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。

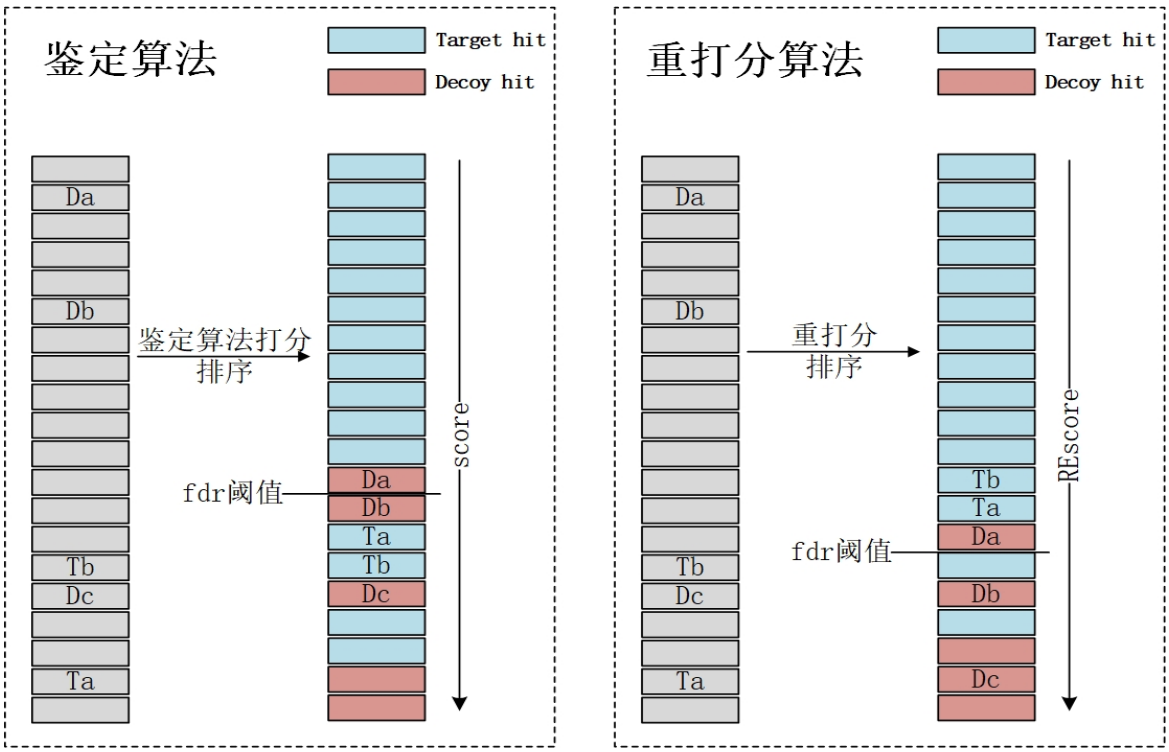


图 3-1 重打分效果:图中可见由于重打分提高了 PrSM 的得分准确性，捕捞了正确命中 target 的 PrSM(Ta,Tb)

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。

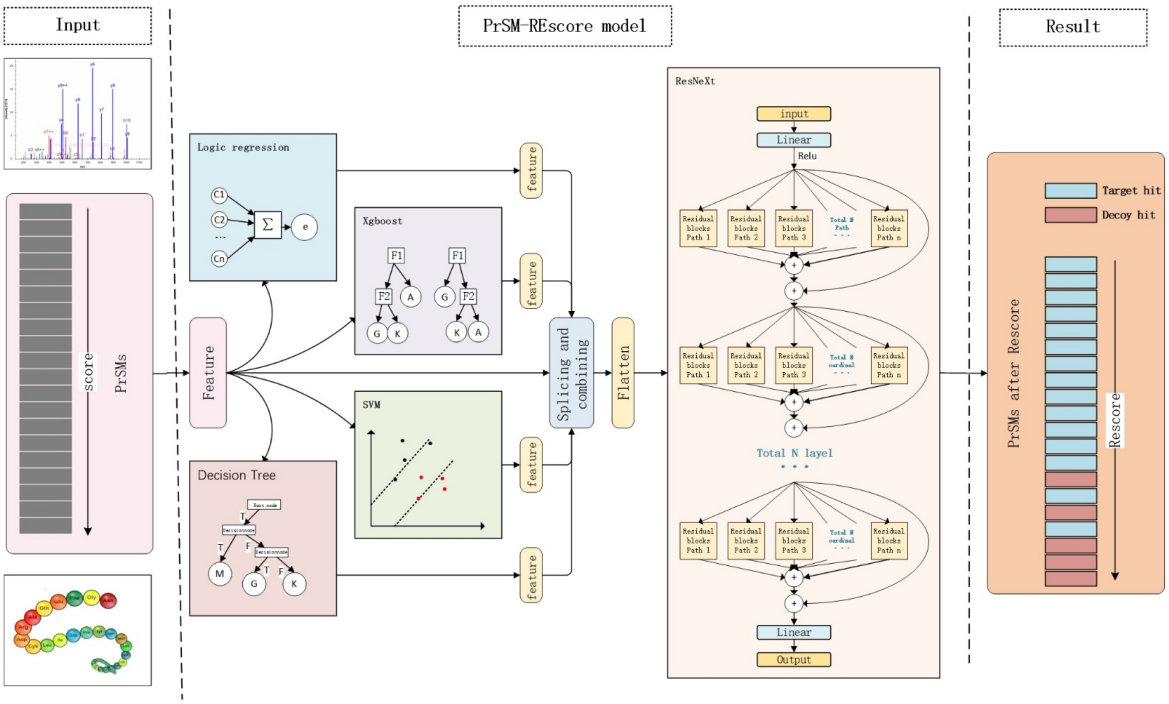


图 3-2 PRSMREscore 模型整体结构图

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。

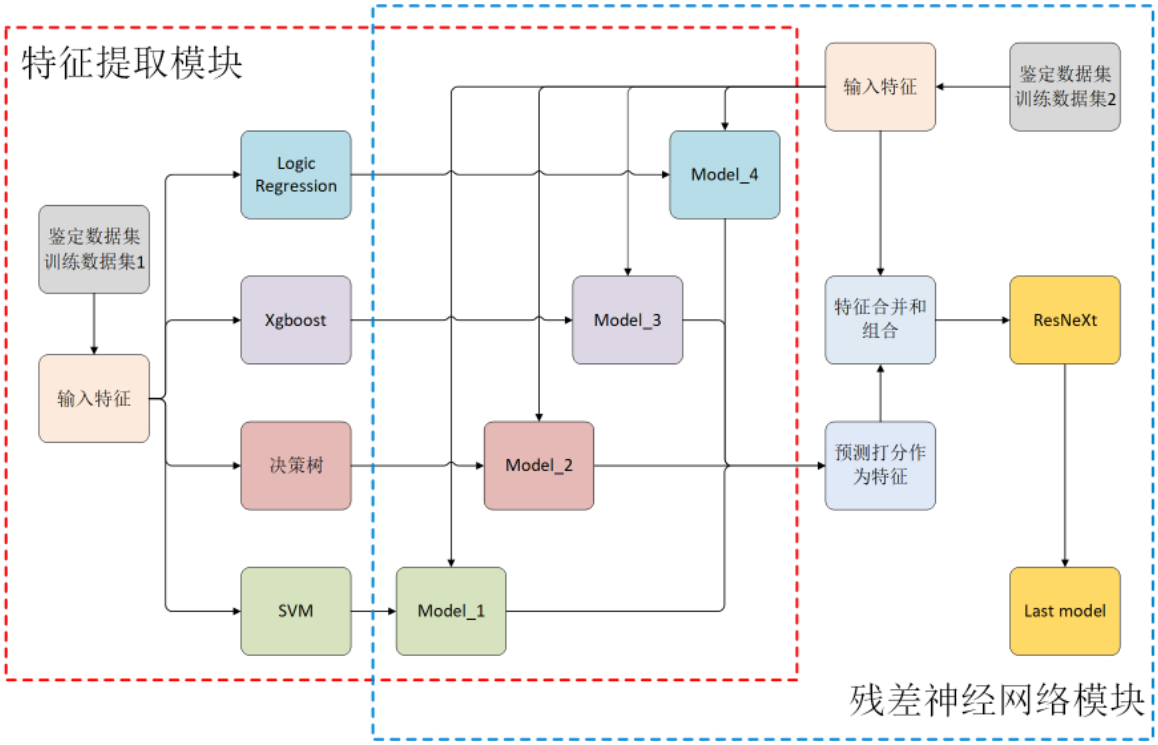


图 3-3 PRSMREscore 两大模块训练

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。

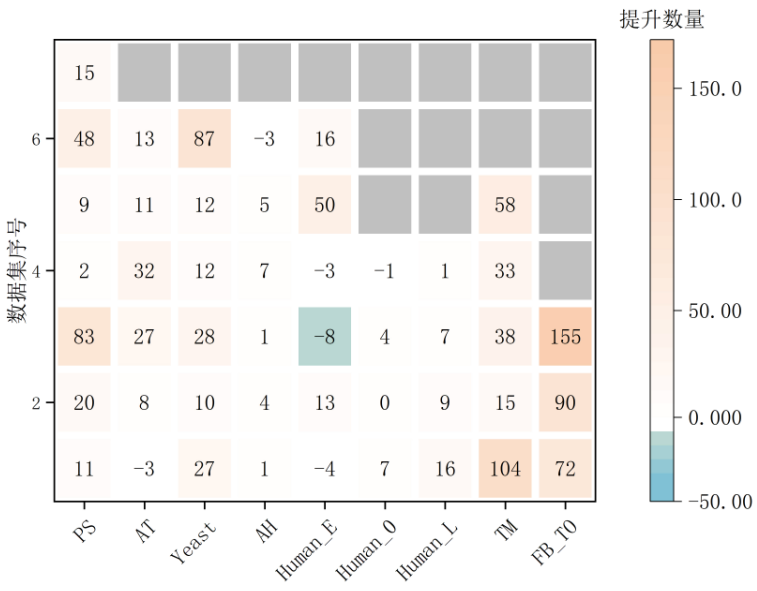


图 3-4 47 个数据集使用 PRSMREscore 算法后的提升数量

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。

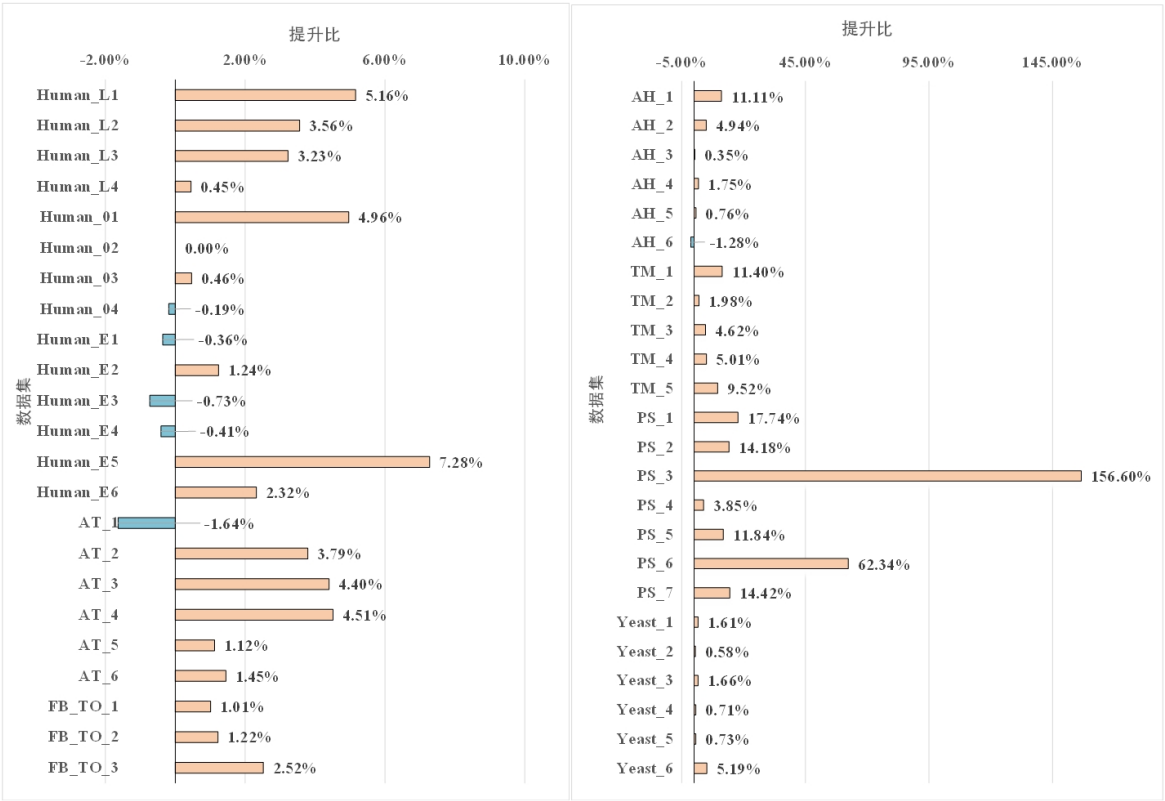
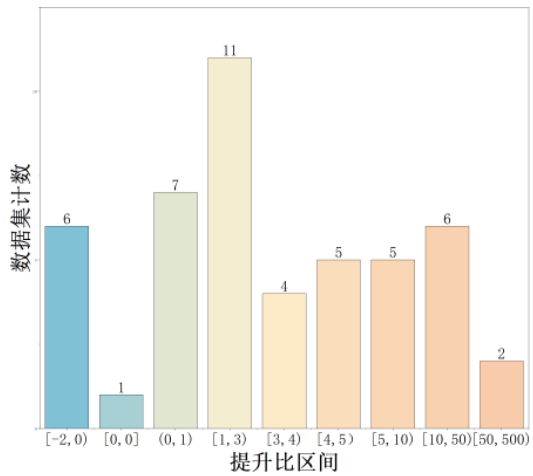


图 3-5 47 个数据集使用 PRSMREscore 算法后的提升比



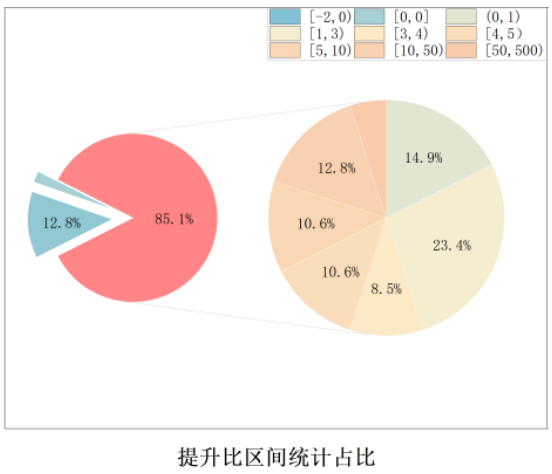


图 3-6 各提升比数据集区间统计　　图 3-7 各提升比的数据集占比

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。

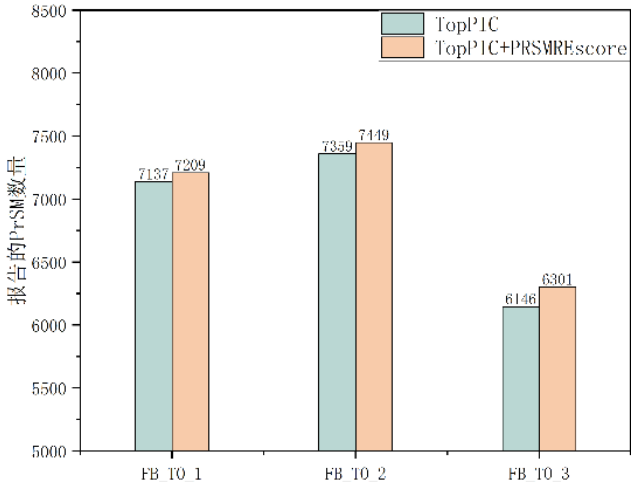


图 3-8 斑马鱼(FB_TO)数据集的提升数量

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。

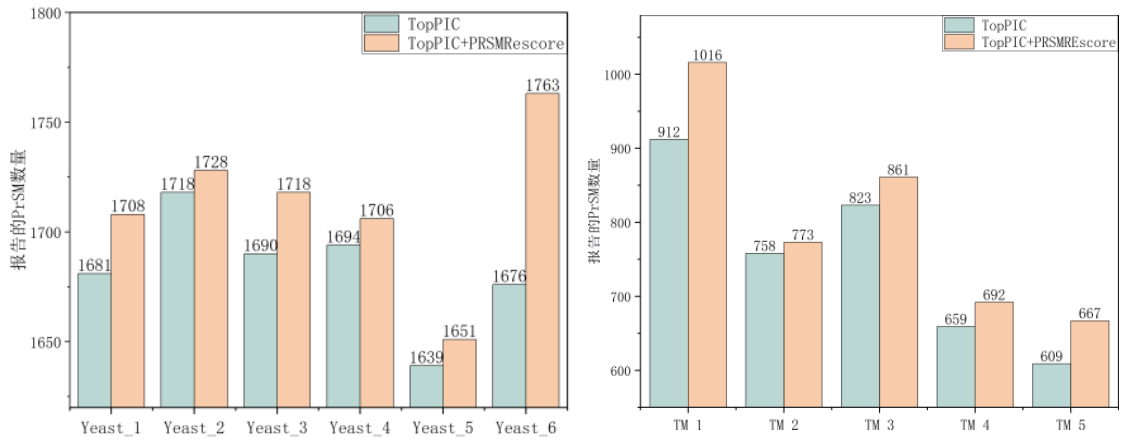


图 3-9 酵母(Yeast)和黄粉虫(TM)数据集的提升数量

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。

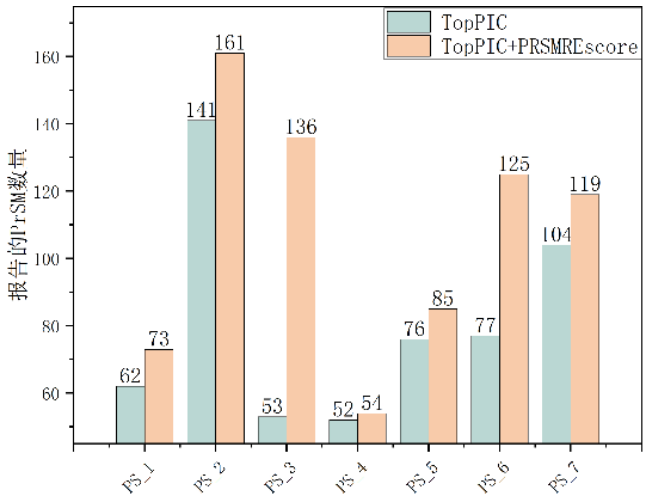


图 3-10 豌豆(PS)数据集的提升数量

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。

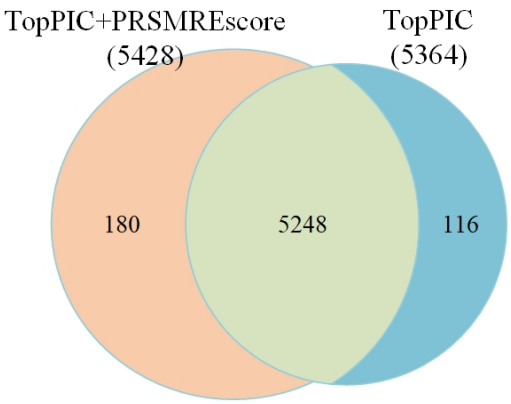


图 3-11 人类直肠癌细胞(Human_E)

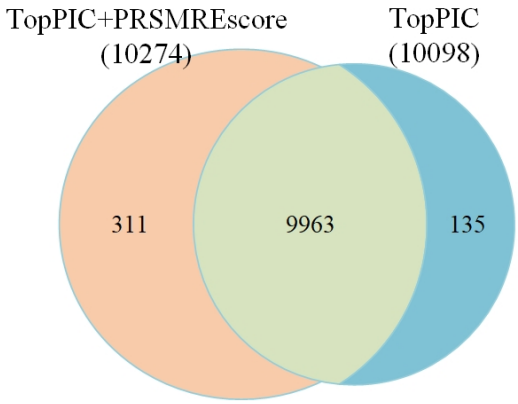


图 3-12 酵母(Yeast)

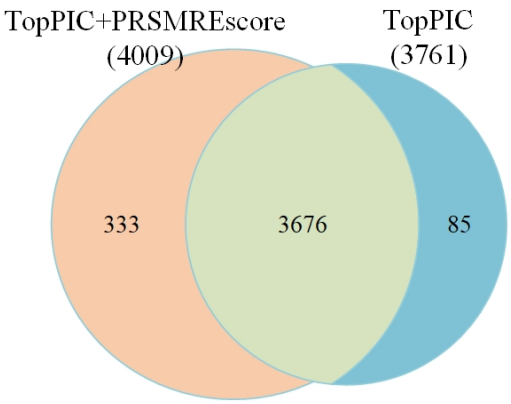


图 3-13 黄粉虫(TM)

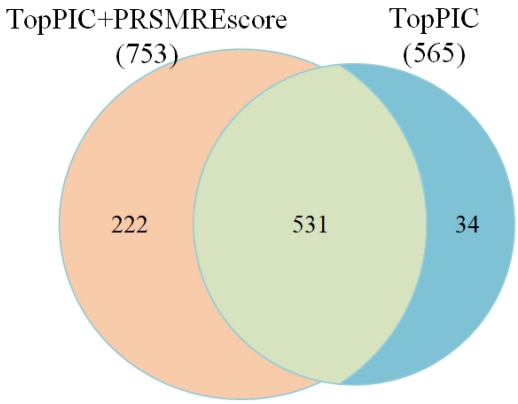


图 3-14 豌豆(PS)

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。

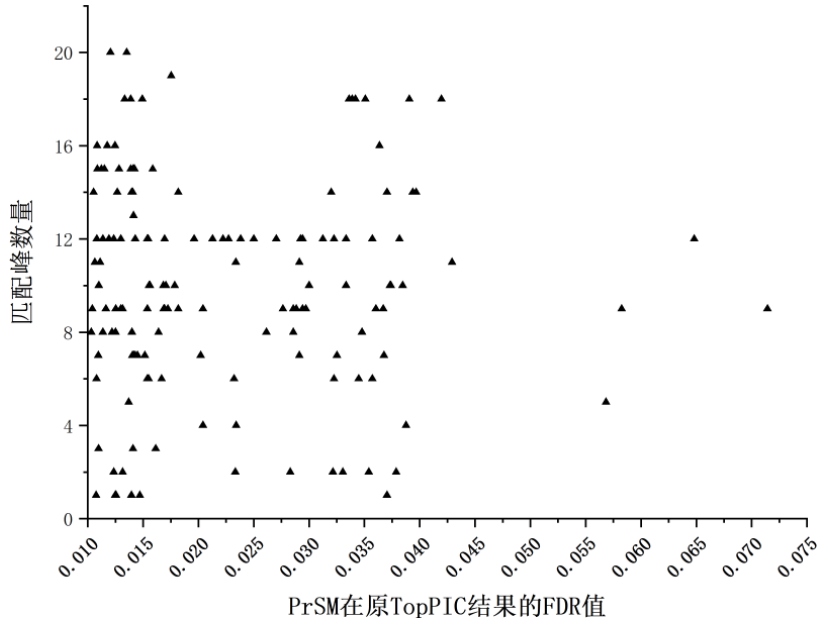


图 3-15 豌豆(PS_1)数据集捕捞 PrSM 的峰匹配数量统计

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。

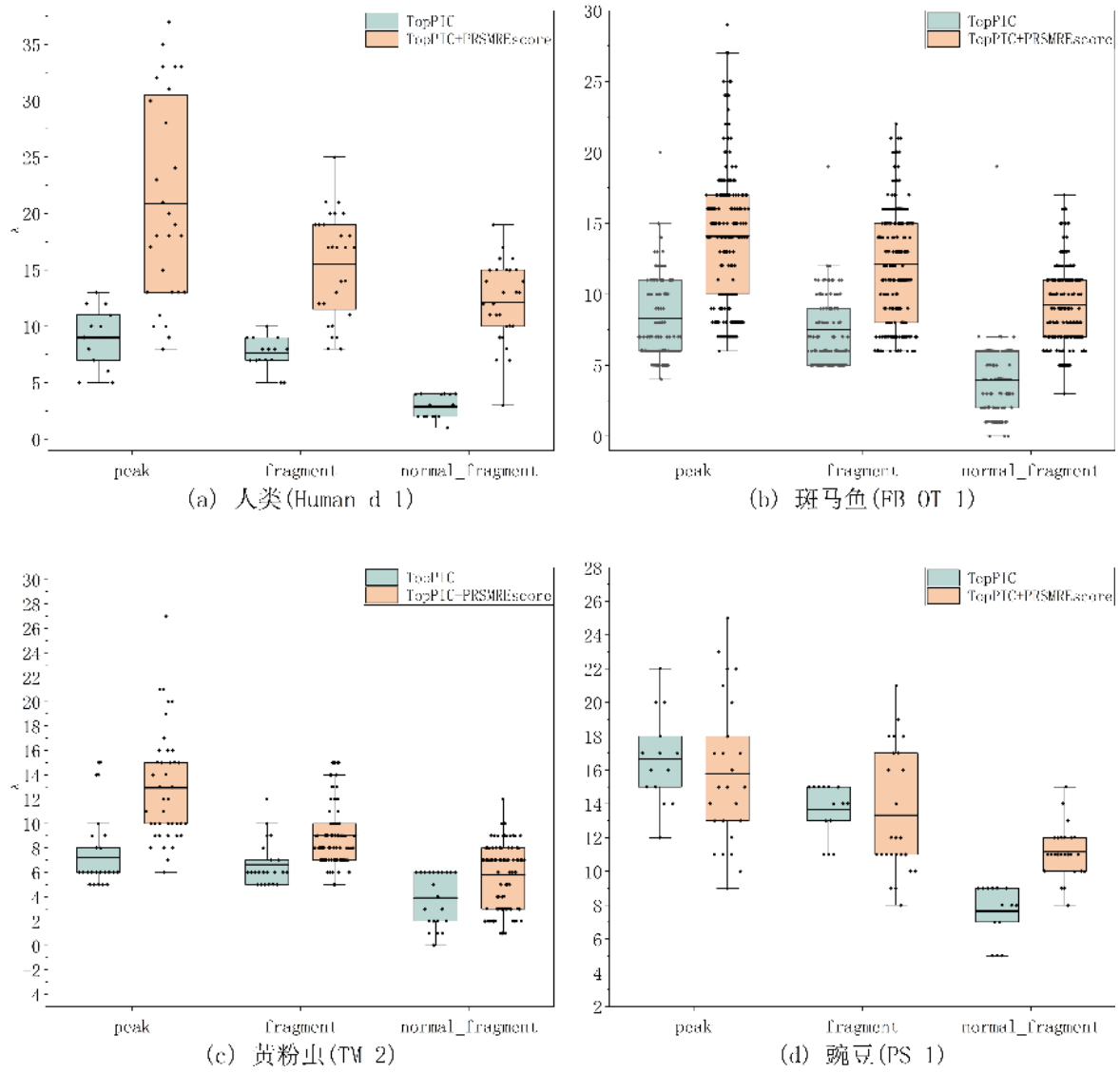


图 3-16 人类、斑马鱼、黄粉虫和豌豆数据集特异性 PrSM 的特征信息对比

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。

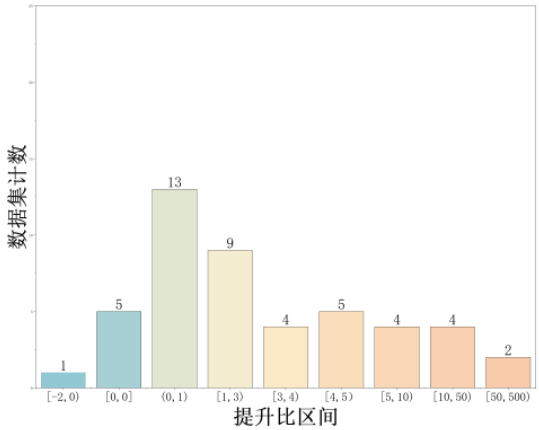


图 3-16 各提升比数据集区间统计

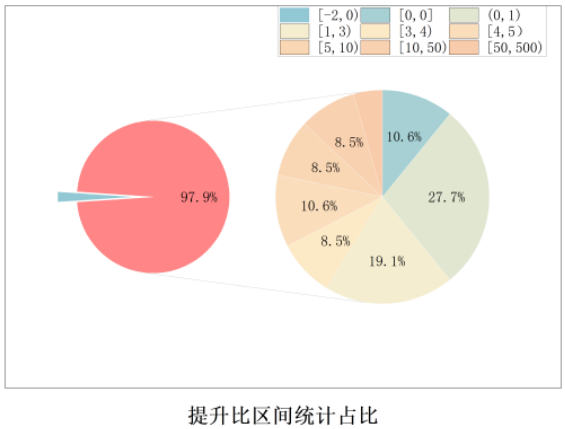


图 3-17 各提升比的数据集占比

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。