Biophysics 生物物理学, 2023, 11(2), 17-29
Published Online May 2023 in Hans. https://www.hanspub.org/journal/biphy
https://doi.org/10.12677/biphy.2023.112002

Hans汉斯

深度神经网络筛选蛋白质组学高置信度定量肽段

郭　欢[1]，何情祖[1,2]，黎玉林[1]，帅建伟[1,2]

[1]厦门大学物理系，福建　厦门
[2]中国科学院大学，国科温州研究院，浙江　温州

收稿日期：2023年3月28日；录用日期：2023年5月9日；发布日期：2023年5月16日

摘　要

质谱分析是蛋白质组学的重要研究方法。数据不依赖获取是一种稳定且复现性高的质谱仪数据采集方式，具有质荷比范围宽广，通量高等特点。DIA-NN是处理DIA蛋白质组学数据的主流定量软件之一。由于DIA-NN分析DIA数据后输出的肽段中存在低置信度肽段，生物学家需要根据肽段碎片离子色谱峰组图(XICs)的相似性来人工筛选出高置信度肽段。人工筛选的任务量大、耗时长，并且筛选标准因人而异，这导致结果具有主观性。本文提出了一种名为MSDeepFilter的算法，它基于深度学习技术，能够自动筛选出高置信度的肽段。MSDeepFilter算法结合压缩激励神经网络和残差网络设计深度学习模型，从XICs中提取特征，以此区分高置信度和低置信度肽段。与传统机器学习模型Adaboosting和支持向量机模型相比，MSDeepFilter模型在基准数据集上的多项分类性能指标均表现更优，测试集AUC值达到了98.7%。这表明MSDeepFilter具有优秀性能，可以替代人工筛选的环节。

关键词

深度学习，DIA蛋白质组学，质谱数据，人工筛选，压缩激励神经网络

A Method for Analyzing DIA-NN Output Peptides Based on Squeeze-and-Excitation Neural Network

Huan Guo[1], Qingzu He[1,2], Yulin Li[1], Jianwei Shuai[1,2]

[1]Department of Physics, Xiamen University, Xiamen Fujian
[2]Wenzhou Institute, University of Chinese Academy of Sciences, Wenzhou Zhejiang

Received: Mar. 28th, 2023; accepted: May 9th, 2023; published: May 16th, 2023

文章引用: 郭欢, 何情祖, 黎玉林, 帅建伟. 深度神经网络筛选蛋白质组学高置信度定量肽段[J]. 生物物理学, 2023, 11(2): 17-29. DOI: 10.12677/biphy.2023.112002

Abstract

Mass spectrometry is an important analytical method of proteomics. Data-Independent acquisition (DIA) is a stable and highly reproducible data acquisition method of mass spectrometer, which has the characteristics of wide range of mass to charge ratio and high throughput. DIA-NN is one of the mainstream quantitative software based on deep learning in the field of DIA proteomics data processing. The output of DIA-NN analysis of DIA data contains low confidence peptides, so biologists need to manually filter out high confidence peptides based on the similarity of peptide fragment ion chromatogram peak profiles (XICs). The task of manual filter is time-consuming, and the filter criteria vary from person to person, leading to subjective results. In this work, we propose an algorithm MSDeepFilter that can automatically filter out high-confidence peptides based on deep learning. The algorithm extracts the features of XICs by a deep learning model designed based on Squeeze-and-Excitation Networks and Residual networks as a way to distinguish high confidence peptides from low confidence peptides. Compared with the traditional machine learning models Adaboosting and Support Vector Machine models, the MSDeepFilter model performs better in several classification performance metrics on the benchmark dataset, with a test set AUC value of 98.7%. This indicates that MSDeepFilter has excellent performance and can replace the manual filtering process.

Keywords

Deep Learning, Data-Independent-Acquisition, Mass Spectrometry Data, Manual Filtering, Squeeze-and-Excitation Networks

1. 引言

蛋白质质谱仪是目前鉴定和定量肽段和蛋白的主要工具。数据依赖获取(Data Dependent Acquisition, DDA)和数据不依赖获取(Data Independent Acquisition, DIA)是常用的两种数据采集方式。鸟枪法实验是典型的 DDA 采集方式[1]，质谱仪会先对肽段母离子进行全扫描，获取一级质谱，然后选择一级质谱中强度排名前 N 的肽段母离子进行碎裂，得到二级质谱。DDA 采集方式获取的一级质谱和二级质谱对应关系清晰，但对母离子强度有很高的依赖性，导致实验结果存在随机性，且低丰度肽段的识别性较差。

相比之下，DIA 采集方式将母离子质荷比区间划分为多个独立窗口，依次将每个窗口内的所有母离子打碎并记录所有碎片子离子的信息作为二级质谱。DIA 采集方式具有实验结果重复性高、数据通量高、灵敏度高等特点，但是由于 DIA 数据中的碎片子离子信号来自于不同母离子，且混合在同一张二级质谱中，导致 DIA 质谱数据的分析极为困难。此外，DIA 质谱数据的准确定量是生物信息学中的基础问题，科学家在此基础上展开一系列具有生物意义的研究，例如基于 SWATH-MS 技术(一种 DIA 质谱技术)对细胞凋亡和坏死性凋亡途径之间串扰的定量理解研究[2]和在 SWATH-MS 质谱数据的基础上进行生物动力学建模[3]。对 DIA 质谱数据的高质量定性定量对这些研究非常重要，因此，开发对 DIA 质谱数据进行定性定量分析的工具一直是研究热点。

分析 DIA 质谱数据，常用的研究方法主要分为两种。一种是依赖库分析法，通过 DDA 实验数据创建文库，将实验输出 DIA 数据中的谱图与文库中的谱图进行匹配，从而完成肽段定量分析，此种研究方法中的谱图库可以是由大量鸟枪法 DDA 实验生成外部库，也可以是直接从 SWATH-MS 数据中的伪质谱文件生成内部库[4]。基于此种研究方法的常用工具有 OpenSWATH [5]、SWATHProphet [6]和 Specter [7]等。

另一种研究方法称为无库分析法，可以省去 DDA 实验。2015 年 Tsou C C 等人提出的 DIA-Umpire [8]利用肽段母离子的碎裂效率小于 100%的特点，计算一级质谱中的肽段母离子和二级质谱中被打碎肽段的相关性，以此找出一级质谱和二级质谱的对应关系，通过这种方式建立伪数据库，实现无库分析。同样基于无库分析法的工具还有 Group-DIA [9]、PIQED [10]和 directDIA (a part of Spectronaut [11])等。

近几年，基于深度学习的无库分析方法逐渐发展起来。2019 年 Bernhard Kuster 和 Mathias Wilhelm 提出的 Prosit 算法，基于 RNN 模型准确预测母离子的理论谱图和驻留时间，从而获得更加精确的质谱鉴定[12]。同年，Tarn 等提出的 DeepNovo-DIA 算法，将深度学习应用于从头测序(de-novo sequencing)法，直接鉴定肽段氨基酸序列[13]。2021 年，厦门大学教授帅建伟和博士何情祖等人提出算法 Ultra-DIA，利用深度变分自动编码器提取离子信号的特征，从而完成对肽段和蛋白进行定性和定量分析[14]，第二年该团队基于该算法进行优化并提出以谱图为中心的算法 Dear-DIA，结合深度变分自动编码器和三重态损失来学习提取的碎片离子色谱图的特征，然后使用 k-means 聚类算法将具有相似特征的碎片聚合到同一类中，从而处理 DIA 质谱数据[15]。同年，厦门大学韩家淮院士和俞容山教授团队提出了基于深度学习 LSTM 的鉴定软件 DreamDIA。该软件随机选取部分谱库结果，归一化其保留时间，并对谱库的保留时间进行拟合，从而对剩余肽段的保留时间进行预测。接着，DreamDIA 结合 LSTM 和全连接网络对输入数据进行打分，选择最优匹配结果[16]。

在这些研究中，于 2020 年由 Markus Ralser 等人开发的 DIA-NN 是将深度学习用于 DIA 蛋白质组学数据处理的集成软件包，开启了蛋白质组学的新篇章[17]。DIA-NN 能够将高通量方法用于可靠，稳健和定量准确的大规模实验。尽管 DIA-NN 通过算法控制假阳性率，但仍然会输出一定比例的低置信度肽段，这使得生物学研究人员在进行下一步精确研究时，需要人工筛选出高置信度肽段。人工筛选的步骤是先通过可视化软件提取碎片子离子色谱峰组(XICs)，然后以六条碎片子离子的色谱峰形相似度为标准筛选出高置信度肽段。目前用于色谱峰可视化的工具有 Skyline [18]、TOPPView [19]、MSSort-DIA^{XMBD} [20]和 DrawAlignR [21]。

其中，于 2022 年李一鸣等人提出的 MSSort-DIA^{XMBD} 作为 OpenSWATH 的最后一步，对肽段的 MS/MS 数据进行可视化和分类肽段母离子[20]。该方法利用 OpenSWATH 输出报告信息对每个肽段和匹配的碎片子离子组进行色谱曲线重构和可视化，并使用深度卷积神经网络对重构后的信息进行数据挖掘，结合双阈值分割策略，自动识别高置信度肽段和低置信度肽段，本质上是针对 OpenSWATH 质谱数据分析软件输出肽段的再筛选，缓解人工检查任务负担。

与 OpenSWATH 相同，DIA-NN 的输出报告中普遍包含数万甚至十几万个肽段，如果一个输出报告中包含 1000 个肽段，检查一个肽段对应的碎片子离子的 XICs 图片花费 30 秒，后期的人工检查环节就达到约 83 个小时，因此人工筛选的过程非常耗时。此外，人工筛选的主观性强，筛选标准不统一，这会导致错误率因人而异。人工检查筛选高置信度肽段如图 1 所示主要分为两步：利用输出报告的信息(母离子的质荷比和保留时间信息以及质谱仪质荷比窗口信息)从质谱原始数据(母离子保留时间附近的一系列二级质谱信息)提取出每个肽段匹配的碎片子离子组的色谱曲线，存储为图片格式；第二步：科研人员观察色谱图片，以六条子离子色谱曲线相似为标准，删除低置信度肽段，筛选出高置信度肽段。

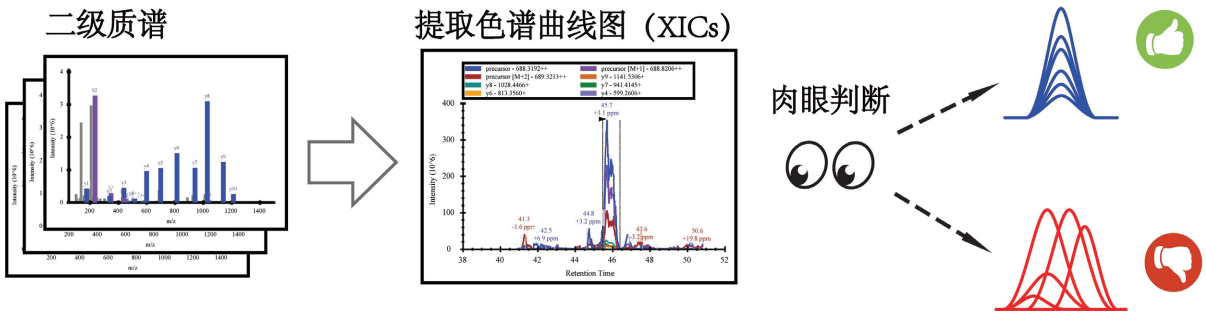


Figure 1. Steps of manual inspection
图 1. 人工检查步骤

在本文中，我们提出了一种基于深度学习的算法 MSDeepFilter，可以对质谱数据软件 DIA-NN 报告的肽段数据进行自动且高速的再次分类筛选，从而筛选出高置信度的肽段，为肽段筛选提供了统一标准，替代了繁琐的人工筛选过程。为了测试该模型的分类性能，我们开发了肽段子离子色谱峰可视化工具 getXIC，将 DIA-NN 报告的肽段人工标记成高置信度与低置信度两类，以此创建了包含 86,443 条肽段的基准测试数据集。我们使用该数据集对 MSDeepFilter 的模型进行训练和测试。测试结果表明，MSDeepFilter 在区别高置信度和低置信度肽段方面表现出色，其可以作为人工筛选 DIA-NN 输出肽段的替代方法。

2. 建立数据集

我们建立了高质量的基准数据集来优化深度学习模型的参数。基准数据集的原始 DIA 质谱数据来自不同物种、不同质谱仪生产厂商的质谱仪。我们使用 DIA-NN 作为主要的定量分析软件，并加入了传统定量工作流 OpenSWATH-PyProphet-TRIC (OSPT)的分析结果来增强深度学习模型的泛化能力。

2.1. 数据集来源

表 1 记录了数据集来源信息，包含数据集名称、质谱仪型号、质谱仪所属公司和所使用的数据分析工作流。按照名称对数据集的基础实验信息进行介绍：

Table 1. Information of dataset.
表 1. 数据集来源信息

数据集名称	质谱仪型号	生产厂商	数据分析软件
Yeast_NN	TTOF 6600	ABSciex	DIA-NN
Human_NN	Fusion Lumos	Thermo Fischer	DIA-NN
E.coli_NN	TTOF 6600	ABSciex	DIA-NN
L929_NN	TTOF 5600	ABSciex	DIA-NN
HYE124_NN	TTOF 5600/6600	ABSciex	DIA-NN
HYE110_NN	TTOF 5600/6600	ABSciex	DIA-NN
Yeast_OSPT	TripleTOF 5600	ABSciex	OSPT workflow
SGS_OSPT	TripleTOF 5600	ABScex	OSPT workflow
Hela_OSPT	Q Exactive HF-X	Thermo Fischer	OSPT workflow
E.coli_OSPT	TTOF 6600	ABSciex	OSPT workflow

Continued

L929_OSPT	TripleTOF 5600	ABSciex	OSPT workflow
BGS_OSPT	Fusion Lumos	Thermo Fischer	OSPT workflow

注：第一列为数据集的名字；第二列为采集该数据集原始 DIA 数据的质谱仪名称；第三列为质谱仪的生产厂商；第四列为处理该数据集的分析软件；其中第二列仪器中的 TTOF 为质谱仪 TripleTOF 简称，Lumos 为质谱仪 Orbitrap Fusion Lumos Tribrid 的简称；第四列分析软件中的 OSPT workflow 为 OpenSWATH-PyProphet-TRIC 的简称。

数据集 Yeast_NN (下载地址：http://www.proteomexchange.org，PXD031160)是由质谱仪 TripleTOF 6600 (生产厂商：ABSciex)获取的酵母样品数据。一级质谱质荷比范围为 400~1250，包含 40 个可变窗口；数据集 Human_NN 是由质谱仪 Orbitrap Fusion Lumos Tribrid (生产厂商：Thermo Fischer Scientific)获取的人类样品数据[22]，一级质谱质荷比范围为 400~1250，包含 30 个可变窗口；数据集 E. coli_NN 是由质谱仪 TripleTOF 6600 获得的大肠杆菌样品[23]，一级质谱质荷比范围为 400~1250，包含 100 个可变窗口；数据集 HYE124_NN 和 HYE110_NN 是由 TripleTOF 5600 或 TripleTOF 6600 质谱仪获得的人类、酵母和大肠杆菌的混合样品数据[24]，一级质谱质荷比范围为 400~1200，包含 32 个或 64 个窗口，窗口大小固定或可变；数据集 L929_NN 是由 TripleTOF 5600 质谱仪以 SWATH 模式获得的小鼠样品数据[9]，一级质谱质荷比范围为 400~1150，包含 100 个可变窗口，一式三份重复样品。这 6 个数据集经过 DIA-NN 处理且人工贴好标签的肽段总量为 42,443，其中高置信度肽段 22,306 个，低置信的肽段 20,137 个。

数据集 Yeast_OSPT (下载地址：http://www.proteomexchange.org，PXD028735)是由质谱仪 TripleTOF 5600 获得的酵母样品数据，一级质谱质荷比范围为 400~1200，包含 64 个固定窗口；数据集 SGS_OSPT 是由质谱仪 TripleTOF 5600 获得 SGS [5]人类样品数据，一级质谱质荷比范围为 400~1200，包含 32 个固定窗口；数据集 Hela_OSPT 是由质谱仪 Q Exactive HF-X (生产厂商：Thermo Fischer Scientific)获得的 HeLa [25]细胞样品数据，一级质谱质荷比范围为 350~1650，包含 45 个窗口；数据集 E. coli_OSPT 是由质谱仪 TripleTOF6600 获得的大肠杆菌样品数据[23]，一级质谱质荷比范围为 400~1250，包含 100 个可变窗口；数据集 L929_OSPT 是由 TripleTOF 5600 质谱仪以 SWATH 模式获得的小鼠样品数据[11]，一级质谱质荷比范围为 400~1150，包含 100 个可变窗口。数据集 BGS_OSPT 是由 Orbitrap Fusion Lumos 质谱仪获得的 BGS [26]小鼠样品数据，一级质谱质荷比范围为 350~1650，包含 40 个窗口；数据集 HYE110_OSPT 和 HYE124_OSPT 是由 TripleTOF 5600 或 TripleTOF 6600 质谱仪获得的人类、酵母和大肠杆菌的混合样品数据[24]，一级质谱质荷比范围为 400~1200，包含 32 或 64 个窗口，窗口大小固定或可变。这 8 个数据集经过 OpenSWATH-PyProphet-TRIC 工作流分析输出结果肽段后，经过人工贴标签的肽段总量为 44,316，其中高置信度肽段 22,744 个，低置信的肽段 21,572 个。

因此，本文建立数据集总样本数为 86,759，其中高置信度肽段 45,050 个，低置信的肽段 41,709 个。

2.2. 软件和参数设置

DIA 质谱数据分析流程为：首先实验人员制备所需样品，然后通过质谱仪采集 DIA 数据，得到原始数据，同时在 uniprot 蛋白质数据库中下载对应物种的蛋白质序列数据库，之后将原始数据和对应物种的蛋白质序列数据库一同输入分析工作流(图 2(A))。工作流 DIA-NN 或 OpenSWATH-PyProphet-TRIC 对数据进行处理，输出报告中包含识别得到的肽段以及对应的量化结果，最后 MSDeepFilter 算法筛选得到高置信肽段(图 2(B))。

目前，DIA-NN 是基于深度学习的 DIA 质谱数据分析领域使用最广泛的工作流之一。OpenSWATH-PyProphet-TRIC 工作流为使用最广泛的传统工作流，结果经过大量实验验证。接下来介绍在数据准备中，两个定量工作流中软件的参数设置。

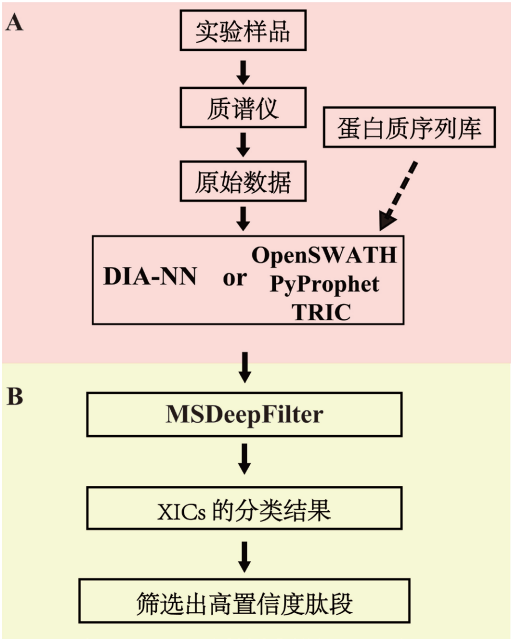


Figure 2. The workflow of DIA-NN & OpenSWATH-PyProphet-TRIC (A) and the workflow of MSDeepFilter (B)
图 2. DIA-NN 和 OpenSWATH-PyProphet-TRIC 的工作流程(A)和 MSDeepFilter 的工作流程(B)

2.2.1. DIA-NN 工作流

由于 DIA 原始文件来源于不同的质谱仪器，导致原始数据文件格式不同，常见的原始格式为.wiff 和.raw 格式。DIA-NN 需要配合 MSConvert (V.3.0.19311)和 Thermo MS File Reader (3.0 SP3)软件来读取这两种格式的数据文件。我们将 DIA 数据和蛋白质序列数据库作为 DIA-NN (version: 1.7.11)的输入，然后设置一级质谱质荷比范围 400~1200，二级质谱质荷比范围 100~1800，设置 FDR 为 1%，其他参数为默认值，最后运行 DIA-NN 获得定量输出报告(图 2(A))。

2.2.2. OpenSWATH-PyProphet-TRIC 工作流程

首先，我们使用 MSConvert (V.3.0.19311)将质谱数据的原始文件转换为 mzXML 文件，并采用 DIA-Umpire 生成伪 DDA 谱图的 mgf 文件。然后，我们以 UniprotKB/Swiss-Prot 数据库为参考库，使用 TPP (Trans-Proteomic Pipeline, V5.1.0)软件包中的 Comet (V2017.01)和 X!Tandem (V2013.06.15.1，模式为 native 和 k-score)搜索引擎对伪 DDA 文件进行数据库搜索，并输出 pep.xml 格式的搜索结果。下一步，我们使用参数设置为-p0.05 -l7 -PPM -OAdPE -dDECOY 的 PeptideProphet 和参数设置为 DECOY = DECOY 的 iProphet 对该搜索结果进行验证和评分。Mayu (V 1.07)用于确定对应于 1%肽 FDR 的 iProphet 概率。通过 1% FDR 筛选得到的肽离子被输入到 SpectraST 中，用于生成定量需要的库文件 sptxt。spectrast2spectrast_irt.py 脚本(可从 http://openswath.org/en/latest/下载)将 sptxt 文件中肽段的保留时间替换为 iRT 时间，其中用于保留时间归一化的 iRT 肽段为内源性肽段。最后，我们使用 spectrast2tsv.py 脚本将 sptxt 文件转换为 tsv 格式，将其转换为 TraML 格式文件，由 OpenSWATH-PyProphet-TRIC 工作流定量分析。

2.3. 数据集构成和预处理

我们得到 DIA-NN 和 OpenSWATH-PyProphet-TRIC 工作流的输出报告后，通过脚本 getXIC.py 从 DIA

原始数据中提取子离子色谱峰组作为 MSDeepFilter 的模型的输入数据(图 2(B))。由于 DIA-NN 输出的碎片子离子的数量不固定，根据经验(传统人工检查以及常用的数据分析软件计算色谱曲线相关性公式时都使用的强度 Top 6 的碎片子离子)，我们提取强度最高的六个作为子离子峰组成员。通过测试 40、55、70、85、90、105、120 等不同曲线长度的数据，在比较分类性能指标后，我们最终决定取 85 个时间点作为时间长度。MSDeepFilter 的输入数据为 6 条长度为 85 的碎片子离子 XICs 构成 6 行 85 列的矩阵。每条 XIC 以强度最高点的时间作为中心点，向前取 42 个时间点的强度数据，向后取 42 个时间长度的强度数据，然后利用 sklearn 库中的 minmax 函数将强度归一化至 0 到 1 之间，以消除数量级影响。归一化后的两类样本数据如图 3 所示。

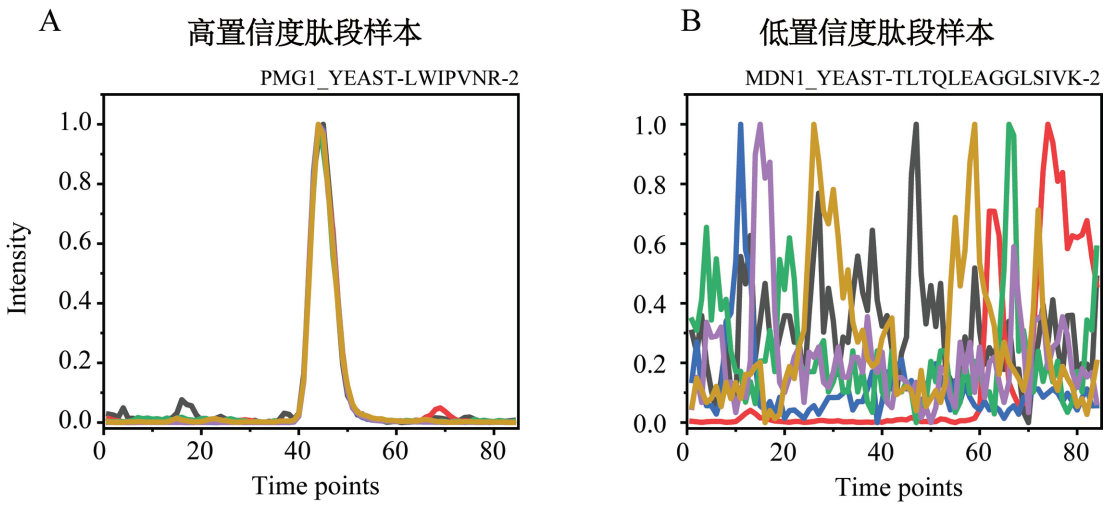


Figure 3. The XICs of High confidence PMG1_YEAST_LWIPVNR and (A) and the XICs of low confidence peptide MDN1_YEAST-TLTQLEAGGLSIVK (B)
图 3. 高置信度肽段 PMG1_YEAST_LWIPVNR 的子离子提取色谱峰组的 XICs 图(A)和低置信度肽段 MDN1_YEAST-TLTQLEAGGLSIVK 的 XICs 图(B)

Minmax 归一化函数公式为：

$$X_{\text{normalization}} = \frac{X_i - \min(X_i)}{\max(X_i) - \min(X_i)}$$

归一化之后对数据进行人工贴标签，数据标签为两类：高置信度肽段和低置信度肽段。贴标签标准为：六条子离子标准化后的色谱峰形状相似则判断为高置信度肽段，反之判断为低置信度肽段。

MSDeepFilter 的基准测试数据集包含 86,443 个肽段，其中高置信度肽段数量为：40,641，低置信度肽段数量为：45,802，将该数据以 6:2:2 的比例随机分为训练集，交叉验证集和测试集，数量信息如表 2 所示。

Table 2. Dataset Information
表 2. 数据集数量信息

数据集	正样本	负样本	总数
训练集	24,385	27,482	51,867
交叉验证集	8128	9160	17,288
测试集(分布一致)	8128	9160	17,288

3. 算法

MSDeepFilter 的模型原理

为了解决分类问题，我们结合残差网络[27]原理、压缩激励模型[28]和自注意力机制模块[29]原理设计出神经网络结构作为 MSDeepFilter 的模型。

为了评估 MSDeepFilter 模型的性能，我们在测试集上应用两种机器学习模型：AdaBoosting [30]和 SVM [31]，对子离子 XICs 进行分类，并比较分类性能，这两个模型的参数都是人工调参得到的。

MSDeepFilter 的神经网络框架基于残差网络，结合压缩激励模型和自注意力机制设计构成，可以视作三个部分连接组成。模型第一部分包含一个卷积核大小为 1 * 1 的卷积层。第二部分为残差结构，残差结构包含两条传播路线，残差结构后并行连接压缩激励模块和自注意力机制模块，第三个部分包含一个为神经元个数为 16，激活函数为 Relu 的全连接层，最后的输出层为神经元个数为 1，激活函数为 Sigmoid 的全连接层，最后输出层输出结果(图 4)。

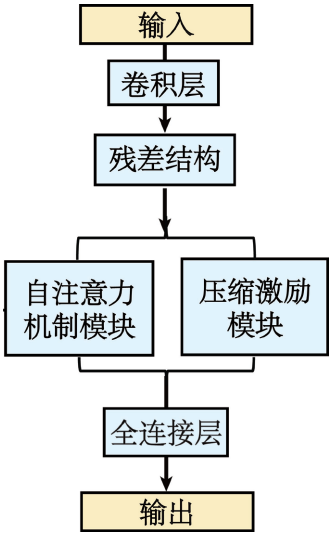


Figure 4. The proposed neural network of MSDeepFilter
图 4. MSDeepFilter 的模型原理

神经网络输入数据为色谱信息，设计为图像类型结构的三维数据，即(1 × 6 × 85)的三维矩阵，输出为一个实数，大小范围为[0, 1]，代表输入肽段为高置信度肽段的预测概率。

MSDeepFilter 模型中，卷积层擅长处理多维数组，自动学习给定数据的潜在空间相关性，残差结构中第二条传播路线为跳跃连接的变形，最大程度的保留了数据的原始信息，减轻由于网络结构导致的信息损失带来的模型退化问题。模型中的压缩激励模块通过显式地建模通道之间的相互依赖性，自适应地重新校准通道特征响应，提高模型的性能。(图 4 中压缩激励模块)。

自注意力机制模块关注数据内部的相关性，使模型聚焦重点信息，有助于提高模型性能(图 4 中自注意力机制模块)。自注意力机制公式为：

$$\mathrm{Attention}\left(X\right)=\mathrm{softmax}\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V$$

$$Q=W^{Q}X,\ K=W^{K}X,\ V=W^{V}X$$

其中的矩阵阵 W^Q、W^K、W^V 分别为 Q、K、V 的权重矩阵，为可训练参数，d_k 为常数，这里我们设置为为通道值，防止 softmax 输入值过高，导致偏导数趋于 0。

我们使用 Adam 优化器训练 MSDeepFilter 模型的神经网络，训练批量大小为 256，参数 beta1 为 0.9，参数 beta2 为 0.999，参数 epsilon 为 1e−5，权重衰减为 1e−5，学习率为 3e−6，训练次数设置为 100，损失函数为二元分类的交叉熵损失。

机器学习模型的输入数据维度均为一维，单个样本为大小为 6 * 85 = 510 的一维数据。

支持向量机(Support Vector Machine, SVM)是一种机器学习方法，可以最大化训练模式和决策边界之间的边距(图 5(A))。我们训练了 sklearn 库里的 sklearn.svm.SVC 模型，将“kernel”参数设置为“rbf”，“gamma”参数设置为 1e−6，参数“C”设置为 1e−6，其他参数均设置为默认值。

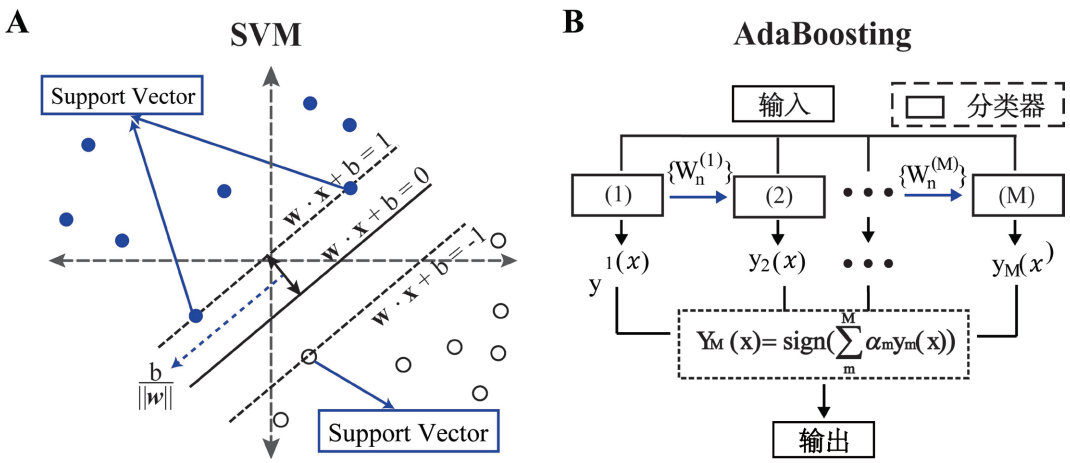


Figure 5. The model of SVM (A) and the model of AdaBoosting (B)
图 5. 支持向量机原理图(A)和 AdaBoosting 原理图(B)

AdaBoosting 属于集成学习中的 boosting 方法，首先，用初始权重训练得到第一个弱学习分类器，根据弱学习器的学习误差率来更新训练样本的权重，提高使第一个弱分类器学习误差率变高的训练样本的权重，让后续的弱分类器更加重视这些使误差率高的样本。然后，用调整权重之后的训练数据集训练第二个弱分类器，重复训练其余的弱学习器，直到弱分类器数量达到预先设置的值。最后，将所有弱分类器组合得到一个性能更好的分类器(图 5(B))。我们训练了 sklearn 库里的 adaboosting 模型，将基础弱分类器设置为决策树，决策树参数最大深(max_depth)设置为 2，内部节点再划分所需最小样本数(min_samples_split)设置为 20，叶节点所需的最小样本数(min_samples_leaf)设置为 5；集成算法设置为“SAMME”，基分类器提升(循环)次数(n_estimators)设置为 200，学习率(learning_rate)设置为 0.8，将其他参数值设置为默认值。

4. 训练测试结果

4.1. MSDeepFilter 的模型训练过程

MSDeepFilter 的深度学习模型的训练损失函数为结合了 Sigmoid 激活函数的二元分类的交叉熵损失，其公式为：

$$L = -\sum_i \text{label}_i * \log(\text{pred}_i) + (1 - \text{label}_i) * \log(1 - \text{pred}_i)$$

其中 pred 为算法输出的预测值，label 是人工标签 0 或者 1。

从图 6 可看到 MSDeepFilter 的模型训练次数在 80 次之后，交叉验证集的损失趋于平稳，最终训练次数为 100 次，训练结果未出现过拟合或欠拟合问题，说明模型训练成功。

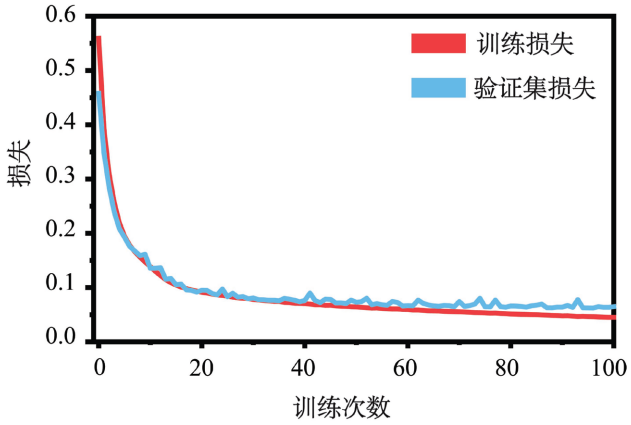
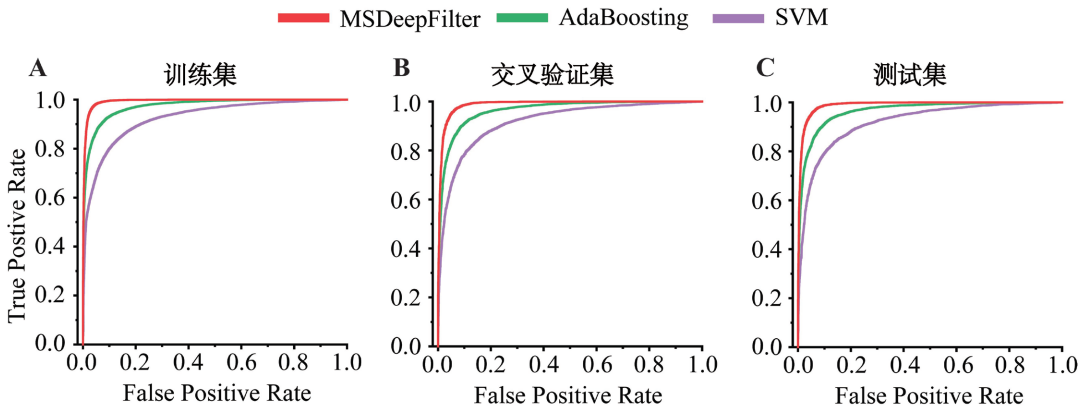


Figure 6. The losses of the proposed networks on the dataset
图 6. 模型训练损失随训练次数变化曲线

4.2. 不同模型在数据集上训练测试过程输出的 ROC 曲线和 P/R 曲线

我们将 MSDeepFilter 的模型、AdaBoosting 和 SVM 应用于训练集、交叉验证集和测试集上，得到 ROC 曲线。ROC 图中，横轴为假阳性率，纵轴为真阳性率，曲线下面积(AUC)越大代表模型分类性能越好。在训练集、交叉验证集和测试集上，MSDeepFilter 的模型的 ROC 曲线包裹住了其他所有模型的 ROC 曲线(图 7)，且它的测试集 AUC、训练集 AUC 和交叉验证集 AUC 分别为 0.9873、0.9979 和 0.9864，为所有模型中对应指标的最高值(表 3)，代表 MSDeepFilter 的模型在训练集、交叉验证集和测试集上的性能表现最好。



注：不同颜色的曲线代表不同模型的性能表现。红色的线代表 MSDeepFilter，绿色的线代表 AdaBoosting，紫色线代表 SVM。(A) 训练集上 3 个模型的 ROC 曲线图；(B) 交叉验证集上 3 个模型的 ROC 曲线图；(C) 测试集上 3 个模型的 ROC 曲线图。

Figure 7. ROC curves of different models
图 7. 模型的 ROC 曲线图

4.3. 模型性能评价指标

我们将三个模型应用于训练集、交叉验证集和测试集上，获取得到的性能指标如表 3 所示。

Table 3. Performance indicators of models
表 3. 模型的性能指标

Metrics\Model	AUC on Testing Set	AUC on Training Set	AUC on Validation Set	Precision	Recall	F1 score	ACC
AdaBoosting	0.9652	0.9735	0.9641	0.9087	**0.9097**	**0.9092**	0.9057
SVM	0.9195	0.9252	0.9175	0.9099	0.7608	0.8286	0.8367
MSDeepFilter	**0.9873**	**0.9979**	**0.9864**	**0.9729**	0.8934	0.9315	**0.9317**

注：AUC on Testing Set 代表模型应用在测试集上得到的 AUC 值，同理，AUC on Training Set 和 AUC on Validation Set 分别代表模型应用在训练和交叉验证集上得到的 AUC 值；黑色加粗代表每个指标里的最高值。

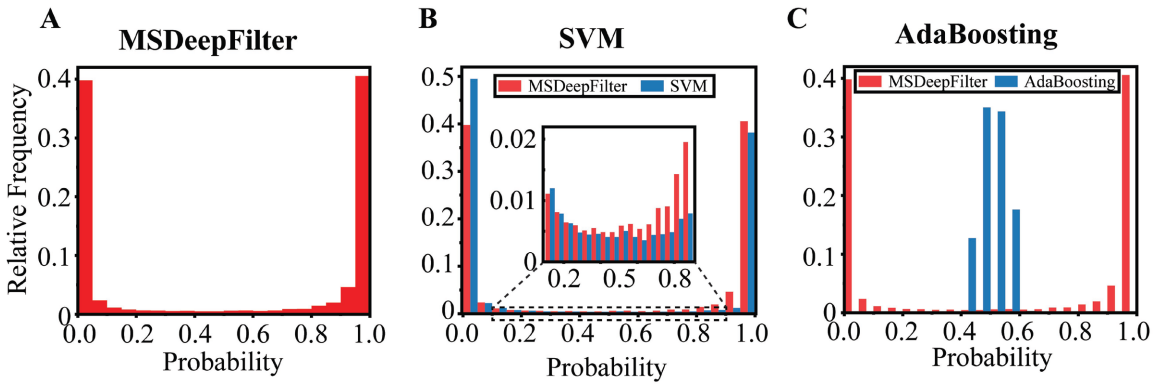
表 3 记录了训练测试中三种模型的七个性能指标。从中可以发现 MSDeepFilter 的七个指标中六个指标高于 0.9，其中测试集 AUC、训练集 AUC、交叉验证集 AUC 三个指标高于 0.95。支持向量机在测试集上的召回率(Recall)、F1 分数、准确率(ACC)分别只有 0.7608、0.8286、0.8367，表现较差；AdaBoosting 算法在三个数据集上的 AUC 值分别为0.9652、0.9735、0.9641，但是精确率和准确率分别为0.9087 和0.9057，低于 MSDeepFilter 的模型表现。

MSDeepFilter 的模型的七个指标中五个指标：测试集 AUC、训练集 AUC、交叉验证集 AUC、精确率和准确率分别为 0.9873、0.9979、0.9864、0.9729、0.9317，均为对应指标中的最高值，且它的召回率和 F1 分数分别为 0.8934 和 0.9315，所以综合考虑多指标，在三个模型中，MSDeepFilter 的模型分类性能最优。

4.4. 模型的概率分布直方图以及对比

二分类问题的结果输出概率分布图中，两侧分布的直方越高，中间部分直方图越低，代表对两类样本的分辨能力越强。

MSDeepFilter 的模型预测概率直方主要分布于两侧，对正负样本具有优秀的分辨能力；支持向量机模型预测概率直方主要分布于两侧，中间分布直方略低于 MSDeepFilter，正负样本区分能力高。但是 0~0.1 区间分布占总数约 51.7%，分布呈现负样本更多，与数据集真实标签正样本分布更多的事实不符(图 8(B))；AdaBoosting 模型概率直方集中于 0.4~0.6 区间(图 8(C))，虽然模型准确率达到 0.9057 (表 3)，但从概率直方分布发现，AdaBoosting 模型对两类样本的分辨能力差(图 8(C))。



注：图里所有红色直方均代表 MSDeepFilter 的概率分布。(A) 红色代表 MSDeepFilter 的分类概率分布直方；(B) 深蓝色为机器学习模型支持向量机(SVM)的分类概率分布直方；(C) 深蓝色为机器学习模型 AdaBoosting 的分类概率分布直方。

Figure 8. Histogram of model probability distribution
图 8. 模型概率分布直方图

综合上述分析，在三个模型中，MSDeepFilter 的模型在区间 0.1~0.9 中的直方分布最少，对两类样本的分辨能力最强，性能表现最好。

5. 讨论与总结

本文基于深度学习设计了神经网络作为算法 MSDeepFilter 的模型，可筛选出 DIA-NN 输出结果中的高置信度肽段。MSDeepFilter 先提取 DIA 实验数据中肽段对应的碎片子离子 XICs 进行归一化处理，然后利用深度神经网络对 XICs 进行分类，最后筛选出高置信度肽段。

在交叉验证数据集和测试集上，MSDeepFilter 的模型性能优于传统机器学习算法 AdaBoosting 和支持向量机 SVM。然而，由于 DIA 数据本身极其复杂和高通量的特点，该模型仍有很多可以完善的地方。例如，数据预处理的方法可以根据不同的质谱仪进行设计；测试数据可以加入更多质谱生产厂商的质谱仪处理的 DIA 数据；神经网络的结构可以设计更多不同类型的网络。

基金项目

本研究由国家自然科学基金(项目编号：12090052，11874310)提供资助。

参考文献

[1] Zhang, Y., Fonslow, B.R., Shan, B., Baek, M.-C. and Yates, J.R. (2013) Protein Analysis by Shotgun/Bottom-Up Proteomics. *Chemical Reviews*, **113**, 2343-2394. https://doi.org/10.1021/cr3003533

[2] Li, X., Zhong, C.Q., Wu, R., Xu, X., Yang, Z.H., Cai, S., *et al*. (2021) RIP1-Dependent Linear and Nonlinear Recruitments of Caspase-8 and RIP3 Respectively to Necrosome Specify Distinct Cell Death Outcomes. *Protein & Cell*, **12**, 858-876. https://doi.org/10.1007/s13238-020-00810-x

[3] Li, X., Zhong, C.Q., Yin, Z., Qi, H., Xu, F., He, Q. and Shuai, J. (2020) Data-Driven Modeling Identifies TIRAP-Independent MyD88 Activation Complex and myddosome Assembly Strategy in LPS/TLR4 Signaling. *International Journal of Molecular Sciences*, **21**, Article 3061. https://doi.org/10.3390/ijms21093061

[4] Zhong, C.Q., Wu, R., Chen, X., Wu, S., Shuai, J. and Han, J. (2019) Systematic Assessment of the Effect of Internal Library in Targeted Analysis of SWATH-MS. *Journal of Proteome Research*, **19**, 477-492. https://doi.org/10.1021/acs.jproteome.9b00669

[5] Röst, H.L., Rosenberger, G., Navarro, P., Gillet, L., Miladinović, S.M., Schubert, O.T., Wolski, W., Collins, B.C., Malmström, J., Malmström, L. and Aebersold, R. (2014) OpenSWATH Enables Automated, Targeted Analysis of Data-Independent Acquisition MS Data. *Nature Biotechnology*, **32**, 219-223. https://doi.org/10.1038/nbt.2841

[6] Keller, A., Bader, S.L., Shteynberg, D., Hood, L. and Moritz, R.L. (2015) Automated Validation of Results and Removal of Fragment Ion Interferences in Targeted Analysis of Data-independent Acquisition Mass Spectrometry (MS) Using SWATHProphet. *Molecular & Cellular Proteomics*, **14**, 1411-1418. https://doi.org/10.1074/mcp.O114.044917

[7] Peckner, R., Myers, S.A., Jacome, A. S.V., Egertson, J.D., Abelin, J.G., MacCoss, M.J., Carr, S.A. and Jaffe, J.D. (2018) Specter: Linear Deconvolution for Targeted Analysis of Data-Independent Acquisition Mass Spectrometry Proteomics. *Nature Methods*, **15**, 371-378. https://doi.org/10.1038/nmeth.4643

[8] Tsou, C.C., Avtonomov, D., Larsen, B., Tucholska, M., Choi, H., Gingras, A.C. and Nesvizhskii, A.I. (2015) DIA-Umpire: Comprehensive Computational Framework for Data-Independent Acquisition Proteomics. *Nature Methods*, **12**, 258-264. https://doi.org/10.1038/nmeth.3255

[9] Li, Y., Zhong, C.Q., Xu, X., Cai, S., Wu, X., Zhang, Y., *et al*. (2015) Group-DIA: Analyzing Multiple Data-Independent Acquisition Mass Spectrometry Data Files. *Nature Methods*, **12**, 1105-1106. https://doi.org/10.1038/nmeth.3593

[10] Meyer, J.G., Mukkamalla, S., Steen, H., Nesvizhskii, A.I., Gibson, B.W. and Schilling, B. (2017) PIQED: Automated Identification and Quantification of Protein Modifications from DIA-MS Data. *Nature Methods*, **14**, 646-647. https://doi.org/10.1038/nmeth.4334

[11] Bruderer, R., Bernhardt, O.M., Gandhi, T., Miladinović, S.M., Cheng, L.Y., Messner, S., *et al*. (2015) Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues. *Molecular & Cellular Proteomics*, **14**, 1400-1410. https://doi.org/10.1074/mcp.M114.044305

[12] Gessulat, S., Schmidt, T., Zolg, D.P., Samaras, P., Schnatbaum, K., Zerweck, J., *et al*. (2019) Prosit: Proteome-Wide Prediction of Peptide Tandem Mass Spectra by Deep Learning. *Nature Methods*, **16**, 509-518. https://doi.org/10.1038/s41592-019-0426-7

[13] Tran, N.H., Qiao, R., Xin, L., Chen, X., Liu, C., Zhang, X., Shan, B., Ghodsi, A. and Li, M. (2019) Deep Learning Enables de Novo Peptide Sequencing from data-Independent-Acquisition Mass Spectrometry. *Nature Methods*, **16**, 63-66. https://doi.org/10.1038/s41592-018-0260-3

[14] 何情祖, 钟传奇, 李翔, 帅建伟, 韩家淮. 数据不依赖获取的质谱数据的深度学习分析方法[J]. 厦门大学学报(自然科学版), 2021, 60(1): 97-103.

[15] He, Q., Zhong, C.Q., Li, X., Guo, H., Li, Y., Gao, M., *et al*. (2022) Dear-DIA^{XMBD}: Deep Autoencoder for Data-Independent Acquisition Proteomics. (Preprint) https://doi.org/10.1101/2022.08.27.505516

[16] Gao, M., Yang, W., Li, C., Chang, Y., Liu, Y., He, Q., Zhong, C.-Q., Shuai, J., Yu, R. and Han, J. (2021) Deep Representation Features from DreamDIA^{XMBD} Improve the Analysis of Data-Independent Acquisition Proteomics. *Communications Biology*, **4**, Article No. 1190. https://doi.org/10.1038/s42003-021-02726-6

[17] Demichev, V., Messner, C.B., Vernardis, S.I., Lilley, K.S. and Ralser, M. (2020) DIA-NN: Neural Networks and Interference Correction Enable Deep Proteome Coverage in High Throughput. *Nature Methods*, **17**, 41-44. https://doi.org/10.1038/s41592-019-0638-x

[18] MacLean, B., Tomazela, D.M., Shulman, N., Chambers, M., Finney, G.L., Frewen, B., *et al*. (2010) Skyline: An Open Source Document Editor for Creating and Analyzing Targeted Proteomics Experiments. *Bioinformatics*, **26**, 966-968. https://doi.org/10.1093/bioinformatics/btq054

[19] Sturm, M. and Kohlbacher, O. (2009) TOPPView: An Open-Source Viewer for Mass Spectrometry Data. *Journal of proteome research*, **8**, 3760-3763. https://doi.org/10.1021/pr900171m

[20] Li, Y., He, Q., Guo, H., Zhong, C.Q., Li, X., Li, Y., Han, J. and Shuai, J. (2022) MSSort-DIA^{XMBD}: A Deep Learning Classification Tool of the Peptide Precursors Quantified by OpenSWATH. *Journal of Proteomics*, **259**, Article ID: 104542. https://doi.org/10.1016/j.jprot.2022.104542

[21] Gupta, S., Sing, J., Mahmoodi, A. and Röst, H. (2020) DrawAlignR: An Interactive Tool for across Run Chromatogram Alignment Visualization. *Proteomics*, **20**, Article ID: 1900353. https://doi.org/10.1002/pmic.201900353

[22] Tatjana, V., Domitille, S. and Jean-Charles, S. (2021) Paraquat-Induced Cholesterol Biosynthesis Proteins Dysregulation in Human Brain Microvascular Endothelial Cells. *Scientific Reports*, **11**, Article No. 18137. https://doi.org/10.1038/s41598-021-97175-w

[23] Midha, M.K., Kusebauch, U., Shteynberg, D., Kapil, C., Bader, S.L., Reddy, P.J., *et al*. (2020) A Comprehensive Spectral Assay Library to Quantify the Escherichia Coli Proteome by DIA/SWATH-MS. *Scientific Data*, **7**, Article No. 389. https://doi.org/10.1038/s41597-020-00724-7

[24] Navarro, P., Kuharev, J., Gillet, L.C., Bernhardt, O.M., MacLean, B., Röst, H.L., *et al*. (2016) A Multicenter Study Benchmarks Software Tools for Label-Free Proteome Quantification. *Nature Biotechnology*, **34**, 1130-1136. https://doi.org/10.1038/nbt.3685

[25] Muntel, J., Gandhi, T., Verbeke, L., Bernhardt, O.M., Treiber, T., Bruderer, R. and Reiter, L. (2019) Surpassing 10000 Identified and Quantified Proteins in a Single Run by Optimizing Current LC-MS Instrumentation and Data Analysis Strategy. *Molecular Omics*, **15**, 348-360. https://doi.org/10.1039/C9MO00082H

[26] Muntel, J., Kirkpatrick, J., Bruderer, R., Huang, T., Vitek, O., Ori, A. and Reiter, L. (2019) Comparison of Protein Quantification in a Complex Background by DIA and TMT Workflows with Fixed Instrument Time. *Journal of Proteome Research*, **18**, 1340-1351. https://doi.org/10.1021/acs.jproteome.8b00898

[27] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[28] Hu, J., Shen, L. and Sun, G. (2018) Squeeze-and-Excitation Networks. *Proceedings of the* 2018 *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7132-7141. https://doi.org/10.1109/CVPR.2018.00745

[29] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., *et al*. (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems*, **30**, 5998-6008.

[30] Freund, Y. and Schapire, R.E. (1997) A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, **55**, 119-139. https://doi.org/10.1006/jcss.1997.1504

[31] Chen, P.H., Lin, C.J. and Schölkopf, B. (2005) A Tutorial on ν-Support Vector Machines. *Applied Stochastic Models in Business and Industry*, **21**, 111-136. https://doi.org/10.1002/asmb.537