DOI: 10.13376/j.cbls/2023170
文章编号：1004-0374(2023)12-1561-09

朱云平，军事科学院军事医学研究院生命组学研究所研究员，博士生导师，国家蛋白质科学中心生物信息学实验室 PI。曾任中国医药生物技术协会生物医学信息技术分会副主任委员，863 重大项目首席专家。研究方向为蛋白质组生物信息学、生物医学大数据挖掘。研发了蛋白质组信息学系列算法，建立的 iProX 是国际学术界公认的蛋白质组数据共享平台。发表 SCI 论文两百余篇；获软件著作权三十余项，中国发明专利 5 项。获中国电子信息科学技术奖一等奖、中国发明协会发明创新奖一等奖，以及国家科技进步创新团队奖、北京市科学技术奖一等奖、中华预防医学科技奖一等奖、军队教学成果一等奖等。

# 蛋白质组学大数据研究进展

刘 祎[1,2,3]，朱云平[2,3]*
(1 北京工业大学环境与生命学部，北京 100124；2 军事科学院军事医学研究院生命组学研究所，国家蛋白质科学中心(北京)，北京 102206；3 北京蛋白质组研究中心，北京 102206)

**摘 要：**随着研究问题的深入和技术的发展，蛋白质组学研究逐渐迈向大数据时代。数据规模的扩大可以为研究人员发现更稳定可靠的结论提供坚实的基础，但也对数据的存储和分析等环节提出了更多的挑战。本文首先介绍了蛋白质组学数据的特点，然后主要从蛋白质组学大数据相关的数据库和分析方法两方面总结目前的研究进展，最后对该领域存在的挑战和机遇进行展望。

**关键词：**蛋白质组学；大数据；数据库；算法；软件
**中图分类号：**G250.74；Q81 **文献标志码：**A

# Advances of big data in proteomics

LIU Yi[1,2,3], ZHU Yun-Ping[2,3]*
(1 Faculty of Environment and Life, Beijing University of Technology, Beijing 100124, China; 2 National Center for Protein Sciences (Beijing), Beijing Institute of Life Omics, Beijing 102206, China; 3 Beijing Proteome Research Center, Beijing 102206, China)

**Abstract:** With the advance of research topics and development of technologies, proteomics research is gradually moving towards the era of big data. The expansion of data scale can provide a solid foundation for researchers to find more stable and reliable conclusions, but it also poses more challenges to data storage and data analysis. In this paper we first introduce the characteristics of proteomics data, then summarize the current research progress from the two aspects of proteomics big data-related databases and analysis methods, and finally look forward to the challenges and opportunities in this field.

**Key words:** proteomics; big data; database; algorithm; software

**收稿日期：**2023-02-01；**修回日期：**2023-03-02
**基金项目：**国家重点研发计划(2021YFA1301603)
***通信作者：**E-mail: zhuyunping@gmail.com

随着信息技术的发展，人类收集、存储和处理数据的能力越来越强，国际数据公司 (International Data Corporation, IDC) 发布的报告 [1] 显示，预计在2025 年，全球每年产生的数据将从 2018 年的 33 ZB 增长到 175 ZB，相当于每天产生 491 EB 的数据。大数据已经成为时代显著的特征，影响了诸多研究和应用领域。在生物医药领域，基因组学大数据研究也推动了整个领域的发展。与基因相比，蛋白质作为生物学过程的直接承担者，能够更准确地反映生理病理过程。蛋白质组学是对蛋白质在时间和空间中的表达、结构和功能的大规模研究。近年来，随着技术发展和研究的深入，蛋白质组学领域研究的规模越来越大；相应地，在数据收集、存储、分析与共享等环节也出现了更多的挑战。本文将从代表性的大规模蛋白质组学研究出发，从数据特点、数据库技术、大数据分析方法这几个方面介绍当前蛋白质组学大数据研究的现状，并对该领域未来的挑战和机遇进行展望。

## 1 代表性的大规模蛋白质组学研究

表 1 [2-9] 列举了部分近年来的大规模蛋白质组学研究，与早期的研究 (2010 年以前 ) 相比，上述研究的规模增长明显，获得了小规模研究难以实现的成果。

除了各个团队独立进行的研究，蛋白质组学领域也不断出现大的合作项目，表 2 列举了部分具有代表性的合作项目。其中，CPTAC 与 ICPC 项目目前都能在 PDC 数据库中查看。迄今为止，PDC 数据库已经收录了 35 TB 的数据。第 21 届国际蛋白质组学大会于 2022 年 12 月 4 日到 8 日在墨西哥坎昆召开，在大会上，贺福初院士做了人类大科学计划的报告，邀请各国科学家参加到人体蛋白质组导航计划 (π-HuB) 中来。该计划预计进行 30 年 (2023–2053)，将完成四大目标：(1) 绘制“空间蛋白质组图”；(2) 追踪以蛋白质组为中心的谱系轨迹；(3) 通过计算建模生成一个元智人虚拟空间；(4) 引导人体远离疾病 / 亚健康，保持健康状态。

这一系列研究预示着蛋白质组学正在朝着大规模、高精度的趋势发展，目标也越来越宏大。可以预见，在未来必将出现更大规模的蛋白质组学研究。随着研究规模的扩大，早期开发的基于小规模样本的蛋白质组学工具或流程出现了处理速度慢、处理效率低及数据共享困难等一系列问题。但要全面认识这些问题，首先需要把握蛋白质组学数据的特点。

## 2 蛋白质组学数据

蛋白质组学数据的特点源于其产出方式，本文在此处先对蛋白质组学分析的基本流程进行简要介绍，然后对蛋白质组学数据的特点进行总结。

### 2.1 蛋白质组学分析流程简介

质谱法是目前蛋白质组学研究的主流技术，其基本流程如图 1，步骤包括样品预处理、质谱分析及数据分析等。样品在经过预处理 ( 包括提取、酶切、除盐等 ) 后进行质谱分析 ( 通常和色谱联用 )，所得数据经过后续分析得到最终结果。蛋白质组学数据包括质谱下机数据及后续分析中产生的所有数据。质谱下机数据处理及后续分析一般包括格式转换、蛋白质鉴定、质量控制、蛋白质定量及后续注释等

**表1 代表性的大规模蛋白质组学研究**

| 研究者 | 日期 | 描述 | 大小(GB) | 参考文献 |
|---|---|---|---|---|
| Jiang等 | 2019 | 早期肝癌研究 | 2 243.1 | [2] |
| Xu等 | 2020 | 肺腺癌研究 | 2 651.2 | [3] |
| Müller等 | 2020 | 跨物种的蛋白质组图谱 | 1 453.1 | [4] |
| Schoof等 | 2021 | 白血病单细胞研究 | 234.6 | [5] |
| Piehowski等 | 2021 | 空间蛋白质组学研究 | 118.1 | [6] |
| Asleh等 | 2022 | 乳腺癌研究 | 394 | [7] |
| Gonçalves等 | 2022 | 949种细胞系 | 3 800.2 | [8] |
| Sun等 | 2022 | 甲状腺结节研究 | 2 660.9 | [9] |

**表2 代表性的蛋白质组合作项目**

| 项目名称 | 启动日期 | 描述 |
|---|---|---|
| HPP | 2003 | 人类蛋白质组计划 |
| CPTAC | 2011 | 临床蛋白质组学肿瘤分析联盟 |
| ICPC | 2016 | 国际癌症蛋白质组学联盟 |
| π-HuB | 2023 | 人体蛋白质组导航计划 |

步骤。

虽然学术界相继提出了 mzXML[10]、mzML[11] 等通用格式，但如表 3 所示，主要的质谱仪器厂商目前均使用各自的商业格式作为仪器输出，这一点增加了蛋白质组数据处理的复杂性。

蛋白质鉴定是鉴定软件从质谱下机数据中识别蛋白质序列的过程，目前常用的软件包括 Mascot[12]、MaxQuant[13]、pFind[14]、MSFragger[15]、PEAKS[16] 及 DIA-NN[17] 等。鉴定算法可以分为数据库搜索策略和从头测序策略。目前数据库搜索是更常见的策略，经过鉴定后，鉴定软件会输出每个下机文件所包含的蛋白质列表。

无论是数据库搜索策略还是从头测序策略，鉴定过程都不能保证正确无误，所以各个软件一般会设置一个阈值，过滤掉得分较低的一些蛋白质序列，这一过程被称为质量控制。也有一些专门用于打分及过滤的质量控制软件。目前最常用的是基于半监督学习的 Percolator[18]。本团队也开发了可以对 Mascot 结果进行质量控制的软件 PepDistiller[19]，由于采用了并行化加速，在速度上优于 Percolator。

在确定包含的蛋白质序列后，定量软件会根据鉴定结果及原始文件中对应的强度值计算每个蛋白的定量值。一些软件如 MaxQuant 和 DIA-NN 整合了鉴定、质量控制与定量过程。也有一些专门的定量软件，如本团队开发的 PANDA[20]。与 MaxQuant 等软件相比，PANDA 定量的准确性更高，且运行速度更快。

在获得蛋白质定量矩阵后，就可以进行后续生物学分析 (GO 分析、聚类分析、差异基因筛选等 )。这一步骤与基因组、转录组等其他组学方法相比并无本质上的区别。

### 2.2　蛋白质组学数据的特点

从以上简介可以看出，质谱数据首先具有信息层级多的特点，要获得有意义的信息需要经过多个步骤的处理，中间结果较多，中间文件格式多样；且目前虽然每个步骤都存在一到两个主流软件，但还没有形成公认的标准处理流程。

另一方面，与基因组测序数据相比，质谱数据具有噪声多、重复性较低的特点。目前通常采用严格的质量控制来减少仪器和实验操作带来的误差，

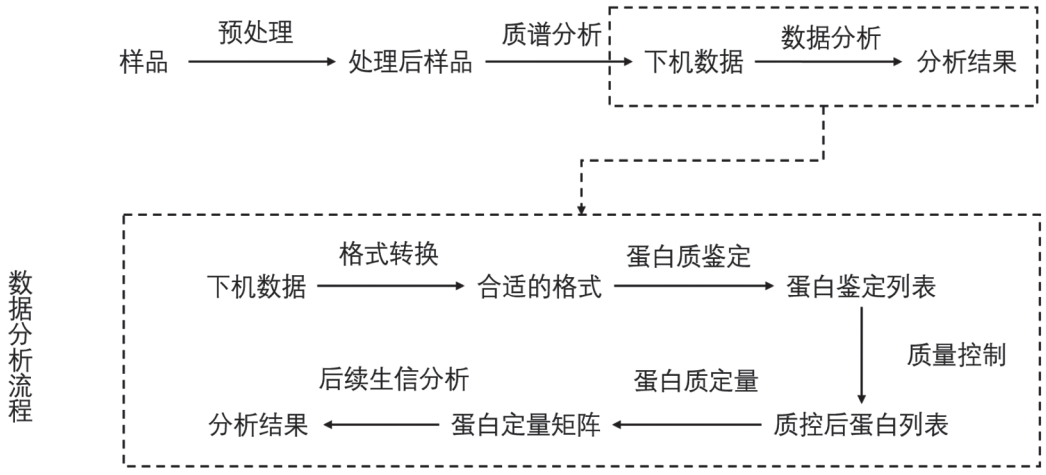


**图1　基于质谱的蛋白质组学分析流程**

**表3　主要质谱仪器厂商及其使用的输出格式**

| 仪器厂商 | 输出格式 |
|---|---|
| ABI | T2D |
| Agilent | MassHunter.d |
| Bruker | Compass.d、YEP、BAF、FID、TDF |
| Sciex | WIFF / WIFF2 |
| Shimadzu | LCD |
| Thermo Scientific | RAW |
| Waters | MassLynx.raw / UNIFI |

缺乏从计算层面对噪声和重复性进行控制的大数据研究。

第三，从数据大小上看，质谱下机数据通常远大于其他数据。1 GB 的原始文件对应的鉴定及定量结果通常不会超过 1 MB。大量的质谱下机数据是目前蛋白质组大数据最直接的体现。不过适用于大数据的分布式数据库、可扩展存储、云计算平台等技术大多能直接应用于存储质谱原始数据的蛋白

质组学数据，无须专门开发特殊的方法，开发人员需要解决的通常是一些细节问题。

在目前的蛋白质组学中，分析质谱原始文件产生的蛋白定量矩阵等数据从规模上还很难称得上大数据，传统的处理方法仍然能满足处理需求，目前也几乎没有研究从大数据的角度分析这一类数据。由于在数据存储等环节目前专门针对蛋白质组数据的研究很少，本文将围绕质谱原始数据的共享和分析这两个成果相对较多的方面介绍目前的研究进展。

## 3 质谱原始数据的共享

数据的共享可以使资源得到有效利用，为了便于数据的共享和重分析，需要在软件和硬件层面提供支持。

在软件层面，数据的标准化是必不可少的。在阿姆斯特丹原则 [21] 及其后续数据质量指标 [22-23] 的推动下，支持快速和开放共享蛋白质组数据的数据政策和指南已经在蛋白质组学界得到实施。在过去十年中，基于质谱的蛋白质组学的许多方面，如数据交换格式、受控词汇表 [24-25] 和报告指南 ( 即 MIAPE[26]) 等都取得了重大进展。人类蛋白质组组织蛋白质组标准委员会 (Human Proteome Organization Proteomics Standards Initiative, HUPO-PSI)[27] 制定并发布了蛋白质组学数据表示的协调标准，包括表示质谱原始数据的 mzML[28] 标准、设计用于报告肽段鉴定结果的 mzIdentML[29] 标准、用于报告定量结果的 mzQuantML[30] 标准和基于文本的格式 mzTab[31]，它可以在简化的概述中显示鉴定和定量结果。所有这些标准格式及其转换工具 [32-35] 都有助于克服蛋白质组学界数据共享的技术挑战。最新的标准化措施中比较重要的是通用质谱标识符 (Universal Spectrum Identifier, USI)，这是一种用于对存放在公共蛋白质组学存储库的数据集中包含的质谱进行重新编码的方法。USI 可以提高谱图证据的透明度 [36]。

在基础设施层面，为了支持数据的共享，Proteome-Xchange 联盟 ( 简称 PX 联盟 ) 于 2006 年成立 [37]。该联盟围绕 PRIDE[38] 和 PeptideAtlas[39]，将之前一些松散的蛋白质组学资源集中了起来。2014 年，美国的 MassIVE 加入 PX 联盟；2016 年，日本的 jPOST 加入了该联盟；2017 年，我国的 iProX[40] 加入该联盟。截至 2019 年 6 月，共有 14 169 个数据集被提交给 ProteomeXchange 联盟成员并获得索引编号 (PXD 编号 )，数据总大小约 1.3 PB。其中，8 638 个数据集 (61%) 已经公开，可供其他研究者下载使用。

本团队开发并负责运行维护的 iProX 于 2017 年 4 月上线，并于 2017 年 11 月加入 PX 联盟。在 2021 年，iProX 又进行了一次较大的版本更新 [41] 来更好地支持大规模蛋白质组学数据的存储与传输。目前的 iProx 可支持 PB 级数据存储、千亿条谱图记录、秒级时延服务能力，满足蛋白质组学数据快速积累带来的需求。截至 2021 年 8 月，已有 1 526 个数据集提交到 iProX，其中公开的数据集 984 个 (64%)，累计数据量 92.42 TB。iProx 由一个具有高扩展性的超融合架构来支持其提交过程，其中，一个 Hadoop 集群用于存储大量的蛋白质组学数据。此外，iProx 还采用了分布式 RESTful 风格的弹性搜索引擎，可在一秒钟内检索数百万条记录。重分析的数据集，例如蛋白质、肽段和谱图数据集，存储在 Hadoop 集群中称为 HBase 的分布式列式存储数据库中，索引则被存储在弹性搜索集群中。iProx 还提供了基于网络和 API 的通用搜索界面，用于获取原始提交数据集的元数据和再分析数据。为了实现独立的高速数据文件传输，基于 Web 和基于 Asepra 的快速上传和下载步骤通过 RESTful 接口重构为独立的传输子服务，即面向服务的架构。搜索元数据以及已识别的蛋白质、肽段和谱图也被包含在子服务中。另外，iProX 的灾备系统和全实时备份站点被设计部署到位于广州 ( 广东，华南地区 ) 的国家超算中心，北京主站点不可用的情况下，广州的备份站点可以在几分钟内接管服务。

ProteomeXchange 联盟的各个成员都不可避免地面临着数据爆发式增长带来的压力。类似的情况大多能从更为成熟的领域 ( 例如电子商务 ) 吸取经验教训，但目前的网络架构无法支撑起 PB 级别的数据传输，一些新技术有可能会被引入来解决大量数据的传输瓶颈。质谱原始数据格式繁多，且无法直接提取有效信息，不利于后续分析。虽然目前数据库一般要求与鉴定结果进行关联，但这种关联比较松散，且用户所使用的鉴定软件各不相同。一种保留足够多信息且压缩比更高的通用格式可能是一种发展趋势，这将有助于数据的整合和重复利用。

## 4 面向大数据的蛋白质组学数据分析方法

面向规模越来越大的蛋白质组学数据，虽然大多数研究仍然采用传统的方法在分析和处理数据，但是也有一些分析方法已经开始着眼于蛋白质组学

大数据。这一系列方法主要体现在两方面：一是为了应对海量的原始数据，基于集群并采用并行化加速方法来处理质谱原始数据；二是基于大量蛋白质组学数据开发的一系列深度学习方法。

### 4.1 并行化加速方法

蛋白质序列的鉴定过程一直是蛋白质组学数据处理的瓶颈，所以基于集群的并行化加速目前主要体现在一系列非商业或商业搜库工具上，数据库搜索策略及从头测序策略都有涉及。

#### 4.1.1 非商业工具

在非商业软件中，开源的肽段数据库搜索软件X!Tandem[42]最早被应用于并行化加速研究。标准的X!Tandem包含一个线程模型，允许它在多个处理器内核上分散单个搜索的工作。Parallel Tandem项目[43]将搜索任务细分为更小的独立子任务在网络节点上运行，然后将多个子结果集合为最终结果。而X!!Tandem项目[44]将标准的X!Tandem模型扩展，允许它在网络中的多个节点分发搜索任务。X!!-Tandem项目基于MPI框架，Pratt等[45]则提出了使用Hadoop MapReduce的MR-Tandem项目。本课题组则与湖南大学李肯立课题组合作开发了SW-Tandem[46]，并在神威太湖之光超级计算机上进行了测试，在多个数据集上进行的实验结果证明SW-Tandem的性能超越X!!Tandem与MR-Tandem。SW-Tandem采用两级并行化机制，即任务级别和线程级别。在任务级别，在处理大规模MS/MS数据集时，根据核心组(Core Group, CG)的总数将光谱数据集划分为适当大小的子数据集。在线程级别，子数据集被连续加载到CG上，此级别的并行化是通过使用称为aThread的特殊加速线程库实现的。

2021年，Haseeb等[47]提出了新的计算框架HiCOPS。该框架通过批量并行超级步骤来构建并行的肽段搜索工作流，其中，超级步骤指的是一组不同的算法和数据通信模块，由所有并行进程异步执行。根据需要，进程之间的同步在每个超级步骤结束时完成。

相较于数据库搜索方式，从头测序方式所要面对的搜索空间更大，计算负担更重。湖南大学李肯立团队相继构建了针对从头测序工具的并行化加速流程MRUniNovo[48]与SWPepNovo[49]。UniNovo[50]是一种采用基于概率模型的从头测序工具，它可以自动过滤掉低质量的谱图。MRUniNovo是利用Hadoop节点并行处理二级谱图数据集的不同部分的工具。MRUniNovo的执行分为两个阶段：第一个阶段MRUniNovo将谱图数据集划分为大小合适的块，并基于Hadoop分布式文件系统将它们分布在多台机器上；第二阶段并行执行Map任务和Reduce任务。SWPepNovo提出了一种高效优化的谱图数据组织方式，以克服内存访问带宽的瓶颈，并提出了一种高度可扩展的内部通信方案，并行化效率超过85%；在设计实现中，还采用了异步任务传递，并提出了一系列有效的优化策略，从而导致与未优化的版本相比，加速了10倍。

#### 4.1.2 商业工具

在商业软件中，目前Mascot和PEAKS都推出了基于集群的版本来应对海量数据的处理。相比于普通版本，集群版本对用户来说在使用方式上基本没有区别，处理速度的提升量取决于集群的大小。

#### 4.1.3 并行化加速方法小结

并行化加速目前仅集中在各种鉴定软件中。实际上，随着蛋白质组学研究规模的扩大，质量控制和后续定量过程也面临着计算上的压力，但目前还很少有研究涉及这些过程的并行化加速。虽然目前传统的蛋白质组学工具还能勉强应对现有研究产生的数据，但对集群的需求相信会越来越高。

### 4.2 深度学习方法

大数据的发展促进了深度学习[51]方法的发展，深度学习技术已经深刻改变了我们的生活方式。深度学习方法对数据极度依赖，得益于蛋白质组学数据规模的增加，各种深度学习方法也在蛋白质组学领域形成突破[52]。这些突破目前主要集中在肽段特征预测任务、理论谱图的预测任务和从头测序任务上。

#### 4.2.1 肽段特征预测

肽段特征预测任务的目的是利用深度学习模型预测肽段的各种特征。这里主要介绍保留时间的预测和肽段可检测性的预测。

在基于质谱的蛋白质组学实验中，样本经过预处理后形成的肽段混合物通常会通过色谱进行分离。肽段的保留时间是指肽段在LC-MS/MS系统中从液相色谱柱上洗脱出来的时间点。在相同的色谱条件下，肽段的保留时间具有高度的重现性，准确预测的保留时间在基于质谱的蛋白质组学中有多种应用，包括提高数据库搜索中肽段鉴定的灵敏度、作为肽段鉴定的质量评估指标或构建谱图库等。肽段保留时间的预测研究可以追溯到1980年[53]，直到今天，这方面的研究仍在持续受到关注。目前，基于深度学习的保留时间预测方法包括DeepRT[54]、

Prosit[55]、DeepMass[56]、Guan 等 [57] 的方法、DeepDIA[58]、AutoRT[59] 和 DeepLC[60] 等。基于深度学习的工具可以根据所使用的神经网络架构类型分为三组：基于循环神经网络 (recurrent neural network, RNN) 的模型、基于卷积神经网络 (convolutional neural network, CNN) 的模型和混合两者的模型，其中 RNN 是大多数工具使用的架构。Prosit 是基于 RNN 的代表性工具。在 Prosit 中，肽序列表示为长度为 30 的离散整数向量，每个非零整数映射到一个氨基酸，并用零填充短于 30 个氨基酸的序列。DeepMass 也是基于 RNN 架构，该模型使用独热编码表示肽段序列，它的网络包括一个 BiLSTM 层和一个 LSTM 层。Guan 等 [57] 提出的 RT 模型类似于 DeepMass，但它使用了两个 BiLSTM 层。DeepRT 和 DeepLC 都使用了基于 CNN 的架构。DeepRT 包含一个嵌入层作为神经网络的第一层。而 DeepLC 使用标准的 CNN 框架，与其他工具相比，DeepLC 的一个独特功能是能够预测训练数据中不存在的修饰肽段的保留时间。这主要是通过使用一种新的基于原子组成的肽编码来实现的。其他模型，例如 DeepDIA 和 AutoRT，在同一网络中结合了 CNN 和 RNN。总的来说，相比传统方法，深度学习方法显著提高了肽段保留时间预测的准确性。

另一个突破是肽段可检测性的预测。质谱检测具有较大的随机性，往往只有很少一部分肽段能被检测到，从而极大地阻碍了对质谱数据进行高精度、大规模的解析。准确预测各肽段的可检测性，将有助于改善蛋白质组学的实验设计和数据分析。DeepDigest[61] 基于 BiLSTM 构建，对肽段可检测性的预测准确性远高于传统的机器学习算法，如随机森林、支持向量机和逻辑斯蒂回归算法。DeepDigest 在训练集上进行 10 折交叉验证的 AUC 介于 0.956 至 0.982 之间。在 11 个独立测试数据集上，DeepDigest 的 AUC 介于 0.849 和 0.978 之间。

除此之外，深度学习模型还可以从序列预测翻译后修饰、预测结构以及预测抗原肽等。这些都属于基于序列对肽段特征进行预测。随着数据量的增加，尤其是高质量标注数据的增加，越来越多的特征都可以通过深度学习模型进行预测。值得一提的是，基于迁移学习等技术构建的 AlphaPeptDeep[62] 可以对保留时间、碰撞截面等多种肽段属性进行预测，其准确性与专用的预测工具相当。利用迁移学习技术和图形用户界面 (GUI)，用户可以比较容易地在自己的数据上进行肽段特征的预测。

#### 4.2.2 理论谱图预测

在典型的基于质谱的蛋白质组学实验中，使用碰撞诱导解离 (collision induced dissociation, CID) 等碎裂后的肽段形成了二级谱图。二级谱图是目前肽段序列鉴定的关键。在数据库鉴定策略中，通常需要先通过序列推测出对应的理论谱图。然而，肽段碎裂的潜在机制很复杂，仍然没有被充分理解。经过大量数据训练的深度学习模型在这一任务上正好体现出很大的优势。目前，已经出现了许多此类模型，如 pDeep[63]、Prosit[55]、DeepMass[56]、MS$^2$CNN[64]、DeepDIA[58]、Predfull[65] 等。一些模型，如 Prosit、DeepDIA，在预测谱图的同时也会进行保留时间的预测。pDeep 由两个 BiLSTM 层和一个全连接输出层组成，它采用独热编码肽段序列和肽段对应的母离子状态作为输入。pDeep2[66] 是 pDeep 的升级版本，通过使用 TensorFlow 中的动态 BiLSTM 克服了原始版本对序列长度的限制。Prosit 和 DeepMass 都是基于 BiRNN 的模型。Prosit 使用 BiGRU 网络，而 DeepMass 使用 BiLSTM 网络。MS$^2$CNN 基于 CNN 而不是 RNN，与上述模型不同，它使用人工构建的特征向量作为输入，而不是直接从肽段序列中学习肽段表示。MS$^2$CNN 中使用的特征包括肽段序列、质荷比 (m/z) 和肽段的理化性质，例如等电点、不稳定指数、芳香性、二级结构分数、螺旋度、疏水性和酸碱性等。用于谱图预测的 DeepDIA 模型混合使用了 CNN 和 BiLSTM，该谱图预测模型类似于 DeepDIA 中用于保留时间预测的模型。Predfull 利用基于残差 CNN 结构的广义序列到序列模型和多任务学习策略来预测肽段序列中所有可能的质荷比的强度，而不是仅预测特定离子 ( 通常为 b 离子与 y 离子 )。深度学习模型已被证明优于传统的机器学习模型和假设驱动的方法。使用深度学习的理论谱库生成将在 DIA 数据分析及靶向蛋白质组学实验中越来越受欢迎。

#### 4.2.3 从头测序工具

深度学习在蛋白质组学领域的另一个突破性应用是从头测序，涌现出了 DeepNovo[67]、SMSNet[68]、PointNovo[69] 和 Casanovo[70] 等工具。DeepNovo 将输入谱图视为图像，将输出肽段序列视为语言。在 DeepNovo 基础上后续还开发了专门针对 DIA 数据的 Deep- Novo-DIA[71]。SMSNet 的深度学习架构与 DeepNovo 中使用的架构类似，它的一项关键创新是使用 Sequence-Mask-Search 策略来提升一些小片段的预测准确性。SMSNet 还在编码器 - 解码器网

络之后引入了一个评分网络来估计预测序列的每个氨基酸的分数，如果分数未达到阈值，就会通过参考库来更新。PointNovo 的主要创新是采用了点云的方式将谱图转化为特征向量，基于点云的架构可以更有效地利用高精度谱图，得到更准确的预测结果。Casanovo 是基于 Transformer 架构 [72] 的工具，在评估 [73] 中得到了更高的评价，体现了 Transformer 架构的优越性。同时，基于深度学习的从头测序工具展现出了优于传统算法或机器学习工具的效果 [73]。这一类工具有望在抗体预测等领域发挥重要作用，在应对一些序列数据不完整的罕见物种或特殊样本时也能体现出不依赖于序列数据的优势。

#### 4.2.4 深度学习方法小结

虽然深度学习方法为蛋白质组学带来了巨大的突破，但在很多方面仍然存在缺陷。例如，虽然基于深度学习的保留时间预测方法准确性大大提高，但对具有修饰的肽段进行预测仍然是一个难题，其准确性远远低于普通肽段保留时间的预测。其余的，如基于深度学习的从头测序算法等也仍然有巨大的提升空间。深度学习可以有效利用大量的数据，这一点和不断增大的蛋白质组学数据相得益彰，采用更大规模的模型有可能会产生“奇迹”。另一方面，随着数据规模的扩大，深度学习方法训练所需的计算资源与计算时间也会随之扩大，深度学习的加速是不得不考虑的问题。而且，如何高效获取高质量数据，尤其是有标注的数据，也是未来值得注意的方向。

## 5 展望

目前，蛋白质组学大数据相关的研究还处于初级阶段，对数据的分析和应用方式相对简单，研究方向也相对较少。由于批次效应等问题，目前的研究很少将大量的数据集联合进行分析，更多地是展示不同的数据集并提供对应的下载渠道。但相信在不久的将来，随着蛋白质组学分析标准化程度的加强和大规模质谱数据去噪算法的发展，大批量数据集的联合分析将成为可能，更加复杂的关联 ( 包括与其他类型数据的关联 ) 将被考虑进来。规模更大的不同研究的联合分析将揭示更多的生物学机制，也会使结论更加可靠。另外，随着精准医学的发展，蛋白质组学研究可能越来越看重时效性，实时的蛋白质组学检测将对数据的收集、传输、分析以及结果汇总提出更高的要求。这一点必然需要硬件和软件的密切配合。除了本文重点讨论的基于质谱的方法，基于抗体的方法也在蛋白质组学领域，尤其是在单细胞蛋白质组研究中有较多应用。除此之外，纳米孔测序等技术也日渐成熟。这些数据与质谱数据有较大的区别，在未来可能会面临这些不同类型数据的汇总和联合分析问题。总之，这是一个充满机会的领域。

## [参 考 文 献]


[1] Reinsel D, Gantz J, Rydning J. The digitization of the world from edge to core [EB/OL]. (2018-11). https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

[2] Jiang Y, Sun A, Zhao Y, et al. Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. Nature, 2019, 567: 257-61

[3] Xu JY, Zhang C, Wang X, et al. Integrative proteomic characterization of human lung adenocarcinoma. Cell, 2020, 182: 245-61

[4] Müller JB, Geyer PE, Colaço AR, et al. The proteome landscape of the kingdoms of life. Nature, 2020, 582: 592-6

[5] Schoof EM, Furtwängler B, Üresin N, et al. Quantitative single-cell proteomics as a tool to characterize cellular hierarchies. Nat Commun, 2021, 12: 3341

[6] Piehowski PD, Zhu Y, Bramer LM, et al. Automated mass spectrometry imaging of over 2000 proteins from tissue sections at 100-μm spatial resolution. Nat Commun, 2020, 11: 8

[7] Asleh K, Negri GL, Spencer Miko SE, et al. Proteomic analysis of archival breast cancer clinical specimens identifies biological subtypes with distinct survival outcomes. Nat Commun, 2022, 13: 896

[8] Gonçalves E, Poulos RC, Cai Z, et al. Pan-cancer proteomic map of 949 human cell lines. Cancer Cell, 2022, 40: 835-49

[9] Sun Y, Selvarajan S, Zang Z, et al. Artificial intelligence defines protein-based classification of thyroid nodules. Cell Discov, 2022, 8: 85

[10] Pedrioli PGA, Eng JK, Hubley R, et al. A common open representation of mass spectrometry data and its application to proteomics research. Nat Biotechnol, 2004, 22: 1459-66

[11] Deutsch EW. Mass Spectrometer output file format mzML. Methods Mol Biol, 2010, 604: 319-31

[12] Perkins DN, Pappin DJC, Creasy DM, et al. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 1999, 20: 3551-67

[13] Cox J, and Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol, 2008, 26: 1367-72

[14] Li D, Fu Y, Sun R, et al. pFind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry. Bioinformatics, 2005, 21: 3049-50

[15] Kong AT, Leprevost FV, Avtonomov DM, et al. MSFragger:

ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nat Methods, 2017, 14: 513-20

[16] Yang W, Chen W, Rogers I, et al. PEAKS Q: software for MS-based quantification of stable isotope labeled peptides [EB/OL]. https://www.bioinfor.com/wp-content/uploads/2017/02/peaksq-softwareforms-basedquantification.pdf

[17] Demichev V, Messner CB, Vernardis SI, et al. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. Nat Methods, 2020, 17: 41-4

[18] The M, MacCoss MJ, Noble WS, et al. Fast and accurate protein false discovery rates on large-scale proteomics data sets with Percolator 3.0. J Am Soc Mass Spectrom, 2016, 27: 1719-27

[19] Li N, Wu S, Zhang C, et al. PepDistiller: a quality control tool to improve the sensitivity and accuracy of peptide identifications in shotgun proteomics. Proteomics, 2012, 12: 1720-5

[20] Chang C, Li M, Guo C, et al. PANDA: a comprehensive and flexible tool for quantitative proteomics data analysis. Bioinformatics, 2019, 35: 898-900

[21] Rodriguez H, Snyder M, Uhlén M, et al. Recommendations from the 2008 International Summit on Proteomics Data Release and Sharing Policy: the Amsterdam principles. J Proteome Res, 2009, 8: 3689-92

[22] Kinsinger CR, Apffel J, Baker M, et al. Recommendations for mass spectrometry data quality metrics for open access data (corollary to the Amsterdam principles). Proteomics, 2012, 12: 11-20

[23] Kinsinger CR, Apffel J, Baker M, et al. Recommendations for mass spectrometry data quality metrics for open access data (Corollary to the Amsterdam Principles). Mol Cell Proteomics, 2011, 10: O111.015446

[24] Mayer G, Montecchi-Palazzi L, Ovelleiro D, et al. The HUPO proteomics standards initiative-mass spectrometry controlled vocabulary. Database, 2013, 2013: bat009

[25] Mayer G, Jones AR, Binz PA, et al. Controlled vocabularies and ontologies in proteomics: overview, principles and practice. Biochimica Biophysica Acta, 2014, 1844: 98-107

[26] Martínez-Bartolomé S, Binz PA, Albar JP. The Minimal Information About a Proteomics Experiment (MIAPE) from the Proteomics Standards Initiative. Methods Mol Biol, 2014, 1072: 765-80

[27] Deutsch EW, Orchard S, Binz PA, et al. Proteomics Standards Initiative: fifteen years of progress and future work. J Proteome Res, 2017, 16: 4288-98

[28] Martens L, Chambers M, Sturm M, et al. mzML—a community standard for mass spectrometry data. Mol Cell Proteomics, 2011, 10: R110.000133

[29] Vizcaíno JA, Mayer G, Perkins S, et al. The mzIdentML data standard version 1.2, supporting advances in proteome informatics. Mol Cell Proteomics, 2017, 16: 1275-85

[30] Walzer M, Qi D, Mayer G, et al. The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics. Mol Cell Proteomics, 2013, 12: 2332-40

[31] Griss J, Jones AR, Sachsenberg T, et al. The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. Mol Cell Proteomics, 2014, 13: 2765-75

[32] Xu Q, Griss J, Wang R, et al. jmzTab: a Java interface to the mzTab data standard. Proteomics, 2014, 14: 1328-32

[33] Perez-Riverol Y, Uszkoreit J, Sanchez A, et al. ms-data-core-api: an open-source, metadata-oriented library for computational proteomics. Bioinformatics, 2015, 31: 2903-5

[34] Qi D, Zhang H, Fan J, et al. The mzqLibrary-an open source Java library supporting the HUPO-PSI quantitative proteomics standard. Proteomics, 2015, 15: 3152-62

[35] Côté RG, Griss J, Dianes JA, et al. The PRoteomics IDEntification (PRIDE) Converter 2 Framework: an improved suite of tools to facilitate data submission to the PRIDE database and the ProteomeXchange consortium. Mol Cell Proteomics, 2012, 11: 1682-9

[36] Deutsch EW, Perez-Riverol Y, Carver J, et al. Universal spectrum identifier for mass spectra. Nat Methods, 2021, 18: 768-70

[37] Deutsch EW, Bandeira N, Sharma V, et al. The ProteomeXchange consortium in 2020: enabling "big data" approaches in proteomics. Nucleic Acids Res, 2020, 48: 1145-52

[38] Perez RY, Csordas A, Bai J, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res, 2019, 47: 442-50

[39] Desiere F, Deutsch EW, King NL, et al. The PeptideAtlas project. Nucleic Acids Res, 2006, 34: 655-8

[40] Ma J, Chen T, Wu S, et al. iProX: an integrated proteome resource. Nucleic Acids Res, 2019, 47: 1211-7

[41] Chen T, Ma J, Liu Y, et al. iProX in 2021: connecting proteomics data sharing with big data. Nucleic Acids Res, 2022, 50: 1522-7

[42] Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics, 2004, 20: 1466-7

[43] Duncan DT, Craig R, Link AJ. Parallel Tandem: a program for parallel processing of tandem mass spectra using PVM or MPI and X!Tandem. J Proteome Res, 2005, 4: 1842-7

[44] Bjornson RD, Carriero NJ, Colangelo C, et al. X!!Tandem, an improved method for running X!Tandem in parallel on collections of commodity computers. J Proteome Res, 2008, 7: 293-9

[45] Pratt B, Howbert JJ, Tasman NI, et al. MR-Tandem: parallel X!Tandem using Hadoop MapReduce on Amazon web services. Bioinformatics, 2012, 28: 136-7

[46] Li C, Li K, Chen T, et al. SW-Tandem: a highly efficient tool for large-scale peptide identification with parallel spectrum dot product on Sunway TaihuLight. Bioinformatics, 2019, 35: 3861-3

[47] Haseeb M, Saeed F. High performance computing framework for tera-scale database search of mass spectrometry data. Nat Comput Sci, 2021, 1: 550-61

[48] Li C, Chen T, He Q, et al. MRUniNovo: an efficient tool for *de novo* peptide sequencing utilizing the hadoop

distributed computing framework. Bioinformatics, 2016, 33: 944-6
[49] Li C, Li K, Li K, et al. SWPepNovo: an efficient *de novo* peptide sequencing tool for large-scale MS/MS spectra analysis. Int J Biol Sci, 2019, 15: 1787-801
[50] Jeong K, Kim S, Pevzner PA. UniNovo: a universal tool for *de novo* peptide sequencing. Bioinformatics, 2013, 29: 1953-62
[51] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521: 436-44
[52] Wen B, Zeng W, Liao Y, et al. Deep learning in proteomics. Proteomics, 2020, 20: 1900335
[53] Meek JL. Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition. Proc Natl Acad Sci U S A, 1980, 77: 1632-6
[54] Ma C, Ren Y, Yang J, et al. Improved peptide retention time prediction in liquid chromatography through deep learning. Anal Chem, 2018, 90: 10881-8
[55] Gessulat S, Schmidt T, Zolg DP, et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nat Methods, 2019, 16: 509-18
[56] Tiwary S, Levy R, Gutenbrunner P, et al. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. Nat Methods, 2019, 16: 519-25
[57] Guan S, Moran MF, Ma B. Prediction of LC-MS/MS properties of peptides from sequence by deep learning. Mol Cell Proteomics, 2019, 18: 2099-107
[58] Yang Y, Liu X, Shen C, et al. *In silico* spectral libraries by deep learning facilitate data-independent acquisition proteomics. Nat Commun, 2020, 11: 146
[59] Wen B, Li K, Zhang Y, et al. Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. Nat Commun, 2020, 11: 1759
[60] Bouwmeester R, Gabriels R, Hulstaert N, et al. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. Nat Methods, 2021, 18: 1363-9
[61] Yang J, Gao Z, Ren X, et al. DeepDigest: prediction of protein proteolytic digestion with deep learning. Anal Chem. 2021, 93: 6094-103
[62] Zeng WF, Zhou XX, Willems S, et al. AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics. Nat Commun, 2022, 13: 7238
[63] Zhou XX, Zeng WF, Chi H, et al. pDeep: predicting MS/MS spectra of peptides with deep learning. Anal Chem, 2017, 89: 12690-7
[64] Lin YM, Chen CT, Chang JM. MS2CNN: predicting MS/MS spectrum based on protein sequence using deep convolutional neural networks. BMC Genomics, 2019, 20: 906
[65] Liu K, Li S, Wang L, et al. Full-spectrum prediction of peptides tandem mass spectra using deep neural network. Anal Chem, 2020, 92: 4275-83
[66] Zeng WF, Zhou XX, Zhou WJ, et al. MS/MS spectrum prediction for modified peptides using pDeep2 trained by transfer learning. Anal Chem, 2019, 91: 9724-31
[67] Tran NH, Zhang X, Xin L, et al. *De novo p*eptide sequencing by deep learning. Proc Natl Acad Sci U S A, 2017, 114: 8247-52
[68] Karunratanakul K, Tang HY, Speicher DW, et al. Uncovering thousands of new peptides with sequence-mask-search hybrid de novo peptide sequencing framework. Mol Cell Proteomics, 2019, 18: 2478-91
[69] Qiao R, Tran NH, Xin L, et al. Computationally instrument-resolution-independent *de novo* peptide sequencing for high-resolution devices. Nat Mach Intell, 2021, 3: 420-5
[70] Yilmaz M, Fondrie WE, Bittremieux W, et al. *De novo* mass spectrometry peptide sequencing with a transformer model [EB/OL]. (2022-06-18). https://www.biorxiv.org/content/10.1101/2022.02.07.479481v2
[71] Tran NH, Qiao R, Xin L, et al. Deep learning enables *de novo* peptide sequencing from data-independent-acquisition mass spectrometry. Nat Methods, 2019, 16: 63-6
[72] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [EB/OL]. (2023-08-02). https://arxiv.org/abs/1706.03762
[73] Beslic D, Tscheuschner G, Renard BY, et al. Comprehensive evaluation of peptide *de novo* sequencing tools for monoclonal antibody assembly. Briefings Bioinform, 2022, 24: bbac542