分析方法 RE-FIGS。该方法在从 SSM 结果中提取相关的特征后, 利用线性判别分析来训练分类模型, 进而实现了对 DI-SPA 质谱数据的定性分析。此外, 我们还发现, 在对 DI-SPA 质谱数据进行循环采集后, 可以利用其产生的重复性特征显著提升肽段定性的数量。最后, 我们利用 FIGS 中的 Fisdec 图谱分解算法对 DI-SPA 质谱数据进行解谱, 并将解谱系数作为对应肽段的定量值。实验结果表明, 与 CsoDIAq 相比, 我们的 RE-FIGS 方法在定性及定量方面均取得了一定的提升。

2. 基于深度学习的 4D-DIA 质谱数据定性分析方法

4D-DIA 质谱数据是一种结合了 TIMS 技术与 DIA 策略生成的 DIA 质谱数据。TIMS 技术的引入不仅大大降低了质谱数据中每张实验图谱的复杂度, 还为其增添了一维离子淌度信息, 从而为后续的定性分析提供了更丰富的数据信息。目前, 针对 4D-DIA 质谱数据的定性分析方法大多依赖于色谱信息, 容易受到不同实验平台间色谱条件差异的影响。因此, 目前需要开发一款能够兼容并处理各个实验室产生的 4D-DIA 质谱数据的通用定性工具。为此, 本文提出了一种深度学习方法, 该方法在不依赖于色谱信息的前提下, 通过提取图谱匹配结果中的离子峰和离子淌度特征来对 4D-DIA 质谱数据进行定性分析。实验结果表明, 与业内领先的商业化蛋白质组学分析工具 Spectronaut 相比, 我们的深度学习方法在肽段鉴定数量方面取得了显著的进步。

3. 基于深度学习的 4D-DIA 质谱数据定量分析方法

4D-DIA 方法的出现显著减少了实验质谱中的离子峰干扰, 从而提升了定量分析的准确性和可靠性。当前, 针对 4D-DIA 质谱数据的定量分析方法大多沿用了传统的色谱峰面积法, 其计算流程包括色谱峰的提取、对齐、打分和积分等步骤。这种定量分析方法较为复杂, 且容易受到色谱梯度变化的影响。为解决上述问题, 本文提出了一种基于深度学习的 4D-DIA 质谱数据定量分析方法。该方法并未使用传统的色谱信息, 而是将图谱匹配结果中的离子峰信息输入到卷积神经网络中进行处理, 并通过构建一个回归模型来预测肽段的定量值, 最后再使用

本文第二个研究内容中的肽段定性结果对定量到的肽段进行过滤，从而得到最终的肽段定量结果。为了验证该定量模型的可靠性，我们对若干组人类血浆数据集进行了测试，并发现深度学习方法在定量准确性和重复性等方面展现出了良好的性能。

本文三个主要研究内容之间的关系如图1.3所示，图中的序号依次表示了这三个主要研究内容。

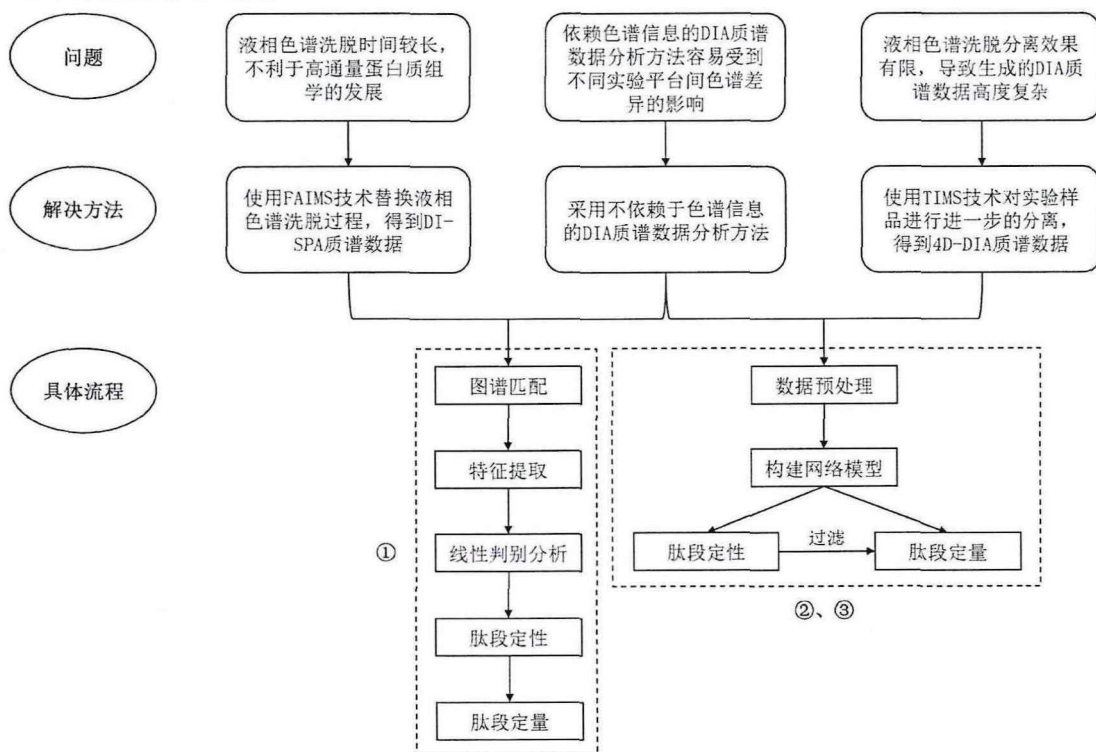


图 1.3 研究内容之间的关系

1.4 内容安排

第一章：绪论

本章首先概述了蛋白质组学和 DIA 质谱数据的研究背景与意义，之后详细介绍了 DIA 质谱数据分析方法的研究现状，最后阐述了本文的主要工作内容。

第二章：针对循环采集的 DI-SPA 质谱数据的定性及定量分析

本章首先对 DI-SPA 质谱数据的研究背景与意义进行了介绍，之后针对循环采集的 DI-SPA 质谱数据，设计了一种基于半监督学习的定性算法，并利用 FIGS 中的 Fisdec 图谱分解算法实现了对 DI-SPA 质谱数据的无标记定量分析。最后，我们通过实验验证了定性及定量算法的准确性，并将实验结果与 CsoDIAq 方法进行了对比。

第三章：基于深度学习的 4D-DIA 质谱数据定性分析方法

本章首先介绍了 4D-DIA 质谱数据及其定性分析方法的相关研究背景，之后

针对 4D-DIA 质谱数据缺少无色谱分析工具这一问题,提出了一种基于深度学习的定性分析方法。最后,为了验证该方法的有效性,我们在人类血浆数据集上进行了定性实验,并将实验结果与著名的商业化蛋白质组学分析工具 Spectronaut 进行了对比。

第四章:基于深度学习的 4D-DIA 质谱数据定量分析方法

本章首先介绍了 4D-DIA 质谱数据定量分析领域的相关研究背景,之后针对现有的定量分析方法依赖于色谱信息这一局限性,提出了一种基于深度学习的 4D-DIA 质谱数据定量分析方法。最后,我们在人类血浆数据集上进行了定量实验,并将实验结果与 Spectronaut 工具进行了对比,从而验证了深度学习方法的可靠性。

第五章:总结与展望

本章是对全文工作的总结,并对后续工作进行了展望。

第2章 针对循环采集的 DI-SPA 质谱数据的定性及定量分析

2.1 研究背景与意义

在蛋白质组学的发展过程中, 色谱技术一直是自下而上的蛋白质组学分析流程中的关键技术^[63]。作为一种高效的分离手段, 它可以和质谱技术相结合, 共同实现对生物样品的质量测定和质谱采集^[64]。色谱技术涵盖了液相色谱、纳米流色谱以及气相色谱等多种类型。其中液相色谱凭借其出色的选择性和灵敏度在蛋白质组学分析中占有重要地位, 但该技术的分离时间相对较长, 这阻碍了高通量蛋白质组学的发展。纳米流色谱虽然具有分辨率高的优点, 但是也存在着分离稳定性不佳、重现性较差以及操作难度较高等问题, 这些因素已成为限制其广泛应用的主要障碍^[65]。为了克服以上这些色谱技术的缺点, Meyer 等人提出了一种全新的 DIA 质谱数据采集方法, 并将其命名为 DI-SPA^[28], 使用该方法生成的质谱数据也被称为 DI-SPA 质谱数据。DI-SPA 质谱数据的采集过程如图2.1所示。与传统的液相色谱-串联质谱技术不同, DI-SPA 使用 FAIMS^[29] 技术替换液相色谱对实验样品进行筛选和分离。FAIMS 是一种气相色谱分离技术, 它能够利用肽段母离子在电场中的迁移行为差异, 从而对混合样品中的肽段进行高效的分离。该技术不仅具有高灵敏度和高分辨率的特点, 同时由于其分离速度明显快于液相色谱, 因此极大地减少了获取质谱数据所需要的总时间。具体来说, 对于采用传统的液相色谱-串联质谱技术制备的 DIA 质谱数据而言, 其采集时间往往是相同情况下 DI-SPA 质谱数据采集时间的 20 倍以上, 采集时间方面的优势也使得 DI-SPA 质谱数据在高通量蛋白质组学的研究中展现出极为广阔的应用前景。

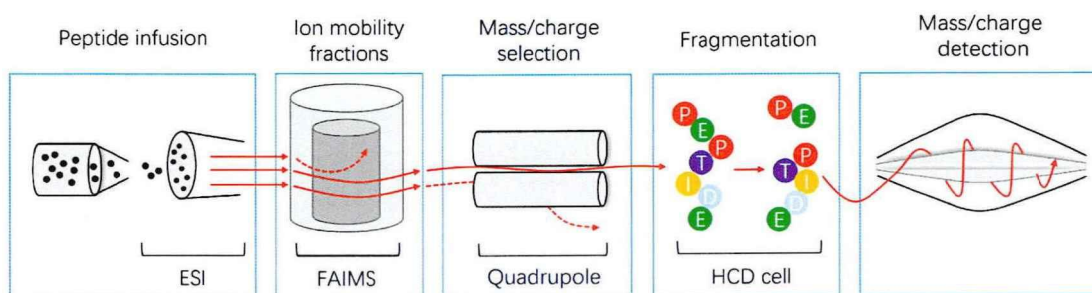


图 2.1 DI-SPA 质谱数据采集过程

然而, 值得注意的是, 与传统的液相色谱-串联质谱技术所生成的 DIA 质谱数据不同的是, DI-SPA 质谱数据没有传统意义上的色谱峰。这一差异源于两种质谱数据在肽段分离过程中存在着本质的区别。在液相色谱洗脱过程中, 由于混

合样品中各个肽段的理化性质存在差异,因此它们会以不同的保留时间从液相色谱仪中依次流出。同时,在液体环境中,每种肽段理论上都会在特定的时间范围内持续地流出,其具体过程如下:在起始阶段,肽段的流出浓度为零。之后,随着肽段开始从液相色谱仪中流出,其浓度也在逐渐上升,并在预期的保留时间点达到峰值。最后,随着肽段完全流出,其浓度也会逐渐降低至零。因此,液相色谱技术可以在保留时间维度产生色谱峰,而色谱峰的时间、形状等相关信息则可以为后续的质谱数据分析提供重要参考。然而,对于 FAIMS 技术而言,它使用一组补偿电压循环地对肽段混合物进行处理,并按照肽段母离子在电场中的迁移率对其进行分离,在不同的补偿电压下,肽段母离子通过的离子量是一个离散且无法预测的值。因此,使用 FAIMS 技术所生成的 DI-SPA 质谱数据没有相应的色谱峰信息。而过去的大多数 DIA 质谱数据分析工具都严重依赖于色谱峰信息,因此它们无法对 DI-SPA 质谱数据进行有效的解析。

现阶段,只提出了少数的方法可以对 DI-SPA 质谱数据进行定性分析。例如,MSPLIT-DIA^[23] 首先将肽段图谱与实验图谱进行匹配,然后使用归一化点积计算图谱匹配 (Spectrum-spectrum matches, SSM) 分数,进而得到肽段鉴定结果。CsoDIAq^[62] 则在 MSPLIT-DIA 的基础上进行了多项算法优化,其中包括通过合并实验图谱来降低时间复杂度,以及引入 MACC (Match count and cosine) 分数等。通过这些改进措施,CsoDIAq 显著增强了对肽段的识别能力。至于 DI-SPA 质谱数据的定量研究方面,CsoDIAq 虽然可以对经过同位素标记的 DI-SPA 质谱数据进行准确的定量分析,然而该方法在处理无标记数据时,其定量准确度较差。考虑到同位素标记的过程耗时较长,因此 CsoDIAq 在实际应用中难以适用于对大规模蛋白质样品的定量分析。

近年来,尽管经过研究证明,DI-SPA 方法可以迅速并有效地对蛋白质样品进行处理和分析,且具备较高的重复性和鲁棒性^[66]。然而,当前 DI-SPA 质谱数据分析工具的发展尚不成熟,这在一定程度上也限制了 DI-SPA 方法在蛋白质组学研究中的广泛应用。为了解决这一难题,当务之急是研发出更为先进和完善的算法,以更有效地处理和分析 DI-SPA 质谱数据,从而推动蛋白质组学研究迈向新的高度。因此,在本章中,我们提出了一种全新的 DI-SPA 质谱数据分析方法 RE-FIGS (Repeat-Enhancing Featured Ion-Guided Stoichiometry)。该方法首先对实验质谱数据和图谱库数据进行预处理,之后利用线性判别分析模型对 SSM 结果进行打分,进而得到肽段定性结果,最后对实验图谱进行解谱以获取对应肽段的定量值。与大多数传统的 DIA 质谱数据分析方法不同,RE-FIGS 彻底摒弃了对色谱信息的依赖,可以对 DI-SPA 质谱数据进行准确的定性和无标记定量分析,从而促进了 DI-SPA 质谱数据在蛋白质组学领域中的发展和应用。

2.2 数据预处理

2.2.1 质谱的采集与处理

为了采集 DI-SPA 质谱数据, 首先需要将实验样品注入电喷雾发射器内进行电离, 待所有肽段母离子在气相中完全分离后, 再将其送入质谱仪中进行 DIA 检测。值得注意的是, 本章中的所有 DI-SPA 质谱数据都是循环采集了多个周期的数据。具体来说, 在电喷雾发射器内, 有 a 个补偿电压用于气相分离, 在每个补偿电压下, 需要进行 b 次 DIA 扫描才能覆盖所有质荷比范围内的肽段母离子。理论上, 只需要一个采集周期 ($a \cdot b$ 次 DIA 扫描) 就可以完成对实验样品的采集。但是, 由于 SSM 结果容易受到实验误差的干扰, 当只对实验样品采集一个周期时, 许多肽段图谱可能只匹配到少量的实验图谱, 此时 SSM 结果的误差可能会变得不可忽略, 这将不利于后续的质谱数据分析。通过对实验样品进行循环采集可以增加 SSM 的数量, 从而减少实验误差对分析结果的整体影响。此外, 我们还发现, 在对实验样品进行循环采集后, SSM 中的一些重复性特征能够得到更有效的提取和利用 (详见下一小节“2.3 定性算法”), 这使得我们的方法在对 DI-SPA 质谱数据进行定性分析时具有更出色的性能和准确度。同时, 由于 DI-SPA 质谱数据使用了速度较快的 FAIMS 技术来替代耗时较长的液相色谱, 其采集速度与传统的 DIA 质谱数据相比得到了 20 倍以上的提升, 因此当我们将采集周期数控制在 20 以内时, 其采集总时间仍然会少于传统的 DIA 质谱数据。

目前市场中存在着多种品牌的质谱仪器, 这些仪器都可以用来采集质谱数据。然而, 不同品牌的质谱仪所生成的原始质谱数据通常采用不同的数据格式。例如, Thermo Fisher Scientific 公司所制备的质谱数据主要采用 .raw 格式, 而 Applied Biosystems 公司所制备的质谱数据则主要采用 .wiff 格式, 这些多样化的数据格式使得研究人员在处理和析质谱数据时可能需要面对兼容性和格式转换问题。为了解决这一问题, Chambers 等人开发了 ProteoWizard^[67] 软件, 该软件中的 MSConvert 工具可以将常见的不同数据格式的质谱数据转换成 .mzML、.mzXML 等格式, 从而完成了对质谱数据格式的统一标准化。在本章的实验中, 我们选择将原始质谱数据统一转换成 .mzML 格式。

对于转换后的质谱数据而言, 每个数据文件都包含了若干张实验质谱。每张实验质谱中都有多个字段来描述自身的特征, 例如扫描号、扫描时间、窗口范围、子离子质荷比和对应的峰强度等。尽管质谱仪一般具有较高的准确性, 但是在采集质谱的过程中也会存在理论误差。对于子离子质荷比这维信息来说, 大部分质谱仪的质量误差 δ 在 15~30ppm (Part per million) 之间。因此, 在实验质谱中, 当离子峰之间的距离小于 δ 时, 这些峰可能源自同一个碎片离子, 也可能是受到了噪声峰的干扰。为了减少这一部分误差对后续分析的影响, 需要先对距离小

于 δ 的离子峰进行合并。峰合并的具体过程如算法2.1所述。对于实验质谱数据 E 中的第 i 张实验质谱 E_i 而言, 我们首先需要获取其离子峰质荷比列表 MZ_i 。随后, 我们计算出 MZ_i 整体右移 δ 个单位后的结果 $\widetilde{MZ}_i$, 并获取 MZ_i 所有元素在 $\widetilde{MZ}_i$ 中的排序位置。显然, 当 MZ_i 中的某些元素两两之间的距离都小于 δ 时, 这些元素的排序位置将是同一个值。之后, 我们计算出那些排序位置相同的离子峰的峰强度均值, 并将其作为合并后峰的峰强度。同时, 我们获取这些离子峰对应的最小质荷比, 并将其作为合并后峰的质荷比, 从而完成了对这些离子峰的峰合并操作。

算法 2.1 峰合并算法

Data: E : 实验质谱
 δ : 质谱仪质量误差
Result: E' : 峰合并后的实验质谱

```

1 for  $E_i$  in  $E$  do
2    $MZ_i = \text{getIonPeakMZ}(E_i)$ ; // 获取  $E_i$  的离子峰质荷比列表  $MZ_i$ 
3    $I_i = \text{getIonPeakIntensity}(E_i)$ ; // 获取  $E_i$  的离子峰强度列表  $I_i$ 
4    $\widetilde{MZ}_i = MZ_i + \delta \cdot MZ_i$ ; //  $\widetilde{MZ}_i$  是  $MZ_i$  整体右移  $\delta$  个单位后的结果
5    $S_i = \text{searchsorted}(\widetilde{MZ}_i, MZ_i)$ ; // 获取  $MZ_i$  所有元素在  $\widetilde{MZ}_i$  中的排序位置
6    $U_i = \text{unique}(S_i)$ ; //  $U_i$  是  $S_i$  去重后的结果
7    $MZ'_i = MZ_i[U_i]$ ;
8   for  $j = 1$  to  $\text{length}(U_i)$  do
9     for  $k = 1$  to  $\text{length}(S_i)$  do
10      if  $U_i[j] == S_i[k]$  then
11        intensity.add( $I_i[k]$ );
12      end
13    end
14    averageIntensity = average(intensity); // 对这些离子峰强度取均值
15     $I'_i.\text{add}(\text{averageIntensity})$ ;
16  end
17   $E'.\text{add}((MZ'_i, I'_i))$ ;
18 end
19 return  $E'$ ;

```

2.2.2 图谱库处理

在依赖于图谱库的 DIA 质谱数据分析方法中, 还需要为用户提供搜库所需要的肽段图谱库。目前常见的图谱库主要是 .mgf 和 .blib 等数据格式, 这些格式的数据都可以使用 Python 进行读取和分析。为了消除不同实验室所制备的图谱库中离子峰强度尺寸对分析结果的影响, 我们对图谱库 P 中的所有肽段图谱进行了强度归一化处理。这一步骤使得每张肽段图谱的峰强度之和均被调整为 1, 从而确保了后续质谱数据定性和定量分析的一致性。具体的归一化计算公式如下:

$$\bar{P}_i = \frac{P_i}{\sum_j P_j} \quad (2.1)$$

其中, P_i 表示图谱库 P 中第 i 张肽段图谱的原始强度向量, p_i^j 表示 P_i 的第 j 个元素, $\bar{P}_i$ 表示 P_i 归一化后的强度向量。

为了对后续的肽段定性结果进行准确的质量控制, 还需要用目标图谱库生成相应的反库, 这也被称为正反库策略^[68]。其中, 正库就是目标图谱库, 而反库则是由正库经过某种方式(如反转、打乱序列等)生成的, 可以用来模拟错误图谱和噪声图谱。在后续的定性分析中, 需要分别对正库和反库进行搜库, 并对定性到的肽段进行打分, 如果正库中部分肽段的打分值较高, 且反库中的肽段得分普遍较低, 则说明所使用的定性方法是可靠的。具体来说, 我们需要根据正库和反库的肽段打分情况来估计肽段定性结果的错误发现率(False discovery rate, FDR), 并按照目前学术界的公认标准, 从正库中筛选出那些 FDR 小于 0.01 的肽段作为最终的肽段定性结果。FDR 的计算公式如下:

$$\text{FDR}(S) = \frac{N_{\text{decoy}}}{N_{\text{target}}} \quad (2.2)$$

其中, N_{target} 和 N_{decoy} 分别表示正库和反库中肽段打分值高于 S 的肽段个数。

在构建反库的过程中, 需要满足以下两个基本条件: 首先, 正库与反库之间不能存在重复的肽段图谱; 其次, 正库和反库中的肽段图谱被错误匹配到的概率应当相同。只有在满足这两个基本条件的前提下, 我们才可以认为正库和反库中那些被错误定性到的肽段数量相同, 从而使得 FDR 可以有效地表示肽段定性结果中的假阳性肽段比例^[69]。目前常见的正反库策略包括目标-诱饵库策略^[70]和陷阱策略^[71]等, 这两种策略均已被证实满足以上两个基本条件。在本章的实验中, 我们采用了目标-诱饵库策略来生成相应的反库, 以便完成后续的 FDR 质量控制。算法2.2描述了生成诱饵库的大致流程。该算法的输入为目标图谱库 T 和最小交换距离 d , 输出为诱饵图谱库 D 。在该算法中, 对于 T 中的肽段图谱 T_i , 我们需要选择一个与其质荷比差值大于 d 且电荷量相同的肽段图谱 T_j , 并记录 T_i 与 T_j 的质荷比差值 Δ 。之后将 T_i 和 T_j 中的离子峰分别左移和右移 Δ 的距离, 并交换 T_i 和 T_j 的质荷比, 从而得到两张伪图谱, 并将其加入到 D 中。最后, 我们将当前处理的 T_i 与 T_j 从 T 中移除, 并重复上述过程, 直到 T 中不再存在符合条件的肽段图谱。

2.3 定性算法

在过去的一些 DI-SPA 质谱数据定性分析方法中(例如 MSPLIT-DIA^[23]和 CsoDIAq^[62]), 需要利用自定义的评分规则来计算 SSM 分数, 并基于该分数来识别图谱库中的肽段。因此, 评分规则的设计对于最终结果的准确性至关重要。然而, 在设计评分规则时, 人们容易受到主观性和局限性的影响, 同时考虑到质

算法 2.2 诱饵库生成算法

Data: T : 目标图谱库
 d : 最小交换距离
Result: D : 诱饵图谱库

```

1 for  $T_i$  in  $T$  do
2    $MZ_i = \text{getPrecursorMZ}(T_i)$ ; // 获取  $T_i$  的肽前体质荷比  $MZ_i$ 
3    $C_i = \text{getCharge}(T_i)$ ; // 获取  $T_i$  的电荷量  $C_i$ 
4    $S_i = \text{getSpectrum}(T_i)$ ; // 获取  $T_i$  的实验图谱  $S_i$ 
5   for  $T_j$  in  $T$  do
6      $MZ_j = \text{getPrecursorMZ}(T_j)$ ;
7      $C_j = \text{getCharge}(T_j)$ ;
8      $S_j = \text{getSpectrum}(T_j)$ ;
9      $\Delta = MZ_i - MZ_j$ ;
10    if  $|\Delta| > d$  and  $C_i == C_j$  then
11       $S_i = S_i - \Delta$ ; // 将  $S_i$  所有的离子峰左移  $\Delta$  的距离
12       $S_j = S_j + \Delta$ ; // 将  $S_j$  所有的离子峰右移  $\Delta$  的距离
13       $D.\text{add}((MZ_j, C_j, S_j))$ ;
14       $D.\text{add}((MZ_i, C_i, S_i))$ ;
15      break; // 跳出 for 循环
16    end
17  end
18   $T.\text{remove}(T_i)$ ;
19   $T.\text{remove}(T_j)$ ;
20 end
21 return  $D$ ;

```

谱数据本身的高度复杂性, 要制定出一个能够适用于所有质谱数据的通用评分规则可能会变得非常困难。为了应对这一难题, 我们针对 DI-SPA 质谱数据, 设计了一种基于半监督学习的定性算法, 并将其命名为 RE-FIGS (Repeat-Enhancing Featured Ion-Guided Stoichiometry) 定性算法。该算法不需要人工设计复杂的 SSM 评分规则, 而是通过线性判别分析模型来对 SSM 结果进行自动的打分。RE-FIGS 定性算法的大致流程如图2.2和算法2.3所示。

首先, 我们利用 CsoDIAq 工具分别得到正库和反库的 SSM 原始结果, 其中包括扫描号、肽段母离子的名称和电荷数、共享峰数量和 MACC (Match count and cosine) 分数等字段信息。在实际的实验中, 一个肽段图谱往往能够匹配到多个实验图谱, 其中也可能混杂了许多干扰图谱, 所以需要对 SSM 结果进行筛选。CsoDIAq 中的研究表明, MACC 分数可以很好地反映 SSM 的质量, 因此我们为每个肽段图谱选取 MACC 分数最高的 SSM。之后, 为了更加准确地描述 SSM 的置信度, 我们提取并计算了 SSM 的若干个相关特征, 例如余弦相似度、均方误差、平均绝对误差和肽段匹配周期数等, 其中肽段匹配周期数表示所有采集周期中肽段图谱匹配的总次数 (每个采集周期最多计数一次)。在上一节中, 我们已经提到本章中的所有 DI-SPA 质谱数据都是经过多次循环采集的数据, 因此肽段匹配周期数的取值范围为 $1 \sim M$, 其中 M 表示 DI-SPA 质谱数据的循环采集次

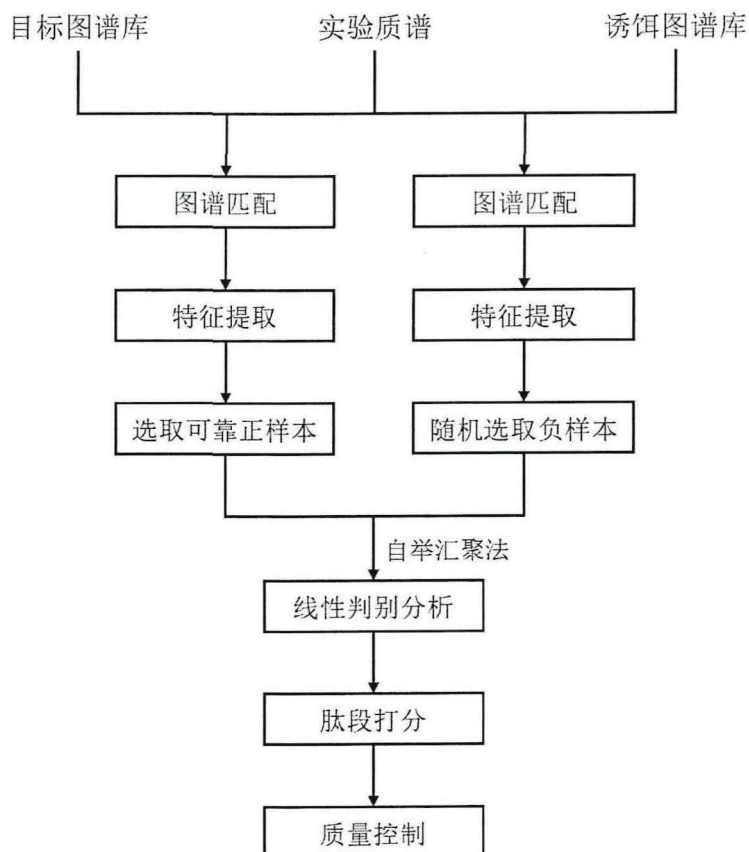


图 2.2 RE-FIGS 定性算法流程图

数, 当且仅当肽段图谱在每个采集周期中至少匹配一次时, 肽段匹配周期数等于 M 。由于图谱匹配具有随机性, 因此某些实验图谱可能会匹配到一些错误的肽段图谱, 但是这些错误的肽段图谱很难重复出现在各个周期下的 SSM 结果中, 因此肽段匹配周期数可以很好地区分正确肽段和错误肽段。

在提取了 SSM 的相关特征后, 需要进行分类模型的训练。在本实验中, 训练集中的负样本可以从反库的 SSM 结果中获取, 但是由于正库中可能存在着与实验样品无关的肽段图谱, 因此正库的 SSM 结果应当被视为未标记样本。这意味着在训练的过程中, 我们需要对这些未标记样本进行特殊处理或过滤, 以确保分类模型的准确性和可靠性。对于这类问题, 我们可以参考半监督学习中的 PU Learning (Positive-unlabeled learning)^[72] 方法, 该方法包含了若干种解决方案, 其中较为常见的一种方法被称为两步法 (Two-step approaches)^[73], 其基本步骤如下:

1. 从未标记样本中选择出可靠的样本进行标记
2. 利用已标记样本和上一步中得到的可靠样本训练一个分类模型, 并应用于新的样本

具体来说, 我们首先会从正库的 SSM 结果中提取一些关键特征, 例如余弦相似度和共享峰数量等。只有当这些特征值满足特定的阈值条件时, 我们才会

算法 2.3 RE-FIGS 定性算法

Data: E : 实验质谱
 T : 目标图谱库
 D : 诱饵图谱库
 t : 训练集正样本特征阈值
 m : LDA 分类器数量

Result: R : 肽段定性结果

```

1 targetSSM = spectralMatch( $E$ ,  $T$ ); // 对  $E$  和  $T$  进行图谱匹配
2 decoySSM = spectralMatch( $E$ ,  $D$ );
3 targetFeature = featureExtraction(targetSSM); // 提取 targetSSM 的有关特征
4 decoyFeature = featureExtraction(decoySSM);
5 positiveSample = thresholdFilter(targetFeature,  $t$ ); // 从 targetFeature 中选取可靠正样本
6 for  $i = 1$  to  $m$  do
7     negativeSample = random(decoyFeature); // 从 decoyFeature 中随机选取负样本
8      $LDA_i$  = trainLDA(positiveSample, negativeSample); // 训练 LDA 分类器
9     LDAList.add( $LDA_i$ );
10 end
11 LDA = averageWeight(LDAList); // 对每个 LDA 分类器的权重值取平均
12 targetScore = peptideScore(LDA, targetFeature); // 对 targetFeature 中的肽段进行打分
13 decoyScore = peptideScore(LDA, decoyFeature);
14  $S_t$  = FDR(targetScore, decoyScore); // 计算 FDR 阈值  $S_t$ 
15  $R$  = filter(targetFeature,  $S_t$ ); // 从 targetFeature 中筛选出打分值高于  $S_t$  的肽段
16 return  $R$ ;
```

将对应的 SSM 作为正样本添加到训练集中, 同时将其标签设置为 1。之后, 我们从反库的 SSM 结果中随机选取相同数量的负样本加入到训练集中, 并为这些负样本赋予标签 0。接下来, 我们利用线性判别分析 (Linear discriminant analysis, LDA)^[74] 来训练分类模型。LDA 的主要思想是将高维数据映射到低维空间, 使得不同类别的数据点在新的空间中尽可能地分开, 而同一类别的数据点则尽可能靠近。在这里, 我们的 LDA 分类器将之前所提取的 SSM 相关特征作为输入, 之后通过计算这些特征的加权和来生成 SSM 的打分值。同时, 为了避免 LDA 分类器出现过拟合的现象, 我们在训练 LDA 分类器时采用了一种被称为自举汇聚法 (Bootstrap aggregating)^[75] 的集成学习方法。具体来说, 我们一共训练了 m 个 LDA 分类器 ($m = 10$), 在训练这些分类模型时, 它们的训练集中的正样本是完全相同的, 而负样本则是从反库的 SSM 结果中随机选取的。之后, 我们获取每个 LDA 分类器计算得到的特征权重值, 并取它们的平均值作为最终的特征权重值, 即:

$$\bar{w}_i = \frac{\sum_{j=1}^m w_i^j}{m} \quad (2.3)$$

其中, $\bar{w}_i$ 表示第 i 个特征的平均权重值, w_i^j 表示第 j 个 LDA 分类器的第 i 个特征的权重值。

因此, SSM 的最终打分值可以表示为:

$$S = \sum_{i=1}^n \bar{w}_i f_i \quad (2.4)$$

其中, SSM 共包含了 n 个特征, f_i 表示 SSM 的第 i 个特征值。

在训练完分类模型后, 我们将正库和反库的 SSM 结果的相关特征输入到模型中进行计算, 得到它们的打分结果。最后, 我们还需要对打分结果进行 FDR 质量控制, 保留正库中打分值高于 FDR 阈值的肽段作为最终的肽段定性结果。由于 FDR 质量控制的相关公式 (公式2.2) 及流程在前文中已经说明, 因此在这里不再赘述。

2.4 定量算法

在过往的 DIA 质谱数据定量分析方法中, 往往通过计算碎片离子的色谱峰面积来得到肽段的定量值, 这类方法的代表工作包括 Spectronaut^[54] 和 DIA-NN^[56] 等。然而, DI-SPA 质谱数据并没有传统的色谱峰信息, 因此无法使用上述方法对其进行定量分析。

FIGS^[61] 是近几年提出的一种 DIA 质谱数据分析方法, 该方法中的 Fisdec 图谱分解算法可以对实验质谱进行准确的图谱分解, 其解谱系数表示相应肽段在某一时间点下的相对丰度。之后, FIGS 会对系数曲线进行修正和积分, 并将积分结果作为肽段的定量值。虽然 FIGS 只能针对传统的有色谱数据构造系数曲线并计算定量值, 因此无法直接应用在 DI-SPA 质谱数据上, 但是其图谱分解过程并不受色谱信息的限制。因此, 我们接下来将利用 Fisdec 图谱分解算法为 DI-SPA 质谱数据设计一种定量算法, 并将此定量算法命名为 RE-FIGS 定量算法。

RE-FIGS 定量算法的具体步骤如下: 首先, 我们从定性阶段的 SSM 最终结果中获取那些被定性到的肽段与相应实验图谱的匹配关系, 其中每个肽段都唯一对应了一个 MACC 分数最高的实验图谱。由于 MACC 分数可以很好地反映 SSM 的置信度, 因此这些实验图谱与相应肽段的相关性较高, 可以对这些实验图谱进行解谱, 并使用解谱系数 (即对应肽段的瞬时相对丰度) 来表示相应肽段的定量值。因此, 之后我们将会利用 Fisdec 图谱分解算法来对这些实验图谱进行解谱。Fisdec 图谱分解算法的大致过程如算法2.4所示。该算法使用动态去卷积 (Dynamic deconvolution) 策略来获取肽段图谱的特征峰, 然后构造特征峰强度方程, 并利用最小二乘法来求解对应的解谱系数, 在这里, 特征峰表示肽段图谱在图谱库中独有的离子峰。最后, 我们使用这些实验图谱的解谱系数来表示对应肽段的定量值, 从而完成了对 DI-SPA 质谱数据的定量分析。

算法 2.4 Fisdec 图谱分解算法

Data: E : 单个实验图谱
 P : 肽段图谱库
 n : 峰匹配阈值

Result: R : 解谱结果

```

1 for  $P_i$  in  $P$  do
2    $F_i = \text{getFeaturedPeak}(P_i, P)$ ; // 计算得到  $P_i$  的特征峰  $F_i$ 
3 end
4 for  $P_i$  in  $P$  do
5   if  $\text{matchPeakNum}(E, P_i) \geq n$  then
6      $M_i = \text{getMatchPeak}(E, F_i)$ ; // 获取  $E$  中与  $F_i$  所匹配的离子峰  $M_i$ 
7      $C_i = \text{leastSquareMethod}(F_i, M_i)$ ; // 利用最小二乘法计算得到解谱系数  $C_i$ 
8      $N_i = \text{getPeptideName}(P_i)$ ; // 获取  $P_i$  对应的肽段名称  $N_i$ 
9      $R.\text{add}((N_i, C_i))$ ;
10  end
11 end
12 return  $R$ ;

```

2.5 实验结果与分析

2.5.1 定性数量对比

为了公正地评估 RE-FIGS 在 DI-SPA 质谱数据上的定性性能,我们直接使用 DI-SPA^[28]论文中提供的部分 DI-SPA 质谱数据集和图谱库进行实验验证,并将其定性结果与 CsoDIAq^[62]工具生成的定性结果进行对比。其中的 DI-SPA 质谱数据集来自于全人类蛋白质组蛋白水解后的肽样品(MCF7 肽样品),图谱库则是一个较为通用的大型图谱库,共包含了超过 120,000 个肽段图谱。

首先,我们使用两组不同分辨率的 DI-SPA 质谱数据来测试 RE-FIGS 和 CsoDIAq 的定性能力,其中的分辨率表示质谱采集过程所使用的质谱仪的分辨率。高分辨率的质谱仪能够更准确地分离和检测质量相近的肽段母离子,并提供更精细的实验图谱。因此,一般来说,质谱仪的分辨率越大,所采集的质谱数据的质量也会越高。对于本实验来说,每组 DI-SPA 质谱数据的分辨率包括 15k, 30k, 60k 和 120k。而两组 DI-SPA 质谱数据之间的差异在于质谱采集过程中的隔离窗口宽度不同,其中第一组数据的隔离窗口宽度为 2 Thomson (Th),第二组数据的隔离窗口宽度为 6 Th。这里的隔离窗口表示质谱仪在进行肽段母离子碎裂时设置的窗口,隔离窗口宽度越小,同一窗口下的肽段母离子越少,从而使得所生成的质谱数据的复杂度也会越低。

在比较 RE-FIGS 和 CsoDIAq 的肽段定性结果之前,我们首先分析了 RE-FIGS 在这些 DI-SPA 质谱数据上的肽段打分情况。图2.3展示了 RE-FIGS 在其中一个样本上的目标库及诱饵库肽段的打分分布情况。当 FDR 等于 0.01 时,对应的肽段分数为 81.842,我们使用该分数作为 FDR 质量控制的分数阈值,如图2.3中的红色虚线所示。在该虚线右侧,所有目标库肽段的 FDR 均小于 0.01,最终这些

肽段会被我们过滤出来作为该样本的肽段定性结果。从图2.3中可以看出，诱饵库中的肽段几乎都分布在红色虚线的左侧，这说明 RE-FIGS 可以较为精准地确定这些与实验样品无关的错误肽段。此外，我们还注意到目标库中分数较低的肽段与诱饵库中的肽段具有相似的打分分布，这表明我们采用的诱饵库能够很好地模拟目标库中的错误肽段。

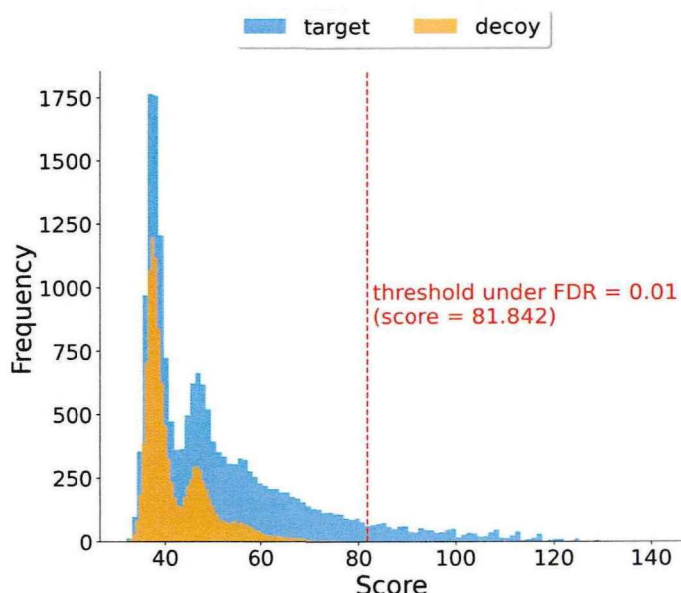


图 2.3 RE-FIGS 肽段打分分布图

除了利用学术界公认的 FDR 指标来对肽段定性结果进行质量控制外，我们还采用了假设检验^[76]方法来验证其可靠性。对于肽段定性结果中某个打分为 s 的肽段 Pep 而言，我们将原假设设定为 Pep 是由于随机误差而被定性到的肽段，备择假设设定为 Pep 是实验样品中真实存在的肽段。之后，我们需要对 P 值进行计算。在假设检验方法中， P 值表示当原假设成立时，观察到当前统计结果或更极端结果的概率。在这里，观测样本 Pep 的打分为 s ，更极端的情况则是 Pep 的打分值大于 s ，因此这里的 P 值表示当原假设为真时， Pep 的打分值大于等于 s 的概率。显然， P 值越小，则表明我们拒绝原假设的理由越充分，即 Pep 为正确肽段这一事件的可信度越高。为了计算 P 值，我们参考了 Granholm 等人发表的论文^[68]，其具体计算公式为：

$$P = \frac{r+1}{n+1} \quad (2.5)$$

其中， r 表示诱饵库中打分值大于等于 s 的肽段数量， n 表示诱饵库中的肽段总数量。

我们在显著性水平设为 0.01 的情况下，利用公式 2.5 对图 2.3 中的样本进行了相关的计算。结果表明，只有当肽段得分 s 的值大于等于 51.833 时，对应的 P 值才会小于等于 0.01。因此，我们可以得到拒绝域为 $\{s \geq 51.833\}$ 。由于在前文中我

们为该样本设定的肽段分数阈值为 81.842, 这一阈值大于拒绝域的下界 51.833, 同时该样本所有被定性到的肽段的打分值均大于肽段分数阈值。因此, 对于该样本肽段定性结果中的每一个肽段而言, 我们都可以拒绝原假设, 从而说明这些肽段都很有可能真实存在于实验样品中, 这在一定程度上也反映出我们的肽段定性结果具备较高的可靠性。

之后, 我们对 RE-FIGS 和 CsoDIAq 的肽段定性结果进行了比较, 如图2.4所示。其中图2.4a展示了 RE-FIGS 和 CsoDIAq 在每个数据点上的肽段定性数量, 可以看出对于绝大多数 ($7/8 = 87.5\%$) 数据点来说, RE-FIGS 的肽段鉴定数量都多于 CsoDIAq。经计算, RE-FIGS 在所有数据点上平均鉴定出 733 条肽段, 比 CsoDIAq 多了 7.9%。之后我们将两种方法的定性结果各自取并集, 再统计两者的交集, 得到韦恩图如图2.4b所示。可以看出, 两种方法鉴定出的肽段集合有着很高的重合度。在 CsoDIAq 的鉴定结果中, 有 94.6% 的肽段也出现在了 RE-FIGS 的定性结果中, 并且 RE-FIGS 比 CsoDIAq 鉴定到了更多的独有肽段 (357 vs. 106)。

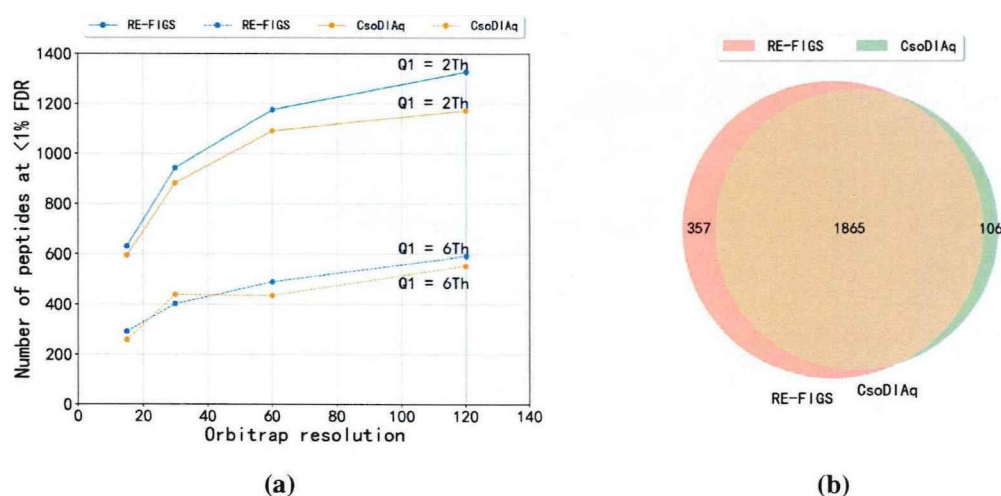


图 2.4 不同分辨率的 DI-SPA 质谱数据肽段定性结果对比

接下来, 我们还分析了一组 100 次重复采集的 DI-SPA 质谱数据, 每个 DI-SPA 质谱数据都来自于同一种 MCF7 肽样品。这一部分实验的定性结果如图2.5所示。从图2.5a中可以看出, 对于几乎所有的数据点 ($98/100 = 98\%$) 来说, RE-FIGS 都比 CsoDIAq 鉴定到更多的肽段。在对 100 次重复实验的肽段定性数量取平均后, RE-FIGS 鉴定出 1883 条肽段, 比 CsoDIAq 多鉴定到 19.3% 的肽段。之后我们统计了 RE-FIGS 和 CsoDIAq 的定性结果中那些曾经多次出现 (在所有重复实验的定性结果中至少出现了 80%) 的肽段, 并绘制了相应的韦恩图如图2.5b所示。在 CsoDIAq 鉴定到的 1133 条肽段中, 有 85.5% 的肽段也出现在了 RE-FIGS 的定性结果中, 并且 RE-FIGS 比 CsoDIAq 多鉴定到 86.6% 的独有肽段。

根据以上的结果, 我们可以看出, 与 CsoDIAq 相比, RE-FIGS 具备更强的

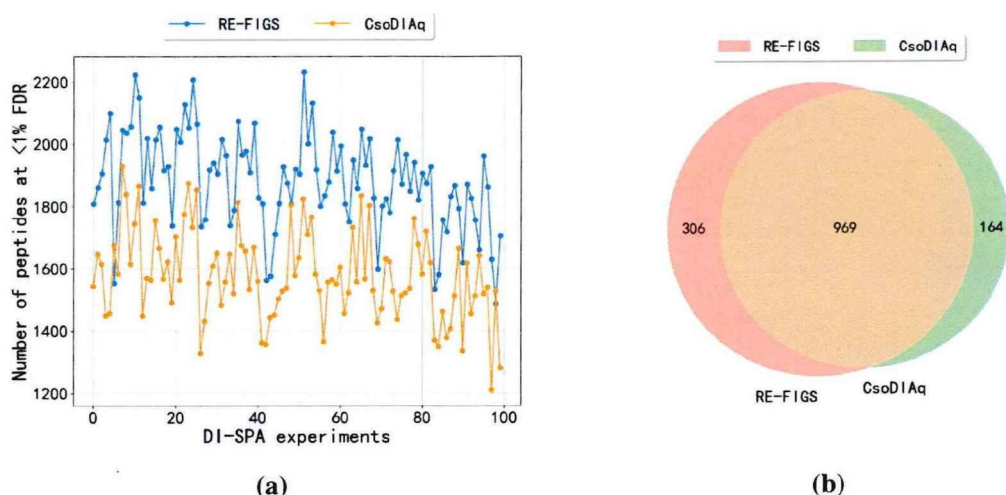


图 2.5 100 次重复采集的 DI-SPA 质谱数据肽段定性结果对比

肽段鉴定能力，其肽段定性结果不仅覆盖了 CsoDIAq 定性到的大多数肽段，而且包含了更多的独有肽段。同时 RE-FIGS 和 CsoDIAq 在定性结果上的高度重合，也侧面验证了 RE-FIGS 定性结果的可靠性。

2.5.2 定性参数研究

在本章的第二小节“2.2 数据预处理”中，我们已经提到本章所有的 DI-SPA 质谱数据都是循环采集了多个周期的数据。这一设计虽然为后续的定性分析提供了更多的特征信息，但也使得质谱采集过程中的时间消耗随之增多。为了进一步探索 DI-SPA 质谱数据的最佳采集参数，以便在较短的实验时间内获得较优的定性性能，我们针对质谱采集过程中的两个参数进行了对照实验，这两个参数分别是质谱数据的采集周期数和质谱仪中隔离窗口的注射时间。为了方便对比，我们同样采用 RE-FIGS 和 CsoDIAq 这两种方法分别对实验质谱数据进行肽鉴定，在定性分析过程中使用的肽段图谱库仍然为 DI-SPA 论文中提供的图谱库数据。值得注意的是，在这一部分的实验中，CsoDIAq 始终使用了全部采集周期下的 DI-SPA 质谱数据进行分析，而 RE-FIGS 则根据具体情况调整了程序运行时的参数，使其最终使用了全部或部分采集周期下的 DI-SPA 质谱数据进行分析。

首先，我们探讨了采集周期数这一参数对定性结果的影响。为此，我们先使用 DI-SPA 论文中提供的 DI-SPA 质谱数据集进行一组对比实验。在该实验中，我们分析了一组不同分辨率的 DI-SPA 质谱数据，这些数据的分辨率分别设置为 15k, 30k, 60k 和 120k，相应的总采集周期数为 3, 3, 3 和 2。这一部分数据的实验结果如图 2.6a 所示，图中的实线表示 RE-FIGS 的肽段定性结果，虚线表示 CsoDIAq 的肽段定性结果。从图 2.6a 中可以发现，当 RE-FIGS 只使用一个周期的质谱数据进行定性分析时，其鉴定的肽段数量略少于 CsoDIAq。然而，随着使用

周期数的增加, RE-FIGS 识别到的肽段数量也在增加。当 RE-FIGS 使用了两个或三个周期的质谱数据后, 其肽段鉴定数量就超过了 CsoDIAq。

接下来, 为了更充分地说明采集周期数对 RE-FIGS 定性结果的影响, 我们在 Hela 细胞样本的 DI-SPA 质谱数据上进行了定性实验。该数据集由上海生命科学院提供, 共包含了 12 个重复采集的质谱数据。针对这批数据, 我们分别使用 RE-FIGS 和 CsoDIAq 对其进行定性分析, 并统计了这两种方法平均鉴定到的肽段数量和至少在定性结果中出现 10 次 (至少出现 80%) 的肽段数量, 实验结果如图 2.6b 所示, 图中的实线和虚线分别表示 RE-FIGS 和 CsoDIAq 的肽段定性结果。从图中可以看出, RE-FIGS 的肽段定性数量与其所使用的质谱数据的周期数呈现正相关关系。值得注意的是, 当使用的周期数不超过 4 时, RE-FIGS 的定性数量随着周期数的增加而呈现出较为明显的增长趋势, 而当周期数超过 4 后, 这种增长趋势就开始逐渐放缓。并且, 在本实验中, RE-FIGS 在所有的数据点上都比 CsoDIAq 鉴定出更多的肽段, 这也充分说明了 RE-FIGS 出色的肽鉴定能力。

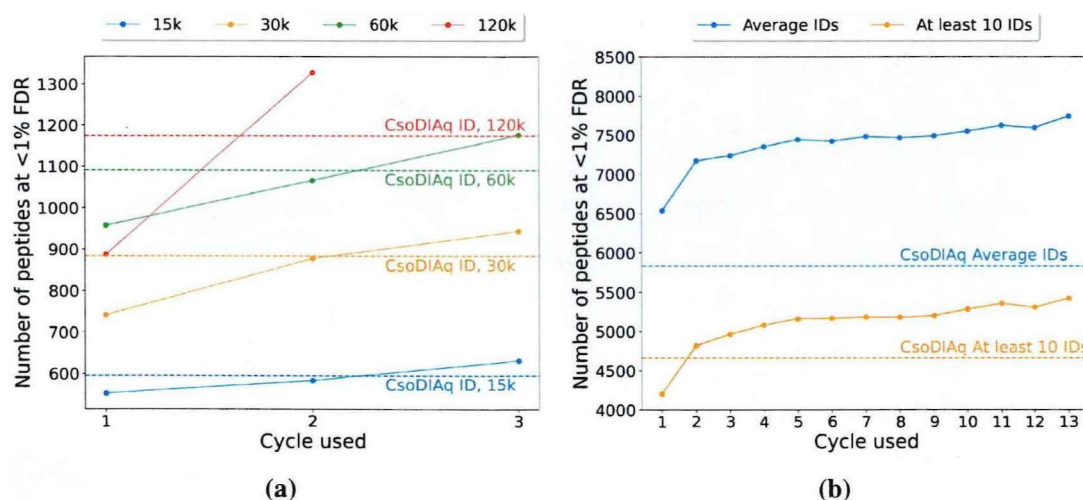


图 2.6 不同采集周期数下的肽段定性结果对比

之后, 我们研究了隔离窗口注射时间这一参数对定性结果的影响。一般来说, 质谱仪中隔离窗口的注射时间越短, 肽段母离子的捕获时间也会越短。较短的肽段母离子捕获时间可能会导致质谱数据的质量下降, 进而对定性结果的准确性产生不利影响。因此, 在实验中需要平衡注射时间和定性性能之间的关系。在这部分实验中, 我们首先使用 RE-FIGS 和 CsoDIAq 对 DI-SPA 论文中提供的一批不同注射时间的 DI-SPA 质谱数据进行定性分析。这批质谱数据共分为两组, 第一组数据的隔离窗口宽度为 1 Th, 第二组数据的隔离窗口宽度为 2 Th。每组 DI-SPA 质谱数据的注射时间包括 30ms, 60ms, 180ms 和 360ms。这两组质谱数据的定性结果如图 2.7a 所示。显然, 从图中可以看出, 随着注射时间的增加, RE-FIGS 和 CsoDIAq 所识别到的肽段数量也在增加。此外, RE-FIGS 在所有数

据点上平均鉴定到 989 条肽段，比 CsoDIAq 多鉴定出 8.6% 的肽段。

接下来，为了进一步验证注射时间与定性结果之间的关系，我们在上海生命科学院提供的三组 DI-SPA 质谱数据上进行了定性实验。这三组数据均来源于 HeLa 细胞样本，其注射时间分别为 30ms、60ms 和 120ms，每组数据都包含 3 个重复采集的质谱数据。对于这些质谱数据，我们使用 RE-FIGS 对其进行定性分析，并记录每个数据的肽段鉴定数量绘制成图 2.7b。从图中可以看出，随着注射时间的增加，RE-FIGS 的肽段鉴定数量也在增加。经计算可知，当注射时间从 30ms 增加到 60ms 后，RE-FIGS 的平均肽段鉴定数量增加了 52.4%；而当注射时间从 60ms 增加到 120ms 后，RE-FIGS 的平均肽段鉴定数量只增加了 10.5%。

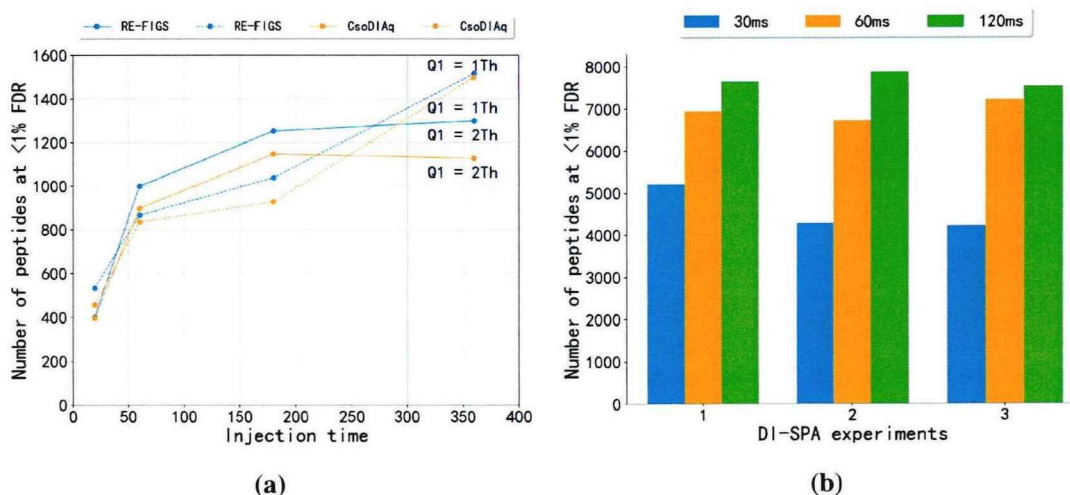


图 2.7 不同注射时间下的肽段定性结果对比

在综合考虑质谱数据的总采集时间和最终定性效果后，对于我们的 HeLa 数据而言，将隔离窗口注射时间设置为 60ms 并选择采集 4 个周期的质谱数据可能是一组较为理想的实验参数。最后，我们测试了 RE-FIGS 在该参数设置下的定性效果，并将其与 CsoDIAq 进行比较。在这里，我们选择之前所采集的注射时间为 60ms 的一组 HeLa 重复样品数据作为实验数据，其中 RE-FIGS 只使用了 4 个周期的质谱数据，而 CsoDIAq 则使用了全部采集周期下的质谱数据（共计 13 个周期）。该实验的定性结果如图 2.8 所示。从图 2.8a 中可以看出 RE-FIGS 在每个数据点上都比 CsoDIAq 鉴定到了更多的肽段，并且在统计了两种方法的平均肽段鉴定数量后发现，RE-FIGS 平均识别出 6694 条肽段，比 CsoDIAq 多出 21.0%。之后我们将 RE-FIGS 和 CsoDIAq 的定性结果各自取交集，再统计两者的交集情况，得到韦恩图如图 2.8b 所示。经计算可知，在 CsoDIAq 鉴定到的肽段中有 76.8% 的肽段也出现在了 RE-FIGS 的定性结果中，同时 RE-FIGS 比 CsoDIAq 多鉴定到了 24.4% 的独有肽段。以上结果表明，在使用我们推荐的采集周期数和注射时间的情况下，RE-FIGS 仍然具备比 CsoDIAq 更强的肽段鉴定能力。

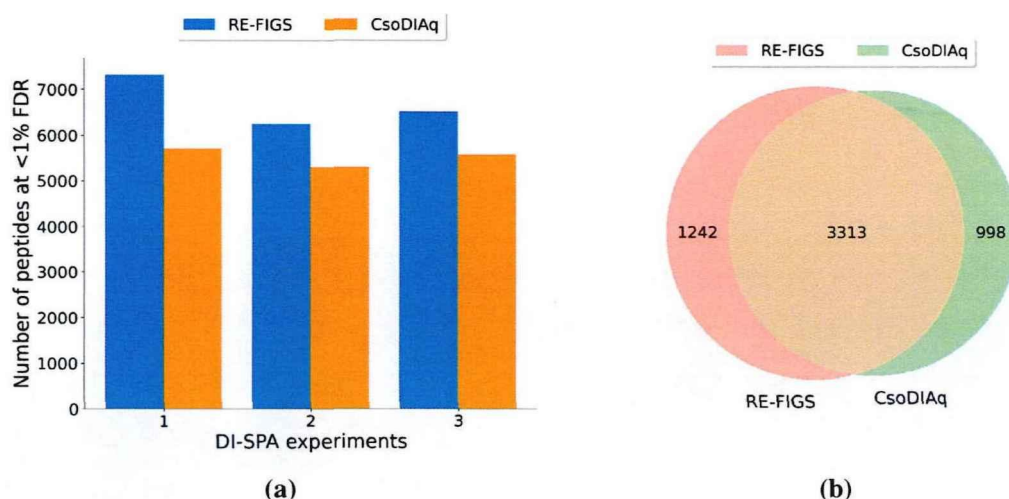


图 2.8 最优实验参数下的肽段定性结果对比

2.5.3 定量性能评估

无标记定量省去了耗时较长的同位素标记过程，提升了大规模蛋白质组学的分析效率，因此在当前的蛋白质组学实验中被广泛地应用。在关于 DI-SPA 质谱数据无标记定量实验的相关研究中，CsoDIAq 可以初步地估计肽段的无标记定量值。下面我们就分别使用 RE-FIGS 和 CsoDIAq 对一些未经过同位素标记的实验样品进行定量分析。

首先，我们在 DI-SPA 论文中的 MCF7 肽样品数据上进行了定量实验，具体所测试的 DI-SPA 质谱数据为其中某次采集的数据，定量过程中使用的肽段图谱库仍然为 DI-SPA 论文中提供的图谱库数据。我们在得到了 RE-FIGS 和 CsoDIAq 的肽段定量结果后，绘制肽段丰度的频数分布直方图，如图2.9所示，图中的横坐标表示肽段的定量值（取以 2 为底的对数后的结果），纵坐标表示肽段定量值落在某一区间内的肽段数量。从图中可以看出 RE-FIGS 的定量结果跨越了 12 个数量级左右，而 CsoDIAq 的定量结果只跨越了 4-5 个数量级，这说明 RE-FIGS 可以定量到丰度范围更广的肽段。

为了评估 RE-FIGS 在 DI-SPA 质谱数据上的定量准确性，我们在一批多物种混合肽段样品数据上进行了定量实验。该数据集由上海生命科学院提供，其采集方式与 LFQbench^[77] 论文中的标准数据集类似，所使用到的生物样品为人类 (Human)、酵母菌 (Yeast) 和大肠杆菌 (E.coli) 肽段的混合物，其中包括样品 A 和样品 B 两种类型的实验样本，每种样本都各自制备了 5 次重复的质谱数据。对于这两种样品而言，它们的肽段总含量完全相同，不同之处则在于两者的肽段混合比例存在差异，其具体的比例参数如表2.1所示。至于定量所需要的肽段图谱库，我们则使用了 LFQbench 论文中所提供的图谱库数据。

我们在使用 RE-FIGS 得到 A、B 样品五次重复样本的定量结果后，需要先

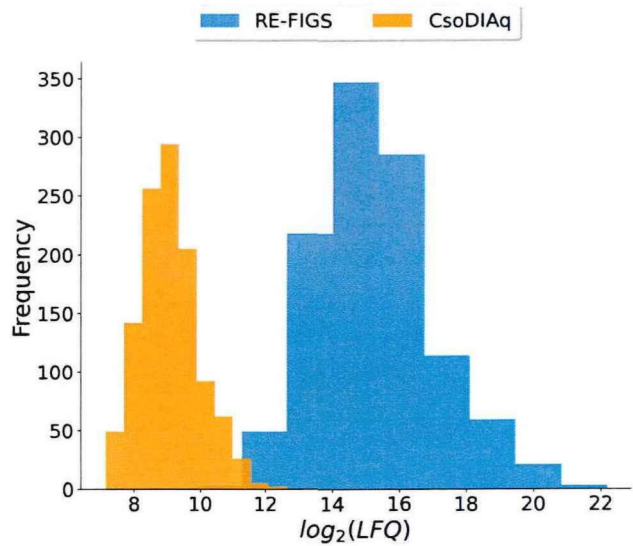


图 2.9 MCF7 肽样品肽段定量结果直方图对比

表 2.1 A、B 样品肽段混合比例表

样品名	人类	酵母菌	大肠杆菌
A	65%	30%	5%
B	65%	15%	20%

对其进行筛选，目的是选出其中定量重复性较高的肽段。筛选过程如下：首先，我们需要获取那些在 A、B 样品的定量结果中都至少出现 4 次的肽段，并分别计算这些肽段在 A、B 样品的定量结果中的变异系数 CV (Coefficient of variation)。 CV 的计算公式如下：

$$CV = \frac{\mu}{\sigma} \tag{2.6}$$

其中， μ 和 σ 分别表示对应数据的均值和标准差。

之后，我们筛选出那些 CV_A 和 CV_B 均小于 0.2 的肽段，并计算这些肽段的对数比 LR (Logarithmic ratio)，其计算公式为：

$$LR = \log_2 \frac{\mu_A}{\mu_B} \tag{2.7}$$

其中， μ_A 和 μ_B 分别表示肽段在 A、B 样品的定量结果中的均值。

最后我们利用计算得到的结果绘制了这些肽段对应的散点图，如图2.10a所示，图中的横坐标表示肽段在样品 B 中的平均定量值的对数，纵坐标表示肽段的对数比，虚线则表示每个物种肽段的理论对数比。根据表2.1可知，理论上来说，对于样品 A 和样品 B 中的同一肽段，当该肽段来自于人类、酵母菌或大肠杆菌时，样品 A 和样品 B 的肽段定量结果的理论对数比分别为 0、1 和-2。从图2.10a中可以看出，大多数肽段都位于理论线附近。而在经过详细计算后可知，人类、酵母菌和大肠杆菌肽段的对数比的中位数分别是 0.108、1.003 和-1.781，这与理论

对数比也较为接近。作为对比实验,我们使用 CsoDIAq 对 A、B 样品也进行了定量分析,并按照和 RE-FIGS 一样的流程绘制了定量结果散点图,如图2.10b所示。从图中可以看出,在 CsoDIAq 的肽段定量结果中,虽然大部分人类肽段都位于理论线附近,但酵母菌和大肠杆菌肽段的对数比误差较大,大多数酵母菌和大肠杆菌肽段已经明显偏离了理论线,部分丰度较低的大肠杆菌肽段的对数比误差甚至已经接近于 2。而在进行详细计算后可知,对于 CsoDIAq 的定量结果而言,人类肽段的对数比的中位数为 0.055,与理论对数比十分接近,但酵母菌和大肠杆菌肽段的对数比的中位数分别为 0.492 和-0.879,这与两者的理论对数比分别有着 0.5 和 1 以上的差距。因此,综合以上结果后,我们可以发现 RE-FIGS 在定量准确性方面显著优于 CsoDIAq。

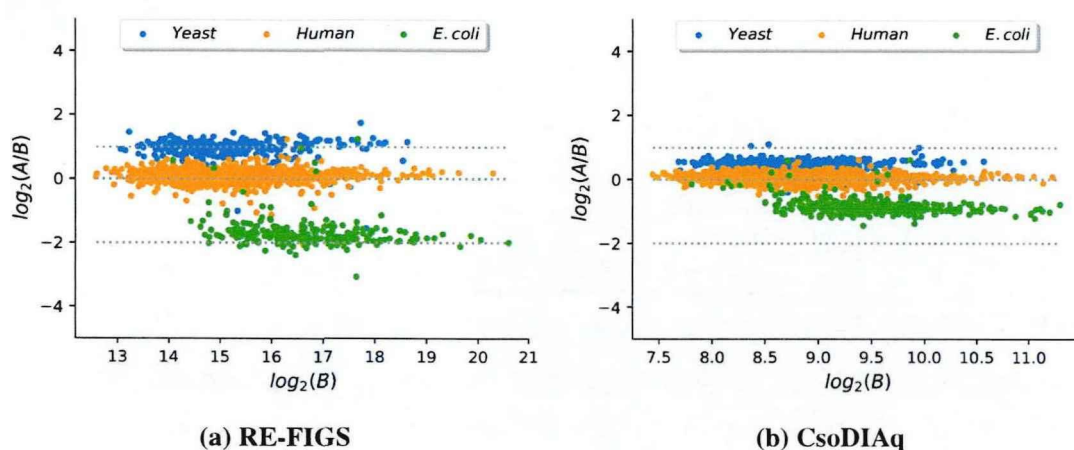


图 2.10 多物种混合肽段样品肽段定量结果散点图对比

为了更细致地评估 RE-FIGS 在 A、B 样品上的定量效果,我们根据肽段在样品 B 中的平均丰度将每个物种的肽段各自平均分成了三等份,并绘制了相应的肽段对数比热力图,如图2.11所示。从图中可以看出,对于人类、酵母菌和大肠杆菌肽段来说,分别有 98.4%、93.7% 和 84.1% 的肽段的对数比误差不超过 0.5,这充分说明了 RE-FIGS 在 DI-SPA 质谱数据的定量分析方面具备较好的准确性。值得注意的是,与前文中 CsoDIAq 的肽段定量结果类似,我们可以发现在 RE-FIGS 的肽段定量结果中,大肠杆菌肽段的定量准确性相对其他两种肽段较差,特别是对于低丰度的大肠杆菌肽段来说,有 26.2% 的肽段的对数比误差都超过了 0.5。结合表2.1中的比例参数来看,我们推测这可能是因为混合样品中,大肠杆菌肽段特别是低丰度大肠杆菌肽段的含量较低,使其在质谱数据中对应的碎片离子容易受到噪声峰和其他物种肽段子离子的影响,从而导致最终的定量准确性偏低。

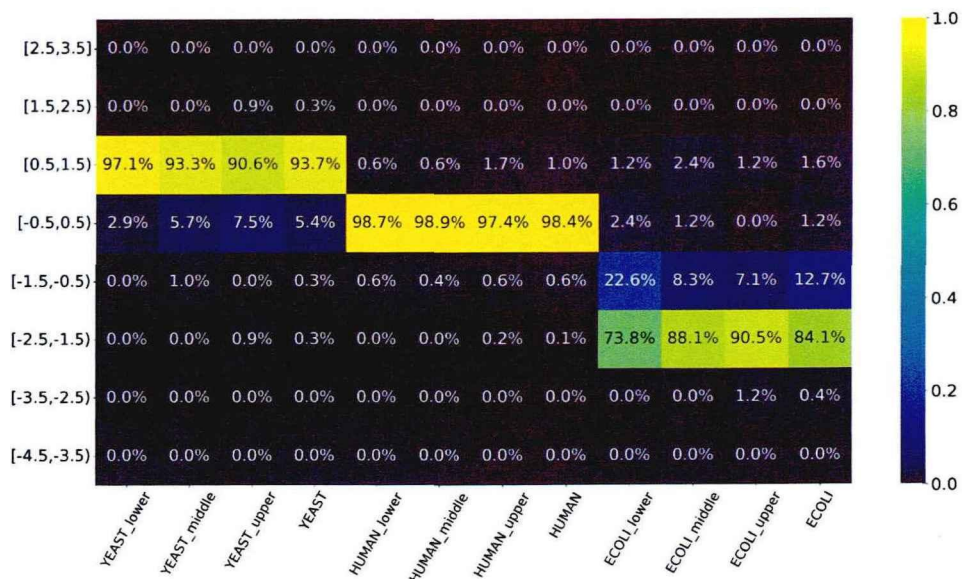


图 2.11 RE-FIGS 肽段定量结果热力图

2.6 本章小结

在传统 DIA 质谱数据的采集过程中，液相色谱洗脱环节的耗时较长，从而阻碍了高通量蛋白质组学的发展。为了克服这一缺点，近年来研究者们提出了 DI-SPA 质谱采集方法。该方法使用分离速度较快的 FAIMS 技术来替代液相色谱，显著提升了蛋白质组学分析的效率。然而，由于 DI-SPA 质谱数据缺少色谱峰信息，因此以往那些依赖于色谱峰信息的 DIA 质谱数据分析工具无法应用在此类数据上。近几年针对 DI-SPA 质谱数据开发的 CsoDIAq 工具虽然可以对其进行处理，但该工具的无标记定量效果还需进一步提升。为了弥补这一不足，在本章中我们提出了 RE-FIGS 方法，该方法可以对 DI-SPA 质谱数据进行更加准确的定性及无标记定量分析。

RE-FIGS 定性算法首先从 SSM 结果中提取了若干种特征信息，之后利用半监督学习的思想获取训练样本，并使用线性判别分析和集成学习方法训练分类模型，最后利用该模型对图谱库中的肽段进行打分，在经过质量控制后得到最终的肽段定性结果。此外，我们还发现，在对 DI-SPA 质谱数据进行循环采集后，可以利用其产生的一些重复性特征显著提高肽段鉴定的数量。RE-FIGS 定量算法则是利用 FIGS 中的 Fisdec 图谱分解算法对 MACC 分数最高的实验图谱进行解谱，并将解谱系数作为肽段的最终定量值。

实验结果表明，RE-FIGS 不仅在定性方面的表现优于 CsoDIAq，而且实现了较为准确的无标记定量分析。此外，由于 FAIMS 的分离速度远远快于液相色谱，所以即使对于循环采集的 DI-SPA 质谱数据而言，其采集总时间依然明显少于传统的 DIA 质谱数据。因此，我们的 RE-FIGS 方法同时兼顾了蛋白质组学分析的性能和速度，从而满足了当前高通量蛋白质组学的发展需求。

第3章 基于深度学习的4D-DIA质谱数据定性分析方法

3.1 研究背景与意义

随着质谱技术的快速发展,液相色谱-串联质谱等质谱联用技术已经成为蛋白质组学分析中一种高效的技术手段。然而,考虑到生物样品中蛋白质种类的错综复杂性,传统的液相色谱-串联质谱技术在处理一些高度复杂的生物样品,例如人类血浆蛋白质样品时,其液相分离效果并不够理想。这种情况导致在随后的DIA质谱分析过程中,所生成的DIA质谱数据极其复杂,且存在着大量不同肽段母离子生成的重叠碎片离子峰,严重影响了后续的DIA质谱数据定性及定量分析过程。此外,在传统的DIA质谱采集过程中,需要先设置一系列相邻的窗口,之后循环地碎裂每个窗口内的所有肽段母离子。在这种情况下,每个肽段母离子在每次循环中只能被采集一次,这也使得传统DIA策略的离子利用率并不高^[78]。针对以上问题,Meier等人提出了一种基于timsTOF Pro质谱平台的DIA方法,并将其命名为4D-DIA (diaPASEF)^[32]。该方法结合了DIA策略和TIMS^[33-34]技术,不仅具备高灵敏度和高通量的优点,还在传统DIA质谱数据三维信息(质荷比、保留时间和峰强度)的基础上增添了一维离子淌度信息,从而为后续的质谱分析提供了更多的特征信息。这一创新方法在不牺牲质谱采集速度的前提下,有效地降低了质谱数据的复杂性,并提升了肽段母离子的利用率。4D-DIA质谱数据的示意图如图3.1所示。

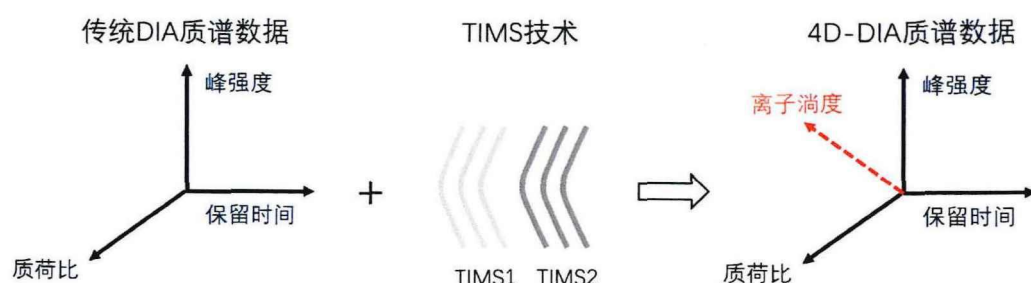


图 3.1 4D-DIA 质谱数据示意图

在4D-DIA质谱数据的采集过程中,实验样品首先会经过蛋白酶水解,随后进入液相色谱仪中进行色谱洗脱,这一步的目的是对实验样品进行初步的分离。之后,从色谱仪中流出的样品将会被导入timsTOF质谱仪中进行进一步的检测。与传统的DIA质谱仪器不同的是,在timsTOF质谱仪中,肽段母离子会先进入一个双串联TIMS装置中进行离子淌度维度上的分离。TIMS装置的核心是一个充满缓冲气体的离子阱,当肽段母离子穿越这片区域时,它们会和该区域中的气体发生碰撞。不同碰撞截面积(Collision cross section, CCS)的母离子会受到不同

程度的阻力,这使得它们在离子阱中的迁移速度也会不同,从而实现了对肽段母离子的分离。同时,通过测量母离子在离子阱中的运行时间,还可以得到母离子的离子淌度值(离子迁移率),这一信息在后续的质谱分析中有着重要的参考价值。

在完成离子淌度维度的分离后,这些肽段母离子会被引入到四级杆质量分析器中进行后续的累积和碎裂操作。具体而言,4D-DIA 方法在四级杆中采用了一种名为并行累积串行碎裂(Parallel accumulation–serial fragmentation, PASEF)^[79]的策略。在这种策略下,仪器会在多个窗口内同时累积母离子,并依次对每个窗口中的母离子进行碎裂和检测,每批累积的母离子都能经过多个窗口的连续扫描,从而有效提高了母离子的利用效率。在每个隔离窗口中,肽段母离子会被碎裂成更细小的片段,检测器能够精准地捕获这些片段,并生成相应的质谱信号。最后,质谱仪对这些信号进行整合处理,从而生成最终的质谱数据。图3.2展示了4D-DIA 质谱数据的大致采集流程。

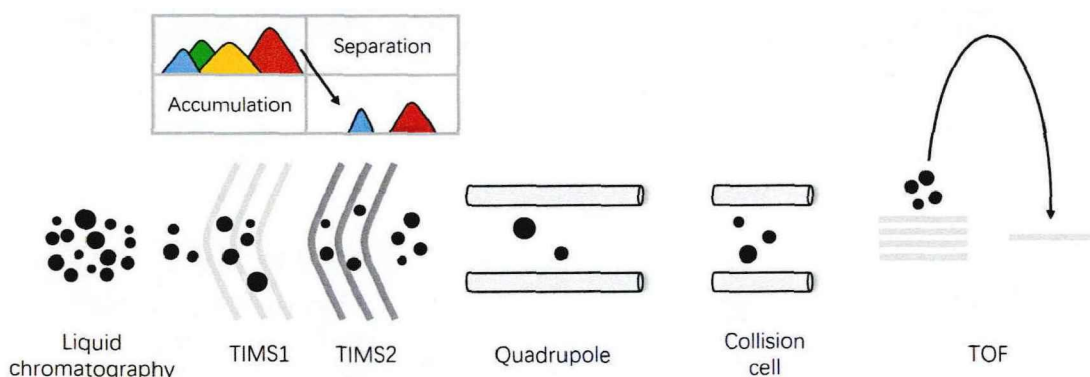


图 3.2 4D-DIA 质谱数据采集流程

目前,针对 4D-DIA 质谱数据的定性分析主要依赖于传统的 DIA 质谱数据分析方法。这类方法往往需要将实验质谱与肽段图谱库中的肽段图谱进行比对,并通过计算图谱匹配分数来评估肽段存在于实验样品中的可能性。在计算图谱匹配分数的过程中,需要考虑诸如色谱峰形状、保留时间、碎片离子峰强度和离子淌度等多种因素。过去常用的大多数 DIA 质谱数据定性分析方法,例如 OpenSWATH^[24]、Spectronaut^[54]和 Skyline^[55]等,都需要从色谱信息中提取相关特征来作为打分的重要依据。这些高度依赖于色谱信息的 DIA 质谱数据定性分析方法在对不同实验室制备的 4D-DIA 质谱数据进行定性分析时,很容易受到不同实验平台间色谱梯度差异的影响,从而为最终的分析结果带来一定的误差。因此,未来仍需提出不依赖于色谱信息的 4D-DIA 质谱数据定性分析方法,以提升 4D-DIA 方法的通用性。

近年来,随着人工智能算法的不断创新和 GPU 计算能力的提升,深度学习在图像识别、个性化推荐、自然语言处理等领域取得了突破性进展^[80-81]。与传

统的机器学习方法相比,深度学习可以自动从海量的训练数据中提取深层特征,因此降低了对特征工程的依赖程度。同时,其错综复杂的网络结构使得模型能够处理更加复杂的非线性关系,进而展现出更强大的学习能力^[82]。如今,深度学习已广泛应用于蛋白质组学领域中,为解决各种实际问题提供了强大的技术支持^[83]。对于 DIA 质谱数据的定性分析问题而言,近几年也提出了一些基于深度学习的分析方法。其中的代表性工作 DIA-NN^[56]首先利用线性分类器为每个肽段迭代地选取最优色谱峰,之后提取了色谱峰中的 73 个相关特征,并将其输入到一个全连接神经网络中进行计算,网络的输出值反映了该肽段存在于实验样品中的概率。最后,该方法会对所有目标肽段与诱饵肽段的评分结果进行质量控制,从而得到最终的肽段定性结果。和一些传统的 DIA 质谱数据分析方法相比,DIA-NN 虽然在肽段鉴定数量方面有了一定的提升,但是仍然没有摆脱色谱信息的限制。因此,鉴于过往 DIA 质谱数据定性分析方法中存在的局限性,在本章中,我们针对 4D-DIA 质谱数据提出了一种基于深度学习的定性分析方法。该方法在对实验质谱数据和图谱库数据进行预处理后,将图谱匹配结果中的离子峰信息和离子淌度信息输入到卷积神经网络分类模型中进行计算,得到对应肽段的打分值,并基于该结果确定最终的肽段定性结果。该方法在整个计算过程中完全没有使用色谱信息,从而有效避免了不同实验平台间色谱梯度差异对分析结果的潜在影响。

3.2 数据预处理

3.2.1 质谱处理

为了获取 4D-DIA 质谱数据,需要使用 Bruker 公司的 timsTOF 系列质谱仪器来对实验样品进行采集。4D-DIA 质谱数据的格式一般为.d 格式,.d 文件是一个目录文件,在该目录下还包含了若干个子文件,这些文件存储了质谱数据的相关信息,如窗口设置、离子峰强度和离子淌度值等信息。与第二章的第二小节“2.2 数据预处理”中所介绍的.raw 格式和.wiff 格式的质谱数据不同,.d 格式的质谱数据不仅包括离子峰强度等信息,还额外包含了离子淌度信息。如果依然使用 ProteoWizard^[67]软件中的 MSConvert 工具来对.d 格式的质谱数据进行转换,那么在同一窗口下,不同离子淌度所对应的二级质谱图将会被合并,这会导致 4D-DIA 质谱数据中的离子淌度信息丢失。在这里,为了读取.d 格式的质谱数据,我们需要在 Python 环境下安装 timspy 标准库,通过调用该库中的函数,可以获取 4D-DIA 质谱数据的相关信息。之后,我们还需要对 4D-DIA 质谱数据中距离过近的离子峰进行峰合并操作,从而减少质谱仪质量误差对后续质谱分析的影响,具体过程与之前的算法 2.1 完全一致。最后,我们将合并后的质谱数据

按照窗口进行划分，并将最终的处理结果保存至文件中。

3.2.2 图谱库处理

Spectronaut^[54]作为业内著名的商业化蛋白质组学分析工具，可以根据输入的实验质谱数据生成对应的肽段图谱库。鉴于在本章的实验中，我们所分析的 4D-DIA 质谱数据全部来源于人类血浆蛋白质样品。因此，为了构建实验所需的图谱库，我们借助 Spectronaut 工具，对上海生命科学院提供的一批来源于人类血浆蛋白质样品的 4D-DIA 质谱数据集进行了搜库处理，得到一个包含 20218 个肽段图谱的图谱库。在该图谱库中，不同肽段图谱的峰数量差异显著，有的仅包含几个峰，而有的则多达数十个。为了方便后续将肽段图谱的离子峰输入到神经网络中进行统一计算，我们需要先对图谱库进行预处理。具体来说，我们筛选出每个肽段图谱中强度最高的 m 个离子峰 ($m = 6$)，并对这些筛选后的离子峰进行强度的归一化处理（详见公式2.1），从而得到一个标准化的目标肽段图谱库。

在得到目标库后，为了对后续的肽段定性结果进行准确的质量控制，我们需要利用目标库来生成相应的反库。在这里，与第二章相同，我们依然采用了目标-诱饵库策略^[70]来生成反库，具体过程详见算法2.2。需要强调的是，在本章的所有实验中，我们都将基于以上的目标库和反库来对 4D-DIA 质谱数据进行定性分析。

3.2.3 质谱的匹配与过滤算法

在处理完质谱数据和图谱库数据后，我们需要先设计质谱的匹配与过滤算法。该算法的目标是从质谱数据中筛选出与肽段图谱最相关的若干张实验图谱，以便于后续生成训练数据和测试数据。质谱匹配的大致过程如算法3.1所示。对于图谱库 P 中的第 i 张肽段图谱 P_i ，我们首先根据 P_i 的肽前体质荷比筛选出 P_i 所在的采集窗口下的实验图谱集合 E_i 。之后，我们将 P_i 与 E_i 中的每张实验图谱进行离子峰的匹配。由于实验质谱的复杂度较高，其中可能存在着大量的噪声峰随机匹配到了图谱库中的某些离子峰，因此，需要预先设置匹配条件来提前过滤出那些不相关的实验图谱。在 P_i 与 E_i 中第 j 张实验图谱 E_{ij} 的峰匹配过程中，我们会统计两者的峰匹配个数，只有当 P_i 与 E_{ij} 的峰匹配个数大于等于 n 时，我们才会记录并保存 P_i 和 E_{ij} 的有关信息，例如两者的离子淌度值和匹配离子峰信息等。否则，我们会选择将 E_{ij} 直接丢弃。考虑到 P_i 中最多只有 6 个离子峰，因此，在这里我们将峰匹配阈值 n 设置为 3。此外，为了方便进行后续的计算，当 P_i 与 E_{ij} 的峰匹配个数小于 6 时，我们会使用 0 来填充两者所缺少的离子峰。

在经过上述的匹配过程后，我们得到了与肽段图谱 P_i 相匹配的实验图谱集合 M_i 。此时， M_i 中可能仍然包含数十甚至上百张实验图谱，并且其中也会存在

算法 3.1 质谱匹配算法

Data: E : 实验质谱
 P_i : 图谱库 P 中的第 i 张肽段图谱
 n : 峰匹配阈值
 δ : 质谱仪质量误差

Result: M_i : 与 P_i 相匹配的实验图谱集合

```

1  $PZ_i = \text{getPrecursorMZ}(P_i)$ ; // 获取  $P_i$  的肽前体质荷比  $PZ_i$ 
2  $IZ_i = \text{getIonPeakMZ}(P_i)$ ; // 获取  $P_i$  的离子峰质荷比列表  $IZ_i$ 
3 for  $j = 1$  to  $\text{length}(E)$  do
4    $W_j = \text{getWindow}(E[j])$ ; // 获取  $E[j]$  的采集窗口  $W_j$ 
5   if  $PZ_i \geq W_j[0]$  and  $PZ_i \leq W_j[1]$  then
6      $E_i.\text{add}(E[j])$ ; // 筛选出  $W_j$  下的实验图谱
7   end
8 end
9 for  $E_{ij}$  in  $E_i$  do
10   $IZ_{ij} = \text{getIonPeakMZ}(E_{ij})$ ;
11   $IZ_{ij}^1 = IZ_{ij} - \delta \cdot IZ_{ij}$ ; //  $IZ_{ij}^1$  是  $IZ_{ij}$  左移  $\delta$  个单位后的结果
12   $IZ_{ij}^2 = IZ_{ij} + \delta \cdot IZ_{ij}$ ; //  $IZ_{ij}^2$  是  $IZ_{ij}$  右移  $\delta$  个单位后的结果
13   $IZ'_{ij} = \text{concatenate}(IZ_{ij}^1, IZ_{ij}^2)$ ; // 将  $IZ_{ij}^1$  和  $IZ_{ij}^2$  拼接在一起
14   $IZ'_{ij} = \text{sort}(IZ'_{ij})$ ; // 对  $IZ'_{ij}$  进行排序
15   $S_{ij} = \text{searchsorted}(IZ'_{ij}, IZ_i)$ ; // 获取  $IZ_i$  所有元素在  $IZ'_{ij}$  中的排序位置
16   $O_{ij} = \text{odd}(S_{ij})$ ; // 筛选出  $S_{ij}$  中的奇数
17  if  $\text{length}(O_{ij}) \geq n$  then
18     $M_i.\text{add}(E_{ij})$ ;
19  end
20 end
21 return  $M_i$ ;

```

着很多与 P_i 无关的实验图谱。如果我们将这些实验图谱全都输入到后续的神经网络中进行计算,不仅会使计算结果受到不相关实验图谱的干扰,还会让神经网络的计算量变得非常巨大。因此,我们需要对集合 M_i 中的实验图谱进行进一步的过滤,只保留其中与 P_i 相关性最高的若干张实验图谱,从而确保计算的准确性和高效性。

在进行质谱过滤操作之前,我们需要先设计肽段图谱和实验图谱之间的相关性计算公式。考虑到对于 M_i 中的第 j 张实验图谱 M_{ij} 而言,如果 P_i 与 M_{ij} 所匹配峰的峰强度相似性越高,则意味着 M_{ij} 由 P_i 对应的肽段母离子碎裂而成的可能性越大,从而说明 P_i 与 M_{ij} 的相关性越高。因此,我们选择用 P_i 与 M_{ij} 所匹配峰的峰强度相似性来衡量两者的相关性。即:

$$\text{Correlation}(P_i, M_{ij}) = \text{Similarity}(P_i, M_{ij}) \quad (3.1)$$

而为了计算 $\text{Similarity}(P_i, M_{ij})$ 的值,我们采用了曼哈顿距离作为评估指标。在这里,为了便于描述,我们使用向量形式来表示 P_i 与 M_{ij} 所匹配峰的峰强度,那么 P_i 的峰强度可以表示为 $(p_i^1, p_i^2, \dots, p_i^6)$, M_{ij} 的对应峰强度可以表示为 $(m_{ij}^1, m_{ij}^2, \dots, m_{ij}^6)$ 。首先,与之前的图谱库处理方式类似,我们仿照公式2.1对 M_{ij} 进

行峰强度的归一化处理得到 $\overline{M}_{ij}$, $\overline{M}_{ij}$ 的对应峰强度可以表示为 $(\overline{m}_{ij}^1, \overline{m}_{ij}^2, \dots, \overline{m}_{ij}^6)$ 。此时, P_i 与 $\overline{M}_{ij}$ 的离子峰强度之和均为 1。考虑到曼哈顿距离与实际相似度之间呈负相关关系, 因此我们用曼哈顿距离的相反数来表示 P_i 与 M_{ij} 所匹配峰的峰强度相似性, 即:

$$\text{Similarity}(P_i, M_{ij}) = -\text{Manhattan}(P_i, \overline{M}_{ij}) = -\sum_{k=1}^6 |p_i^k - \overline{m}_{ij}^k| \quad (3.2)$$

之后, 为了获取与肽段图谱 P_i 最相关的若干张实验图谱, 我们根据公式3.1和公式3.2计算了 P_i 与 M_i 中每张实验图谱的相关性, 最终只保留了相关性最高的 s 张实验图谱。值得说明的是, 当 P_i 的图谱匹配数量小于 s 时, 我们会选择使用零图谱 (图谱中的所有信息都设置为 0) 来填充缺少的图谱。最后, 在综合考虑了实验表现和计算效率后, 我们将 s 设置为 6。质谱过滤算法的大致过程如算法3.2所述。

算法 3.2 质谱过滤算法

Data: P_i : 图谱库 P 中的第 i 张肽段图谱
 M_i : 与 P_i 相匹配的实验图谱集合
 s : 保留实验图谱个数

Result: R_i : 与 P_i 最相关的实验图谱集合

```

1 if length( $M_i$ ) >  $s$  then
2   for  $M_{ij}$  in  $M_i$  do
3      $C_i$ .add(Correlation( $P_i, M_{ij}$ )); // 计算并保存  $P_i$  与  $M_{ij}$  的 Correlation
4   end
5    $A_i$  = argsort( $-C_i$ ); // 获取  $C_i$  降序排列后的索引值
6    $I_i$  =  $A_i[1:s]$ ; // 获取  $A_i$  前  $s$  个元素
7    $I_i$  = sort( $I_i$ ); // 对  $I_i$  进行排序
8    $R_i$  =  $M_i[I_i]$ ; // 筛选出索引位于  $I_i$  中的实验图谱
9 else
10   $R_i$  =  $M_i \cup (s - \text{length}(C_i)) \cdot \text{zeroSpectrum}$ ; // 使用零图谱填充缺少的图谱
11 end
12 return  $R_i$ ;

```

3.2.4 训练数据与测试数据的生成

在对定性神经网络模型进行训练之前, 我们需要先获取合适的训练数据集。研究表明, 多样且丰富的训练数据对于神经网络的学习来说至关重要, 它们能够帮助神经网络更好地捕捉到输入数据中的各种模式和特征, 同时可以有效地防止模型出现过拟合现象, 并提升模型在测试集上的准确性和可靠性^[84]。为此, 我们使用了上海生命科学院提供的若干批 4D-DIA 质谱数据作为训练集。这些质谱数据均来源于人类血浆样品, 共计 89 个数据文件, 其中包括 5 组数据, 每组数据都是多个重复采集的质谱数据。训练集的具体组成如表3.1所示。

之后, 我们需要基于以上数据集来生成相应的训练数据。在依赖于肽段图谱库的质谱数据定性分析问题中, 问题求解的目标是确定肽段图谱库中的哪些肽

表 3.1 定性模型训练集组成

组号	数据集名称	质谱数据个数
1	LCH_MN_144-Plasma	8
2	LCH_Plasma	12
3	Evosep_CCS_MN_Plasma	44
4	CCS_Plasma_trypsin	5
5	H15150_Plasma-GXJ	20

段存在于质谱数据中,这本质上是一个二分类问题。因此,为解决该问题,我们需要训练一个分类模型,并且在训练数据中应当同时包含正样本和负样本,其中正样本表示对应的肽段存在于实验质谱中,其标签值被设定为1;而负样本则表示对应的肽段不存在于实验质谱中,其标签值被设定为0。这样的标签设置使得神经网络的输出会保持在0到1之间,且网络的输出值越接近于1,则说明对应肽段存在于实验质谱中的可能性越大。

为了获取训练数据中的正样本,我们需要利用现有的4D-DIA质谱数据分析工具来对训练数据集进行定性分析,从而确定这些数据集中存在了哪些肽段。考虑到Spectronaut工具在4D-DIA质谱数据定性分析方面有着较高的准确性,因此,我们使用了Spectronaut来对表3.1中的89个4D-DIA质谱数据进行定性分析,从而得到了每个质谱数据的肽段定性结果,其中定性所使用的图谱库正是前文中所介绍的目标图谱库。之后,我们利用质谱匹配算法3.1和质谱过滤算法3.2,计算出Spectronaut定性结果中每条肽段所对应的实验图谱集合(与肽段图谱最相关的 s 张实验图谱),并将这些肽段图谱和实验图谱集合的有关信息,包括离子淌度值和匹配离子峰信息等,保存至文件中,作为训练数据中的正样本数据备用。至于负样本数据的获取,由于诱饵库中的肽段均为人为构造出的肽段,它们必定不存在于实验样品中。因此,我们可以直接将诱饵库中的全部肽段作为训练数据中的负样本。在此过程中,我们同样需要将这些肽段及其对应的实验图谱集合的有关信息保存至文件中,以备后续的使用。

同样,在对测试集进行测试之前,我们也需要生成相应的测试数据。在这里,我们只介绍测试数据的生成方式,至于实验中具体使用到的测试集,我们会在本章后续的“3.4 实验结果与分析”小节中进行介绍。由于后续的质量控制环节要求我们在对目标库中的肽段进行打分的同时,还需要对诱饵库中的肽段进行打分,最后再结合两者的打分结果,筛选出那些错误发现率(False discovery rate, FDR)小于0.01的目标库肽段作为最终的定性结果。因此,我们需要同时为目标库和诱饵库生成测试数据。具体来说,我们将目标库与诱饵库中的全部肽段以及它们相应的实验图谱集合的有关信息分别保存至文件中,从而生成了目标库和诱饵库的测试数据。

3.3 定性模型构建

在完成数据的预处理后,我们首先需要根据具体的数据和任务特点来选择最合适的神经网络模型。目前,在深度学习中包含了许多种模型结构,例如循环神经网络 (Recurrent Neural Network, RNN)^[85]、Transformer^[86]、卷积神经网络 (Convolutional Neural Network, CNN)^[87]和图神经网络 (Graph Neural Network, GNN)^[88]等。以下是对这些模型结构的具体介绍:

在处理序列数据时,RNN 是一种常用的深度学习模型。与传统的前馈神经网络不同,RNN 的隐藏层不仅接收当前时间点的输入数据,还接收前一个时间点的隐藏层输出,这种循环连接机制使得 RNN 在处理序列数据时能够记忆并利用之前的信息。通过迭代和更新其内部状态,RNN 能够捕捉到序列数据中的时间依赖关系,从而有效地处理诸如语音、文本和时间序列等具有时序特性的数据。而 Transformer 作为一种基于自注意力 (Self-Attention) 机制的深度学习模型,同样适用于处理序列数据。其核心思想是通过自注意力机制,让模型能够关注输入序列中的不同片段,从而捕获序列中元素之间的依赖关系。与 RNN 相比,Transformer 展现出了更强大的并行计算能力,同时能够更全面、更精准地捕获序列中的长距离依赖关系。这些显著优势使得 Transformer 在自然语言处理和语音识别等领域中备受青睐,并引发了广泛的研究和应用。

在图像处理方面,CNN 则具有十分明显的优势。通过卷积操作,CNN 能够在图像的局部区域内有效地捕捉相关特征,并利用层次化的网络结构对这些特征进行逐步地抽象和组合,从而深入挖掘图像中的深层信息。此外,CNN 的权值共享和局部连接等特性,使其在处理图像数据时表现出模型参数少、计算效率高以及鲁棒性强等突出优点。

而在处理图结构数据时,往往采用 GNN 这一深度学习模型。GNN 能够捕获图中节点之间的依赖关系和结构信息,并通过迭代的方式更新节点的特征表示,从而学习到图的整体结构特性。与传统的 CNN 和 RNN 不同,GNN 不受限于规则的网格结构数据或序列数据,其在处理社交网络和推荐系统等领域的图数据时,具有较高的效率和准确性。

在本实验中,我们的训练数据和测试数据由多组肽段图谱和实验图谱构成,每组数据中都包含了一张肽段图谱和 s 张实验图谱的离子峰强度及离子淌度等信息。尽管同一组数据中的不同图谱之间往往存在相似性,但它们并不具有明显的时序关系或上下文关系,同时,同一张图谱中的不同离子峰之间也不存在较强的关联性。因此,那些适用于处理序列数据的深度学习模型,如 RNN 和 Transformer 等,在此场景下并不是一个较为理想的选择。考虑到这些图谱数据由多个一维向量组成,这种网格结构的数据形式与图像数据较为类似,因此我们认

为在本实验中更适合采用 CNN 作为后续计算的神经网络模型。通过对每组数据中的肽段图谱和多张实验图谱进行卷积操作, CNN 可以有效地提取出它们之间的相似性关系。此外, 由于这些图谱数据的相关信息都是长度统一的向量, 并不包含复杂的图结构关系, 因此在这里我们也无需采用 GNN。综合以上分析后, 为了在数据解析与特征提取等方面达到最佳效果, 我们最终选择了 CNN 作为 4D-DIA 质谱数据定性分析的神经网络模型。

之后, 我们需要详细设计 CNN 定性模型的网络结构。在这里我们考虑到, 在训练数据和测试数据中, 肽段图谱及其对应的实验图谱的离子峰强度和离子淌度值是两个至关重要的信息。从理论上来说, 如果肽段图谱与实验图谱所匹配峰的峰强度相似性越高, 则说明该实验图谱由该肽段图谱对应的肽段母离子碎裂产生的可能性越大, 进而表明对应肽段存在于实验样品中的概率越大。而肽段图谱和实验图谱之间的离子淌度差值则反映了两者在离子淌度维度上的匹配情况。一般情况下, 如果肽段图谱和实验图谱的离子淌度差值越小, 则意味着两者之间存在关联的可能性越大, 从而说明对应肽段存在于实验样品中的概率越大。因此, 我们选择将训练数据中每张肽段图谱及其相应的实验图谱集合的离子峰强度和离子淌度值作为 CNN 定性模型的原始输入。鉴于每张图谱的离子峰强度都是长度为 6 的一维向量 (离子峰数量为 6), 而离子淌度值则是一个单一的数值。因此, 为了方便 CNN 定性模型后续进行统一计算, 我们将肽段图谱和实验图谱的离子淌度值均重复了 5 次, 从而也将其扩展成长度为 6 的一维向量。

随后, 我们将肽段图谱和实验图谱集合的离子峰强度向量以及离子淌度向量各自输入到多个卷积层中, 这些卷积层能够逐步提取出输入数据中不同层次的特征, 从而提高模型对输入数据的理解能力。此外, 多卷积层的设计还能通过非线性激活函数引入更多的非线性因素, 从而使得模型能够学习到更复杂的映射关系。接下来, 我们将这些各自提取出来的深层特征拼接在一起, 再次进行卷积处理, 这一步的目的是提取它们之间的深层交互信息。值得注意的是, 在每次进行卷积计算的过程中, 我们都会对卷积结果应用 ReLU 激活函数, 该函数的计算方式非常简单, 它会将正值直接映射, 将负值置为零。ReLU 激活函数不仅为卷积层引入了非线性关系, 还使得一部分神经元的输出为 0, 这样做增强了网络的稀疏性, 有助于降低过拟合风险并提高模型的泛化能力。之后, 我们将卷积层的输出结果展平为一维向量, 然后将其输入到三个连续的全连接层中。这些全连接层会逐步汇总和整合之前提取到的所有特征, 并输出最终的结果。通过堆叠多个全连接层, 模型能够逐步将低级特征转化为更高级别的特征, 从而进一步提升其泛化性能。在此过程中, 对于前两个全连接层, 我们仍然采用了 ReLU 激活函数来对输出结果进行处理, 从而增强特征的非线性表达能力。而对于最后一个全连接层, 我们则采用了 Sigmoid 激活函数将输出结果映射到 0 到 1 的范围内, 以

便于对结果进行二分类判定, 输出结果越接近于1则表明对应肽段存在于实验样品中的可能性越大。Sigmoid 激活函数的具体表达式如下:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (3.3)$$

此外, 我们在每个卷积模块以及除最后一层外的每个全连接层之后都进行了 dropout 操作。该操作可以随机地将部分神经元失活, 从而有效防止模型对训练数据中的噪声或某些特定特征产生过度依赖, 进而降低过拟合的风险。CNN 定性模型的网络结构如图3.3所示。

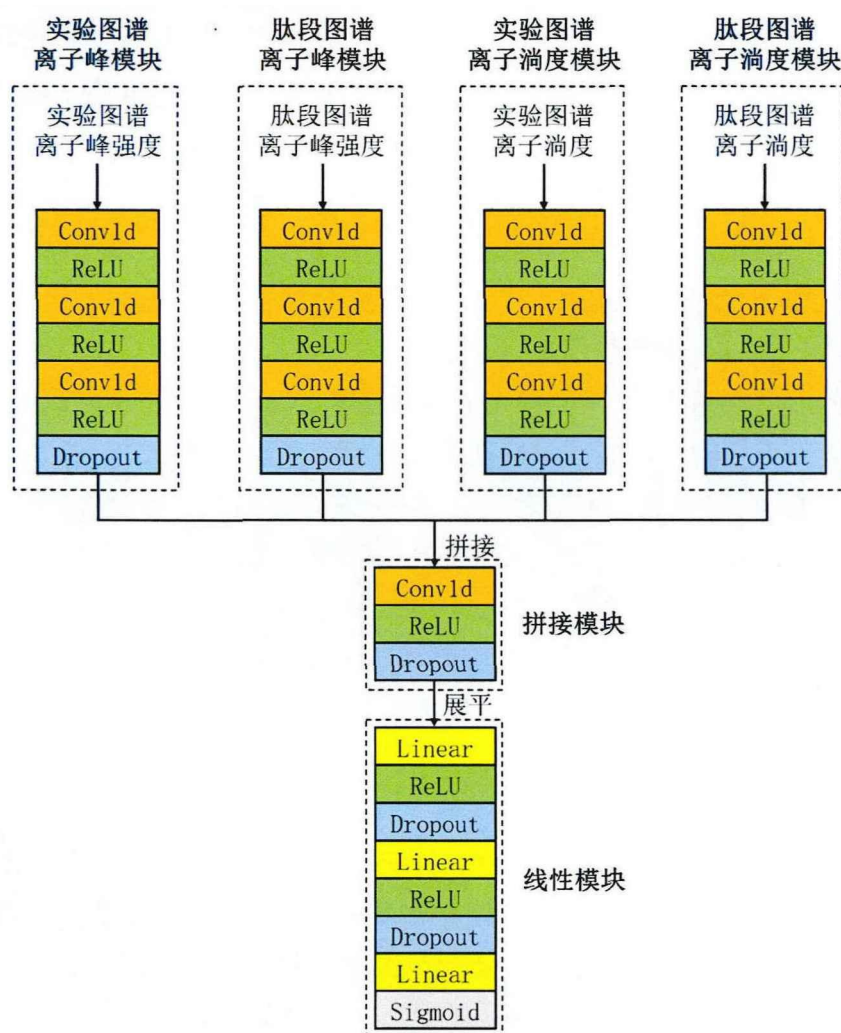


图 3.3 CNN 定性模型网络结构示意图

在完成 CNN 定性模型的网络结构设计后, 我们还需要对神经网络中的一些超参数进行设置, 这里的超参数可以大致分为网络结构参数和训练参数两种类型。以下是对它们的具体介绍:

CNN 定性模型的网络结构参数主要包括各个网络层的通道数、卷积核的大小和填充等。表3.2展示了神经网络结构中的部分参数信息。其中, b 表示批量大小, 即神经网络在每次训练中处理的样本数量, s 表示实验图谱集合中的图谱数

量, m 表示每张肽段图谱的离子峰数量。考虑到 CNN 定性模型的输入数据由多个一维向量组成, 因此在模型的每个卷积层中, 我们都采用了一维卷积核来进行卷积计算。在表3.2的前四个模块中, 我们将三个卷积层的卷积核个数分别设置为 64、128 和 256, 通过不断增加卷积核的个数可以使得模型在每一个卷积层中都能学习到更多的特征信息, 同时较多的卷积核个数也增强了模型的表示能力。之后, 考虑到在拼接模块中, 输入数据的通道数较多, 因此我们将这里的卷积核个数设置为 256, 从而减少了后续模型的计算量。最后, 我们将线性模块中的三个线性层的输出神经元个数分别设置为 256、64 和 1, 从而让模型能够对输入的特征进行逐步的压缩和降维, 并最终输出二分类判定的概率。此外, 考虑到同一张图谱中的不同离子峰之间并没有较强的关联性, 因此对于每个卷积层, 我们都将一维卷积核的大小设置为 1, 填充设置为 0, 步长设置为 1。这样做不仅有助于降低模型的复杂度, 而且避免引入了不必要的空间依赖关系。

表 3.2 CNN 定性模型网络结构参数

模块	网络层类型	输入大小	输出大小
实验图谱离子峰模块	Conv1d	(b, s, m)	$(b, 64, m)$
	Conv1d	$(b, 64, m)$	$(b, 128, m)$
	Conv1d	$(b, 128, m)$	$(b, 256, m)$
肽段图谱离子峰模块	Conv1d	$(b, 1, m)$	$(b, 64, m)$
	Conv1d	$(b, 64, m)$	$(b, 128, m)$
	Conv1d	$(b, 128, m)$	$(b, 256, m)$
实验图谱离子淌度模块	Conv1d	(b, s, m)	$(b, 64, m)$
	Conv1d	$(b, 64, m)$	$(b, 128, m)$
	Conv1d	$(b, 128, m)$	$(b, 256, m)$
肽段图谱离子淌度模块	Conv1d	$(b, 1, m)$	$(b, 64, m)$
	Conv1d	$(b, 64, m)$	$(b, 128, m)$
	Conv1d	$(b, 128, m)$	$(b, 256, m)$
拼接模块	Conv1d	$(b, 256 \times 4, m)$	$(b, 256, m)$
线性模块	Linear	$(b, 256 \times m,)$	$(b, 256,)$
	Linear	$(b, 256,)$	$(b, 64,)$
	Linear	$(b, 64,)$	$(b, 1,)$

CNN 定性模型的训练参数可以分成数值型训练参数和非数值型训练参数两类。其中数值型训练参数主要包括学习率、迭代次数和批量大小等, 这些参数是否合理对于模型的训练至关重要。以学习率为例, 较大的学习率可以使模型在训练初期快速收敛到较优的解, 但可能会导致模型在训练过程中出现震荡现象, 甚至难以收敛到最优解。而较小的学习率虽然可以提升模型在训练过程中的稳定性, 但可能会导致训练过程变得非常缓慢, 甚至出现欠拟合现象。因此, 在设置学习率时, 需要综合考虑计算资源、训练时间和模型性能等多个因素。为了找到较为理想的训练参数组合, 我们首先会根据模型在验证集上的表现, 为所有训练参数设置一个相对合理的初始值。之后, 我们采用手动调参的方法, 每次只调

整个训练参数,并观察其对模型性能的影响,进而确定每个训练参数的较优取值范围。接下来,我们采用了网格搜索方法来寻找最优参数组合。具体来说,我们需要根据每个训练参数的较优取值范围,为其设置若干个取值点,之后遍历所有可能的训练参数组合,并评估每个组合在验证集上的性能,从中选出最优的参数组合。最后,我们还可以对得到的训练参数组合进行手动地微调,以进一步提升模型性能。此外,为了提高网络的收敛速度,我们还采用了学习率调度器来动态调整学习率,即在训练初期使用较大的学习率以加快网络的训练,之后随着训练的进行逐渐减小学习率,从而使得模型能够在较短的时间内收敛到较优的解。而 CNN 定性模型的非数值型训练参数则主要包括优化器类型和损失函数类型等。在优化器的选择上,考虑到 Adam^[89] 优化器具有计算效率高、收敛速度快和适应性强等优点,因此我们采用了 Adam 优化器来自动优化模型在训练过程中的内部参数。而在损失函数的选择方面,鉴于该模型为二分类预测模型,因此我们使用了常见的二元交叉熵损失函数 (Binary Cross Entropy Loss, BCELoss) 来作为网络模型的损失函数,其具体公式为:

$$\text{BCELoss} = -[y \log_2 p + (1 - y) \log_2(1 - p)] \quad (3.4)$$

其中 y 表示训练数据的标签,对于正样本而言,其值为 1,而对于负样本而言,其值为 0。 p 表示模型的预测输出值,其范围为 (0, 1)。

最后,为了得到最优模型,我们从训练数据中划分了 20% 的数据作为验证集。每当神经网络完成一次迭代训练后,我们都会在验证集上计算其分类准确率,并最终选择在验证集上分类准确率最高的模型作为最终模型。在这里,分类准确率的计算公式为:

$$\text{Accuracy} = \frac{\sum_{i=1}^m n_i}{m} \quad (3.5)$$

$$n_i = \begin{cases} 1, & \text{如果 } |y_i - p_i| < 0.5, \\ 0, & \text{如果 } |y_i - p_i| \geq 0.5. \end{cases} \quad (3.6)$$

其中 m 表示验证集中的样本个数, y_i 表示第 i 个样本的标签, p_i 表示第 i 个样本的网络预测值。

本章实验中所使用的部分实验设备及环境配置如表 3.3 所示。

在构建完定性模型后,我们将数据预处理阶段生成的测试数据输入到模型中进行计算,分别得到目标库和诱饵库的肽段打分结果。之后利用前文中的公式 2.2 对目标库中的肽段进行质量控制,只保留那些 FDR 小于 0.01 的肽段作为对应实验样品的肽段定性结果。

表 3.3 实验设备及环境配置

设备/环境	配置
操作系统	Ubuntu 22.04.2
GPU 型号	NVIDIA GeForce RTX 3090
CUDA 版本号	12.0
驱动版本号	525.125.06
Python 版本号	3.7.13
PyTorch 版本号	1.7.1

3.4 实验结果与分析

3.4.1 同源血浆数据集定性结果

首先,我们对上海生命科学院提供的一组人类血浆样品数据进行了定性分析。这组数据与训练数据集同源,都是对同一实验室的相似样品进行采集得到的,共包含了8个重复采集的4D-DIA质谱数据,其数据集名称为LCH_Plasma。为了验证CNN定性模型的性能,我们还使用了Spectronaut工具与其进行定性结果的对比。

在比较CNN和Spectronaut的肽段定性结果之前,我们首先分析了CNN在LCH_Plasma数据集上的肽段打分情况。图3.4展示了CNN在其中一个样本上的目标库及诱饵库肽段的打分分布情况。当FDR等于0.01时,对应的肽段分数为0.060,我们使用该分数作为FDR质量控制的分数阈值,如图3.4中的红色虚线所示。在该虚线右侧,所有目标库肽段的FDR均小于0.01,最终这些肽段会被我们过滤出来作为该样本的肽段定性结果。从图3.4中可以看出,几乎所有的目标库肽段的打分值要么在1附近,要么在0附近,这一现象反映出CNN模型可以较为明确地判断出目标库中的肽段是否存在于实验样品中,而极少出现模棱两可的情况(对应于图中那些打分值居中的目标库肽段)。而诱饵库中肽段的打分值几乎全部在0附近,则说明CNN模型可以较为精准地确定这些与实验样品无关的错误肽段。

除了利用FDR指标来对肽段定性结果进行质量控制外,我们还采用了假设检验^[76]方法来进一步验证其可信度。对于肽段定性结果中某个打分值为 s 的肽段 Pep 而言,我们将原假设设定为 Pep 是由于随机误差而被定性到的肽段,与之相对的备择假设设定为 Pep 是实验样品中真实存在的肽段。之后,我们需要对 P 值进行计算。在假设检验方法中, P 值表示在原假设为真的前提下,观察到当前统计结果或更极端结果的概率。在这里,观测样本 Pep 的打分值为 s ,更极端的情况则是 Pep 的打分值大于 s ,因此这里的 P 值表示当原假设为真时, Pep 的打分值大于等于 s 的概率。显然, P 值越小,则表明我们拒绝原假设的理由越充分,即 Pep 为正确肽段这一事件的可信度越高。为了计算 P 值,我们参考了Granhholm等人发表的论文^[68],其具体计算方法如前文中的公式2.5所示。

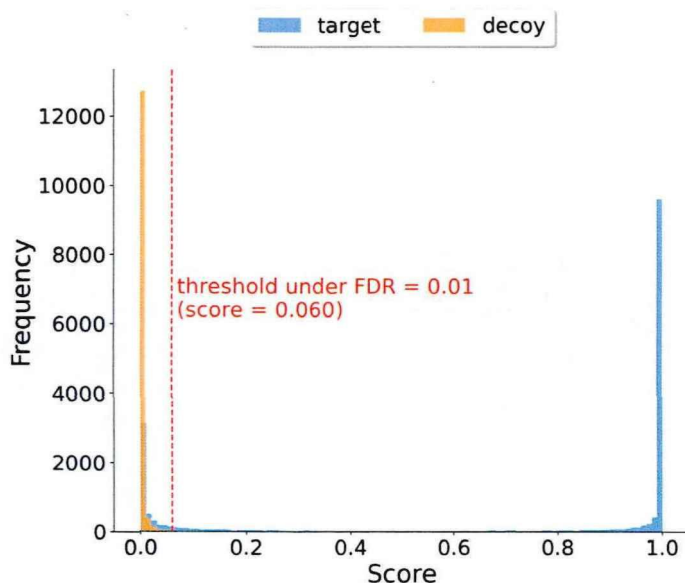


图3.4 CNN定性模型肽段打分分布图

我们在显著性水平设为 0.01 的情况下, 利用公式2.5对图3.4中的样本进行了相关的计算。计算结果表明, 只有当肽段得分 s 的值大于等于 0.043 时, 对应的 P 值才会小于等于 0.01。因此, 我们可以得到拒绝域为 $\{s \geq 0.043\}$ 。鉴于在前文中我们为样本设定的肽段分数阈值为 0.060 (大于拒绝域的下界 0.043), 且该样本所有被定性到的肽段的打分值均大于肽段分数阈值。因此, 对于该样本肽段定性结果中的每一个肽段, 我们都有足够的证据拒绝原假设, 从而说明这些肽段都很有可能真实存在于实验样品中, 这在一定程度上也反映出我们的肽段定性结果具备较高的可靠性。

之后, 我们对 CNN 和 Spectronaut 的肽段定性结果进行了比较, 如图3.5所示。从图3.5a中我们可以看出, 对于每个样本而言, CNN 的肽段定性数量都明显高于 Spectronaut。在这 8 次重复实验的肽段定性结果取平均后, CNN 平均鉴定到 12204 条肽段, 是 Spectronaut 平均肽段鉴定数量的 2.905 倍。之后, 我们统计了 CNN 和 Spectronaut 定性结果中那些重复性较高的肽段 (在所有重复实验的定性结果中至少出现了 80%), 并绘制了相应的韦恩图, 如图3.5b所示。在 Spectronaut 多次鉴定到的 3607 条肽段中, 有 94.4% 的肽段也被 CNN 多次定性到, 同时 CNN 比 Spectronaut 多鉴定到 7691 条独有的肽段。

考虑到这 8 个样本都来源于同一种实验样品, 并且它们在采集过程中的实验参数完全一致, 因此理论上来说这 8 个样本的肽段定性结果应当非常接近。为了验证 CNN 定性结果的重复性, 我们筛选出那些在全部实验的定性结果中都出现的肽段, 并计算了这些肽段在每次实验的定性结果中的占比。作为对照, 我们对 Spectronaut 的肽段定性结果也做了同样的处理。两种方法的统计结果如图3.6所示, 可以看出对于所有的样本来说, CNN 的重复肽段占比都超过了 Spectronaut。

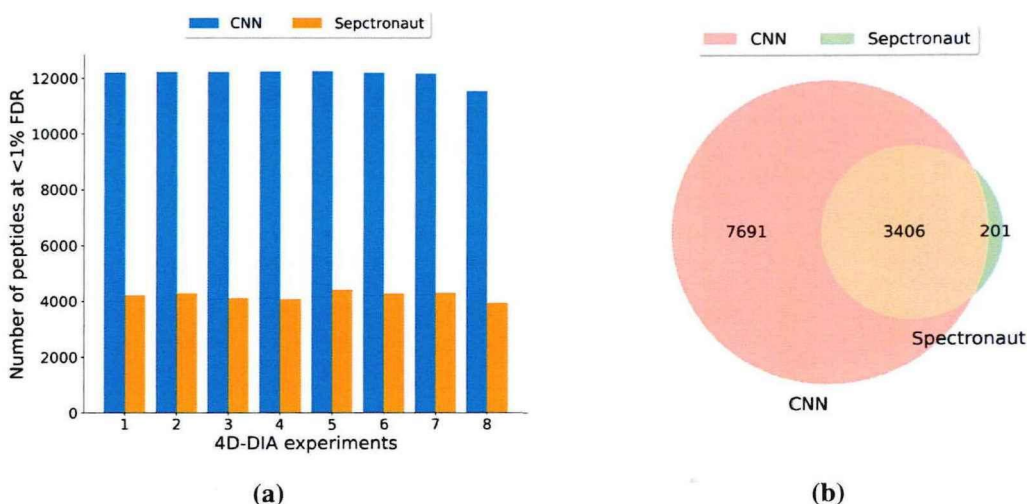


图 3.5 LCH_Plasma 数据集肽段定性结果对比

经统计后可知, CNN 共定性到 10152 条出现 8 次的肽段, 平均比率为 83.2%, 而 Spectronaut 只鉴定到 3144 条出现 8 次的肽段, 平均比率为 74.8%。显然, 对于 LCH_Plasma 数据集而言, CNN 定性结果的重复性更好。

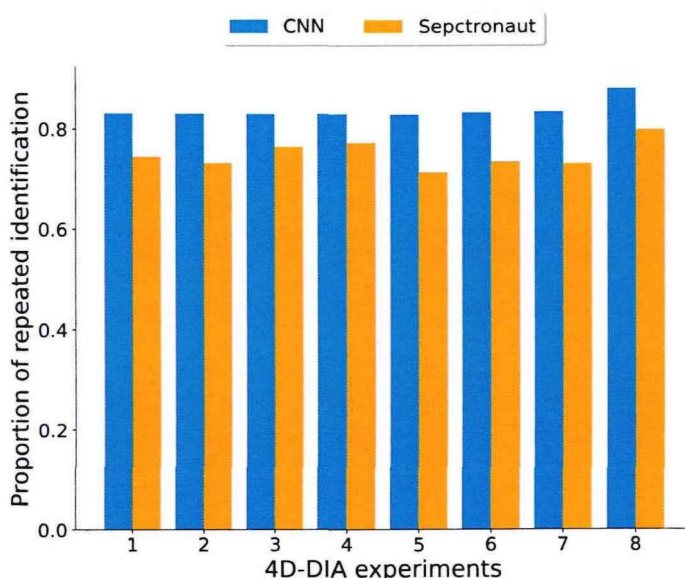


图 3.6 LCH_Plasma 数据集肽段定性重复性对比

从以上的实验结果中, 我们可以看出 CNN 在同源血浆数据集上的定性表现突出, 不仅在肽段鉴定数量上显著地优于 Spectronaut, 同时也具备较高的可靠性和重复性。

3.4.2 公共血浆数据集定性结果

考虑到上一小节的测试集与 CNN 定性模型训练集的相似性较高, 其实验结果可能难以反映出 CNN 定性模型的泛化性能。为此, 我们分别使用 CNN 和

Spectronaut 对三组公共血浆数据集进行了定性分析, 这些数据集分别由不同的实验室制备而来, 并且均已被相关科研人员上传至 EBI PRIDE Archive 数据库中, 对应的数据集编号分别为 PXD030327^[90]、PXD036608^[91]和 PXD040205^[92]。

首先, 我们根据 CNN 和 Spectronaut 的肽段定性结果绘制了一个箱型图, 如图3.7所示。从图中可以看出, 对于每个公共血浆数据集而言, CNN 的肽段鉴定数量都明显超过了 Spectronaut。经计算后可知, CNN 在 PXD030327、PXD036608 和 PXD040205 这三个数据集上的平均肽段鉴定数量分别为 9579、9441 和 9556 条。与 Spectronaut 相比, CNN 的平均肽段鉴定数量分别提高了 2.272 倍、2.200 倍和 2.355 倍, 这充分表明 CNN 在这些公共血浆数据集上具有更加出色的肽段鉴定能力。

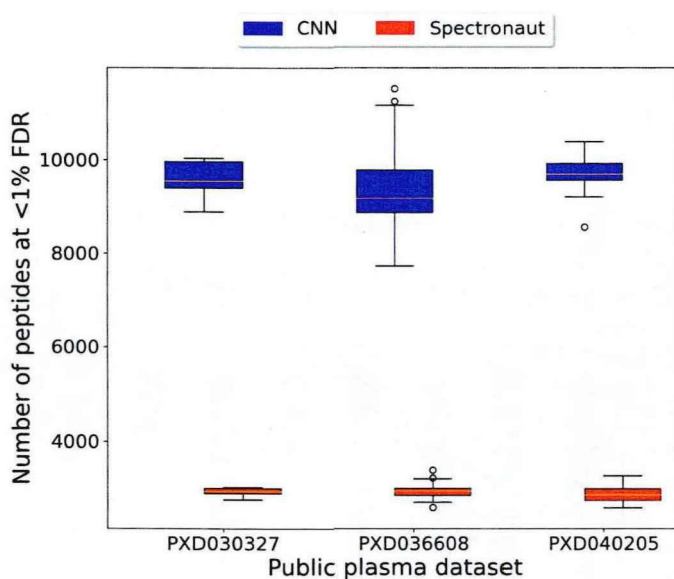


图 3.7 公共血浆数据集肽段定性结果对比

之后, 我们分别对 CNN 和 Spectronaut 在三个公共血浆数据集上的定性结果进行了并集处理, 并针对每个数据集绘制了相应的韦恩图, 如图3.8所示。对于这三组数据集而言, 在 Spectronaut 鉴定到的所有肽段中, 至少有 93.6% 的肽段也出现在了 CNN 的定性结果中, 同时 CNN 还比 Spectronaut 多鉴定到超过 8000 条独有肽段。

在 PXD030327 数据集中, 有 8 个质谱数据是重复采集的实验数据。接下来, 我们就利用这 8 个质谱数据的定性结果来验证 CNN 在公共血浆数据集上的定性重复性。与之前的处理方法类似, 我们从 CNN 的肽段定性结果中筛选出那些在 8 次实验中都被定性到的肽段, 并计算这些肽段在每次实验的定性结果中的占比。之后, 作为对照组, 我们对 Spectronaut 的定性结果也进行同样的处理。两种方法的统计结果如图3.9所示。CNN 共定性到 7115 条出现 8 次的肽段, 平均比率为 76.0%, 而 Spectronaut 共定性到 2259 条出现 8 次的肽段, 平均比率为 77.4%。从

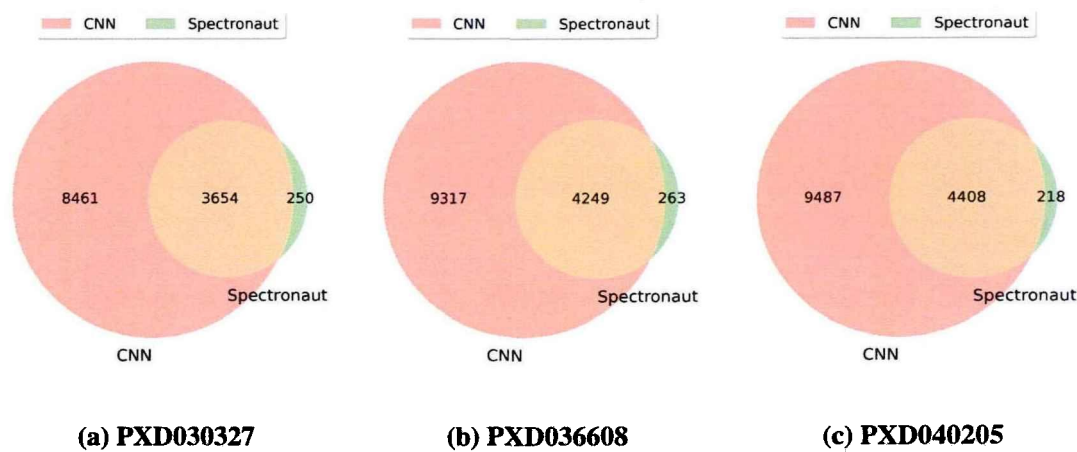


图 3.8 公共血浆数据集肽段定性结果韦恩图

结果中可以看出，在 PXD030327 数据集上，CNN 和 Spectronaut 在定性重复性方面的表现基本相当。两者的重复肽段平均比率均超过了 0.75，这表明两种方法都具备良好的定性重复性。

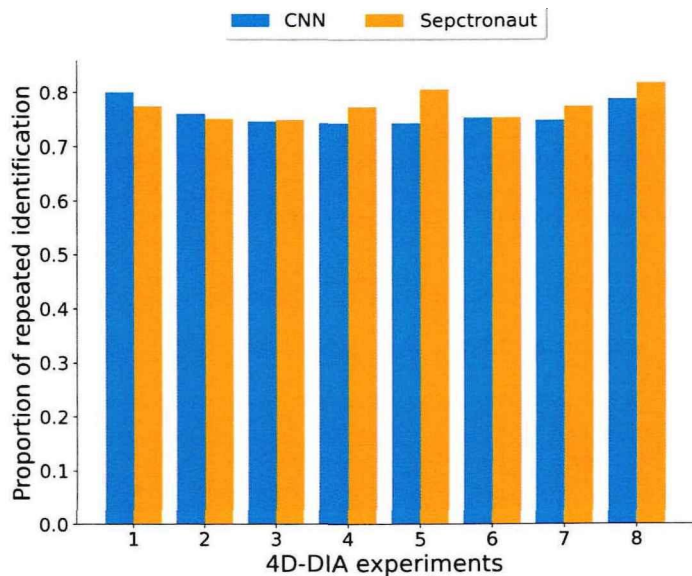


图 3.9 PXD030327 数据集肽段定性重复性对比

以上实验结果表明，CNN 在公共血浆数据集上依然具备较优的定性性能，其肽段定性结果不仅覆盖了 Spectronaut 定性到的绝大多数肽段，还比 Spectronaut 多鉴定到 2 倍以上数量的肽段。此外，CNN 在公共血浆数据集上也表现出了较高的定性重复性，可以在每次重复实验中稳定地鉴定到其中的大多数肽段。

3.5 本章小结

传统的液相色谱-串联质谱技术在处理复杂生物样品时,面临分离效果不佳和质谱数据复杂度过高等问题。为解决这些问题,Meier等人提出了4D-DIA质谱采集方法。该方法通过结合TIMS技术和DIA策略,不仅降低了质谱数据的复杂度,还增添了一维离子淌度信息。目前针对4D-DIA质谱数据的定性分析方法大多依赖于色谱信息,容易受到不同实验平台间色谱差异的影响。为解决该问题,在本章中我们提出了一种基于深度学习的定性分析方法,该方法可以在不依赖于色谱信息的前提下,对4D-DIA质谱数据进行准确的定性分析。

在我们的深度学习定性分析方法中,首先需要对实验质谱和图谱库进行预处理,获取每个肽段图谱最相关的实验图谱集合,并利用Spectronaut工具的定性结果为训练数据赋予标签值。之后,我们构建并训练了一个CNN分类模型,该模型能够从输入数据中提取出离子峰及离子淌度信息的有关特征,并根据这些特征对相应的肽段进行打分。最后,我们对目标库和反库的肽段打分结果进行质量控制,从而得到最终的肽段定性结果。

为了验证CNN模型的性能,我们在若干批人类血浆数据集上进行了实验验证,并将CNN的定性结果与Spectronaut进行对比。结果表明,我们的CNN模型不仅在肽段鉴定数量方面有着更加出色的表现,同时具备较好的定性重复性。

第4章 基于深度学习的 4D-DIA 质谱数据定量分析方法

4.1 研究背景与意义

随着液相色谱-串联质谱技术的快速发展,基于色谱峰面积的无标记定量方法已逐渐成为 DIA 质谱数据定量分析的主流手段。在这类方法中,研究人员一般需要对色谱峰执行提取、对齐、评分和积分等一系列操作,最后利用色谱峰在保留时间维度上的峰面积来作为肽段定量的依据。与传统的谱图计数法相比,色谱峰面积法具有更高的准确性和稳定性,尤其适用于对复杂样品中低丰度肽段的定量。目前已经提出了很多种基于色谱峰面积法的 DIA 质谱数据定量分析方法,例如 Spectronaut^[54]和 DIA-NN^[56]等。这些方法的定量准确性已经通过实验得到了充分验证,并在当前的蛋白质组学研究中得到了广泛应用。然而,传统 DIA 质谱数据中的色谱信息很容易受到色谱实验条件的影响,而不同实验室的色谱梯度又存在差异性,这导致生成的质谱数据在保留时间维度上并不一致,这种差异性削弱了以上这些依赖于色谱信息的 DIA 质谱数据定量方法的通用性。

近年来,4D-DIA 质谱采集方法的出现与广泛应用,为蛋白质组学领域带来了突破性的进展。该方法通过结合 DIA 策略、TIMS 技术以及 PASEF 策略等先进技术,显著提高了蛋白质定量的准确性、重复性和灵敏度,从而为蛋白质组学领域的研究提供了强大的技术支持^[32]。目前,针对 4D-DIA 质谱数据的定量分析仍然沿用了传统的 DIA 质谱数据定量分析方法,因此也同样受限于对色谱信息的依赖。鉴于前文所提及的色谱信息带来的局限性问题,在本章中,我们提出了一种基于深度学习的 4D-DIA 质谱数据定量分析方法。该方法首先对实验质谱数据和肽段图谱库数据进行预处理,然后将实验质谱和肽段图谱的离子峰信息输入到一个卷积神经网络回归模型中进行计算,并使用模型输出的结果来表示对应肽段的定量值,最后再利用第三章中的肽段定性结果对定量到的肽段进行过滤,从而得到最终的肽段定量结果。与传统的色谱峰面积法相比,该方法的流程更为简洁,并且可以在不使用色谱信息的情况下,对 4D-DIA 质谱数据进行准确且可靠的定量分析。

4.2 数据预处理

4.2.1 质谱的匹配与过滤算法

首先,我们需要对原始的实验质谱数据和图谱库数据进行处理,其中所使用的图谱库数据与第三章完全相同。处理过程中的大致步骤包括质谱数据的读取和峰合并、图谱库数据的归一化以及生成相应的诱饵库等,具体细节详见第三

章中的“3.2.1 质谱处理”和“3.2.2 图谱库处理”。之后，我们开始设计质谱的匹配与过滤算法，该算法的目标是从质谱数据中筛选出与肽段图谱最相关的实验图谱集合，以便于后续生成训练数据和测试数据。其中质谱的匹配过程与第三章中的算法3.1完全一致，因此在这里不再赘述。

在完成质谱匹配后，对于图谱库 P 中的第 i 张肽段图谱 P_i ，我们得到了与其相匹配的实验图谱集合 M_i 。在集合 M_i 中，每张实验图谱与 P_i 的匹配峰个数都大于等于峰匹配阈值 n ($n=3$)。接下来，我们需要对 M_i 中的实验图谱进行进一步的过滤，目的是从中选出与 P_i 最相关的若干张实验图谱。与第三章中的相关性计算公式3.1不同，对于质谱数据的定量问题而言，我们在计算 P_i 与 M_i 中的第 j 张实验图谱 M_{ij} 的相关性时，除了要考虑 P_i 与 M_{ij} 所匹配峰的峰强度相似性外，还应当考虑 M_{ij} 中所匹配峰的整体峰强度大小。这是因为如果 M_{ij} 所匹配峰的整体峰强度越大，则说明 P_i 在 M_{ij} 的保留时间点附近碎裂的可能性越大，进而表明 P_i 与 M_{ij} 相关的几率越大。更重要的是，从理论上来说，如果 M_{ij} 的确是由 P_i 对应的肽段母离子参与碎裂而成的，那么 M_{ij} 的整体峰强度大小与 P_i 对应的肽段母离子的丰度应当呈正相关关系。因此，实验图谱中峰强度的准确性会直接影响到后续的肽段定量结果。考虑到峰强度较大的离子峰受到噪声峰的干扰相对较小，所以我们更倾向于筛选出那些整体峰强度较大的实验图谱，从而减少网络模型的最终预测误差。

在经过上述的分析后，本章我们在计算 P_i 与 M_{ij} 的相关性时综合考虑了 P_i 与 M_{ij} 所匹配峰的峰强度相似性以及 M_{ij} 的整体峰强度大小这两个因素。对于前者而言，我们沿用了曼哈顿距离作为度量的标准， $\text{Similarity}(P_i, M_{ij})$ 的具体计算方式详见前文中的公式3.2。至于后者，我们可以通过计算 M_{ij} 所匹配到的峰强度之和来量化这一指标，其计算公式为：

$$\text{OverallPeak}(M_{ij}) = \sum_{k=1}^6 m_{ij}^k \quad (4.1)$$

其中， m_{ij}^k 表示 M_{ij} 所匹配到的第 k 个离子峰的强度。

考虑到 $\text{OverallPeak}(M_{ij})$ 的数量级远远超过 $\text{Similarity}(P_i, M_{ij})$ ，因此，为了同时使用这两个信息来计算 P_i 与 M_{ij} 的相关性，我们采用最小-最大值归一化公式对 $\text{Similarity}(P_i, M_{ij})$ 和 $\text{OverallPeak}(M_{ij})$ 进行处理。最小-最大值归一化公式如下：

$$X' = \frac{X - \text{Min}(X)}{\text{Max}(X) - \text{Min}(X)} \quad (4.2)$$

其中 $\text{Max}()$ 和 $\text{Min}()$ 分别为取最大值函数和取最小值函数。

在进行归一化处理后， $\text{Similarity}(P_i, M_{ij})$ 和 $\text{OverallPeak}(M_{ij})$ 的取值范围均在 0 到 1 之间。显然， P_i 与 M_{ij} 的相关性应当与 $\text{Similarity}(P_i, M_{ij})$ 和 $\text{OverallPeak}(M_{ij})$

成正相关。因此，我们使用如下公式来计算 P_i 与 M_{ij} 的相关性：

$$\text{Correlation}(P_i, M_{ij}) = \text{Similarity}(P_i, M_{ij}) + \text{OverallPeak}(M_{ij}) \quad (4.3)$$

之后，我们根据上述公式计算了 P_i 与 M_i 中每张实验图谱的相关性，最终只保留了相关性最高的 s 张实验图谱 ($s = 6$)。当 M_i 中的图谱数量小于 s 时，我们则使用零图谱（图谱中的所有信息都设置为 0）来填充缺失的实验图谱。整个质谱过滤流程的算法描述可以参考前文中的算法3.2。

4.2.2 训练数据与测试数据的生成

为了完成定量模型的训练，我们需要先获取相应的训练数据集。为此，我们采用了上海生命科学院提供的一组 4D-DIA 质谱数据作为训练数据。这组数据是通过使用 timsTOF 质谱仪对同一批人类血浆样品进行多次质谱采集而获得的，总共包含 20 个不同的质谱数据，其数据集名称为 LCH_Plasma。由于在质谱采集过程中，所有实验参数均保持不变，并且这些质谱数据来源于同一个生物样品，因此理论上来说，这些质谱数据所包含的肽段种类和含量应当非常接近。

对于定量问题而言，我们需要预测出质谱数据中所包含的肽段的定量值，这本质上是一个回归问题。因此，我们需要训练一个回归模型，模型的输出就代表肽段的定量值。为了获取训练数据集的标签，我们使用定量准确性较高的 Spectronaut 工具对以上 20 个 4D-DIA 质谱数据进行了定量分析，得到了相应的肽段定量结果，其中所使用到的图谱库与第三章中的目标图谱库相同。考虑到 Spectronaut 的定量结果中可能仍然存在部分定量不准确的肽段，因此还需要对其进行进一步的筛选，从中获取那些定量准确性较高的肽段，从而减少后续的计算误差，筛选的具体过程如算法4.1所示。在这里，为了方便描述，我们将前 10 个质谱数据记作样品 A，后 10 个质谱数据记作样品 B。首先，我们获取那些在 A、B 样品的定量结果中都全部出现的肽段，并分别计算这些肽段在 A、B 样品的定量结果中的变异系数 CV (Coefficient of variation)， CV 的计算方法如前文中的公式2.6所示。之后，我们筛选出那些 CV_A 和 CV_B 均小于 0.1 的肽段，并计算出这些肽段的对数比 LR (Logarithmic ratio)，其计算方法详见公式2.7。最后，考虑到所有肽段的理论对数比为 0（所有质谱数据均来源于重复样本采集），我们只保留了那些对数比的绝对值小于 0.5 的肽段，并将这些肽段对应的 Spectronaut 定量结果作为训练数据的标签保存下来。

在得到训练数据的标签后，我们使用前文中的质谱匹配与过滤算法计算出那些定量准确性较高的肽段所对应的实验图谱集合，并将这些肽段图谱和实验图谱的有关信息，包括肽段名称和匹配离子峰信息等，保存至文件中，作为训练数据备用。关于测试数据的生成，我们只需将目标图谱库中的所有肽段及其对应

算法 4.1 定量模型训练集标签获取算法

Data: Q_A : 样品 A 的 Spectronaut 定量结果
 Q_B : 样品 B 的 Spectronaut 定量结果
Result: Q_{AB} : 经过滤后准确性较高的 Spectronaut 定量结果

```

1  $P_A = \text{getPeptide}(Q_A)$ ; // 获取  $Q_A$  对应的肽段集合  $P_A$ 
2  $P_B = \text{getPeptide}(Q_B)$ ;
3  $P_{A \cap B} = P_A \cap P_B$ ;
4  $C_A = \text{counter}(Q_A)$ ; // 统计  $Q_A$  中肽段的出现次数
5  $C_B = \text{counter}(Q_B)$ ;
6 for  $p$  in  $P_{A \cap B}$  do
7   if  $C_A[p] == 10$  and  $C_B[p] == 10$  then
8      $P_{AB}^1.add(p)$ ;
9   end
10 end
11  $CV_A = \text{computeCV}(Q_A, P_{AB}^1)$ ; // 计算  $P_{AB}^1$  中肽段在  $Q_A$  中的 CV 值
12  $CV_B = \text{computeCV}(Q_B, P_{AB}^1)$ ;
13 for  $p$  in  $P_{AB}^1$  do
14   if  $CV_A[p] < 0.1$  and  $CV_B[p] < 0.1$  then
15      $P_{AB}^2.add(p)$ ;
16   end
17 end
18  $LR_{AB} = \text{computeLR}(Q_A, Q_B, P_{AB}^2)$ ; // 计算  $P_{AB}^2$  中肽段的 LR 值
19 for  $p$  in  $P_{AB}^2$  do
20   if  $|LR_{AB}[p]| < 0.5$  then
21      $P_{AB}^3.add(p)$ ;
22   end
23 end
24  $Q_{AB} = (\text{filter}(Q_A, P_{AB}^3), \text{filter}(Q_B, P_{AB}^3))$ ; // 获取  $P_{AB}^3$  中肽段在  $Q_A$  和  $Q_B$  中的定量结果
25 return  $Q_{AB}$ ;

```

的实验图谱集合的有关信息保存至文件中即可。至于本章实验中所使用的具体测试集，我们将会在“4.4 实验结果与分析”小节中进行介绍。

4.3 定量模型构建

质谱数据定量分析任务是获取质谱数据中存在的所有肽段的定量值。由于在第三章中，我们已经实现了对 4D-DIA 质谱数据的定性分析，因此，在本章中，我们只需要对图谱库中的所有肽段进行定量值的预测即可，而无需关心该肽段是否存在于质谱数据中。至于最终的肽段定量结果，我们可以在定量模型输出结果的基础上使用定性模型的肽段定性结果来进行过滤。而为了实现 4D-DIA 质谱数据的定量分析，我们需要构建一个回归模型。与第三章中的定性模型类似，在这里我们仍然采用了 CNN 作为 4D-DIA 质谱数据定量分析的神经网络模型。

之后，我们开始设计定量模型的网络结构。从理论上来说，在不考虑色谱信息的情况下，肽段的定量值应当主要依赖于其肽段母离子碎裂产生的碎片离子离

子峰强度信息,而与其离子淌度等信息关系不大。因此,在定量模型中,我们只需要利用训练数据中的离子峰强度信息即可。具体来说,我们首先将肽段图谱和实验图谱集合的离子峰强度向量输入到多个卷积层中进行卷积处理。这些卷积层可以逐步提取出输入数据的有关特征,从而使得模型能够学习到更高层次的特征表示,进而提高模型的表达能力和泛化能力。接下来,我们将各自提取到的特征拼接起来,再次输入到一个卷积层中进行计算,这一步的目的是提取这些特征之间的深层交互信息。此外,在每次完成卷积计算后,我们都会对卷积结果应用 ReLU 激活函数,该函数会将正值直接映射,将负值置为零。ReLU 激活函数不仅为神经网络引入了非线性关系,还使得一部分神经元的输出为 0,这一机制有效增强了网络的稀疏性,从而降低了模型的过拟合风险。之后,我们将卷积层输出的结果展平为一维向量,然后将其输入到多个全连接层中,并在最后一个全连接层后输出对应肽段的定量值预测结果。这些全连接层通过权重矩阵将前一层的输出转换为更高层次的特征,进而增强了模型的回归预测能力。在每个全连接层中,我们都会使用 ReLU 激活函数来对输出结果进行处理,这样做不仅增强了特征的非线性表达能力,而且确保了最终的网络预测结果一定是一个非负的数值,从而满足了肽段定量值非负这一前提。此外,我们在所有卷积模块以及除最后一层外的每个全连接层后都进行了 dropout 操作,该操作可以随机地将部分神经元失活,从而降低模型出现过拟合现象的风险。CNN 定量模型的网络结构如图4.1所示。

在完成 CNN 定量模型的网络结构设计后,我们还需要对神经网络中的一些超参数进行设置,这里的超参数可以大致分为网络结构参数和训练参数两种类型。以下是对它们的具体介绍:

CNN 定量模型的网络结构参数主要包括各个网络层的通道数、卷积核的大小和步长等。表4.1展示了神经网络结构中的部分参数信息。其中, b 表示批量大小,即模型在每次训练中处理的样本数量, s 表示实验图谱集合中的图谱数量, m 表示每张肽段图谱的离子峰数量。考虑到 CNN 定量模型的输入数据由多个一维向量组成,因此在模型的每个卷积层中,我们都采用了一维卷积核来进行卷积计算。在表4.1的前两个模块中,我们将三个卷积层的卷积核个数分别设置为 64、128 和 256,通过不断增加卷积核的个数可以让模型在每一个卷积层中都能捕获到更多的特征信息。之后,考虑到在拼接模块中,输入数据的通道数较多,因此我们将这里的卷积核个数设置为 256,从而降低了模型后续计算的复杂度。而在最后的线性模块中,我们将三个线性层的输出神经元个数分别设置为 256、64 和 1,这样做可以让模型逐步降低输入特征的维度,并最终输出对应肽段的定量值预测结果。此外,鉴于同一张图谱中的不同离子峰之间并不具备较强的关联性,因此对于每个卷积层,我们都将一维卷积核的大小设置为 1,填充设置为 0,步

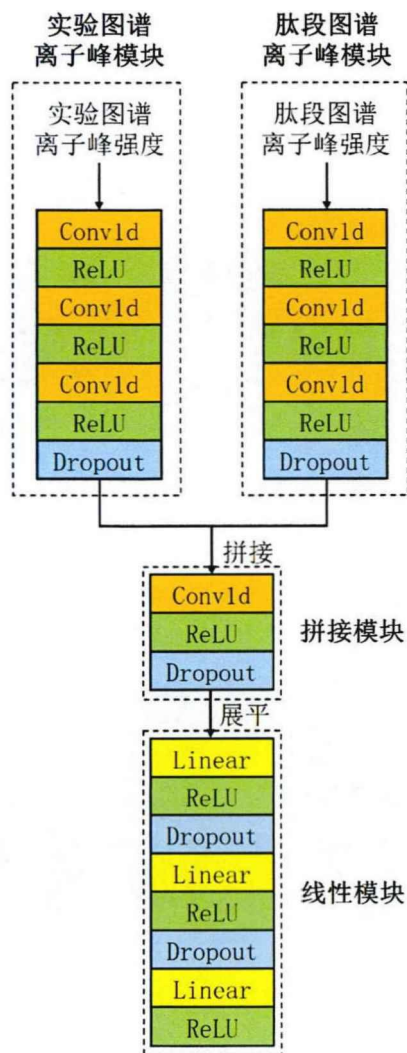


图 4.1 CNN 定量模型网络结构示意图

长设置为 1。这种设置不仅降低了模型的复杂度，还有效避免了不必要的空间依赖关系，从而提高了模型的效率和准确性。

对于 CNN 定量模型的训练参数而言，我们可以将其分成两大类：数值型训练参数和非数值型训练参数。其中数值型训练参数主要包括学习率、迭代次数和批量大小等，这些参数对于模型至关重要，它们可以直接影响模型的训练速度和最终性能，其调整策略与“3.3 定性模型构建”小节中介绍的调参方法完全一致，因此在这里不再赘述。至于非数值型训练参数则主要包括优化器类型和损失函数类型等。在优化器的选择上，与第三章类似，考虑到 Adam^[89] 优化器具有适应性强和收敛速度快等优点，因此我们仍然采用了 Adam 优化器来进一步优化定量模型的内部参数。而在损失函数的选择方面，由于该模型为回归模型，所以我们采用了常用的均方误差损失函数 (Mean Squared Error Loss, MSE Loss) 来作为网

表 4.1 CNN 定量模型网络结构参数

模块	网络层类型	输入大小	输出大小
实验图谱离子峰模块	Conv1d	(b, s, m)	$(b, 64, m)$
	Conv1d	$(b, 64, m)$	$(b, 128, m)$
	Conv1d	$(b, 128, m)$	$(b, 256, m)$
肽段图谱离子峰模块	Conv1d	$(b, 1, m)$	$(b, 64, m)$
	Conv1d	$(b, 64, m)$	$(b, 128, m)$
	Conv1d	$(b, 128, m)$	$(b, 256, m)$
拼接模块	Conv1d	$(b, 256 \times 2, m)$	$(b, 256, m)$
线性模块	Linear	$(b, 256 \times m,)$	$(b, 256,)$
	Linear	$(b, 256,)$	$(b, 64,)$
	Linear	$(b, 64,)$	$(b, 1,)$

络模型的损失函数，其具体公式为：

$$\text{MSELoss} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (4.4)$$

其中 N 为样本数量， y_i 为第 i 个样本的标签值， $\hat{y}_i$ 为第 i 个样本的网络预测值。

最后，为了获得最优模型，我们从训练数据中抽取了 20% 的数据作为验证集。每当网络完成一轮迭代训练后，我们都会在验证集上计算其相对误差，并最终选择在验证集上相对误差最小的模型作为最终模型。在这里，相对误差的计算公式为：

$$\text{RelativeLoss} = \sum_{i=1}^m \frac{|y_i - \hat{y}_i|}{y_i} \quad (4.5)$$

其中 m 表示验证集中的样本个数， y_i 表示第 i 个样本的标签值， $\hat{y}_i$ 表示第 i 个样本的网络预测值。

本章实验中所使用的硬件和软件环境与第三章完全一致，相关配置信息如前文中的表3.3所示。

在构建完定量模型后，我们将数据预处理阶段生成的测试数据输入到模型中进行计算，得到目标库的肽段定量结果。之后利用第三章中的 CNN 定性模型获取对应的肽段定性结果，并利用该结果对 CNN 定量模型定量到的肽段进行过滤，从而得到对应实验样品的最终肽段定量结果。

4.4 实验结果与分析

4.4.1 同源血浆数据集定量结果

首先，我们对一组与训练数据集同源的人类血浆样品数据进行了定量分析。这组测试数据集由上海生命科学院提供，共包含 8 个重复采集的 4D-DIA 质谱数据，其数据集名称为 LCH_MN_144-Plasma。为了评估 CNN 定量模型的准确性，

我们将前4个实验样本记作样品A，后4个实验样本记作样品B。首先，我们获取那些在样品A和样品B中都至少被定量到3次的肽段。之后，我们还需要进行CV过滤操作，其目的是筛选出那些定量准确性较高的肽段。具体来说，我们会分别计算这些肽段在样品A和样品B中的变异系数 CV ，并保留 CV_A 和 CV_B 均小于0.2的肽段。最后，我们计算出这些过滤后的肽段的对数比，并将其与理论对数比进行比较。此外，我们还使用了定量准确性较高的Spectronaut工具进行对比实验，确保在相同的条件下对这组测试数据集进行定量分析及实验结果统计。

经统计，在进行CV过滤之前，CNN共定量到10486条重复性较高的肽段（在样品A和样品B中都至少被定量到3次），而Spectronaut只定量到4641条重复性较高的肽段，对应的韦恩图如图4.2a所示。从图中可以看出CNN覆盖了Spectronaut超过96%的定量结果。而在进行CV过滤后，根据图4.2b可知，CNN和Spectronaut各定量到5035条和2201条肽段，两种方法经CV过滤后的肽段得率分别为48.0%和47.4%，同时在Spectronaut的定量结果中有50.8%的肽段也出现在了CNN的定量结果中。以上结果说明，CNN在肽段定量数量方面显著优于Spectronaut，且两者肽段定量结果的准确性相当。

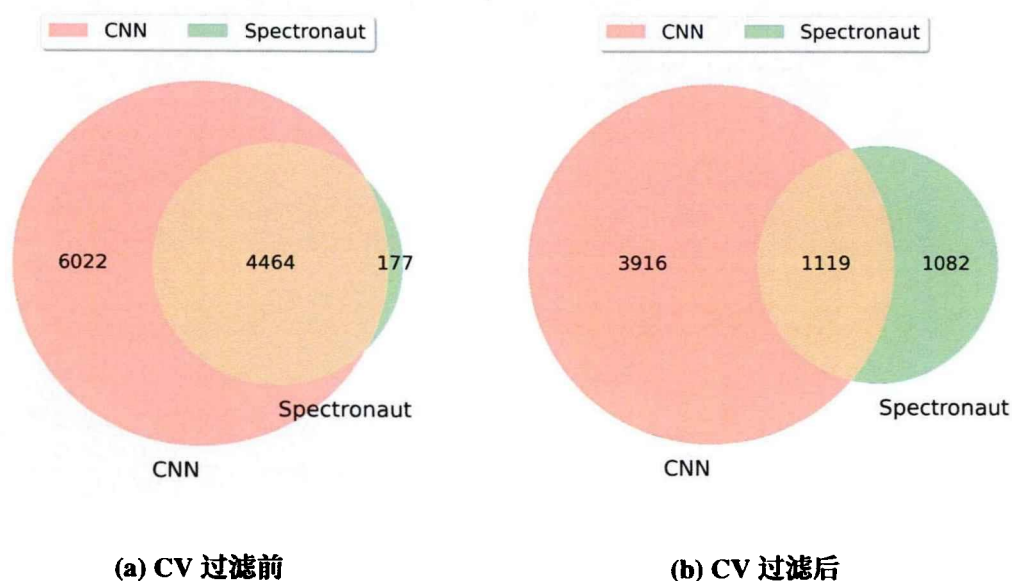


图 4.2 LCH_MN_144-Plasma 数据集肽段定量结果韦恩图

之后，我们根据CNN和Spectronaut经CV过滤后的肽段定量结果的对数比，绘制了相应的散点图，如图4.3所示。图中的横坐标表示肽段在样品B中的平均定量值的对数，纵坐标表示肽段的对数比，虚线则表示肽段的理论对数比。考虑到所有实验数据均为重复采集的质谱数据，因此理论对数比应当为0。从图4.3中可以看出，CNN和Spectronaut定量到的大多数肽段都较为均匀地分布在理论线

附近。在经过详细计算后可知,对于CNN和Spectronaut经CV过滤后的肽段定量结果而言,两者的对数比的中位数分别为0.015和-0.008,这与理论对数比都十分接近。

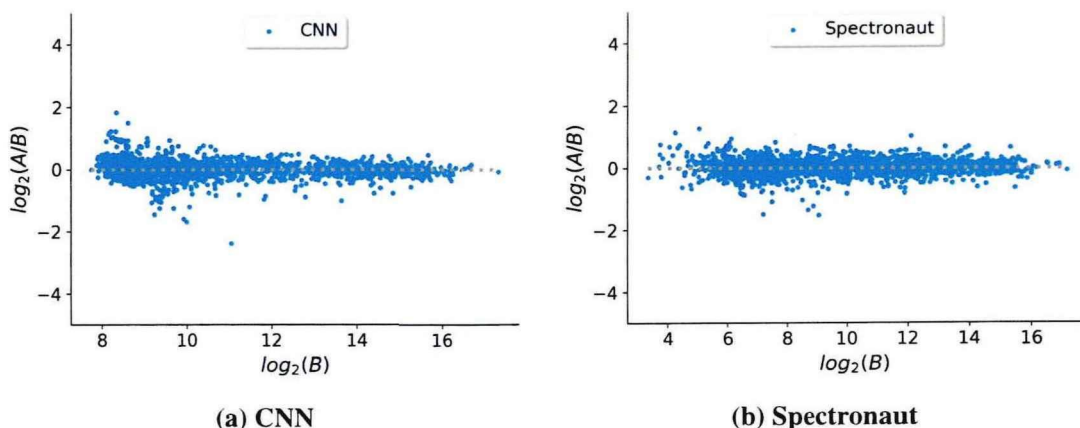


图4.3 LCH_MN_144-Plasma数据集肽段定量结果散点图对比

为了更细致地评估CNN和Spectronaut在A、B样品上的定量准确性,我们根据肽段在样品B中的平均定量值大小将CV过滤后的肽段平均分成了三份,并绘制了相应的肽段对数比热力图,如图4.4所示。从图中可以看出,在CNN的定量结果中有54.2%的肽段的对数比误差小于0.1,有97.7%的肽段的对数比误差小于0.5。相比之下,在Spectronaut的定量结果中,对数比误差小于0.1和0.5的肽段分别占了所有肽段的39.0%和95.0%。此外,在CNN的定量结果中,对数比误差小于0.1的低、中、高丰度肽段的占比分别为66.2%、53.2%和43.1%,而对于Spectronaut的定量结果来说,在相同的误差阈值下,低、中、高丰度肽段的占比分别为31.2%、43.4%和42.4%,这些数值均低于CNN的对应占比。从上述结果中我们可以看出,对于这些定量准确性较高的肽段而言,与Spectronaut相比,CNN展现出了更低的对数比误差,这一点在低丰度肽段上尤其突出,这进一步说明了CNN具有较高的定量准确性。

此外,我们还对CNN和Spectronaut在每个重复样本上的定量相关性进行了评估。对于这8个重复样本,我们筛选出那些在8次定量结果中全部出现的肽段,并计算这些肽段在每对定量结果上的相关系数,绘制相应的热力图如图4.5所示。在这里,我们采用了常用的皮尔逊相关系数作为统计指标,它可以用来衡量两个变量之间的线性相关程度,其计算公式为:

$$r(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (4.6)$$

其中, $r(x, y)$ 表示变量 x 和变量 y 之间的相关系数, x_i 和 y_i 分别表示 x 和 y 第 i 个元素的值, $\bar{x}$ 和 $\bar{y}$ 分别表示 x 和 y 的平均值, n 表示 x 和 y 的元素个数。

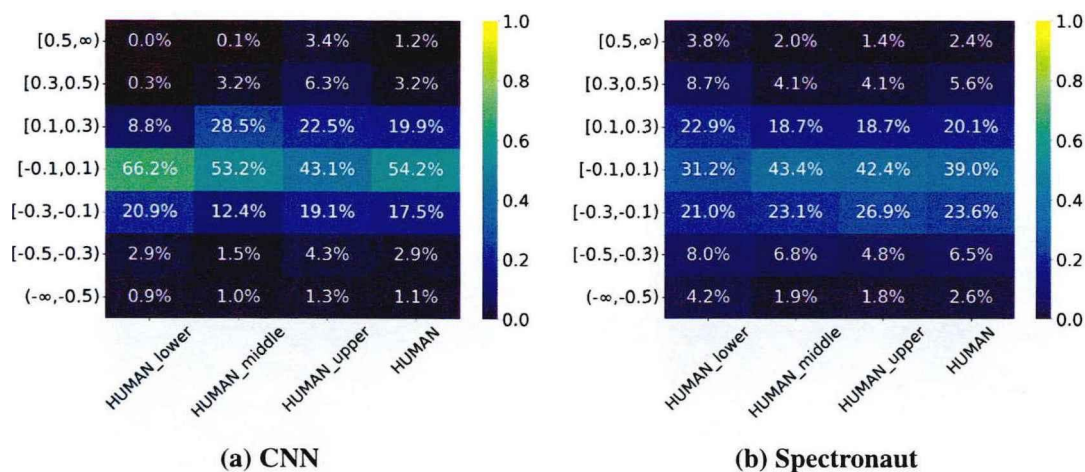


图 4.4 LCH_MN_144-Plasma 数据集肽段定量结果热力图对比

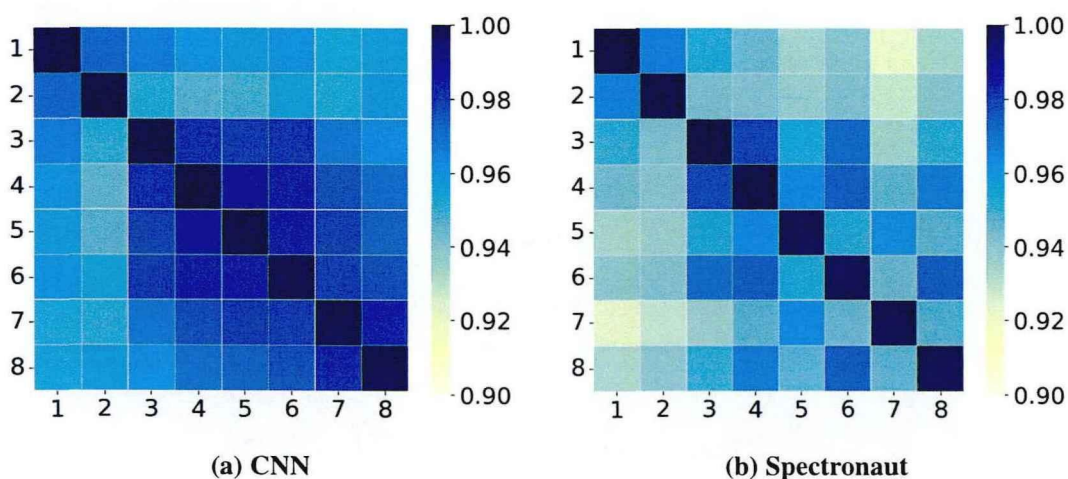


图 4.5 LCH_MN_144-Plasma 数据集肽段定量重复性对比

从图4.5中我们可以看出，与 Spectronaut 相比，CNN 在每对定量结果上的相关系数普遍更高。经统计可知，CNN 和 Spectronaut 相关系数的均值分别为 0.971 和 0.955。这一结果表明，在 LCH_MN_144-Plasma 数据集上，CNN 表现出了更加出色的定量重复性。

综合以上实验结果可知，和 Spectronaut 相比，CNN 定量模型在与训练数据集同源的 LCH_MN_144-Plasma 数据集上可以定量到更多的肽段。而在对两者的定量结果进行同样的筛选得到准确性较高的肽段后，CNN 的定量准确性和重复性也都略优于 Spectronaut。

4.4.2 公共血浆数据集定量结果

为了进一步验证 CNN 定量模型的泛化性能，我们分别使用 CNN 和 Spectronaut 对一组重复采集的公共血浆数据集进行了定量分析测试。该数据集来自于

其他实验平台,共包含8个重复采集的4D-DIA质谱数据,可以从EBI PRIDE Archive数据库中下载得到,对应的数据集编号为PXD040205^[92]。与之前的统计方法相同,我们将前4个实验样本记作样品A,后4个实验样本记作样品B,并从肽段定量结果中筛选出那些重复性较高(在样品A和样品B中都至少被定量到3次)且 CV_A 和 CV_B 均小于0.2的肽段,计算相应的对数比。

对于该数据集而言,在进行CV过滤前,CNN和Spectronaut分别定量到8571条和2545条重复性较高的肽段,两者对应的韦恩图如图4.6a所示。从图中,我们可以看出CNN覆盖了Spectronaut超过90%的定量结果。而在进行CV过滤后,两种方法的定量结果如图4.6b所示。CNN和Spectronaut分别定量到5232条和1694条肽段,两种方法经CV过滤后的肽段得率分别为61.0%和66.6%,同时在Spectronaut的定量结果中有58.0%的肽段也出现在了CNN的定量结果中。以上实验结果表明,和Spectronaut相比,CNN在PXD040205数据集上仍然可以定量到更多的肽段,且两者的定量准确性相当。

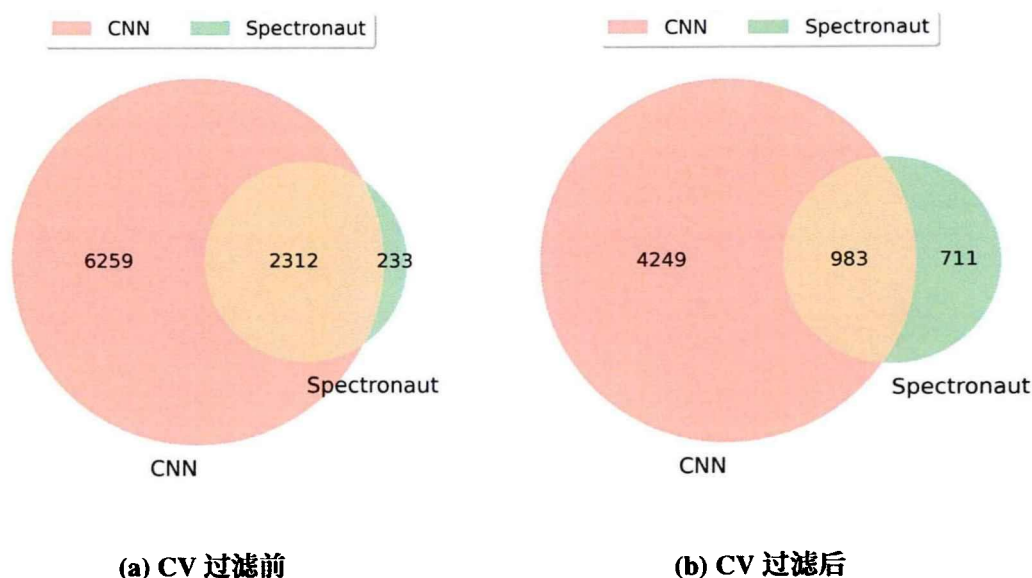


图 4.6 PXD040205 数据集肽段定量结果韦恩图

之后,我们根据CNN和Spectronaut在该数据集上经CV过滤后的肽段定量结果的对数比,绘制了相应的散点图,如图4.7所示。图中的横坐标表示肽段在样品B中的平均定量值的对数,纵坐标表示肽段的对数比,虚线则表示肽段的理论对数比。由于所有实验数据均为重复采集的质谱数据,因此理论对数比应当为0。从图4.7中可以看出,CNN和Spectronaut定量到的大部分肽段都较为均匀地分布在理论线附近。在进行详细计算后可知,对于CNN和Spectronaut经CV过滤后的肽段定量结果而言,两者的对数比的中位数分别为0.001和0.005,这与理论对数比都十分接近。

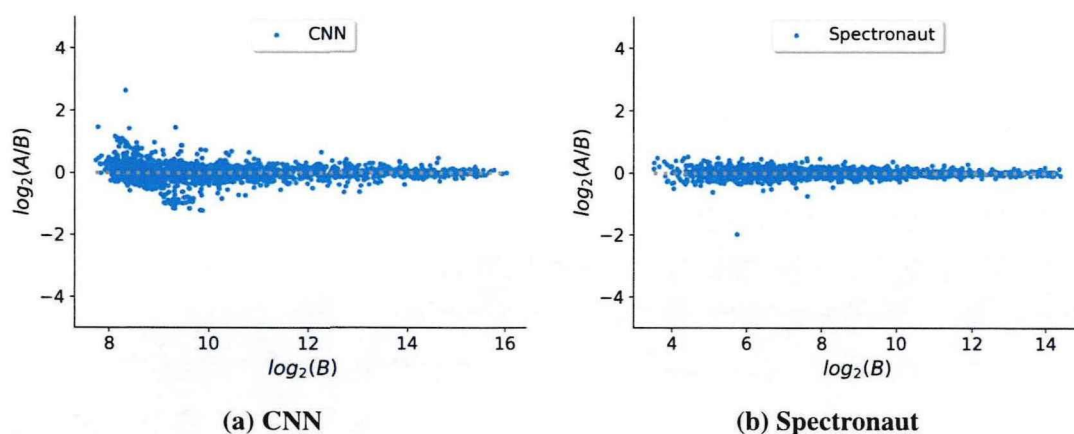


图 4.7 PXD040205 数据集肽段定量结果散点图对比

而在绘制了相应的热力图后，我们发现 CNN 仍旧表现出了较高的定量准确性。根据图4.8中的结果可知，在 CNN 和 Spectronaut 经 CV 过滤后的定量结果中，分别有 55.6% 和 66.1% 的肽段的对数比误差小于 0.1。此外，在 CNN 的定量结果中，对数比误差小于 0.1 的低、中、高丰度肽段的占比分别为 67.3%、53.0% 和 46.6%，而对于 Spectronaut 的定量结果来说，在相同的误差阈值下，低、中、高丰度肽段的占比分别为 47.9%、68.3% 和 82.1%。以上结果表明，CNN 和 Spectronaut 在整体对数比误差方面表现相近，且 CNN 在低丰度肽段上的对数比误差更低。

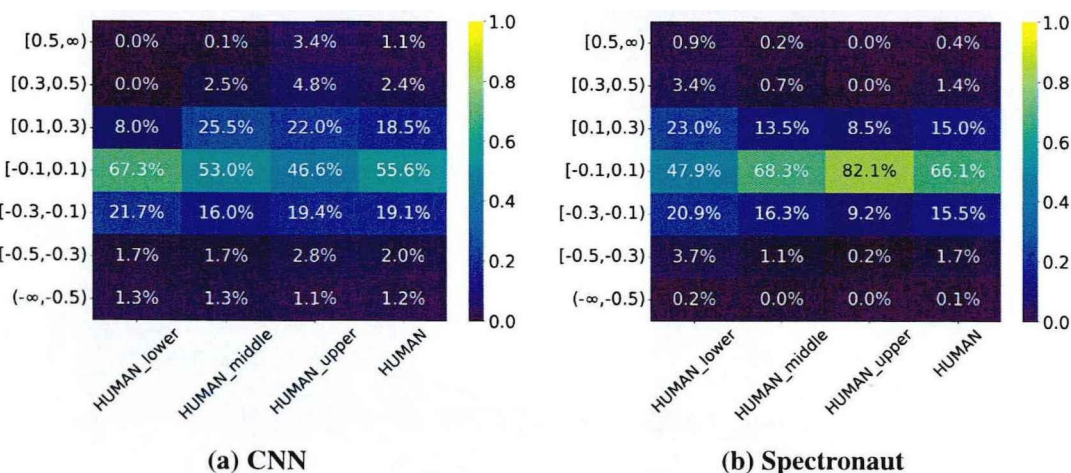


图 4.8 PXD040205 数据集肽段定量结果热力图对比

最后，我们同样对 CNN 和 Spectronaut 的定量重复性进行了评估，结果如图4.9所示。CNN 和 Spectronaut 在每对定量结果上的相关系数均高于 0.95，两者的相关系数均值分别为 0.976 和 0.984，这说明 CNN 模型在 PXD040205 数据集上同样具备较高的定量重复性。

以上实验结果表明，对于 PXD040205 数据集这一来自于其他实验平台的公共血浆数据集而言，CNN 定量模型不仅比 Spectronaut 定量到了更多的肽段，同

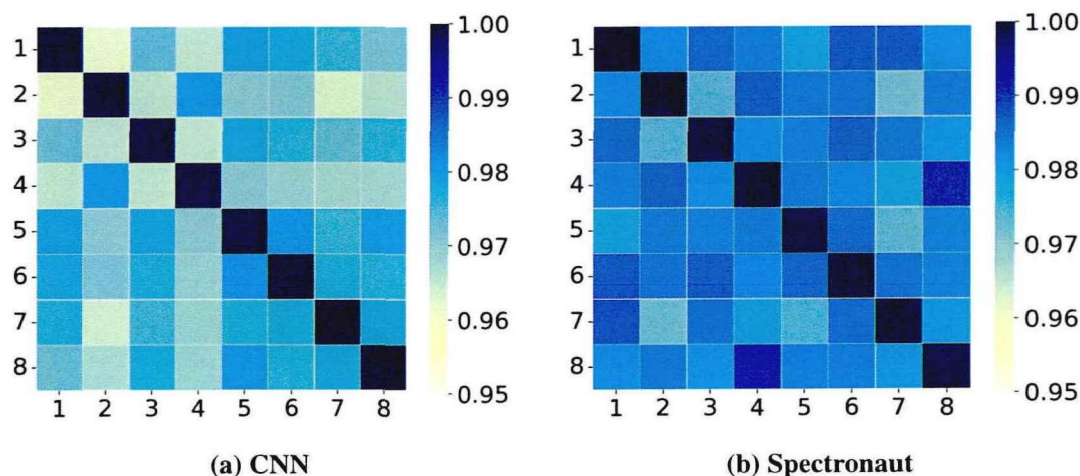


图 4.9 PXD040205 数据集肽段定量重复性对比

时还具备与其相当的定量准确性和重复性，这充分反映出 CNN 定量模型优异的定量性能。

4.5 本章小结

近年来，4D-DIA 方法的出现为蛋白质组学研究带来了重大突破。它通过结合 TIMS 技术和 DIA 策略等先进技术，显著提高了蛋白质定量的准确性、重复性和灵敏度。然而，当前针对 4D-DIA 质谱数据的定量分析大多沿用了传统的色谱峰面积法。这类方法严重依赖于色谱信息，容易受到不同实验平台间色谱梯度差异的影响。为了克服这一局限性，在本章中我们提出了一种基于深度学习的 4D-DIA 质谱数据定量分析方法。该方法能够在不使用色谱信息的情况下，对 4D-DIA 质谱数据进行精准可靠的定量分析。

在本章提出的深度学习定量分析方法中，我们首先对实验质谱和肽段图谱库进行预处理，得到每张肽段图谱最相关的实验图谱集合，并利用定量准确性较高的 Spectronaut 工具获取训练集的标签，从而生成对应的训练数据。之后，我们设计了一个 CNN 回归模型，该模型通过提取输入数据中离子峰信息的相关特征，实现了对相应肽段定量值的准确预测。最后，我们使用上一章中 CNN 定性模型的肽段定性结果对定量到的肽段进行过滤，从而获取最终的肽段定量结果。

为了验证 CNN 定量模型的准确性，我们在人类血浆样品数据集上进行了定量实验验证，并与 Spectronaut 的定量结果进行对比。结果表明，我们的 CNN 定量模型不仅可以定量到更多肽段，同时具有较好的定量准确性和重复性。

第5章 总结与展望

5.1 工作总结

在蛋白质组学的发展历程中,质谱技术发挥了至关重要的作用。该技术利用高精度的质量测量手段,可以和液相色谱等分离技术相结合,对实验样品中的蛋白质进行细致准确的质谱采集和分析,从而实现对蛋白质的鉴定和定量。在质谱采集的过程中,DIA策略是一种常用的二级质谱采集方法。与DDA策略相比,DIA策略具备高通量、高重复性和高灵敏度等显著优势,特别适用于大规模样本的蛋白质组学分析。但不可忽视的是,当前传统的DIA质谱数据仍面临液相色谱洗脱时间较长和谱图复杂度较高等难题。为了应对这些挑战,近年来科研人员开始探索并采用一些新型的DIA质谱数据采集方法,如DI-SPA和4D-DIA等。然而,现有的DIA质谱数据分析方法大多都严重依赖于色谱信息,容易受到不同实验平台间色谱差异的影响,难以对DI-SPA质谱数据和4D-DIA质谱数据进行高效且通用的分析。鉴于此,本文主要针对这两种新型的DIA质谱数据,提出了若干种无色谱分析方法,其主要研究内容包括以下三个方面:

1. 针对循环采集的DI-SPA质谱数据的定性及定量分析方法

在传统的DIA质谱数据采集流程中,液相色谱洗脱环节因其较长的分离时间成为高通量蛋白质组学发展的瓶颈。为应对这一挑战,研究者们提出了DI-SPA方法,该方法采用FAIMS技术替换液相色谱技术,大幅缩减了质谱采集时间。然而,由于DI-SPA质谱数据缺少色谱峰信息,因此大部分传统的DIA质谱数据分析工具无法直接对其进行解析。尽管已有CsoDIAq等方法可以对DI-SPA质谱数据进行分析,但其定性及定量效果仍存在一定的不足。鉴于此,本文提出了一种名为RE-FIGS的DI-SPA质谱数据分析方法。该方法首先提取SSM结果中的多个特征信息,之后利用半监督学习和线性判别分析等方法构建分类模型。该模型可以对肽段图谱库中的肽段进行准确的打分,进而得到可靠的肽段定性结果。此外,我们还发现,在对DI-SPA质谱数据进行循环采集后,可以利用其中的一些重复性特征大幅提升肽段鉴定数量。同时,由于FAIMS技术的分离速度远远快于液相色谱,因此即使对于循环采集的DI-SPA质谱数据而言,其质谱采集时间仍然远远少于传统的DIA质谱数据。在定量方面,RE-FIGS采用了FIGS中的Fisdec图谱分解算法对MACC分数最高的实验图谱进行解谱,并将解谱系数作为肽段的定量值,从而实现了对DI-SPA质谱数据的定量分析。实验结果表明,与CsoDIAq相比,RE-FIGS在定性及定量方面均取得了一定的提升。

2. 基于深度学习的4D-DIA质谱数据定性分析方法

传统的液相色谱-串联质谱技术在处理人类血浆样本等复杂生物样品时,常

因分离效果不佳而出现质谱数据复杂度过高的问题。为此, Meier 等人创新性地提出了 4D-DIA 质谱采集方法, 该方法结合了 TIMS 技术和 DIA 策略, 不仅降低了每张实验图谱的复杂度, 同时还为质谱数据增添了一维离子淌度信息。然而, 现有的 4D-DIA 质谱数据定性分析方法大多依赖于色谱信息, 容易受到不同实验平台间色谱梯度差异的影响。为解决该问题, 本文提出了一种基于深度学习的 4D-DIA 质谱数据定性分析方法。该方法在不依赖于色谱信息的前提下, 实现了对 4D-DIA 质谱数据的定性分析。在该方法中, 我们首先需要对实验质谱和肽段图谱库进行预处理, 筛选出每张肽段图谱最相关的实验图谱集合, 并利用 Spectronaut 的肽段定性结果为训练数据赋予标签值。之后, 我们构建并训练了一个卷积神经网络分类模型, 该模型能够自动提取输入数据中的离子峰和离子淌度特征, 并基于这些特征计算相应肽段的打分值, 从而获得最终的肽段定性结果。最后, 为了验证 CNN 定性模型的性能, 我们在多批人类血浆数据集上进行了实验, 并与 Spectronaut 的定性结果进行了对比。实验结果表明, CNN 定性模型不仅在肽段定性数量方面表现得更为出色, 而且具有较好的定性准确性和重复性。

3. 基于深度学习的 4D-DIA 质谱数据定量分析方法

当前, 针对 4D-DIA 质谱数据的定量分析, 大多仍依赖于传统的色谱峰面积法。尽管此类方法具有较高的灵敏度, 但不同实验平台间的色谱条件差异却限制了其通用性。为了克服这一局限性, 我们提出了一种基于深度学习的 4D-DIA 质谱数据定量分析方法。这一方法在完全摒弃色谱信息的情况下, 依然能对 4D-DIA 质谱数据进行准确可靠的定量分析。在该方法中, 我们首先对实验质谱和肽段图谱库进行预处理, 为每张肽段图谱匹配最相关的实验图谱集合, 并利用 Spectronaut 的肽段定量结果获取训练集的标签。接下来, 我们构建并训练了一个卷积神经网络回归模型, 它能从输入数据的离子峰信息中提取出相应特征, 进而准确预测对应肽段的定量值。最后, 我们利用 CNN 定性模型的肽段定性结果对 CNN 定量模型定量到的肽段进行了过滤, 得到最终的肽段定量结果。为了验证 CNN 定量模型的性能, 我们在人类血浆数据集上进行了定量实验, 并与 Spectronaut 进行了全面对比。实验结果表明, 我们的 CNN 定量模型不仅可以定量到更多的肽段, 并且在定量准确性和重复性方面也展现出了优异的表现。

5.2 未来工作展望

以下是本文工作的未来展望:

1. 深入挖掘 DI-SPA 质谱数据中的补偿电压信息

尽管 DI-SPA 质谱数据不包含传统的色谱峰信息, 但 FAIMS 技术的应用也为其引入了一维补偿电压信息。然而, 考虑到肽段母离子在不同补偿电压下的通

过量是一个离散且难以预测的值,因此,本文在对 DI-SPA 质谱数据进行定性及定量分析时并没有使用到补偿电压信息。不过,目前我们尚不清楚同一肽段母离子对应的碎片离子在补偿电压维度上是否存在关联性。倘若这种关联性确实存在,那么我们便可以利用这维信息对 SSM 结果进行更准确的评分,从而增强肽段鉴定的可靠性。至于定量方面,由于 DI-SPA 质谱数据不具备峰形状的强度模式,因此我们没有对图谱解谱后的系数进行修正和平滑处理,这使得我们的 RE-FIGS 定量方法更容易受到噪声峰的干扰。同样,如果我们能够根据补偿电压信息获取相关实验图谱之间的强度关系,那么便可以利用这一点对定量结果进行修正,进而提升肽段定量的准确性。因此,未来还需要对 DI-SPA 质谱数据中的补偿电压信息进行更深入的挖掘,从而进一步提升定性和定量的性能。

2. 优化深度学习中的数据预处理过程

在对 CNN 定性模型和 CNN 定量模型进行训练和测试之前,我们都需要对实验质谱数据和肽段图谱库数据进行预处理,其目的是确定每张肽段图谱最相关的实验图谱集合,从而减少无关图谱对后续模型计算的干扰。在本文中,我们利用自定义的图谱匹配和过滤算法来进行预处理操作,其中涉及到匹配峰强度相似性和峰强度之和的计算等。这种预处理方法需要人为的先验知识来设计相应的计算公式,难以保证为所有肽段图谱选出的实验图谱集合都是与其最相关的。因此,在后续工作中,可以考虑对预处理过程进行优化,例如采用神经网络方法对实验质谱进行筛选,并针对不同的肽段图谱动态保留不同数量的实验图谱。

3. 进一步完善 CNN 定量模型

在人类血浆数据集上的定量实验结果表明,我们的 CNN 定量模型可以对 4D-DIA 质谱数据进行准确的定量分析,其定量重复性及 CV 过滤后的肽段得率均与定量准确性较高的 Spectronaut 工具相当。然而,通过比较 CNN 定量模型和 Spectronaut 的肽段定量结果散点图,我们可以看出, CNN 定量模型的肽段定量丰度范围大约为 8 个数量级,而 Spectronaut 则可以达到 12 个数量级左右。在对两者的定量结果进行更细致的比较分析后,我们发现出现上述现象的原因在于, CNN 定量模型在低丰度肽段上的定量值相较于 Spectronaut 偏高,这可能是由于 CNN 定量模型对占比较少低丰度肽段的学习并不充分,从而导致其在低丰度肽段上的定量结果偏差较大。因此,未来还需要在优化数据预处理过程的基础上,对 CNN 定量模型的训练数据和网络结构等方面进行进一步的完善,提高其在低丰度肽段上的定量准确性,进而提升该模型的肽段定量丰度范围。

参 考 文 献

- [1] WILKINS M R, PASQUALI C, APPEL R D, et al. From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis[J]. *Bio/technology*, 1996, 14(1): 61-65.
- [2] ASLAM B, BASIT M, NISAR M A, et al. Proteomics: technologies and their applications[J]. *Journal of chromatographic science*, 2017, 55(2): 182-196.
- [3] SHRUTHI B S, VINODHKUMAR P, et al. Proteomics: A new perspective for cancer[J]. *Advanced biomedical research*, 2016, 5.
- [4] MONTI C, ZILLOCCI M, COLUGNAT I, et al. Proteomics turns functional[J]. *Journal of proteomics*, 2019, 198: 36-44.
- [5] SCHUBERT O T, RÖST H L, COLLINS B C, et al. Quantitative proteomics: challenges and opportunities in basic and applied research[J]. *Nature protocols*, 2017, 12(7): 1289-1294.
- [6] MELBY J A, ROBERTS D S, LARSON E J, et al. Novel strategies to address the challenges in top-down proteomics[J]. *Journal of the American Society for Mass Spectrometry*, 2021, 32(6): 1278-1294.
- [7] GILLET L C, LEITNER A, AEBERSOLD R. Mass spectrometry applied to bottom-up proteomics: entering the high-throughput era for hypothesis testing[J]. *Annual review of analytical chemistry*, 2016, 9: 449-472.
- [8] DUPREE E J, JAYATHIRTHA M, YORKEY H, et al. A critical review of bottom-up proteomics: the good, the bad, and the future of this field[J]. *Proteomes*, 2020, 8(3): 14.
- [9] CRISTOBAL A, MARINO F, POST H, et al. Toward an optimized workflow for middle-down proteomics[J]. *Analytical chemistry*, 2017, 89(6): 3318-3325.
- [10] STEGER M, DEMICHEV V, BACKMAN M, et al. Time-resolved in vivo ubiquitinome profiling by dia-ms reveals usp7 targets on a proteome-wide scale[J]. *Nature communications*, 2021, 12(1): 5399.
- [11] HANSEN F M, TANZER M C, BRÜNING F, et al. Data-independent acquisition method for ubiquitinome analysis reveals regulation of circadian biology[J]. *Nature Communications*, 2021, 12(1): 254.
- [12] AEBERSOLD R, MANN M. Mass spectrometry-based proteomics[J]. *Nature*, 2003, 422(6928): 198-207.
- [13] ZHOU J, YIN Y. Strategies for large-scale targeted metabolomics quantification by liquid chromatography-mass spectrometry[J]. *Analyst*, 2016, 141(23): 6362-6373.
- [14] AEBERSOLD R, MANN M. Mass-spectrometric exploration of proteome structure and func-

- tion[J]. *Nature*, 2016, 537(7620): 347-355.
- [15] HONOUR J W. Development and validation of a quantitative assay based on tandem mass spectrometry[J]. *Annals of clinical biochemistry*, 2011, 48(2): 97-111.
- [16] NEILSON K A, ALI N A, MURALIDHARAN S, et al. Less label, more free: approaches in label-free quantitative mass spectrometry[J]. *Proteomics*, 2011, 11(4): 535-553.
- [17] WU C C, MACCOSS M J. Shotgun proteomics: tools for the analysis of complex biological systems[J]. *Curr Opin Mol Ther*, 2002, 4(3): 242-250.
- [18] DOERR A. Dia mass spectrometry[J]. *Nature methods*, 2015, 12(1): 35-35.
- [19] EGERTSON J D, MACLEAN B, JOHNSON R, et al. Multiplexed peptide analysis using data-independent acquisition and skyline[J]. *Nature protocols*, 2015, 10(6): 887-903.
- [20] VIDOVA V, SPACIL Z. A review on mass spectrometry-based quantitative proteomics: Targeted and data independent acquisition[J]. *Analytica chimica acta*, 2017, 964: 7-23.
- [21] SEGER C, SALZMANN L. After another decade: Lc-ms/ms became routine in clinical diagnostics[J]. *Clinical biochemistry*, 2020, 82: 2-11.
- [22] CHEN C, HOU J, TANNER J J, et al. Bioinformatics methods for mass spectrometry-based proteomics data analysis[J]. *International journal of molecular sciences*, 2020, 21(8): 2873.
- [23] WANG J, TUCHOLSKA M, KNIGHT J D, et al. Msplit-dia: sensitive peptide identification for data-independent acquisition[J]. *Nature methods*, 2015, 12(12): 1106-1108.
- [24] RÖST H L, ROSENBERGER G, NAVARRO P, et al. Openswath enables automated, targeted analysis of data-independent acquisition ms data[J]. *Nature biotechnology*, 2014, 32(3): 219-223.
- [25] BONDARENKO P V, CHELIUS D, SHALER T A. Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography- tandem mass spectrometry[J]. *Analytical chemistry*, 2002, 74(18): 4741-4749.
- [26] BLEIN-NICOLAS M, ZIVY M. Thousand and one ways to quantify and compare protein abundances in label-free bottom-up proteomics[J]. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 2016, 1864(8): 883-895.
- [27] GACHUMI G, PURVES R W, HOPF C, et al. Fast quantification without conventional chromatography, the growing power of mass spectrometry[J]. *Analytical Chemistry*, 2020, 92(13): 8628-8637.
- [28] MEYER J G, NIEMI N M, PAGLIARINI D J, et al. Quantitative shotgun proteome analysis by direct infusion[J]. *Nature methods*, 2020, 17(12): 1222-1228.
- [29] PURVES R W, PRASAD S, BELFORD M, et al. Optimization of a new aerodynamic cylindrical faims device for small molecule analysis[J]. *Journal of The American Society for Mass*

- Spectrometry, 2017, 28(3): 525-538.
- [30] BILBAO A, VARESEO E, LUBAN J, et al. Processing strategies and software solutions for data-independent acquisition in mass spectrometry[J]. *Proteomics*, 2015, 15(5-6): 964-980.
- [31] EWING M A, GLOVER M S, CLEMMER D E. Hybrid ion mobility and mass spectrometry as a separation tool[J]. *Journal of Chromatography A*, 2016, 1439: 3-25.
- [32] MEIER F, BRUNNER A D, FRANK M, et al. diapasef: parallel accumulation–serial fragmentation combined with data-independent acquisition[J]. *Nature methods*, 2020, 17(12): 1229-1236.
- [33] FERNANDEZ-LIMA F, KAPLAN D, PARK M. Note: Integration of trapped ion mobility spectrometry with mass spectrometry[J]. *Review of Scientific Instruments*, 2011, 82(12).
- [34] FERNANDEZ-LIMA F, KAPLAN D A, SUETERING J, et al. Gas-phase separation using a trapped ion mobility spectrometer[J]. *International Journal for Ion Mobility Spectrometry*, 2011, 14: 93-98.
- [35] SHARIFANI K, AMINI M. Machine learning and deep learning: A review of methods and applications[J]. *World Information Technology and Engineering Journal*, 2023, 10(07): 3897-3904.
- [36] SHINDE P P, SHAH S. A review of machine learning and deep learning applications[C]// 2018 Fourth international conference on computing communication control and automation (ICCUBEA). IEEE, 2018: 1-6.
- [37] MEYER J G. Deep learning neural network tools for proteomics[J]. *Cell Reports Methods*, 2021, 1(2).
- [38] GROSSMANN J, ROSCHITZKI B, PANSE C, et al. Implementation and evaluation of relative and absolute quantification in shotgun proteomics with label-free methods[J]. *Journal of proteomics*, 2010, 73(9): 1740-1746.
- [39] BLEIN-NICOLAS M, XU H, DE VIENNE D, et al. Including shared peptides for estimating protein abundances: A significant improvement for quantitative proteomics[J]. *Proteomics*, 2012, 12(18): 2797-2801.
- [40] GERSTER S, KWON T, LUDWIG C, et al. Statistical approach to protein quantification[J]. *Molecular & cellular proteomics*, 2014, 13(2): 666-677.
- [41] GILLET L C, NAVARRO P, TATE S, et al. Targeted data extraction of the ms/ms spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis[J]. *Molecular & Cellular Proteomics*, 2012, 11(6).
- [42] MESSNER C B, DEMICHEV V, BLOOMFIELD N, et al. Ultra-fast proteomics with scanning swath[J]. *Nature biotechnology*, 2021, 39(7): 846-854.
- [43] MUN D G, RENUSE S, SARASWAT M, et al. Pass-dia: a data-independent acquisition ap-

- proach for discovery studies[J]. *Analytical chemistry*, 2020, 92(21): 14466-14475.
- [44] TING Y S, EGERTSON J D, PAYNE S H, et al. Peptide-centric proteome analysis: an alternative strategy for the analysis of tandem mass spectrometry data[J]. *Molecular & Cellular Proteomics*, 2015, 14(9): 2301-2307.
- [45] TING Y S, EGERTSON J D, BOLLINGER J G, et al. Pecan: library-free peptide detection for data-independent acquisition tandem mass spectrometry data[J]. *Nature methods*, 2017, 14(9): 903-908.
- [46] WEISBROD C R, ENG J K, HOOPMANN M R, et al. Accurate peptide fragment mass analysis: multiplexed peptide identification and quantification[J]. *Journal of proteome research*, 2012, 11(3): 1621-1632.
- [47] TSOU C C, AVTONOMOV D, LARSEN B, et al. Dia-umpire: comprehensive computational framework for data-independent acquisition proteomics[J]. *Nature methods*, 2015, 12(3): 258-264.
- [48] LI Y, ZHONG C Q, XU X, et al. Group-dia: analyzing multiple data-independent acquisition mass spectrometry data files[J]. *Nature methods*, 2015, 12(12): 1105-1106.
- [49] ZHANG F, GE W, RUAN G, et al. Data-independent acquisition mass spectrometry-based proteomics and software tools: a glimpse in 2020[J]. *Proteomics*, 2020, 20(17-18): 1900276.
- [50] YANG Y, LIU X, SHEN C, et al. In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics[J]. *Nature communications*, 2020, 11(1): 146.
- [51] BROSCHE M, YU L, HUBBARD T, et al. Accurate and sensitive peptide identification with mascot percolator[J]. *Journal of proteome research*, 2009, 8(6): 3176-3181.
- [52] ENG J K, JAHAN T A, HOOPMANN M R. Comet: an open-source ms/ms sequence database search tool[J]. *Proteomics*, 2013, 13(1): 22-24.
- [53] LAM H, DEUTSCH E W, EDDER J S, et al. Development and validation of a spectral library searching method for peptide identification from ms/ms[J]. *Proteomics*, 2007, 7(5): 655-667.
- [54] BRUDERER R, BERNHARDT O M, GANDHI T, et al. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues*[s][J]. *Molecular & Cellular Proteomics*, 2015, 14(5): 1400-1410.
- [55] PINO L K, SEARLE B C, BOLLINGER J G, et al. The skyline ecosystem: Informatics for quantitative mass spectrometry proteomics[J]. *Mass spectrometry reviews*, 2020, 39(3): 229-244.
- [56] DEMICHEV V, MESSNER C B, VERNARDIS S I, et al. Dia-nn: neural networks and interference correction enable deep proteome coverage in high throughput[J]. *Nature methods*, 2020, 17(1): 41-44.

- [57] DISTLER U, KUHAREV J, NAVARRO P, et al. Label-free quantification in ion mobility-enhanced data-independent acquisition proteomics[J]. *Nature protocols*, 2016, 11(4): 795-812.
- [58] CHOI H, FERMIN D, NESVIZHSHII A I. Significance analysis of spectral count data in label-free shotgun proteomics[J]. *Molecular & cellular proteomics*, 2008, 7(12): 2373-2385.
- [59] GEIS-ASTEGGIANTE L, OSTRAND-ROSENBERG S, FENSELAU C, et al. Evaluation of spectral counting for relative quantitation of proteoforms in top-down proteomics[J]. *Analytical chemistry*, 2016, 88(22): 10900-10907.
- [60] PECKNER R, MYERS S A, JACOME A S V, et al. Specter: linear deconvolution for targeted analysis of data-independent acquisition mass spectrometry proteomics[J]. *Nature methods*, 2018, 15(5): 371-378.
- [61] FANG Y, LI Q R, ZHANG Z H, et al. Figs: Featured ion-guided stoichiometry for data-independent proteomics through dynamic deconvolution[J]. *Journal of Proteome Research*, 2021, 20(8): 4131-4138.
- [62] CRANNEY C W, MEYER J G. Csodiaq software for direct infusion shotgun proteome analysis [J]. *Analytical Chemistry*, 2021, 93(36): 12312-12319.
- [63] NEVEROVA I, VAN EYK J E. Role of chromatographic techniques in proteomic analysis[J]. *Journal of Chromatography B*, 2005, 815(1-2): 51-63.
- [64] GAO M, QI D, ZHANG P, et al. Development of multidimensional liquid chromatography and application in proteomic analysis[J]. *Expert Review of Proteomics*, 2010, 7(5): 665-678.
- [65] MOUSSA A, DERIDDER S, BROECKHOVEN K, et al. Computational fluid dynamics study of the dispersion caused by capillary misconnection in nano-flow liquid chromatography[J]. *Analytical Chemistry*, 2023, 95(37): 13975-13983.
- [66] BEKKER-JENSEN D B, MARTÍNEZ-VAL A, STEIGERWALD S, et al. A compact quadrupole-orbitrap mass spectrometer with faims interface improves proteome coverage in short lc gradients[J]. *Molecular & Cellular Proteomics*, 2020, 19(4): 716-729.
- [67] CHAMBERS M C, MACLEAN B, BURKE R, et al. A cross-platform toolkit for mass spectrometry and proteomics[J]. *Nature biotechnology*, 2012, 30(10): 918-920.
- [68] GRANHOLM V, NOBLE W S, KALL L. On using samples of known protein content to assess the statistical calibration of scores assigned to peptide-spectrum matches in shotgun proteomics [J]. *Journal of proteome research*, 2011, 10(5): 2671-2678.
- [69] ELIAS J E, GYGI S P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry[J]. *Nature methods*, 2007, 4(3): 207-214.
- [70] CHENG C Y, TSAI C F, CHEN Y J, et al. Spectrum-based method to generate good decoy libraries for spectral library searching in peptide identifications[J]. *Journal of proteome*

- research, 2013, 12(5): 2305-2310.
- [71] FENG X D, LI L W, ZHANG J H, et al. Using the entrapment sequence method as a standard to evaluate key steps of proteomics data analysis process[J]. BMC genomics, 2017, 18(2): 1-9.
- [72] LI F, DONG S, LEIER A, et al. Positive-unlabeled learning in bioinformatics and computational biology: a brief review[J]. Briefings in bioinformatics, 2022, 23(1): bbab461.
- [73] KABOUTARI A, BAGHERZADEH J, KHERADMAND F. An evaluation of two-step techniques for positive-unlabeled learning in text classification[J]. Int. J. Comput. Appl. Technol. Res, 2014, 3(9): 592-594.
- [74] DU X, YANG F, MANES N P, et al. Linear discriminant analysis-based estimation of the false discovery rate for phosphopeptide identifications[J]. Journal of proteome research, 2008, 7(6): 2195-2203.
- [75] KIM H, LIM Y. Bootstrap aggregated classification for sparse functional data[J]. Journal of Applied Statistics, 2022, 49(8): 2052-2063.
- [76] CURRAN-EVERETT D. Explorations in statistics: hypothesis tests and p values[J]. Advances in physiology education, 2009, 33(2): 81-86.
- [77] NAVARRO P, KUHAREV J, GILLET L C, et al. A multicenter study benchmarks software tools for label-free proteome quantification[J]. Nature biotechnology, 2016, 34(11): 1130-1136.
- [78] LUDWIG C, GILLET L, ROSENBERGER G, et al. Data-independent acquisition-based swath-ms for quantitative proteomics: a tutorial[J]. Molecular systems biology, 2018, 14(8): e8126.
- [79] MEIER F, BECK S, GRASSL N, et al. Parallel accumulation–serial fragmentation (pasef): multiplying sequencing speed and sensitivity by synchronized scans in a trapped ion mobility device[J]. Journal of proteome research, 2015, 14(12): 5378-5387.
- [80] DONG S, WANG P, ABBAS K. A survey on deep learning and its applications[J]. Computer Science Review, 2021, 40: 100379.
- [81] POUYANFAR S, SADIQ S, YAN Y, et al. A survey on deep learning: Algorithms, techniques, and applications[J]. ACM Computing Surveys (CSUR), 2018, 51(5): 1-36.
- [82] ALZUBAIDI L, ZHANG J, HUMAIDI A J, et al. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions[J]. Journal of big Data, 2021, 8: 1-74.
- [83] WEN B, ZENG W F, LIAO Y, et al. Deep learning in proteomics[J]. Proteomics, 2020, 20(21-22): 1900335.
- [84] GONG Z, ZHONG P, HU W. Diversity in machine learning[J]. Ieee Access, 2019, 7: 64323-64350.
- [85] YU Y, SI X, HU C, et al. A review of recurrent neural networks: Lstm cells and network

- architectures[J]. Neural computation, 2019, 31(7): 1235-1270.
- [86] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
- [87] ALBAWI S, MOHAMMED T A, AL-ZAWI S. Understanding of a convolutional neural network[C]//2017 international conference on engineering and technology (ICET). Ieee, 2017: 1-6.
- [88] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model[J]. IEEE transactions on neural networks, 2008, 20(1): 61-80.
- [89] KINGMA D P, BA J. Adam: A method for stochastic optimization[A]. 2014: 273-297.
- [90] FERDOSI S, STUKALOV A, HASAN M, et al. Enhanced competition at the nano-bio interface enables comprehensive characterization of protein corona dynamics and deep coverage of proteomes[J]. Advanced Materials, 2022, 34(44): 2206008.
- [91] RYAN F J, NORTON T S, MCCAFFERTY C, et al. A systems immunology study comparing innate and adaptive immune responses in adults to covid-19 mrna and adenovirus vectored vaccines[J]. Cell Reports Medicine, 2023, 4(3).
- [92] SZYRWIEL L, GILLE C, MÜLLEDER M, et al. Fast proteomics with dia-pasef and analytical flow-rate chromatography[J]. Proteomics, 2024, 24(1-2): 2300100.