湖南師範大學
HUNAN NORMAL UNIVERSITY

专业学位硕士论文

基于深度学习的蛋白质变体表征重打分算法研究

Research on Rescoring Algorithm for Proteoform Identification Based on Deep Learning

专　业　类　别　085400 电子信息

专　业　领　域　085404 计算机技术

研　究　生　姓　名　杨　琛

导师姓名、职称　钟坚成　教授

实践导师姓名、职称　李大静　讲师

湖南师范大学学位评定委员会办公室

二零二四年 五 月

分 类 号_____ 密级 公开
学校代码 10542 学号 202170293868

基于深度学习的蛋白质变体表征重打分算法研究
Research on Rescoring Algorithm for Proteoform Identification Based on Deep Learning

研 究 生 姓 名 杨 琛
指导教师姓名、职称 钟坚成 教授
学 科 专 业 计算机技术
研 究 方 向 生物信息学

湖南师范大学学位评定委员会办公室

二零二四年五月

摘　要

蛋白质变体是蛋白质经过结构性变异形成的产物，许多人体疾病与特定的蛋白质变体相关。质谱技术是诊断这些疾病的关键方法之一，其通过鉴定蛋白质变体来提供信息。自顶向下的蛋白质组学方法能够提供最全面的蛋白质变体信息，但每条质谱的鉴定都需要对蛋白质序列库进行全面搜索，因此计算量巨大。而且完整大分子蛋白质的表征结果——蛋白质变体-质谱匹配(Proteoform Spectrum Matches, PrSM)在排序准确率等方面存在偏低的问题。

蛋白质变体鉴定常用的错误发现率（False Discovery Rate, FDR）控制方法是通过 Target-decoy 搜索策略来实现。本文提出了一种基于机器学习和深度学习技术的蛋白质变体鉴定后处理方法：PRSMREscore 重打分模型。同时，本文对 PRSMREscore 进行了基于 Target-decoy 搜索策略的优化，研究了相应的优化损失函数。

本文提出了一种基于残差神经网络的 PrSM 重打分算法，并将其命名为 PRSMREscore，其包含特征提取模块和重打分模块两部分。特征提取模块基于集成学习思想应用了机器学习的经典算法，以实现特征的提取，并巧妙地整合了各模型的优势。该模块具备良好的可扩展性，能够并行增加更多的机器学习算法，从而在结构上具备了优秀的灵活性。重打分模块主要采用了深度学习模型残差神经网络，利用残差块的连接机制解决了普通神经网络的退化问题，确保在保留原始特征的情况下进行有效的信息传递。在 47 个测试数据集上的实验结果显示，超过 85%的数据集获得了提升，其中一些数据集鉴定结果数量提升比率达到了倍数级。这些结果充分显示了 PRSMREscore 算法作为蛋白质变体鉴定后处理方法的优势。

本文提出了一种基于 PRSMREscore 算法模型的损失函数优化方法。其中考虑了 Target-decoy 搜索策略中必然存在的 PrSM 结果得分排名和 FDR 控制阈值。将这些指标纳入深度学习模型并作为排名优化理论的限制条件。设计了保证顶部数据的精确度在设定阈值以上再最大化召回率的排名优化损失函数，其精度要求指标使用 Target-decoy 策略的阈值。在这一优化方法中，使用了代理凸函数作为原始问题的下界，以规避原始问题函数离散不连续的缺点，并最终进行凸优化求解。将优化损失函数结合 PRSMREscore 形成了

PRSMREscore+算法。实验结果显示，PRSMREscore+算法相比PRSMREscore有性能优势，证明了该优化方法的有效性。

关键词：蛋白质变体，后处理方法，深度学习，重打分

ABSTRACT

Protein variants are products formed by structural variations of proteins, many of which are associated with specific human diseases. Mass spectrometry technology is a key method for diagnosing these diseases, providing information by identifying protein variants. Top-down proteomics methods can provide the most comprehensive information about protein variants, but each spectrum identification requires a comprehensive search of protein sequence libraries, resulting in massive computational requirements. Additionally, there are issues with the accuracy of ranking complete protein identification results (Protein Spectrum Matches, PrSM).

The commonly used method for controlling false discovery rates (FDR) in protein variant identification is through the Target-decoy search strategy. This paper proposes a post-processing method for protein variant identification based on machine learning and deep learning techniques: the PRSMREscore re-scoring model. Additionally, this paper optimizes PRSMREscore based on the Target-decoy search strategy and studies the corresponding optimization loss functions.

This paper proposes a re-scoring algorithm based on protein variant characterization results, named PRSMREscore, which consists of two parts: a feature extraction module and a re-scoring module. The feature extraction module adopts classical algorithms of machine learning and deep learning to extract features, cleverly integrating the advantages of various models. This module has good scalability, allowing for the parallel addition of more machine learning and deep learning frameworks, thereby possessing excellent flexibility in structure. The re-scoring module primarily employs a deep learning model called residual neural network, which addresses the degradation

problem of ordinary neural networks by utilizing the connection mechanism of residual blocks, ensuring effective information transmission while retaining the original features. Experimental results on 47 test datasets demonstrate that over 85% of the datasets achieved significant improvements, with the highest improvement reaching an order of magnitude. These results fully demonstrate the advantage of the PRSMREscore algorithm as a post-processing method for protein variant identification.

Furthermore, this paper proposes a method for optimizing the loss function based on the PRSMREscore algorithm model. It considers the inevitable ranking of PrSM results and FDR control thresholds in the Target-decoy search strategy. These indicators are incorporated into the deep learning model as constraints for ranking optimization theory. A ranking optimization loss function is designed to ensure that the accuracy of the top data exceeds the set threshold while maximizing recall, with the accuracy requirement indicator using the FDR threshold of the Target-decoy strategy. In this optimization method, a surrogate convex function is used as a lower bound for the original problem to circumvent the discontinuity of the original problem and ultimately perform convex optimization. Combining the optimized loss function with PRSMREscore forms the PRSMREscore+ algorithm. Experimental results demonstrate that on large-scale datasets, this algorithm can achieve significant advantages.

Keywords : Protein Variants, Post-processing Methods, Deep Learnin, Rescore

目　录

第 1 章 绪论

1.1 研究背景和意义

蛋白质是生命的物质基础，是生物体内执行多种生命功能的关键分子。与蛋白质先天遗传变异相比，蛋白质分子结构和状态的变化(即蛋白形式)与人体疾病的病理变化更直接相关。故表征蛋白形式对医学领域具有实际意义。例如，心脏疾病的发展与重要肌丝蛋白的磷酸化变化密切相关[1]。在人体细胞调节复杂特征方面，研究人员已经表明剪接变异体起着关键作用[2]。此外，蛋白质变体的表征有助于进行疾病与蛋白质之间的相关性分析[3]，帮助确定药物开发的潜在靶点[4]，为药物开发提供重要参考。蛋白质变体表征工作涉及识别和定位蛋白形式中的一级结构改变(PSA)，专业术语一般称之为主要结构性变异(Primary Structure Alteration, PSA)。PSA 揭示了蛋白质变体的特点，这是疾病诊断和治疗的关键[5]。目前的变体形式中主要存在几种 PSA。主要分为五类：1)序列突变； 2)固定翻译后修饰(Post-translational Modification, PTM)；3)可变翻译后修饰；4)末端截断；5)未知质量偏移。而且这五类变化并不存在互斥关系和数量限制。所以蛋白基因序列的变体形式变化会受到 PSA 数量、PSA 发生的位置和 PSA 组合位置顺序等单一因素或多因素组合的影响。一个基因组或蛋白基因序列可以产生难以想象多的变体数量。例如，人类组蛋白 H3.1 上的蛋白质变体的理论数量可达到 40 万亿[6-7] 。因此，蛋白质变体鉴定这一难题引起了很多研究者关注。

蛋白质变体表征研究目前有了一定的发展。蛋白质变体表征分析有三种类型的 MS 方法(质谱法)，包括自底向上(Bottom-up, BU)蛋白质组学、中间向下(Middle-down, MD)蛋白质组学和自顶向下(Top-down, TD)蛋白质组学。目前，蛋白质分离技术[8]和高通量技术的快速发展，使研究人员能够以低成本代价从生物样品中获得完整的蛋白质高精度串联质谱数据，这推动了自顶向下 TD 蛋白质组的发展。基于自顶向下的方法通过使用蛋白质分离技术如反相液相色谱(Reversed-phase Liquid Chormatography, RPLC)从生物样品中分离完整的蛋白质[8]，然后通过 LC 串联 MS(LC-MS/MS)分析完整的蛋白质来鉴定蛋白质形式。可以在完美的条件下获得蛋白质的氨基酸序列以及 PTM 的完整信息，而无需对样品进行酶前消化，这为蛋白质形式的鉴定提供了先验知识[9]。

PrSM(Proteoform Spectrum Match) 蛋白质变体-质谱匹配[10]是质谱蛋白质组学中常用的术语，用于描述质谱实验中质谱数据与特定肽段的匹配结果。质谱仪器会产生大量的质谱数据，包括肽段的质荷比(m/z)值和对应的信号强度等

信息。PrSM 即表示质谱数据中的特定肽段与已知的蛋白质序列进行匹配的结果。而确定这个匹配关系的工作通常由鉴定算法来完成，自底向上蛋白质组学和自底向上蛋白质组学[11-12]在各种实验算法结尾通常以输出 PrSM 来承载鉴定信息。PrSM 能够帮助研究人员鉴定蛋白质组中的蛋白质、确定蛋白质的变体、探索蛋白质的功能等，为生物学研究提供重要的数据支持。所以 PrSM 的准确性和可靠性对于质谱蛋白质组学研究至关重要。

近年来，深度学习技术的快速发展为蛋白质变体研究提供了新的机遇[13]。深度学习技术作为一种强大的机器学习方法，能够从大规模生物信息数据中学习蛋白质的特征表示，实现对复杂生物学问题的高效预测和分析。如叶等人[14]开发了一种测试时训练范式，该范式调整预训练模型以生成特定于实验数据的模型帮助在自底向上的蛋白质组学中肽鉴定。李等人[15]提出了一种用于肽的从头测序的结合了卷积神经网络和递归神经网络的深度学习模型，学习串联质谱、片段离子和肽序列模式的特征，提高了氨基酸识别准确率。Carter 等人[16]使用了基于 m/z、同位素强度和碎片模式等特征的机器学习进行 CCS 预测从而能识别秀丽隐杆线虫突变株的未知代谢产物。通过机器学习技术，研究人员可以更加全面地理解蛋白质变体的结构和功能，揭示其在生物学和疾病机制中的作用。传统的特征提取和模式识别方法通常依赖于人工设计的特征和规则，缺乏对复杂数据中潜在模式的全面把握。而深度学习技术具有适应性强、泛化能力强、处理复杂数据的能力强等特点，能够克服传统方法的局限性，为蛋白质变体研究提供新的思路和方法。

1.2 国内外研究历史以及现状

除了生物湿实验外，基于质谱法的蛋白质变体形态分析目前的 MS 方法主要有三种：自底向上(Bottom-up, BU)蛋白质组学、中间向下(Middle-down, MD)蛋白质组学和自顶向下(Top-down, TD)蛋白质组学。在十年前，BU 霰弹枪蛋白质组学被广泛应用于蛋白质鉴定研究[9]。随着实验设备和科技进步，由于 TD 蛋白质组学能够获得完整的蛋白质信息这一独特优势，目前自顶向下面向质谱技术来鉴定蛋白质已成为蛋白质组学研究中的重点[17]。

1.2.1 自底向上蛋白质变体鉴定算法

大多数自底向上蛋白质组学分析使用蛋白酶将蛋白质消化成具有可预测末端的肽[19]。然后在 MS/MS 仪器中分析肽，并使用它们的质荷比和预测序列来推断关于样品中蛋白质的信息。其是从分析完整蛋白质消化物中的肽开始，并使用蛋白质数据库来表征该肽来源的开放阅读框架。

自底向上的蛋白质组学利用了肽相对于蛋白质的优势：肽较为容易通过反相液相色谱法[18]分离，并产生良好的离子信号[20]，并以更可预测的方式进行片段化。

这是一种稳定健壮的方法，能够进行高通量分析，从复杂的裂解物中鉴定和定量数千种蛋白质[21]。如今，使用数据依赖采集(Data Dependent Acquisition, DDA)工作流程的自底向上的方法是蛋白质组学的核心技术。这些简单的工作流程也被称为枪式蛋白质组学，产生了大量的蛋白质鉴定列表，并用于解决当今大多数可用的、复杂的、完整的蛋白质组，包括人类蛋白质组的初稿[22-23]。

自底向上蛋白质组学的标志是蛋白酶消化的广泛使用，这有其缺点。胰蛋白酶是喷枪方法的黄金标准，用于全球蛋白质组机器数据库中约 96%的沉积数据集[24]。胰蛋白酶是一种非常有效的蛋白酶，具有高催化活性，在 C 末端产生具有碱性精氨酸或赖氨酸的肽，这是碰撞诱导解离(Collision Induced Dissociation, CID)串联质谱分析的理想选择[25]。尽管有效，但当使用胰蛋白酶时，56%的生成肽≤6 个残基，因此太小，无法通过 MS 鉴定[26]。从小肽推断出的有限序列信息通常足以分配蛋白质簇，但不总是足以鉴定蛋白形式。尽管可以通过测量质量变化来进行，但在自底向上的蛋白质组学中，在没有先验知识的情况下进行蛋白质异构体和 PTM 鉴定是极其有限的。

1.2.2 自顶向下蛋白质变体鉴定算法

在十年前，用于分析蛋白质形态的 TD 蛋白质组学在蛋白质样品制备[27-28]、液相色谱分离和数据分析等方面存在许多困难和挑战。故 BU 蛋白质组学被广泛应用于蛋白质变体鉴定研究。随着蛋白质分离技术和高通量技术的快速发展，研究人员能够以低成本的代价从生物样品中获得完整蛋白质的高精度串联质谱数据，这促进了 TD 蛋白质组学的发展。

自顶向下蛋白质变体鉴定算法是一种用于鉴定整个蛋白质的方法，其基本原理是将完整的蛋白质分析为一系列的碎片，然后通过质谱技术来鉴定这些碎片的序列和修饰状态。首先，从生物样品中提取出目标蛋白质，并对其进行化学或生物学处理，以得到所需的样品。将处理后的蛋白质样品送入质谱仪进行测量，生成原始质谱数据。这些数据通常以质谱图的形式呈现，其中包含了质子/离子信号的质荷比和强度信息。其次，对原始质谱数据进行解卷积处理，以将质谱信号从离子混叠中分离出来，得到更准确的质谱峰。再次，凭借从解卷积后的质谱数据中提取的特征，例如质荷比、峰面积、峰高度等进行鉴定搜索。将提取的特征与已知的蛋白质数据库进行匹配，以确定质谱峰对应的蛋白质序列和修饰状态。根据匹配结果，对蛋白质序列和修饰状态进行分析和验证，确定蛋白质的变体及其相关信息。最后，对鉴定结果进行评估和验证，包括误差分析、置信度评估等，以确保鉴定结果的准确性和可靠性。

自顶向下蛋白质变体鉴定算法通常涉及复杂的数据处理和分析过程，需要结合质谱技术、生物信息学和统计学等多个领域的知识和方法。它可以提供更全面、

更准确的蛋白质变体信息，对于理解生物学功能和疾病机制具有重要意义。

1.2.3 结合机器学习技术和深度学习技术

近年来，机器学习技术和深度学习技术在蛋白质变体鉴定研究领域发挥了重要作用，推动了该领域的快速发展。越来越多的研究论文涉及到了这些技术的应用。例如，刘等人[29]提出了一种基于卷积神经网络(CNN)[30]的方法，通过该方法能够在解卷积过程中更准确地识别并提取蛋白质质谱数据中的包络信息，从而为蛋白质变体鉴定提供了更为可靠的工具和方法。在自底向上蛋白质变体鉴定领域，Sven 等人[31]提出了一个将机器学习与肽鉴定完全融合的搜索引擎，通过这一方法能够提高了开放搜索的 PrSM 置信度。此外，Kevin L. Yang 等人[32]提出了基于深度学习的肽特性预测，如 LC 保留时间、离子强度和 MS/MS 光谱，以及附加功能重新评分肽与谱的匹配。这些研究成果不仅拓展了本文对蛋白质变体鉴定技术的理解，还为未来的研究和应用提供了重要的参考和借鉴。

1.3 本文研究内容

蛋白质变体鉴定是一个极其复杂问题， 一条蛋白质序列经过几种 PSA 组合变化，再加上 PSA 发生的序列位置选点，理论上能产生万亿级别数量的蛋白质变体。在蛋白质组学中，使用 MS 质谱法鉴定和定量肽、蛋白质和蛋白变体是一种主要方法。在十年以前自底向上的蛋白质变体鉴定算法应用的较为广泛。但随着当今技术发展，质谱数据分辨率等高速进步。自底向上蛋白质变体鉴定技术因有以低成本从生物样品中获得完整蛋白质的高精度串联质谱数据的优点获得了很大的发展。在该领域已经涌现出多种不同的算法，以满足对蛋白质变体鉴定和分析的需求。然而，在蛋白质变体鉴定这一领域，仍然存在一些挑战，那就是最终鉴定结果的准确性问题。PrSM 是蛋白质变体鉴定最终的匹配结果。鉴定算法按照内置的打分规则给 PrSM 打分。于是所有鉴定结果即 PrSM 集合可以按照得分进行排序然后报告高可信度 PrSM。然而，在最终的排序结果中，必定存在大量假阳性 PrSM 得到了高评分而影响了最终结果。基于此，本文利用经典机器学习和深度学习技术，提出一种新颖的重打分模型，以解决当前在自顶向下蛋白质变体鉴定领域中存在的鉴定结果 PrSM 排序的准确性问题。通过该模型，本文旨在改善蛋白质变体鉴定的准确性和效率，为蛋白质组学研究提供新的解决方案。具体内容如下

1）自底向上蛋白质变体鉴定后处理方法研究。本文综合运用了经典机器学习和深度学习技术，提出了一种名为 PRSMREscore 的重打分算法。该算法以热门鉴定算法 TopPIC 为基线，首先，对其输入和输出文件进行解析，并提取其中的关键信息作为输入数据。随后，将机器学习模型如 Logic Regression[33]、XGBoost[34]

和 SVM[35]等作为弱分类器进行特征提取，进一步加工原始特征数据，以获取更丰富的特征信息。最终，利用所提取的特征信息使用深度学习技术进行重打分预测，对已有的蛋白质变体鉴定结果进行进一步优化和重评分。PRSMREscore 凭借机器学习方法能自动从大型复杂数据中学习特征和模式，无需手工进行特征研究。克服传统算法未能全面挖掘和利用数据特征信息的局限性而得到比原本基线模型更多的 PrSM。

2）针对自顶向下蛋白质变体鉴定后处理方法中深度学习模型的打分排名优化 loss 函数研究。本文总结了关于排名优化目标函数的理论。并根据蛋白质变体鉴定中 FDR 控制方法和 PRSMREscore 算法模型的需要，提出了以优化排名为目标的 loss 函数并提供完整的数学证明过程。该 loss 函数旨在继续优化 PRSMREscore 重打分算法。其主要目的是提高在 Target-decoy 搜索策略下，排名在阈值以上的命中 target 的 PrSM 得分 rank 排名。本文将结合该 loss 函数的 PRSMREscore 算法记为 PRSMREscore+。最终 TopPIC 基线模型使用 PRSMREscore+算法重打分后能输出更多的 PrSM，进一步挖掘蛋白质变体鉴定研究工作中的关键生物信息。

1.4 本文组织结构

本文围绕自顶向下蛋白质变体鉴定算法的后处理过程进行研究，具体结构如下：

第一章是绪论。首先，本文从蛋白质变体鉴定的研究背景和意义开始入手，详细地讲述了当前有蛋白质变体鉴定的国内外研究现状，并分别介绍了自底向上和自顶向下的蛋白质变体鉴定方法。然后，介绍了一些目前已有的使用机器学习和深度学习帮助蛋白质变体鉴定研究的技术。最后，介绍了本文的研究内容和具体章节安排。

第二章为介绍相关鉴定算法背景和实验数据基础。首先，介绍整个自顶向下蛋白质变体鉴定质谱法的整体流程。然后，着重介绍与本文中所提出算法密切相关的 Target-decoy 搜索策略。最后，介绍了领域内已经提出的一些结合机器学习和深度学习技术的相关研究。

第三章介绍了基于经典机器学习和深度学习的 PrSM 重打分算法。详细介绍了模型的输入特征，并剖析了 PRSMREscore 模型的详细模型结构。介绍了实验环境和数据集来源及处理。最后在不同跨物种的数据集上，设置了大量的对比实验，证实了所提算法的有效性，并对本章工作进行了总结。

第四章介绍了基于 PrSM 重打分算法 PRSMREscore 的排名优化损失函数的理论和研究。主要内容包括该算法的构建条件、函数推理过程和最终的对比试验

结果。实验结果主要内容是对比分析使用了排名优化损失函数的 PRSMREscore+ 算法和未使用优化损失函数 PRSMREsocre 算法及基线模型的差异。最后对本章工作进行了总结。

第 2 章 蛋白质变体鉴定研究及相关理论

2.1 蛋白质变体鉴定和重打分过程

蛋白质变体鉴定的基本流程包括以下步骤：首先，将生物样品经质谱仪处理，产生原始质谱数据文件。这些文件记录了样品中蛋白质的质谱信息。然后，对原始数据文件进行解卷积处理[37]，这一步骤旨在去除质谱数据中的复杂峰和重叠峰，以提高后续分析的准确性。解卷积后，得到的文件包含了更清晰、更易分析的质谱信息。最后，将解卷积后的文件作为输入，运行蛋白质变体鉴定算法软件进行表征。这些算法会分析质谱数据，并尝试将其与已知的蛋白质序列或变体匹配，最终以确定样品中存在的蛋白质及其变体的类型和特征。如图 2-1 是鉴定算法 TopPIC[10]的整体输入文件的变化过程。点 raw 是质谱仪产生的原始二进制文件，经过软件 TopFD[38]后解卷积，生成了二级质谱[36]的点 msalign 格式文件。最后经过鉴定软件 TopPIC 得到最终的鉴定结果，鉴定的输出结果通过 xml 格式显示。

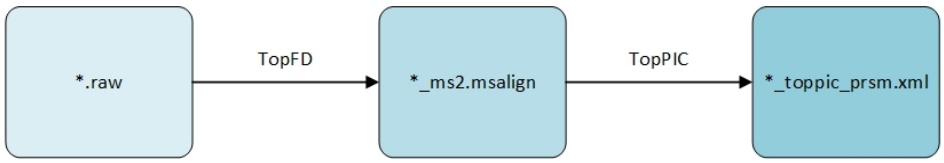


图 2-1 TopPIC 鉴定过程中文件格式变化

重打分过程属于蛋白质变体鉴定流程里面的后处理方法。后处理方法是指在使用某种算法或模型进行数据处理或预测之后，对结果进行进一步处理或修正的一系列技术或策略，且独立于原本的鉴定流程。图 2-2 是重打分算法的处理过程。

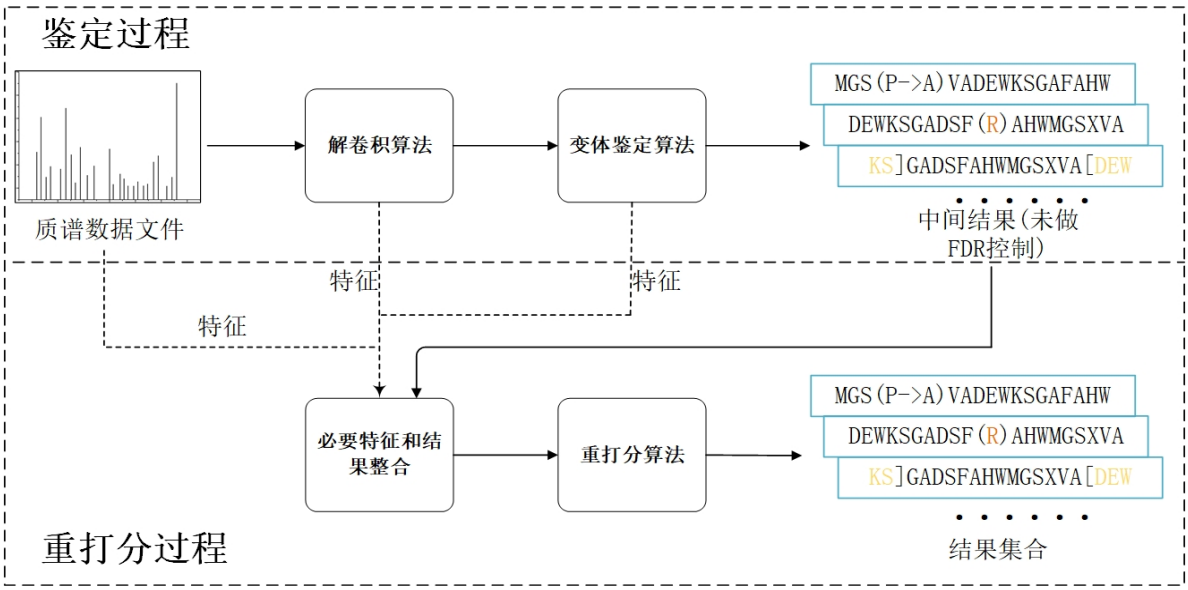


图 2-2 重打分算法的处理过程。

本文将在整个鉴定过程中出现的对重打分有关联影响的特征信息解析提取出来。输入进设计好的重打分算法重新预测得分，经过同样 FDR 控制后得到了更为准确的结果集合。

以下为蛋白质鉴定过程的详细介绍：

1)生物样品经质谱仪产生原始质谱数据：生物蛋白样品经过适当的制备处理后，例如消化、纯化等，获得了可用于质谱分析的样品。这些样品进入质谱仪，经过质谱仪处理产生的是原始数据文件。原始的质谱数据文件，记录了样品离子的信号强度和质量/电荷比信息，以及相关的时间和参数信息。这些文件是后续数据处理和分析的基础。这些数据通常以原始数据文件(如.raw、.mzML 等格式)的形式保存在质谱仪的计算机或外部存储设备中。生成的原始质谱数据文件可以通过相应的数据处理软件进行处理和分析，以提取有用的信息并生成质谱图谱。这些软件可以执行去噪声、峰识别、质量校准、质谱匹配等操作，从而得到更加准确和可靠的结果。

2)质谱数据的解卷积：质谱解卷积是一种数据处理技术，用于恢复质谱图谱中隐藏的原始信号。在质谱实验中，由于仪器和环境等因素的影响，质谱数据往往会受到噪声、畸变或者模糊的影响，导致实验结果的失真或者模糊。而且质谱数据中也会存在一些重叠峰需要辨别。而解卷积的目的便是解决这些问题。其通过一些处理方式将经过变换或者模糊处理的质谱数据还原成原始的信号，以便后续的分析和解释。解卷积技术可以应用于不同类型的质谱数据，包括质子质谱(MS)、核磁共振质谱(NMR)等。在 MS 质谱中，解卷积通常用于提高峰的分辨率、识别峰的形状和区分峰与峰的重叠，以便更准确地识别和定量目标分子。在核磁共振质谱中，解卷积可以帮助恢复被噪声遮盖的信号，提高谱图的清晰度和可读性。解卷积方法通常基于数学模型和信号处理算法，包括傅立叶变换[39]、小波变换[40]、最小二乘法[41]等。这些方法可以通过分析质谱数据的频谱特征和噪声分布，来推导出原始信号的估计值，并进行反演操作，从而还原出质谱图谱中隐藏的信息。

3)使用解卷积后的文件进行蛋白质变体鉴定：蛋白质变体鉴定或蛋白质变体表征是指确定样品中存在的蛋白质变体的过程。在蛋白质组学研究中，常常会发现同一种蛋白质在不同生理或病理状态下可能存在多种变体[42-43]，这些变体可能是由于基因突变、剪切异构体、后转录翻译修饰等导致的。蛋白质变体鉴定和表征旨在确定样品中存在的蛋白质变体的类型、结构和数量，通常通过质谱分析等技术来实现。当前蛋白质变体鉴定方法主要是基于两种不同的蛋白质组学方式来进行研究：自底向上(Bottom-up, BU)和自顶向下(Top-down, TD)。

2.2 自顶向下蛋白质变体鉴定算法

基于 Top-down(TD)质谱技术在蛋白质鉴定面临的主要挑战在于谱图比对的精准性，以及质谱与蛋白质序列的匹配过程中所引发的耗时问题。由于完整蛋白质中存在五种主要结构性变异类型，每种变异类型及修饰位点的多样性，进而导致一个蛋白质可能对应多种蛋白质变体的组合，造成了鉴定难度的爆炸性增长。本节旨在介绍针对 TD 质谱技术的蛋白质变体鉴定方法。目前针对 TD 质谱技术的蛋白质变体鉴定方法可大致分为两类：基于衍生蛋白质数据库的方法和基于 PSA 盲搜的方法。

衍生蛋白质数据库是基于蛋白质数据库扩展和衍生而来的，利用 TD 质谱数据进行搜索。该方法从生物数据库中提取结合常见的特定修饰信息，生成包括序列变体库和翻译后修饰的蛋白质变体库。由于衍生蛋白质变体数据库中含有正确的蛋白质变体序列，因此能够加速蛋白质变体鉴定过程，将二级质谱与扩展的蛋白质变体序列进行匹配。然而，随着蛋白质复杂性的增加，衍生蛋白质变体数据库中的蛋白质变体数量呈指数增长，导致内存空间无法有效控制。为了解决这一问题，常见的方法是通过排除一些不常见的蛋白质变体来减少数据库大小，以控制内存空间。然而，这种方法虽然能排除不常见的蛋白质变体，但也可能将一些正确的蛋白质变体排除在外。尽管存在这些局限性，由于其速度优势，目前仍广泛应用。包括 ProSight[44]、MascotTD[45]、BUPIDTop-Down[46]、ProteinGoggle[47]、MetaMorpheus[48]、TDPortal[49]等算法都采用了该方法。

复杂的蛋白质变体具有多种类型和数量的 PTM(翻译后修饰)，因此扩展蛋白质变体数据库的规模不断增大，导致需要更多的存储空间。为了解决这些问题，发展出了 PSA 盲搜方法。与扩展蛋白质变体数据库方法不同，PSA 盲搜方法不需要构建蛋白质变体样本库，而是直接比较实验质谱与理论谱(由蛋白质参考序列构建)的相似性，从而找出与实验质谱相匹配的序列。PSA 盲搜方法不仅可以减少存储空间的需求，还能提高蛋白质变体鉴定算法的性能。目前，PSA 盲搜方法已被广泛应用于各种蛋白质变体鉴定工具中，例如 MS-TopDown[50]、MSAlign+[51]、MS-Align-E[52]、MASH Suite Pro[53]、TopPIC、SPECTRUM[54]、pTop[55]、TopMG[56]、MSPathFinder[57]、Twister[58]、HomMTM[59]、PIITA[60]等。不同的 PSA 盲搜方法具有各自的优势，能够发现不同类型的 PTM。例如，TopPIC 能够识别具有末端截断和未知 PSA 的蛋白质变体，而 MSPathFinder 则能够识别具有可变 PTM 和末端截断的蛋白质变体。而 TopMG 则是功能最全面的方法，支持所有 PSA 类型的蛋白质变体识别，但分析时间大幅度加长，不适用大量数据的快速分析。尽管 PSA 盲搜类方法能够考虑所有可能的翻译后修饰，但随着 PTM 类型数量的增加，搜索空间变得非常庞大，搜索过程也变得非常耗时[61]。

2.3 Target-decoy 策略

目标-诱饵(Target-decoy)搜索策略是一种常用于质谱数据分析中的方法[62]，用于评估鉴定结果的可信度。

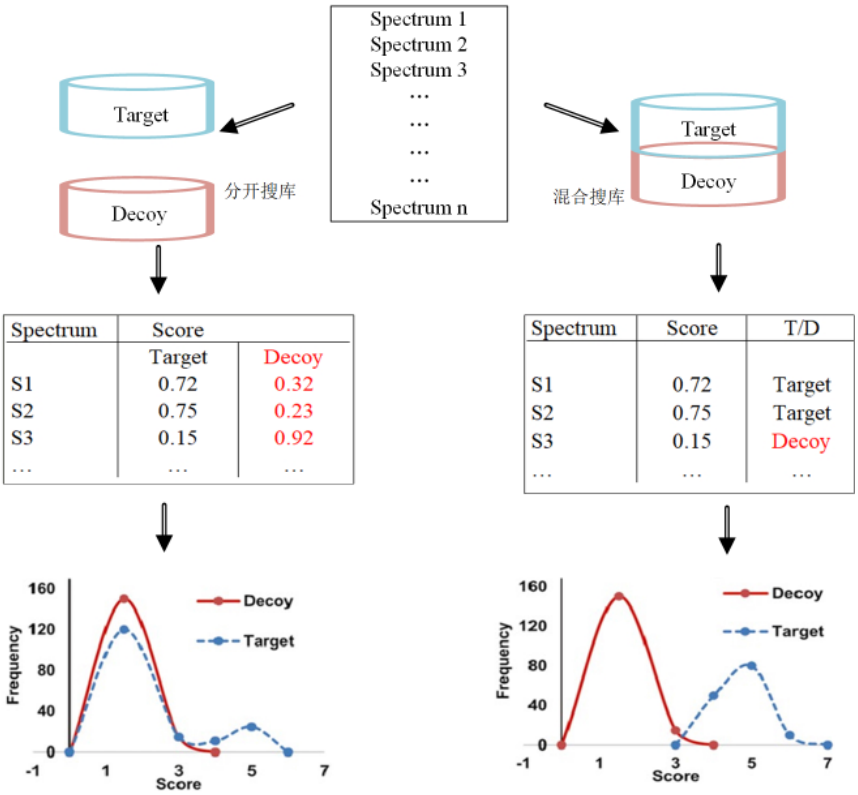


图 2-3 两种搜库方式

该策略的基本原理是通过使用真实蛋白质序列创造和伪造“诱饵”序列进行质谱匹配搜索，以模拟真实和假阳性鉴定之间的差异。搜库方式也分为两种，将诱饵序列和目标序列整合搜库，和分开进行搜库。前者是将诱饵数据库和目标数据库进行混合搜索。这相当于扩大了鉴定算法的匹配空间，导致鉴定算法因为匹配的数量级增大而增长运行时间[63]。而后者是分开搜库，分开搜库时所有高于打分阈值的匹配目标序列库和诱饵序列库的 PrSM 都被用来计算最终的错误发现率(FDR)。图 2-3 显示了两者的区别。

混合搜库允许目标序列库的 PrSM 和诱饵序列库的 PrSM 相互竞争，从而淘汰掉部分诱饵序列库的 PrSM，因此有研究人员推荐使用混合搜库策略[64]。分开搜库或混合搜库的优劣并没有最终结论。从用户的角度来讲，两者都有各自的优势，一般可以按照鉴定算法的推荐选择适宜的搜库方式。如鉴定算法 TopPIC 是使用混合搜库，以此为例，这类搜库一般会经过下面的 5 个步骤：

1)准备目标和诱饵数据库：首先，需要准备一个包含真实蛋白质序列的目标数据库和一个包含伪造蛋白质序列的诱饵数据库。创建诱饵序列的方法有很多。诱饵数据库中的序列通常是通过对目标数据库中的序列进行随机化或以其他方式

生成的，大致可分为反转、混编和随机三种类型，确保其与真实序列具有相似的长度和氨基酸组成，但不包含在样品中。将目标数据库和诱饵数据库合并成一个单一的综合数据库。

2)质谱搜索和匹配：使用质谱数据对综合数据库进行搜索，以鉴定质谱中存在的蛋白质序列。搜索算法将质谱数据与数据库中的所有序列进行比对，寻找与质谱峰对应的蛋白质。这一步一般由鉴定算法执行。

3)鉴定结果排名和筛选：通常情况下，鉴定算法会根据质谱数据与序列的匹配程度给出一个得分或置信度，然后按照得分或置信度将鉴定结果进行排名。在排名过程中，目标序列(真实的待鉴定生物蛋白质)和诱饵序列(假序列)的鉴定结果会被排列在一起，并根据其得分或置信度进行排序，使得得分高的序列排在前面。

4)FDR 计算：在排名的基础上，通过计算错误发现率来评估鉴定结果的可靠性。通常，FDR 是按照排名计算的，即在一系列不同排名阈值下计算 FDR。在结果排名之中选择一条 PrSM，将排名在此之前匹配目标序列和诱饵序列的 PrSM 数量分别进行统计，然后使用这两个数据计算在该阈值下的 FDR。FDR 计算的公式为：

$$fdr_{prsm(i)} = \frac{decoy_i}{target_i} \tag{2.1}$$

其中$decoy_i$代表排名在 prsm(i)之前的所有匹配 decoy 序列的 PrSM 数量之和。$target_i$则代表排名在 prsm(i)之前的所有匹配 target 序列的 PrSM 数量之和。

5)根据阈值报告结果：根据实验需求和性能要求，可以调整排名阈值控制报告出的 PrSM 数量。选择适当的阈值后，按照 Target-decoy 策略会根据排名顺序将在 FDR 小于阈值的所有匹配目标序列的结果报告出来。通常情况下，会将结果排名中的靠前部分报告出来。

目标-诱饵搜索策略的主要优势在于能够估计鉴定结果的可靠性，提供对假阳性鉴定的控制并且它几乎可以适用于所有的搜库引擎和鉴定方法。通过模拟诱饵序列与真实序列之间的差异，该方法能够更准确地评估鉴定结果，并减少假阳性的数量。因此，目标-诱饵搜索策略在质谱数据分析中得到了广泛的应用，尤其是在蛋白质鉴定和定量方面。

2.4 相关深度学习模型

残差网络(Residual Network, ResNet)是一种深度神经网络结构，旨在解决深度神经网络训练过程中的梯度消失和梯度爆炸等问题。它于 2015 年由何凯明[65]等人提出，并在 ImageNet 图像分类挑战赛中取得了令人瞩目的成绩。ResNet 引入了残差连接，使得可以训练更深的网络结构，并在图像分类、目标检测、语

义分割等领域取得了显著的成果。其核心思想是通过在网络中添加跨层的直接连接来传递捷径或跳跃连接，使网络能够更有效地学习到残差信息。这种残差连接使得网络可以更轻松地学习到恒等映射或近似的恒等映射。ResNet 的基本单元是残差块，其中包含两个主要分支：主路径用于学习特征表示，而残差路径则直接将输入信息传递给输出，或通过一些变换后传递给输出。残差块通常采用卷积层、批归一化层和激活函数等组件构成，以提高网络的表征能力。在 ResNet 中，残差块可以堆叠在一起形成深层网络，同时由于残差连接的引入，网络的训练变得更加稳定和容易。此外，ResNet 还采用了全局平均池化层和 softmax 分类器作为最后的输出层，用于进行图像分类等任务。

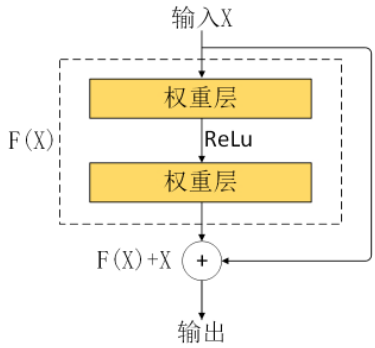


图 2-4 残差块示意图

ResNeXt[66]是微软亚洲研究院提出的一种深度卷积神经网络，它是对 ResNet 的扩展。

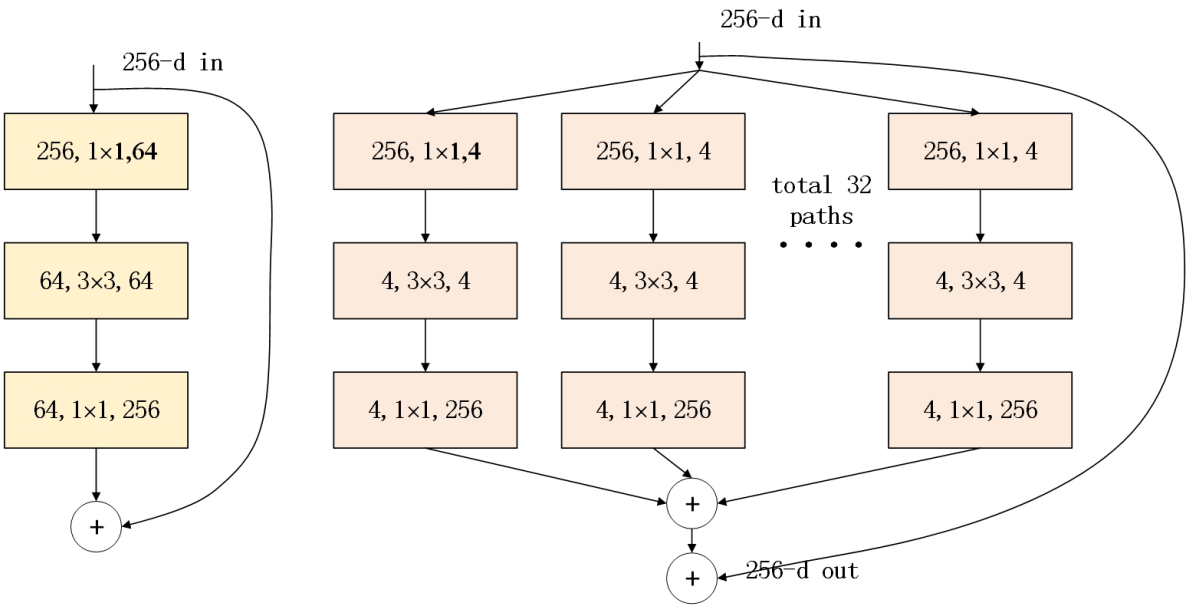


图 2-5 ResNet(左)和 ResNeXt(右)结构示例

ResNeXt 通过并行的方式构建更深的神经网络，在各个分支中采用残差连接来解决梯度消失和梯度爆炸等问题，使得可以训练出更深的网络。ResNeXt 引入了“cardinality”的概念，表示每个分支的通道数，例如 ResNeXt-50 含有 50 个

分支，每个分支通道数为 4。这种设计使得网络更加灵活，适应不同的输入数据，并且用较少的参数实现更高性能。ResNeXt 的核心是“split-transform-merge”(STM)单元，包含多个子网络，每个子网络具有相同的结构和参数，但输入不同的子集。STM 单元是一个高效的多分支卷积模块，能够学习不同的特征表示，并且也能通过残差连接解决梯度问题。

总的来说，ResNeXt 相比 ResNet 具有更好的泛化性能和更高的训练速度。

2.5 本章小结

本章全面介绍了蛋白质变体鉴定领域的研究现状和主流方法。首先，本文详细介绍了蛋白质变体鉴定的整体工作流程，包括自顶向下和自底向上两种主要方法的基本原理和步骤。然后，重点讨论目前自顶向下蛋白质变体鉴定算法常用的两种搜索匹配方式衍生蛋白质数据库鉴定和 PSA 盲搜方法。本章还介绍了蛋白质变体领域内控制鉴定结果假阳性的 Target-decoy 策略，详细介绍并说明该策略通过在鉴定过程中引入诱饵序列来评估鉴定结果的可靠性。最后介绍了相关深度学习技术框架，详细介绍了其基本原理和优点。

尽管一些现有算法在蛋白质变体鉴定方面取得了一定进展，但 Target-decoy 策略的使用，一定会使得鉴定算法混淆部分得分不高的真实 PrSM 和得分过高的匹配“诱饵”序列的 PrSM。故可以使用重打分算法等后处理手段解决该问题。利用机器学习和深度学习技术进行领域内后处理算法的研究则具有广阔的发展前景和潜力。

第 3 章 基于残差神经网络的 PrSM 重打分算法

3.1 引言

目前，蛋白质质谱分析领域普遍采用 Target-decoy 搜索策略来提高鉴定结果的可靠性。其中，混合搜库策略允许目标序列库和诱饵序列库的 PRSM 相互竞争，以淘汰部分诱饵序列库的 PRSM。然而，研究表明，这种竞争机制可能导致高打分值的目标序列库的 PRSM 被更高打分值的诱饵序列库的 PRSM 竞争掉，从而增加假阳性率。因此，选择分开搜库还是混合搜库策略并无定论，而可根据具体情况选择合适的搜库方式。同时，随着机器学习技术的发展，可以利用机器学习技术弥补搜索策略的不足。本章介绍了一种基于深度学习模型的重打分模型 PRSMREscore，作为鉴定算法的后处理过程。PRSMREscore 模型涵盖了对鉴定结果的筛选、重打分、结果整合和统计、假阳性控制以及结果可视化等多个方面，使用并扩充了鉴定算法的原始输入和输出信息，最终给鉴定结果进行重打分，发掘出因为 Target-decoy 搜索策略而遗漏的 PrSM 结果。这有助于提高鉴定结果的准确性和可靠性，为蛋白质质谱分析领域的发展提供重要支持。

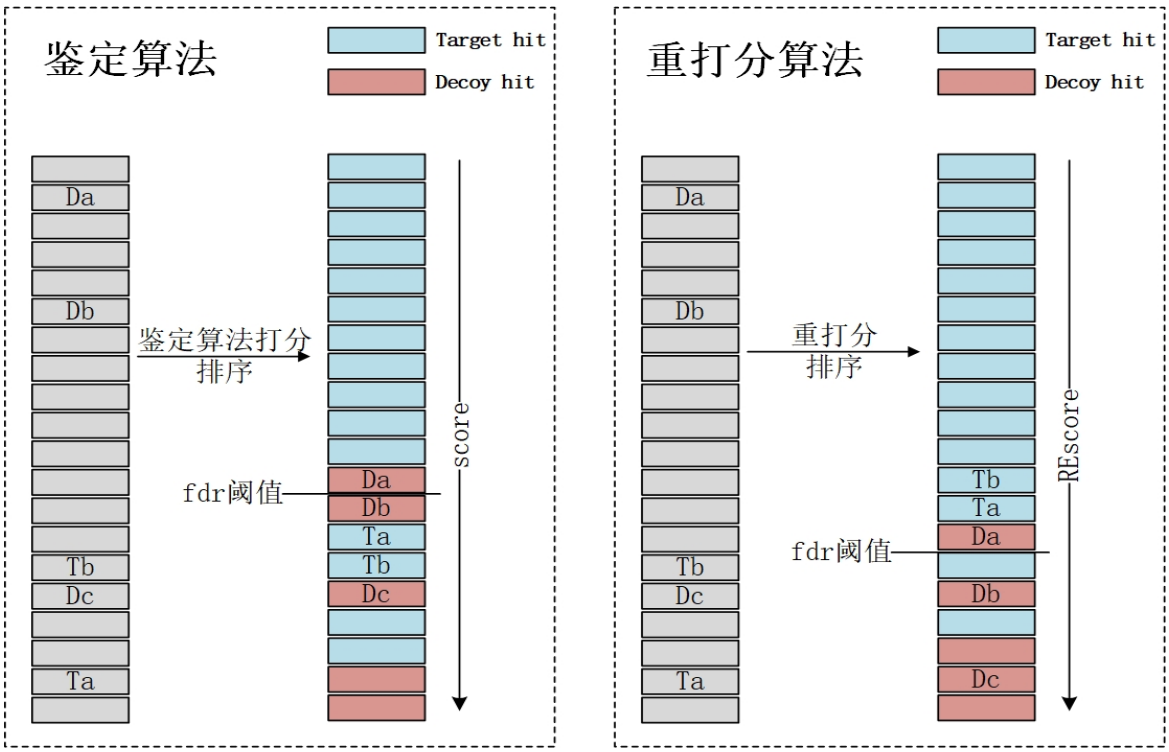


图 3-1 重打分效果:图中可见由于重打分提高了 PrSM 的得分准确性，捕捞了正确命中 target 的 PrSM(Ta,Tb)

3.2 基于残差神经网络的 PrSM 重打分方法

经过鉴定算法以后，会得到一个 PrSM 结果集合。本文设定这个集合为 C_x={ (s_1, n_1) ,(s_2, n_2) , (s_3, n_3) , (s_4, n_4), ...,(s_n, n_n) }。其中s_i表示实验获得的二级谱图， n_i表示蛋白质数据库中的一条蛋白质序列，(s_i, n_i)表示集合 C 中的第 i 条 PrSM。通常，集合 C 还会伴随着一个的得分向量 X= $(x_1, x_2, x_3, \dots, x_n)^T$。其中表示第 i 个 PrSM 的得分，这个一般由鉴定算法确定给出，如 TopPIC 的 E-value、P-value。鉴定结果的重打分目的在于为所有结果进行重打分，得到一个所有结果的新得分向量 Y= $(y_1, y_2, y_3, \dots, y_{n+m})^T$。其中 m 代表鉴定算法报告出来但被 Target-decoy 策略所遗弃的所有结果数量。经过重新打分后，得到的新集合C_y在包含了C_x集合的同时还能扩充更多 PrSM 进去。

下面将具体描述本章节中提出的基于机器学习技术和深度学习技术的算法 PRSMREscore，该算法的结构如图 3-1 所示，除了输入和结果以外，他的主体部分由两部分组成：(1)提取特征模块，充分发挥各种算法的打分能力，进一步从原始输入数据中提取初步的特征和模式。(2)残差神经网络模块，该模块针对前一模块得到不同的特征，融合组合代表 PrSM 的新特征。这样既能引入第前一模块预测的信息，同时保留原始数据的特征。

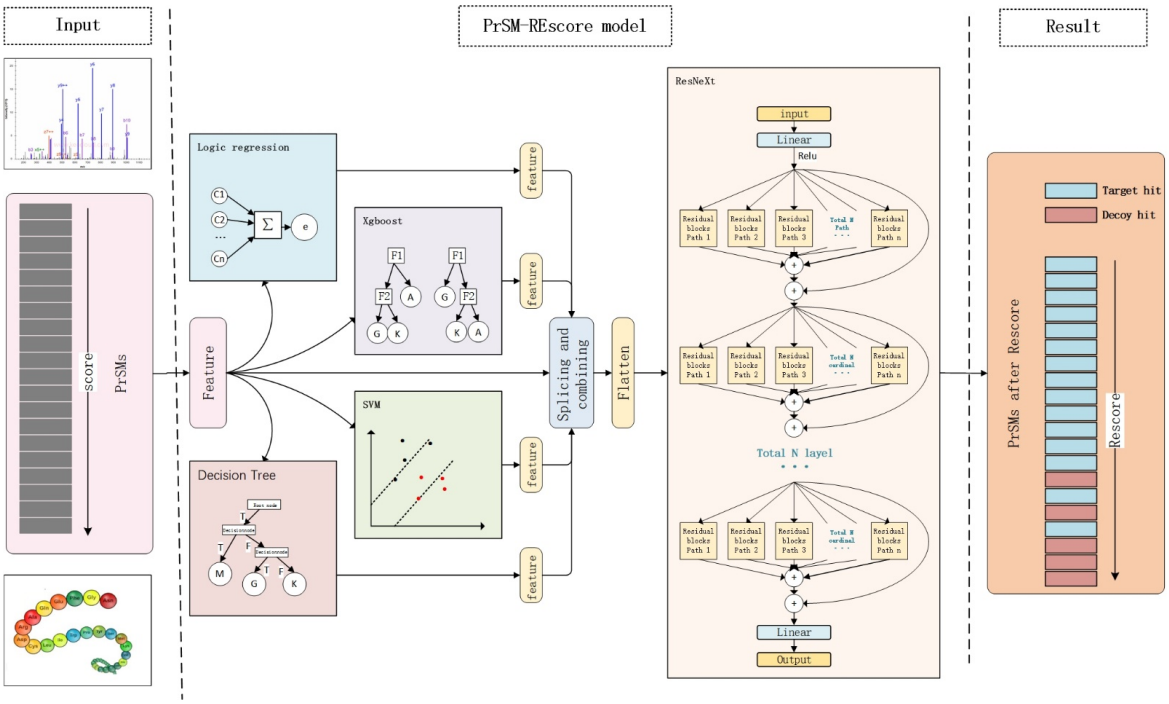


图 3-2 PRSMREscore 模型整体结构图

图 3-2 也可看成是 PRSMREscore 模型的预测流程，图 3-3 是 PRSMREscore 的训练过程。训练过程按照模型结构有两步，第一步使用部分训练数据将特征提

取模块的机器学习模型 Logic Regression、SVM、XGBoost 和决策树模型分别训练进行训练。第二部使用另外部分训练数据先输入第一步得到的机器学习模型，将得到的预测分数作为特征与原始的特征组合合并。再集成输入进 ResNeXt 残差神经网络进行训练，这样就能训练好最终的残差网络部分。

本节之后将按照特征介绍、模型结构介绍和数据预处理的步骤来安排。

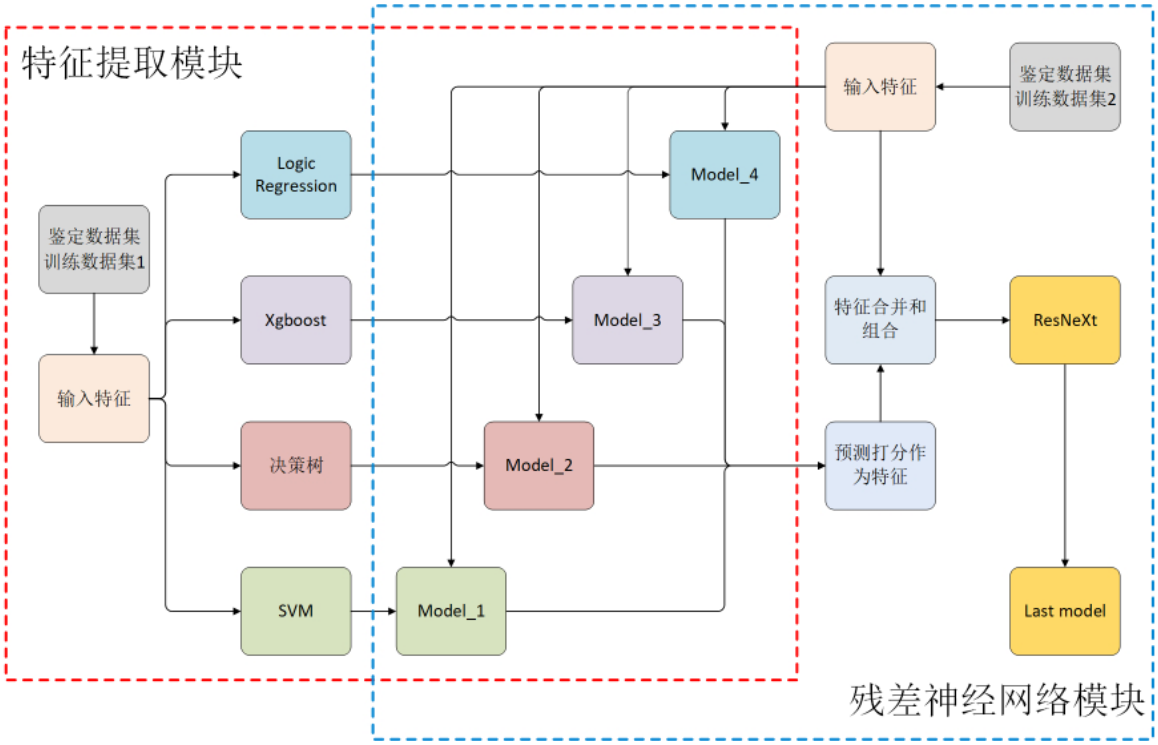


图 3-3 PRSMREscore 两大模块训练

3.2.1 特征选取

蛋白质变体鉴定领域涌现了很多的鉴定算法。但在蛋白质变体鉴定的领域内，其最终结果的所描述的信息都一致。即都会将匹配正确的 PrSM 信息报告出来。本文的后处理过程，依托重打分的策略。使得本文并不关注鉴定算法所负责的研究工作。本文只使用机器学习的技术和统计概率知识，根据鉴定算法输出的所有信息，给范围内所有的 PrSM 进行重新打分。所以，在这些鉴定算法输出的所有信息中，选择关键的信息是一个重点。经过实验，本文以下选择了一些指数作为本文模型的输入特征数据。

(1)匹配峰的数量：在蛋白质质谱分析中，“匹配峰的数量”通常指的是实验测得的质谱图中的峰与理论质谱图中的峰之间的匹配数量。这是一种用于比较实验数据和理论预测之间的一致性的度量。这涉及到将实验中测得的质谱图与已知蛋白质的理论质谱图进行比较。理论质谱图是通过计算蛋白质的氨基酸序列，预测离子片段的质荷比，然后生成的理论质谱。而实验中得到的质谱图是通过质谱仪测

量样品离子化后的质谱信号。“匹配峰的数量”表示实验质谱图中的峰有多少与理论质谱图中的峰相对应。匹配的标准通常基于质荷比误差和信号强度等因素。更多匹配意味着实验数据与理论预测更一致，这有助于确定蛋白质的序列和结构。总而言之，匹配的峰数越多，一般是代表表征鉴定越准确的。

(2)匹配质谱碎片数量和一般匹配质谱碎片数量：“匹配碎片数量”是指实验测得的质谱中的离子片段与理论质谱中的离子片段相匹配的数量。这是一种用于评估蛋白质鉴定准确性的度量。蛋白质质谱实验通常会产生一系列离子片段，这些片段反映了蛋白质的氨基酸序列和其可能的修饰。这些实验片段与已知蛋白质的理论片段进行比较，理论片段是通过计算蛋白质氨基酸序列并预测其在质谱中产生的离子片段而得到的。“匹配碎片数量”表示实验中测得的离子片段有多少与理论质谱中的离子片段相对应。匹配通常基于质荷比误差、离子片段类型和信号强度等因素。更多的匹配碎片数量通常意味着实验数据与理论预测更一致，从而提高了蛋白质鉴定的可靠性。“一般匹配碎片数量”含义与“匹配碎片数量”是相似的，都指的是实验质谱中的离子片段与理论质谱中的离子片段相匹配的数量。区别可能在于一般匹配碎片数量的表述更广泛，包括更多类型的离子片段，如 b 离子、y 离子、内离子、b++离子等等。这一度量考虑了所有可能的匹配，而不仅仅是特定类型的碎片。这样的综合匹配数量提供了更全面的蛋白质鉴定信息，因为不同类型的碎片可以提供关于蛋白质序列的不同方面的信息。

(3)可变和未知 PTM 数量：PTM(Post-translational Modification 翻译后修饰)改变蛋白质的结构、功能和活性。包含了磷酸化，甲基化等。对 PTM 的研究对于理解蛋白质功能、调控细胞过程等方面都具有重要的意义。在质谱分析中，检测到的修饰模式和位置信息有助于确定蛋白质的功能和调控机制。如果一个蛋白质有多个修饰位点，这些修饰可能影响质谱图中的峰的位置和强度。因此，了解变异修饰的位点数量有助于更精确地解释和鉴定蛋白质的质谱数据。

(4)原本的 rank 指标或者说是原本结果的最终评分：对于所有使用了 Target-decoy 方法控制 FDR 的鉴定算法。都必然会存在一个 rank 排名，这其实原本就代表了鉴定算法对所报告的 PrSM 的确定性和可信度度量。鉴定算法 TopPIC 是通过采用谱概率计算方法从而得出所以 PrSM 的错误期望值(e_value)。根据这个期望值，从小到大排列，然后计算出整个结果集在 Target-decoy 策略下的 FDR 值。故一个 PrSM 的错误期望值越小，代表这个 PrSM 更可信。TopPIC 还提供了另外一个概率指标 p_value。这个指标代表 PrSM 的不可信度概率。这个指标基本是与 e 值进行正相关的，但是 p_value 是概率值，即 p 值的取值范围在 0-1 之前，所有不会向 e 值那样变得很大(10 的 300 次方)。故一般而言，当 PrSM 的 p 值等于 1 的时候，代表鉴定算法认为这条 PrSM 不可信。本文将以上两者纳入了本文的模型，其中对于 e 值进行了取对数处理，据本文已发现的数据，

e 值的区间范围大致为10^{-50}到10^{300}之间。使用对数处理后，能够将 e 值指标映射到合理区间。

3.2.2 提取特征模块

PRSMREscore 的前半部分是特征的进一步挖掘过程。在这个阶段，本文利用了多种机器学习算法，包括逻辑回归、XGBoost、决策树和支持向量机(SVM)，对特征进行了初步的打分预测。逻辑回归是一种经典的分类算法，通过对特征进行线性组合来进行分类预测。XGBoost 是一种梯度提升树模型，能够处理非线性关系并具有很强的泛化能力。决策树是一种基于树结构的分类模型，通过构建树形结构来进行决策和分类。而 SVM 则是一种基于间隔最大化的分类方法，能够有效处理高维数据和非线性数据。

本文使用鉴定算法 TopPIC 所报告的 PrSM 结果作为例子。TopPIC 鉴定算法是一个通过自顶向下的 MS 进行高通量蛋白质组全蛋白质形态识别和表征的软件工具，它集成了蛋白质过滤、光谱比对、E 值计算和贝叶斯模型的算法，用于表征未知氨基酸突变和 PTM。

本文实验将 3.2.1 节的报告指标作为特征，每一条 PrSM 必会生成一个特征向量 $x \in R^n$。将特征向量 x 分别输入进各个机器学习模型进行打分预测。按机器学习扩充了的模型数量会得到一个分数向量$S_x = (score_1, score_2, score_3, \ldots, score_n)$。其中结合 n 个模型则分数向量$S_x$有 n 维。然后，既符合人们常理也符合机器学习科学性的结论是，正例所得到的打分是较反例高的。而使用多个机器学习模型的目的正是能够综合利用它们各自的优势，从而更准确地预测蛋白质变体鉴定结果的得分。

将每个机器学习模型得到的打分向量S_x与一开始的初始特征向量 x 组合拼接，形成最终的特征向量X_{last}，PRSMREscore 前半部分的任务就完成了。目的旨在为后续的重打分模型提供更为准确和可靠的输入特征。

3.2.3 残差神经网络

通过 PRSMREscore 提取特征模块得到了一条 PrSM 的最终特征X_{last}。这组特征就直接通过残差神经网络得到最终的重打分结果。

首先，ResNet(Residual Network)相比普通神经网络的优势在于其能够更有效地训练深层网络，并且具有更好的性能表现。这主要归功于 ResNet 引入的残差连接(Residual Connection)机制。在蛋白质变体鉴定的后处理过程中，通常会涉及到对大量蛋白质序列匹配结果进行深度学习模型的训练和预测。由于蛋白质变体鉴定涉及复杂的数据和特征，需要构建更深的神经网络来提取和表示这些信息。然而，普通的深层神经网络可能会遇到梯度消失和梯度爆炸等问题，导致训

练困难和性能下降。这时候，使用 ResNet 就能够更好地应对这些问题，使得网络更容易训练，并且具有更好的泛化能力。

然后，相比 ResNet，本文的模型选取直接使用 ResNeXt。ResNeXt 相比 ResNet 在一些方面表现更优秀。ResNeXt 是在 ResNet 的基础上进行扩展和改进的，主要引入了“cardinality”(基数)的概念，即通过增加并行的分支来构建更深的网络。这种设计使得 ResNeXt 在提高模型性能的同时，能够更有效地利用有限的参数。 本文经实验最终选择使用 ResNeXt 作为最后的重打分预测模型。大大加快深度学习模型的训练时间。

3.3 实验环境及数据预处理

本章的实验使用的是 Windows 平台的机器，软件平台使用 Pytorch 深度学习框架，其中，使用 Python 作为编程语言。本章具体的实验环境细节和使用到的库版本如表 3-1 所示：

表 3-1 本章实验环境细节

软/硬件	型号/参数
中央处理器(CPU)	AMD Ryzen 7 5800X 8-Core Processor
深度学习框架	Pytorch1.10.1
平台	Python3.6.15
Python 包 sklearn	0.0.post4
Python 包 xml	1.0.1
Python 包 numpy	1.19.5
Python 包 joblib	0.11
Python 包 XGBoost	1.4.2

3.3.1 数据集

数据集均从欧洲生物信息学研究所[67]网站下载的 raw 质谱源文件。本文使用的数据集来自不同物种，涵盖的物种包括了动植物两个大类。使用的数据集物种有斑马鱼，人类，黄粉虫，麝香小鼠，豌豆，拟南芥和酵母。数据集的大小按照原本鉴定算法 TopPIC 报告的 PrSM 数量，大小范围为 9 到 12000 条 PrSM。一共使用了 59 个数据集，其中的 12 个被划分为了训练集。剩下的 47 个数据集作为测试集。表 3-2 是各个数据集的信息。实验使用的 12 个训练集来自两个物种，分别是斑马鱼和人类。其余数据集的物种信息也都已在表中列出。每一个物种存在多个数据集。本文便以表中的物种简写加序号进行指代。如 AH_1 和 AH_2 分别代表物种是肌肉麝香小鼠的第一个、第二个数据集。

表 3-2 本章实验数据集

训练集/测试集	物种信息	数据集代号简写	包含数据集个数
训练集	斑马鱼	FB_CB	3
训练集	智人血浆	Human_H	4
训练集	智人血浆	Human_T	5
测试集	智人血浆	Human_L	4
测试集	智人血浆	Human_0	4
测试集	智人直肠癌细胞	Human_E	6
测试集	斑马鱼	FB_TO	3
测试集	肌肉麝香小鼠	AH	6
测试集	黄粉虫	TM	5
测试集	豌豆	PS	7
测试集	拟南芥	AT	6
测试集	酵母	Yeast	6

3.3.2 数据集预处理

本文实验的所有数据均从质谱仪生成的.raw 文件开始预处理的，在整个蛋白质变体表征的流程中。本文会清楚的介绍全部数据处理流程使用的工具和数据格式。所有用到的软件名称、软件版本和数据处理参数都将列在附表。以下是三个预处理使用到的工具，MSconvert(版本：3.0.23054-b585bc2)，TopFD(版本：1.6.2)，TopPIC(版本：1.6.2)。

(1)使用 MSconvert[68]将下载的经过质谱仪生成的二进制原始.raw 转化为 mzXML 文件。MSconvert 是一种开源工具。使用其将.raw 文件转化为.mzXML 格式的文件。mzXML 文件包含了一次实验中所有的分子片段的质谱图,并且包含了一些实验的基本数据。其核心数据是峰谱图。MSconvert 软件参数中，peakPicking 设置为 vendor msLevel=1.2; scanSumming 中各设置分别为 precursorTol=0.1，scanTimeTol=120 ，ionMbilityTol=5，sumMs 1=0。其余参数使用默认参数。

(2)使用 TopFD[38]进行解卷积。将.mzXML 文件经过解卷积算法转化为.msalign 文件。解卷积的软件，本文使用的是 TopFD。处理完后会生成相应的特征文件和可视化文件。Msalign 文件里面含有生物样本经过质谱仪产生的所有碎片离子的电荷量,值荷比和道尔顿等信息。TopFD 软件参数中,Use EnvCNN for scoring 设置为 True。其余使用软件默认参数。

(3)使用 TopPIC 软件进行鉴定。TopPIC 是刘[10]等人开发的自顶向下蛋白质变体表征的开源工具。输入.msalign 文件，相应生物蛋白质序列库和修饰列表即可

得到蛋白质变体鉴定的最终结果——PrSM。所报告的每一条 PrSM 都标注了该蛋白质变体的所有信息。TopPIC 软件参数中，Max variable PTM number 设置为 3。Mass errortolerance 设置为 15。其余设置为默认。需要特殊说明的是，在 TopPIC 的 FDR 设置中，训练集的 FDR 阈值设为 0.1。而测试集设置为 1。其目的是避免测试集的结果存在鉴定算法的 FDR 信息而影响重打分过程。

3.3.3 深度学习模型参数设置

本文认为在实际实验中，参数和批次大小的限制取决于实验平台的资源配置。本文残差网络 ResNeXt 的优化器使用的是 Adam Optimizer，训练 100 个 epoch，初始学习率设置为 0.1，分别在第 50 个 epoch 和第 80 个 epoch 时，学习率衰减为原来的十分之一，将其批次大小设置成 100。实验中，残差网络 ResNeXt 的 cardinality 设为 8。每个残差块里面包含三个线性层，神经元个数分别为 32，64，64。激活函数使用的是 Relu。

3.4 实验结果分析

本章节所提出的的 PRSMREscore 算法，已经在 47 个测试数据集上进行实验。且这些输入数据是跨物种的，这体现了模型能够不受物种限制的优点。本文根据原本鉴定算法鉴定后输出 PrSM 数量大小将数据集分为了三类。在原本鉴定算法报告的 PrSM 数量在 5000 以上，认为这个数据集属于大型数据集；鉴定算法报告的 PrSM 数量在 300 以内，认为是小型数据集；鉴定算法报告的 PrSM 数量在 5000 以下 300 以上，认为是中型数据集。

实验的对比基线模型是TopPIC算法。本文因为类别属于后处理过程，故需要介绍提升数量和提升比例作为性能度量指标。在这里，本文定义介绍一些指标性的概念。

对一个数据集来说，定义一个数据集的提升数量是使用PRSMREscore算法重打分后，报告的PrSM数量减去原本的鉴定算法报告的PrSM总数，本文将其记为ΔNum。公式如下：

$$\Delta Num = PrSMS_{REscore} - PrSMS_{TopPIC} \tag{3.1}$$

其中$PrSM_{TopPIC}$代表原本鉴定算法TopPIC报告的PrSM总数。$PrSMS_{REscore}$代表鉴定结果使用PRSMREscore重打分后的PrSM总数。

再定义数据集的提升比，其公式如下：

$$\Delta Ratio = \frac{PrSMS_{REscore} - PrSMS_{TopPIC}}{PrSMS_{TopPIC}} \tag{3.2}$$

其中原本$PrSMs_{TopPIC}$即原本的鉴定算法 TopPIC 报告的 PrSM 总数。故公式可以理解为使用本文重打分模型后，多报告的 PrSM 数量占原本的鉴定算法报告的 PrSM 总数比例多少。

3.4.1 提升情况分析

本文将蛋白质变体鉴定算法 TopPIC 的结果使用 PRSMREscore 算法重打分。在分别给 47 个不同物种的数据集结果进行重打分后，本文将属于同一物种的数据集的 PrSM 总数进行了累加，其整体结果如表 3-1 和表 3-2 所示。在实验的所有物种数据上，TopPIC+PRSMREscore 的结果都是最优秀的。所有的原始结果使用 PRSMREscore 算法经过了 20 轮训练并重打分并预测结果的平均。20 轮结果提升比的方差为 0.004374461。这显示本文的结果较为稳定。

表 3-1 人类血浆、人类直肠癌细胞、斑马鱼和肌肉麝香小鼠数据集上的 PrSM 数量

方法	Human_L	Human_0	Human_E	FB_TO	AH
TopPIC	1003	2184	5364	20642	1662
TopPIC+PRSMREscore	1036	2194	5428	20959	1677

表 3-2 黄粉虫、豌豆、拟南芥和酵母数据集上的 PrSM 数量

方法	TM	PS	AT	Yeast
TopPIC	3761	565	3593	10098
TopPIC+PRSMREscore	4009	753	3681	10274

为了将 47 个数据集的提升情况完整的表述，图 3-4 显示了每一个数据集的提升数量。

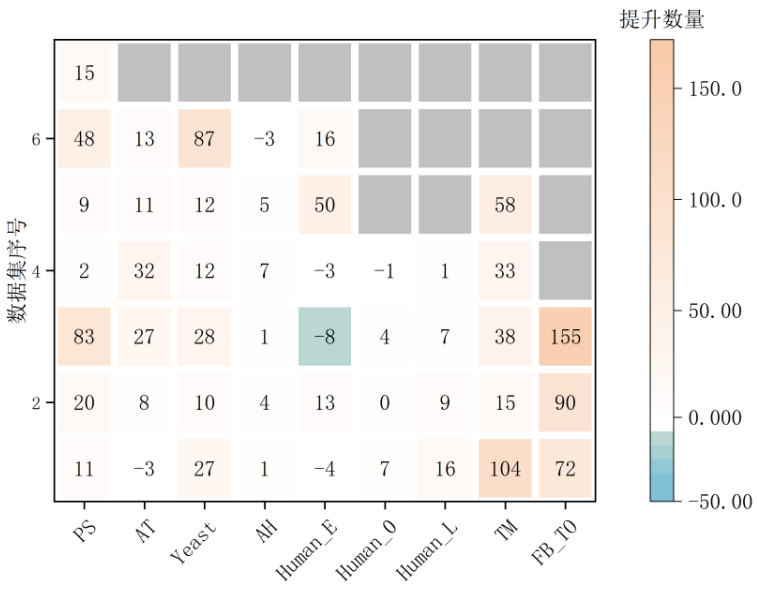


图 3-4 47 个数据集使用 PRSMREscore 算法后的提升数量

在图 3-4 中横坐标代表数据集的物种，纵坐标代表数据集的序号。图中显示，绝大部分的数据集，使用 PRSMREscore 重打分后，都得到了提升效果，这证明算法对于绝大多数数据集都是有效的。

继续分析，从另外一个提升比的角度展现结果。本文统计使用 PRSMREscore 重打分后，这 47 个数据集的提升比例，结果如图 3-5 所示。在所有的结果之中，除了 AT_1 等数据集得到了最高不超过 2%的降低以外，其他的数据集都获得了一定的提升比。

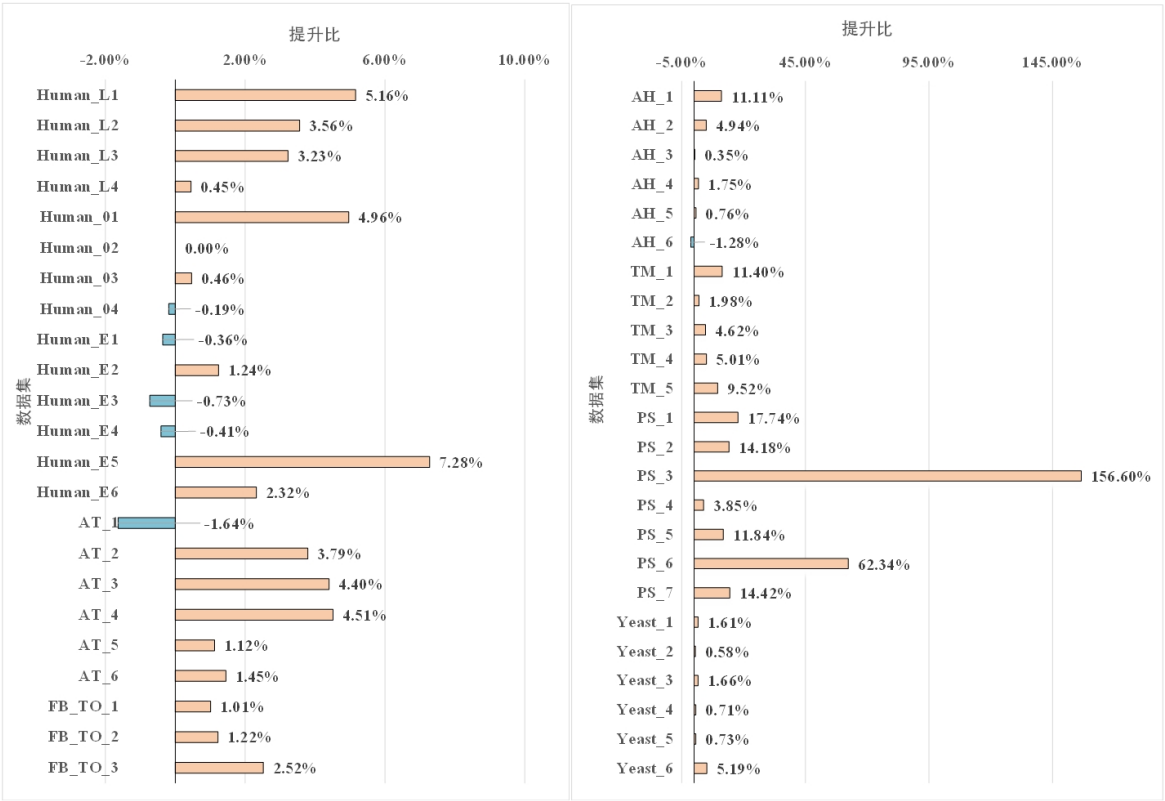
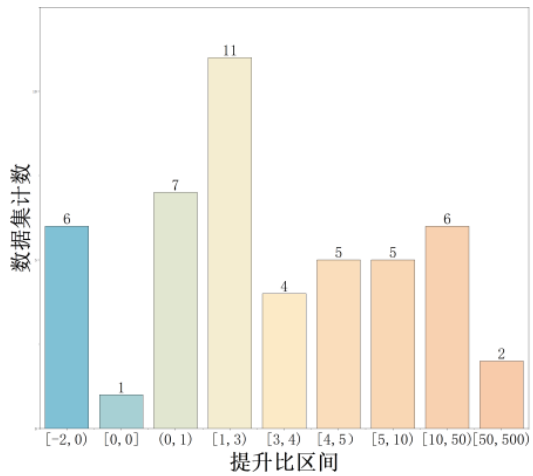


图 3-5 47 个数据集使用 PRSMREscore 算法后的提升比



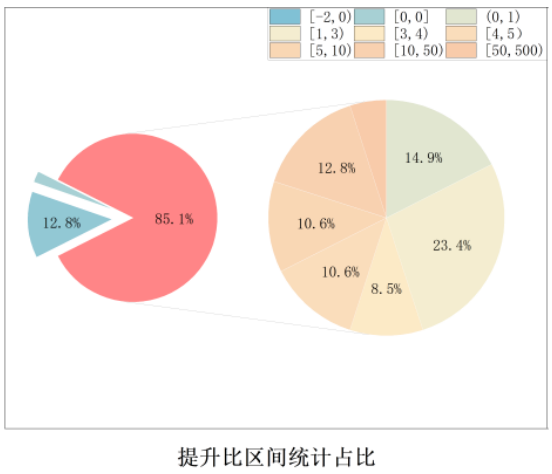


图 3-6 各提升比数据集区间统计　　图 3-7 各提升比的数据集占比

本文按提升比的大小进行区间统计，结果如图 3-6 和图 3-7。在这 47 个数据集中，其中提升比不为负的数据集有 41 个，数量占比约为 87.2%。本文将包含 0 提升的数据集在 47 个数据集所占的比例称之为非降比。在图 3-7 的复合饼图也可以清晰看出，这 47 个数据集中，使用本文的算法重打分后有提升的数据有 40 个，在总数据集个数中占比达到了 85.1%。这表明本文的方法在实际场景中对大多数数据集都能够有效地改进原本鉴定算法的预测结果。而且在有提升情况的数据集之中，提升比在 5%以上的数据集有 13 个，数据集个数占比约为 27.6%。提升比在 10%以上的数据集有 8 个，数据集个数占比约为 23.4%。这种提升比的数据集的占比对于解决实际问题，特别是对于提高模型在真实世界应用中的可用性具有十分重要的意义。

接下来具体分析相关数据集。通过图 3-8 大型数据集的结果分析后看出，在超过 5000+PrSM 的数据集(FB-T0)上，使用 PRSMREscore 重打分后结果的提升数量是极其可观的。这三个大型数据集，每个数据集原本经过鉴定算法报告的 PrSM 都在 5000 以上。使用本文的重打分算法后，PrSM 的报告数量分别都得到了 72，90 和 155 提升。

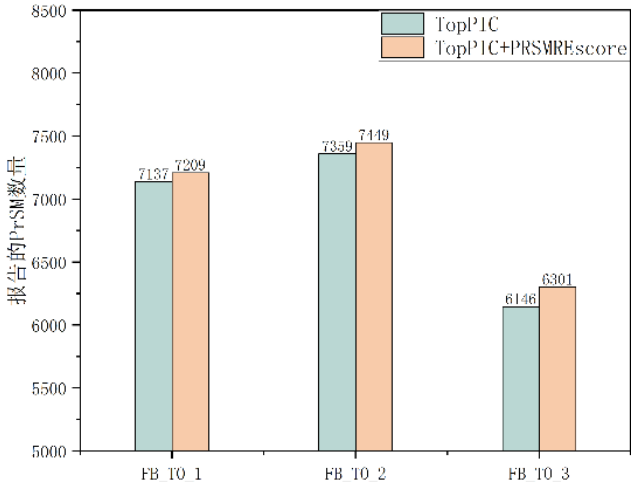


图 3-8 斑马鱼(FB_TO)数据集的提升数量

而在一些中型数据集中，使用 PRSMREscore 重打分后也能够获得非常好的提升比例效果，图 3-9 是酵母(Yeast)和黄粉虫(TM)的数据集，图 3-9 左是 6 个酵母数据集的结果，其提升数量分别为 27、10、28、12、12 和 87。其中提升数量最小的数据集是 Yeast_2，提升数量最大的为 Yeast_6。

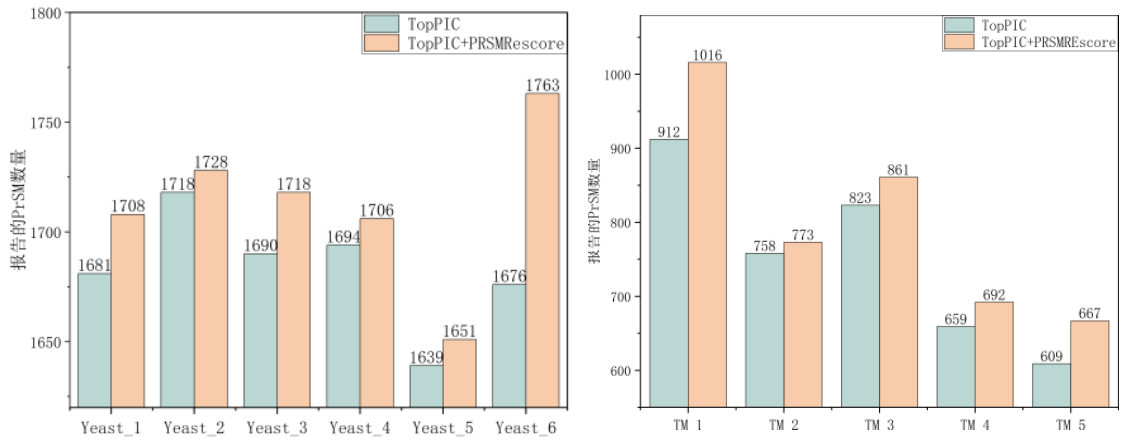


图 3-9 酵母(Yeast)和黄粉虫(TM)数据集的提升数量

另外一个中型数据集黄粉虫可以看出，除了数据集 TM_2 以外，其余大部分数据集的结果都拥有 30 条以上的 PrSM 数量的提升。其中 TM_1 和 TM_5 增加的幅度最大，都分别提升了 104 和 58 条 PrSM 的报告数量。

图 3-10 是物种为植物的小型的数据集豌豆，通过柱状图可以看出：相比不同于大型和中型数据集在提升数量上的优势，小型数据集的性能提升在提升比例上的优势更为明显。在除去数据集 PS_4，其他数据集都能达到 10%的提升率。其中 PS_3 数据集达到了 156%的提升比例，这是所有数据集中最高的提升比例。

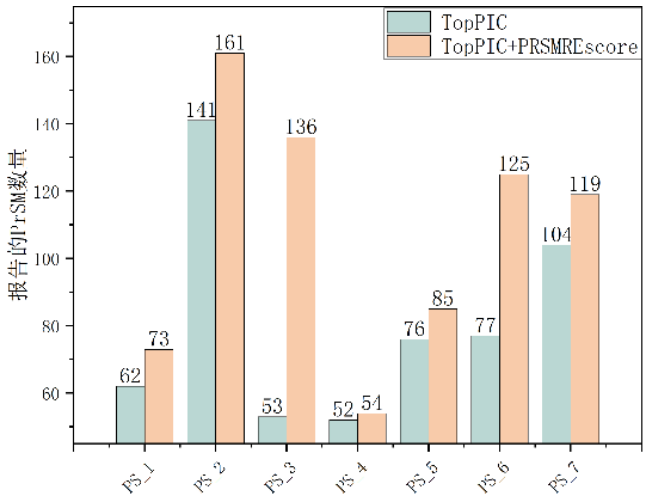


图 3-10 豌豆(PS)数据集的提升数量

3.4.2 交叠分析

本文采用了交叠分析方法来比较本文的重打分结果与原本结果之间的重叠情况。本文收集了四组数据并计算了四组数据之间的元素交集，以及每组数据的元素总数。这四组数据以同一物种同一生物取样为总体，即以统计数个同类数据集的总数。这能够量化 PRSMREscore 重打分结果与现有研究结果之间的相似性。本文的覆盖率定义为：

$$rate_{coverage} = \frac{PrSMS_{Common}}{PrSMS_{TopPIC}} \tag{3.3}$$

其中$PrSMS_{Common}$代表原本鉴定算法输出的所有结果和经过 PRSMREscore 重打分后输出的所有结果中相同的 PrSM 总数，所以$rate_{coverage}$代表这部分 PrSM 结果在经过重打分后，还能被报告出来的 PrSM 数量占原本鉴定算法结果总数之比。

图 3-11 至图 3-14 分别显示了四个物种人类、酵母、黄粉虫和豌豆数据集 PrSM 结果的交叠情况。如图 3-13 黄粉虫数据交叠图能明显看出，数据集使用 PRSMREscore 重打分后，多输出的 PrSM 数量远远大于原本鉴定算法的结果数量。

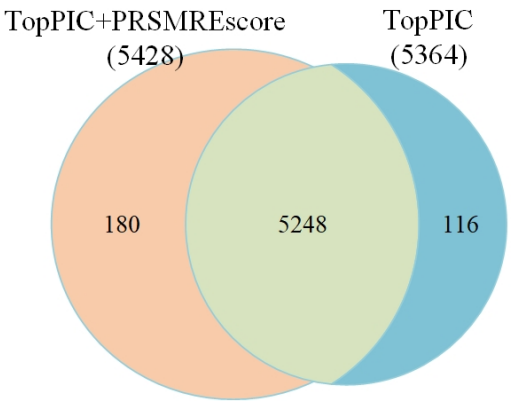


图 3-11 人类直肠癌细胞(Human_E)

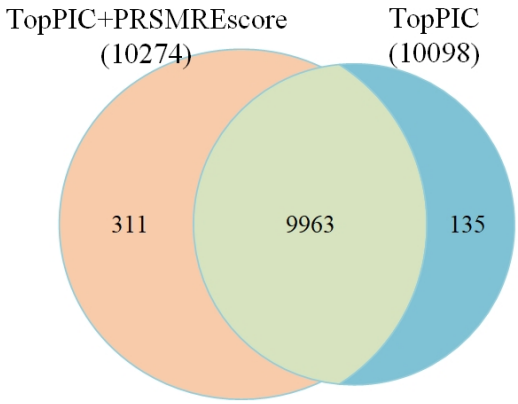


图 3-12 酵母(Yeast)

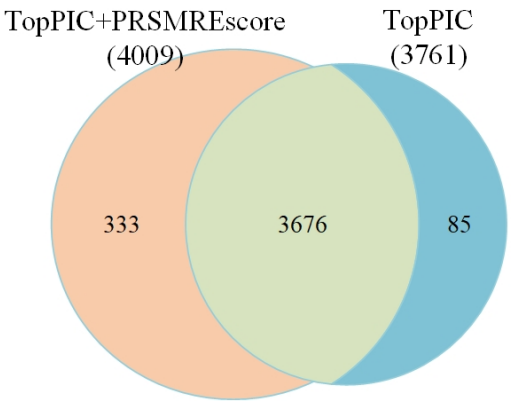


图 3-13 黄粉虫(TM)

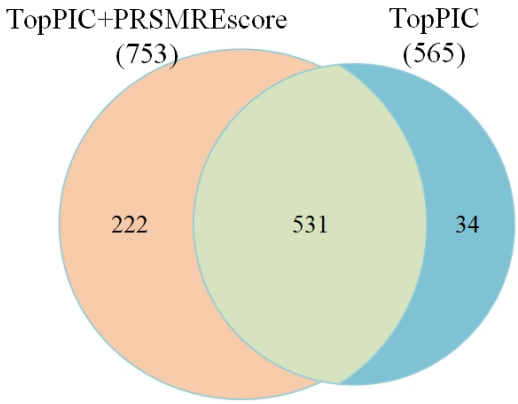


图 3-14 豌豆(PS)

在本文的交叠分析中，人类、酵母、黄粉虫和豌豆四个物种取样内的结果覆盖率分别约为 97.8%、98.6%、97.7%和 93.9%。研究结果在很大程度上覆盖了现有研究的结果。这些共同的结果可能代表了领域内的一些核心特征或普遍存在的现象。这样的结果显示使用 PRSMREscore 模型不会造成鉴定结果的失真。这能够体现模型不会对原本的鉴定结果进行否定，表现出了重打分模型该有的科学性。继续对豌豆数据集与其他三个数据集进行深入比较分析，可以发现算法模型在较大的数据集上都有十分卓越的性能，在三个 PrSM 结果大于 1000 的数据集

上都达到了 97%以上的覆盖率。这一发现表明，PRSMREscore 算法在处理大型数据集时具有显著优势，而且使用重打分算法的数据集规模大小与覆盖率之间存在着正向关联。最后从图 3-11 至图 3-14 中可知，人类、酵母、黄粉虫和豌豆数据集使用 PRSNREscore 重打分后得到的结果总数都比原有结果更多。其中算法模型重捕捞后独有的 PRSM 数量比原本独有的结果数量也高出了至少 50%。

于此，本小节继续分析重打分模型中独有的结果。如图 3-15 显示。

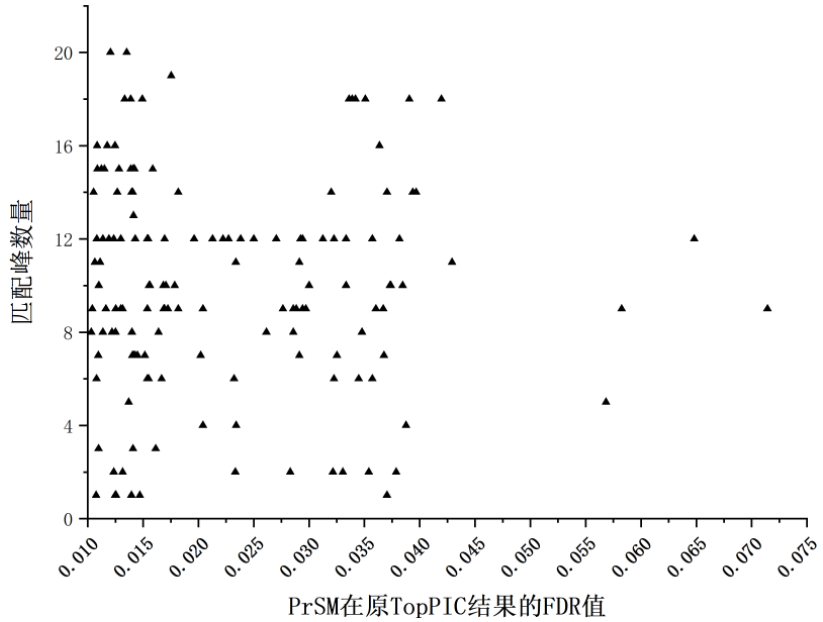


图 3-15 豌豆(PS_1)数据集捕捞 PrSM 的峰匹配数量统计

图 3-15 是豌豆(PS_3)使用 PRSMREscore 后独有输出结果峰匹配情况。PS_3 原鉴定算法输出 53 条 PrSM，使用 PRSMREscore 后输出 136 条 PrSM，覆盖率为 100%。从输出结果的 FDR 值分析，将原本鉴定重打分后，PRSMREscore 不仅将几条原本鉴定算法遗弃的 FDR 值在 0.05 以上的 PrSM 重捕捞。而且还输出了很多 FDR 值在 0.01 至 0.045 的 PrSM。这些 PrSM 在 TopPIC 中都是因为 FDR 值在阈值 0.01 以外而被截去的数据。从输出结果的匹配特征上分析。大部分输出 PrSM 的匹配峰一般以 8-16 个为主。故 PRSMREscore 算法对于匹配峰数在 8 个以上的 PrSM 作用明显。

3.4.3 特异性 PrSM 分析

TopPIC 和 PRSMREscore 的分别打分后，所有输出的 PrSM 能通过交叠分析看出分为了三类：TopPIC 打分输出的独有 PrSM、PRSMREscore 打分输出的独有 PrSM 和两者都输出了的 PrSM。本节主要分析 TopPIC 和 PRSMREscore 独有 PrSM 的具体情况，如图 3-16 所示。

图中四个数据集的情况能明显看出，PRSMREscore 输出的 PrSM 在峰的匹配数量，质谱碎片匹配数量和一般质谱碎片匹配数量上都比 TopPIC 输出的 PrSM

多。这证明了 PRSMREscore 的打分可信度是要高于 TopPIC 的。

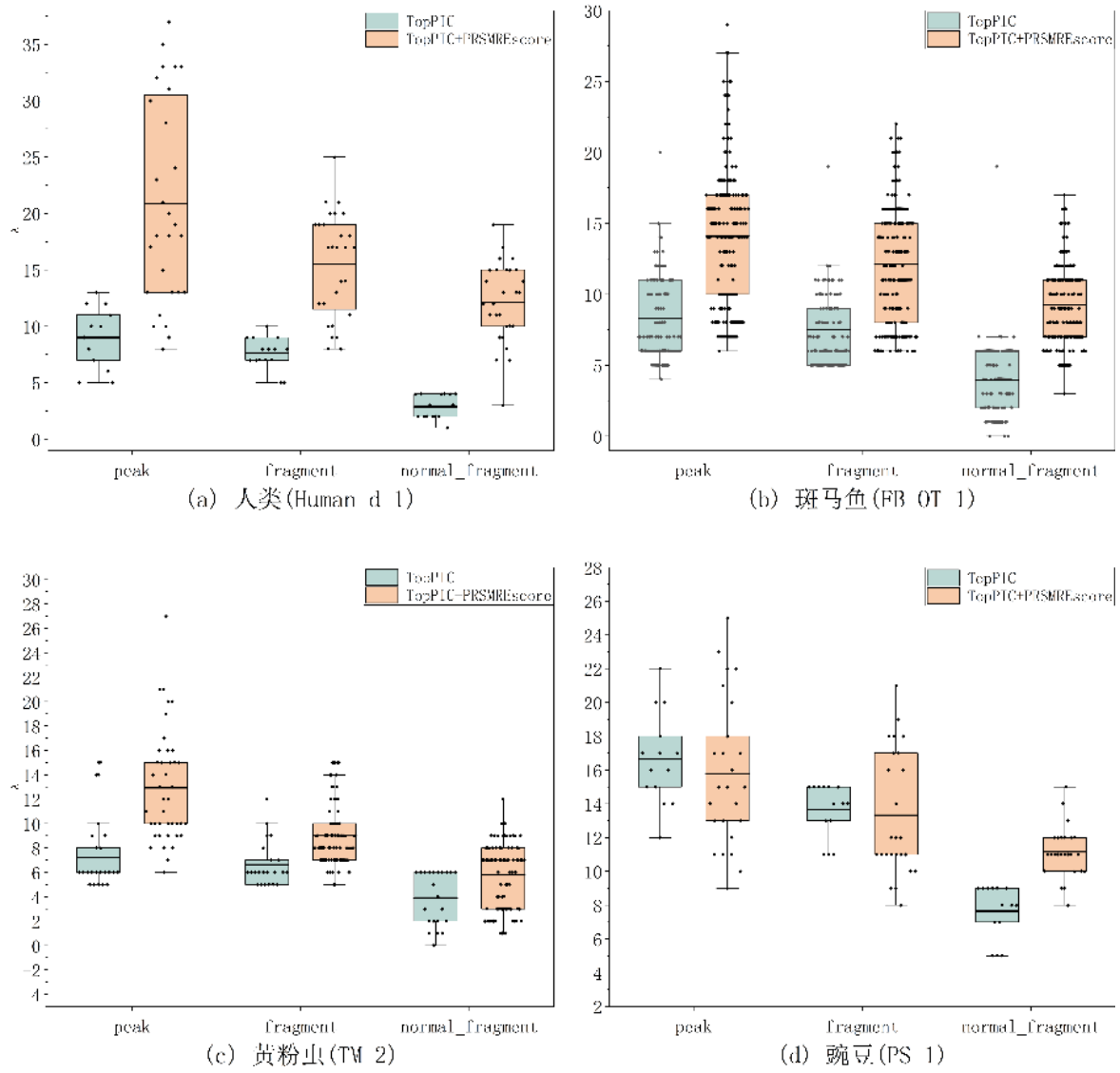


图 3-16 人类、斑马鱼、黄粉虫和豌豆数据集特异性 PrSM 的特征信息对比

3.4.4 融合打分策略

原本经过鉴定算法得出的高可信度 PrSM 其实在本文的重打分模型里面是最好分辨的，故这一类 PrSM 的得分其实就会与原本得分一样是正相关的，在 rank 中的位置其实是没有太大出入的。故本文的重打分算法的主要作用目标，其实是在原本 rank 排名中 FDR 值在 0.01 附近的 PrSM 集合。而如果本文的算法能够将在目标区间内的匹配 target 的 PrSM 给与足够的得分补偿，其实也能够达到提升这些 PrSM 排名的作用。因此可以引入一个灵活的算术运算过程，例如可以将鉴定算法 TopPIC 原本的 rank 分数 e_value 和经过重打分后的分数进行组合处理，这种组合操作不仅仅可以简单的相加或相乘，也可以通过一系列运算获得了更加复杂和综合的结果。以下是将两者进行简单线性组合的结果，由于 e_value

是从小到大排列，而得分是从大到小排列。在进行试验相关系数后，本文简单的将本文的得分减去 e_value。目标公式如下，即相当于一个简单的线性映射。形成了一个新的分数，本文将此称为最终分数。

$$Score_{last} = Score_{RE} - Score_{TopPIC} \tag{3.3}$$

其中，$Score_{TopPIC}$是鉴定算法 TopPIC 给 PrSM 的评分，$Score_{RE}$是 PRSMREscore 给 PrSM 的评分。$Score_{last}$可作为 PrSM 的最终得分进入 Target-decoy 策略排序。实验要求并不限定于只能使用减法，本文按照理论使用减法手段作为映射变换示例。

另外一种打分策略能够很好的减少小数据集抖动的情况，不过会损失部分提升数量和提升率。但是这对于假阳性要求较高的情况下，是一种极其优秀的策略。充分发挥了鉴定算法的表征能力又使用了本文重打分模型的捕捞能力。结果如图 3-16 和图 3-17 所示。

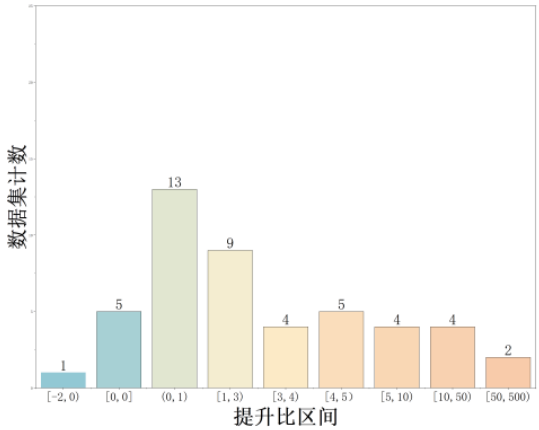


图 3-16 各提升比数据集区间统计

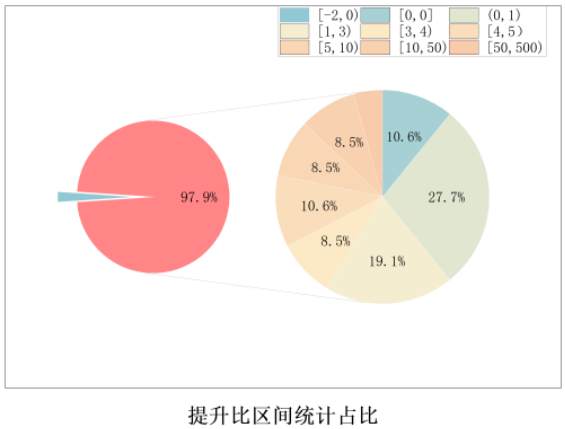


图 3-17 各提升比的数据集占比

虽然得到的高提升比数据集的数量虽然降低了一些。但在 47 个数据集中，只有 1 个以内的非提升，除了其中有 5 个数据集是 0 提升 0 降低的情况下，剩下的占比 87%左右的数据集都能够获得提升。最重要的是，使用这种较为折中的处理方式后，整个数据集的非降比(有提升的和 0 提升)已经达到 97%以上。这证明任意一个数据集使用 PRSMREscore 重打分后有 97%以上的概率是不会造成 PrSM 结果减少的。这非常适合一些要求高精度的应用场景。

3.5 本章小结

本章提出了一个基于经典机器学习和深度学习技术的自底向上蛋白质变体鉴定的重打分算法PRSMREscore，它使用整合经典机器学习算法Logic Regression、XGBoost、决策树、SVM和残差神经网络ResNeXt模型对热门蛋白质变体鉴定TopPIC的报告结果进行重打分。使得数据集能够再捕捞一些因为

Target-decoy被遗漏的PrSM结果。最后对本章的PRSMREscore模型的实际实验结果做了详细分析。分别从实验数据结果的总体提升情况，大中小型数据集的提升情况和原本鉴定结果的覆盖度等多角度分析了使用PRSMREscore模型后和原鉴定算法的对比情况，以此证明提出模型的有效性。从实验结果可发现，模型算法能在跨物种学习的基础上，应用至广泛的鉴定算法之后进行后处理。且除去小数据集的抖动，在大中型数据集上势必能取得很好的效果。为蛋白质变体鉴定的准确性、可靠性等方面贡献了显著的实际意义。

第 4 章 目标-诱饵策略下的排名损失函数优化算法研究

4.1 绪论

现代检索系统通常由底层机器学习模型驱动，而且最大化分类准确性是一个非常重要的指标。此类系统的目标是给输入的一组特征或者数据进行正确分类打分，分数越高意味着机器学习模型对这组特征或数据越有把握。但现实中，还存在于一些其他的目标。比如说银行智能化系统接待 100 个客人的时候，如果里面存在欺诈者，银行会因为更加关注单一顾客的真伪准确度而拒绝掉低分的真实顾客。这就是一种追求真阳性为目标的例子。现实中还有一些以其他目标为目的机器学习和深度学习系统。这些特殊目标影响着机器学习模型的参数更新和抉择。当前有很多基于排名的性能指标来评估机器学习系统[69] 。而在蛋白质变体鉴定过程中，广泛使用的控制 FDR 指标的手段便是使用 Target-decoy 搜索策略。Target-decoy 搜索策略在设计时便已经有排名和分数的相关定义。Target-decoy 搜索策略要求报告出的 PrSM 的 FDR 必须小于阈值，而这个阈值是受到排名的影响。这其实就属于一个排序问题，要求相关样本的排序高于不相关样本。因此，以此为要求能设计一个适合优化排名的损失函数。

目前，根据相关综述[70]，二元分类问题的框架方法分为三种：排名法、关注排名顶部精确度的方法和优化 Neyman-Pearson[71]判据的框架方法。排名法的主要目的是尽可能的将阳性相关样本放在首位。其为每个样本分配一个数字分数，并根据该分数进行排名。通常，只考虑将超过阈值分数的样本预测为正例。关注顶部精确度的方法类似于排名算法。但它并没有对最相关的样本进行排名，而是只最大限度地提高了这些顶级样本的准确性(相当于最大限度地减少了错误分类)。Neyman-Pearson(奈曼-皮尔逊)问题关注于最小化了 II 型误差(原假设为假，接受原假设)，同时保持了 I 型误差(原假设为真，拒绝原假设)一定是小于等于既定阈值。

4.2 Top 数据指标优化算法研究

在蛋白质变体鉴定流程中，所有鉴定的最后一个环节是报告在 Target-decoy 策略下阈值以上的 PrSM 结果。Target-decoy 由于其定义有排名的要求。在结合机器学习技术进行后处理的过程中，排名优化理论刚好能有优秀的实验基础的。即可以在 Target-decoy 搜索策略的允许阈值下，使用优化函数优化重打分模型，使得匹配 target 目标序列的 PrSM 正例尽可能的排名在前，匹配 decoy 伪序列的 PrSM 负例尽可能的排名在后。主要思想是关注顶部 top 数据的真阳性，在保

持顶部数据的精度(Target-decoy 搜索策略中的 FDR 阈值)情况下，最大化召回率。因为在本文重打分算法框架内，PrSM 的排名的绝对正确并不一定是第一目标，而是尽可能的将正确匹配 target 序列的 PrSM 推至 FDR 阈值前。