scientific data

Check for updates

OPEN

ARTICLE

Preventing Proteomics Data Tombs Through Collective Responsibility and Community Engagement

Uladzislau Vadadokhau *et al.*[#]

Public proteomics repositories now host vast amounts of mass spectrometry data, yet much of it remains difficult to reuse, risking "data tombs" that are open access but not practically re-analyzable. In spring 2025, a graduate-level course at the University of Helsinki tasked six student teams with reanalyzing six projects from the Proteomics Identification Database (label-free quantification only) using a common R-based workflow (rpx, mzR, QFeatures, DEP/MSqRob2/limma/OmicsQ packages) that was shared across all teams. The teams reproduced identification, optional quantification, normalization, imputation, and differential expression analyses, and compared the outcomes to the original studies. As expected, systemic barriers recurred across cases: (i) no sample and data relationship format for proteomics metadata in any of the cases; (ii) missing details regarding decoy sets for false discovery rate assessment; (iii) proprietary-only outputs or software (e.g., Thermo.msf, Progenesis) that impeded open reanalysis in interoperable, community-standard formats; (iv) missing data-independent acquisition spectral libraries or protein sequences database files (FASTA); (v) absent or vague normalization/imputation/statistical parameters; (vi) inconsistent file naming; and (vii) insufficient biological/technical replication in at least one project. These shortcomings yielded large discrepancies in the analysis results (e.g., 13,068 vs. 4,923 proteins; 108 vs. 11 differentially expressed proteins), and, in one instance, a highlighted protein lacked robust support in the deposited identifications. We observed that reproducibility in mass spectrometry-based proteomics hinges less on instruments than on transparent metadata, open formats, and executable analysis provenance. We propose that data creators provide a minimum re-analysis package, including raw data and open formats, community standards, basic quality control summaries, data-independent acquisition spectral libraries, and complete parameter/code sets with pinned versions or containers. Moreover, we recommend repository-level nudges toward making such packages mandatory. This educational exercise simultaneously trains the students as well as stress-tests the community data practices to prevent proteomics "data tombs".

Introduction

In spring 2025, the University of Helsinki offered a graduate-level course *DPBM-230 Mass Spectrometry-Based Proteomics: Concepts, Methods, and Data*, which provided a comprehensive overview of proteomics, with a particular focus on mass spectrometry (MS) and data analysis methodologies. The course encompassed core techniques for separations, identification, and quantification of proteins and peptides, delivered through a combination of expert-led lectures, laboratory tours, and hands-on computational training. Participants benefited from insights shared by both internal faculty and invited international experts and visited the university's proteomics core facilities to observe advanced and diverse MS platforms, including Orbitrap, MALDI-TOF, and Q-TOF systems. A central pedagogical component of the course was a capstone project, in which student teams engaged with publicly available datasets from the ProteomeXchange Consortium, aiming to reproduce and critically evaluate the findings of published proteomics studies.

[#]A full list of authors and their affiliations appears at the end of the paper.

The ProteomeXchange Consortium is a globally coordinated initiative established to standardize data submission and dissemination in the proteomics field[1]. It promotes open data practices and ensures broad accessibility to high-quality proteomics datasets by integrating several major international repositories, including Proteomics Identification Database (PRIDE)[2], PeptideAtlas[3], Mass Spectrometry Interactive Virtual Environment (MassIVE)[1], jPOST[1,4], iProx[5], and Panorama Public[6]. During the course, students accessed datasets through *ProteomeCentral*, the consortium's central portal for browsing public ProteomeXchange entries (PXD). This access allowed students to download authentic experimental and published datasets and attempt their replication, as well as assess data quality and reproducibility. ProteomeXchange resources, including PRIDE, adhere to the FAIR principles (Findable, Accessible, Interoperable, and Reusable)[2,7]. FAIRness emphasizes not just data availability but also clarity, structure, and context, enabling others to locate, access, and reuse datasets without ambiguity. By embedding these principles into data generation and sharing practices, researchers can transform deposited datasets from static archives into dynamic assets that support reproducibility, integration, and long-term scientific progress.

Reproducibility and transparency remain fundamental pillars of scientific integrity, yet they pose unique challenges in proteomics due to the intricate and multi-step nature of experimental workflows[8–12]. Consistency in results depends on stringent control across key stages, including experimental design and sample preparation, fractionation and chromatography, MS data acquisition, peptide/protein identification and quantification, and downstream statistical and pathway analyses[13–15]. Each of these stages introduces potential variability that may affect reproducibility. The capstone project was intentionally designed to highlight that reproducibility in proteomics is not solely dependent on experimental design or instrumentation but is also critically influenced by a continuum of computational decisions and analytical conditions. By engaging with real-world datasets and attempting to replicate complex proteomic workflows, students were encouraged to critically evaluate how upstream computational variation influences downstream data interpretation.

In this article, we summarize six case studies resulting from student-led capstone projects, each involving the re-analysis of a distinct dataset from the ProteomeXchange repository. These projects collectively highlight the latent potential of archived proteomics data to be reimplemented, reinterpreted, and scrutinized for reproducibility. Through these re-analyses, students confronted both the strengths and limitations of current practices in proteomics data sharing and curation. We showed that adhering to the FAIR principles is not enough, and their execution must be controlled before the data is deposited into proteomics databases. Drawing from the findings and reflections of these case studies, we propose a set of community-oriented recommendations for researchers. We also think that other parties involved in the proteomics research, such as publishers, reviewers, data curators, and repository developers, can use these recommendations to strengthen submission guidelines, improve metadata validation, and implement policies that make datasets more transparent, interoperable, and ready for reproducible re-analysis. These recommendations emphasize the need for continued commitment to data transparency, standardized metadata, and long-term data stewardship, critical elements in preventing valuable proteomics datasets from becoming inaccessible or unusable "data tombs" and in fostering a more reproducible and collaborative scientific ecosystem.

Course Structure and Pedagogical Framework

The structure of this course was intentionally designed to integrate a diverse range of instructional modalities, combining lectures, interactive workshops, and immersive laboratory experiences to facilitate deep learning and student engagement. The course unfolded over several weeks through thematically grouped modules, which included basic principles of chromatography and MS for protein analysis, components of MS instruments, data-dependent (DDA) and -independent (DIA) acquisition techniques, and advanced MS methods such as label-based proteomics and phosphoproteomics. The hands-on part of the course was dedicated to retrieving the data from public repositories, data preprocessing and visualization, as well as statistical analysis. Guest lectures were held in four sessions, featuring international and local experts who introduced students to advanced topics in proteomics and data science. Regular course lectures continued in three sessions, building foundational knowledge in MS workflows and analytical approaches. A sequence of hands-on R programming workshops was conducted in six sessions, where students applied theoretical concepts to practical computational tasks, including data preprocessing, visualization, and differential expression (DE) analysis using real proteomics datasets. Lab tours during two sessions provided students with firsthand exposure to cutting-edge MS instrumentation and workflows at the proteomics core facilities. The course culminated in the "Race Day" event, which included a special seminar, capstone presentations, and peer-evaluated awards, which was a celebration of learning, synthesis, and scientific communication.

Pedagogically, the course was grounded in active learning principles, with a strong emphasis on formative and summative self-assessment, collaboration, and reflective practice. Each session began with an individual quiz administered through Socrative.com, designed to reinforce pre-read materials such as review articles and supplemental chapters, as well as test knowledge obtained during the lectures and solve some questions by using the discussed R packages[16]. This was followed by peer discussion quizzes, where students were assigned to randomly generated small groups using PickerWheel.com, promoting collaborative problem-solving and exposing participants to diverse perspectives[17]. These groupings were reshuffled nearly every session to enhance group dynamics and equitable participation. Students were also required to complete session-specific feedback forms, which were used to refine instructional strategies and harmonize group teaching iteratively. The use of continuous interactive assessment through Socrative allowed instructors to monitor engagement and comprehension in real time. Self-assessment components, embedded within the quiz discussions and project reflections, fostered metacognitive awareness and personal accountability[18]. Collectively, the blended structure of lectures, hands-on practice, guided exploration, and reflective evaluation created a learner-centered environment that supported both conceptual understanding and skill development in proteomics.

Capstone Projects

The capstone project served as the central evaluative component of this course (Fig. 1), assessed through a norm-referenced framework[19,20] and culminating in a final "Race Day" presentation. Designed to integrate theoretical understanding with hands-on application, the project required student teams (three members per group) to reanalyze a publicly available proteomics dataset from the ProteomeXchange repository. Leveraging the R programming environment and key Bioconductor packages (e.g., rpx, mzR, QFeatures, and DEP), students followed a structured workflow that reflected a typical MS-based proteomics analysis pipeline. This included dataset acquisition, generation, and filtering of peptide-spectrum match (PSM) objects, peptide and protein identification, and optional steps such as quantification, normalization, imputation, statistical testing, and data visualization.

A central objective was to compare team-generated results with those reported in the original publication, allowing students to assess the reproducibility of biological findings and to identify potential sources of experimental and computational variability. The four-week project followed sequential milestones: initial data acquisition and preprocessing (Weeks 1-2), statistical and downstream analysis (Week 3), and final report submission and oral presentations (Week 4). Peer feedback was embedded into the Race Day format to promote deeper engagement[21]. Each presentation concluded with a structured discussion led by an assigned opponent team. In addition, all teams assessed each other's performance as both presenters and opponents using a rubric table (Table 1) that emphasized analytical depth and clarity of scientific communication[22]. These peer evaluations were used to nominate the best-performing team, reinforcing the importance of collaborative learning and reflective assessment.

Pedagogically, the capstone project aimed to cultivate both technical competence and critical thinking in the context of real-world proteomics data analysis. Beyond computational proficiency, the experience fostered metacognitive development through structured reflection, formative peer review, and iterative problem-solving[23,24]. The use of a transparent, norm-referenced evaluation rubric ensured fairness and consistency across groups, supporting the development of core skills in data interpretation, reproducibility, and scientific communication. In the following section, we summarize the six team projects to illustrate the educational value and research potential of reanalyzing publicly available proteomics datasets and to reflect on the broader role of data repositories in advancing biomedical research.

We provide a detailed reflection on every step in the workflow, including challenges that were faced by students and how data creators could improve the transparency and reusability of their data. In this paper, we consider ProteomeXchange as a data storage and not a processing tool storage. Therefore, proteomics data repositories should not be considered as an independent and only independent source of all data that is needed to reproduce. Some of the processing tools can be deposited in special repositories (e.g., GitHub, BitBucket, etc.). Hence, to reproduce the author's analysis, students considered all available sources of information (publications, supplementary files and data and code availability statements). However, the main focus is data available on the ProteomeXchange. Therefore, anything that is supposed to be published on ProteomeXchange but was missing from there is considered missing. At the first stage of the capstone projects, teams selected the project of their preference from PRIDE submissions (Fig. 1a). The selection criteria included (1) availability of identification results in mzIdentML (.mzID) format and (2) an experimental design enabling protein abundance comparison between experimental groups. In addition, students could select projects addressing biological questions of their interest. Therefore, the data selection process was not designed to target exclusively irreproducible or non-transparent cases. Instead, it aimed to simulate a realistic scenario in which non-experts in proteomics attempt to reuse public datasets, for example, to validate their own findings.

Due to the limited duration of the course, we decided to focus only on label-free quantitation experiments (LFQ), the most utilized approach in MS-based proteomics. To facilitate the process of selecting the project from PRIDE, students used the Filezilla FTP client to access the PRIDE server. It allowed them to search for projects where identification results were submitted (.mzID,.mzTab), so-called "complete submission" (Table 2). However, teams reported that a number of projects that could have been used lacked identification results in Human Proteomics Organization Standard Initiative (HUPO-PSI) formats (mzIdentML, extension.mzID or mzTab extension.mzTab) and instead contained identifications in other formats, outputs of various proteomics search software (termed as "partial submission" in PRIDE)[25,26]. While these file formats could be used, retrieving the information from community standards representations is more straightforward, as it does not require special software or conversion (mzTab is a tab-delimited file; for mzIdentML files, parsers in R and Python exist), especially for data users outside of the proteomics field. For example, among the case studies, two of them (case 3 and case 6) had proteomics software search results in.msf format, in addition to identification results (.mzID), and could be opened only with Proteome Discoverer (PD) software (proprietary). One potential solution for this would have been the use of the Thermo MSF parser, which would have introduced an additional processing step in the workflow[27]. Even if you had software to open that file, there are cases where there is no forward compatibility between files and software (e.g., Spectronaut 17.1 does not open files saved in Spectronaut 19.0). Currently, almost all proteomics search software tools allow the extraction of the results in HUPO-PSI formats. It could happen that the extraction of the results in these formats is not set at the step of creating parameters for the search software, and then it is not possible to obtain them after the search. This point raises the issue of data FAIRness, where data creators should have a plan and oversee the data reusability and be familiar with submission guidelines, or seek help when planning the data sharing. In other words, the data creators should always think about their data from the perspective of the data user. On the other hand, the term "complete" data submission is not related to completeness in view of data users; rather, this term is used to label submissions in which data can be parsed and integrated for visualization in PRIDE. In our view, the terminology should be changed to remove misunderstanding from the user's side. All in all, the first step did not cause challenges for students to retrieve the data using the rpx R-package.

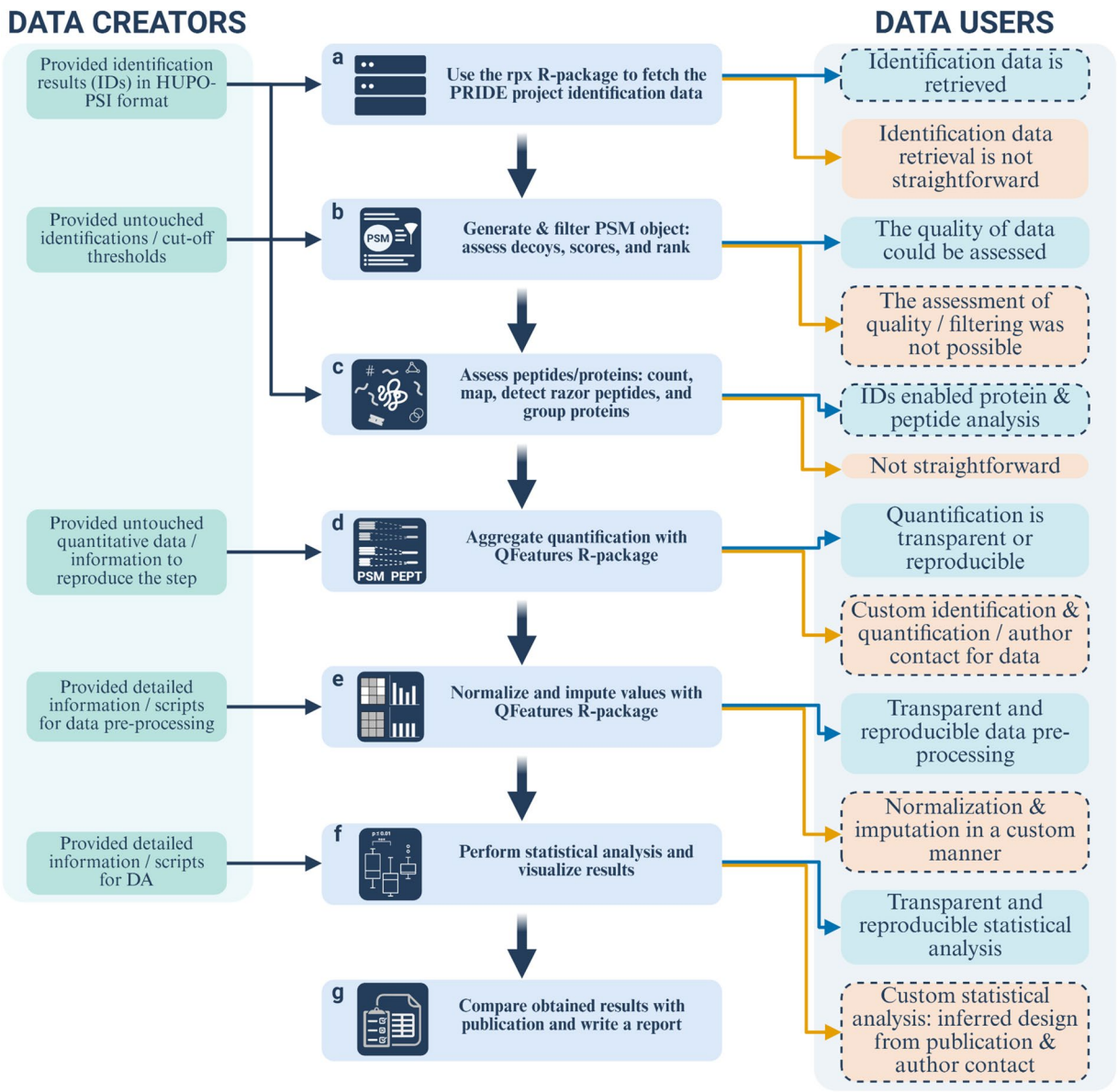

Fig. 1 Capstone Project workflow and the PRIDE data submission influence on it. An R-based pipeline for reanalyzing mass spectrometry proteomics data, covering dataset acquisition, preprocessing, statistical analysis, visualization, and comparison with published results (shown in the center, which was applied during the proteomics course). To the left of the workflow is the data creator's side, illustrating their actions related to data submission to PRIDE. To the right is the data user's side, representing their experience in data reanalysis. The color of the arrows indicates the impact of the data creator's actions: blue signifies that the condition was fulfilled, resulting in a positive outcome for the data user; yellow indicates that the condition was not fulfilled, thereby hindering or complicating data transparency and reusability. The data user's side outlines the possible paths users can take depending on the data creator's actions. Dashed outlines around the data user's boxes indicate the specific path taken by participants during the course.

In the next stages, students were asked to create a PSM object in the R environment from retrieved identification results (Fig. 1b,c). The created object was then used to assess the decoy hits and FDR, identification scores and ranks, and finally, apply filtering to the data and count peptides/proteins and map them to each other. The main issue with these case studies is the absence of details regarding decoy sets in all projects. While decoy sequences are not used for the downstream analysis, they play an important role in calculating FDR for the protein identifications. As a result, the quality of protein identification could not be properly assessed. Defining a cutoff based solely on identification scores may lead to weak identifications in downstream analyses; therefore, the scores of normal identifications should be evaluated relative to those of decoy identifications. Data creators are encouraged to upload unfiltered data to proteomics repositories and provide filtering criteria that were applied in their downstream analysis. This is especially important if the goal is to check the reproducibility of the results.

Score	Presenting team				Opponent team	
	Slide Quality	Time Management	Comparison with Original Publication	Handling of Questions	Relevance of Questions	Constructive Feedback
1	Incoherent, poorly formatted	Exceeded or insufficient	No comparison attempted	Avoided or deflected	Off-topic or unclear	Vague, irrelevant, or absent
2	Minimal structure, low clarity	Minor overruns or gaps	Surface-level or incorrect	Unclear or hesitant	Partially related	General comments, little depth
3	Functional but unpolished	Mostly within limits	Basic comparison, lacks insight	Adequate, some confusion	Mostly relevant	Identifies issues, lacks solutions
4	Clear, well-structured	Well-paced with purpose	Strong analytical comparison	Confident and precise	Clearly relevant	Specific and actionable
5	Professional and compelling	Optimal and efficient	Deep, critical, and contextual	Insightful and integrative	Sharp and purposeful	Enhances work through insight

Table 1. Rubric table for peer-assessment of students on the race day.

PXD (type of submission), year	Acquisition type, instrument	Software used for the search	Project focus	Organism/material	Ref.
Case study 1 PXD061542 (Complete), 2025	DIA, Orbitrap Exploris 480	EncyclopeDIA	To understand the contribution of fibrosis to kidney function decline in nephronophthisis. MS-based proteomics was utilized to assess differences in the proteome of wild-type and Fbxw7-deleted cells.	Mus musculus/cell culture	36,37
Case study 2 PXD060654 (Complete), 2025	DDA, Orbitrap Eclipse	Proteome Discoverer	To characterize the proteome of the Arabidopsis leaf apoplast during immune response induced by flg22	Arabidopsis thaliana/leaf	38,39
Case study 3 PXD045844 (Complete), 2025	DDA, Orbitrap Eclipse	Proteome Discoverer	To characterize the properties of locus coeruleus neurons in male and female mice and the differences between them. MS-based proteomics was used to record protein expression profiles, among other methods applied	Mus musculus/brain slices	40,41
Case study 4 PXD045506 (Complete), 2025	DDA, Orbitrap Fusion	Byonic	To assess MGAT5 as a biomarker to predict patients' responses to anti-PD-1 immunotherapy. MS-based proteomics was utilized to analyse glycopeptides	Homo sapiens/paraffinized tissues	42,43
Case study 5 PXD058779 (Complete), 2025	DIA, SYNAPT G2-Si	Progenesis QI	To compare the proteome of the frontal cortex of patients with Down syndrome and Alzheimer's disease with early-onset Alzheimer's disease	Homo sapiens/post-mortem brain tissue	44,45
Case study 6 PXD054726 (Complete), 2024	DDA, Orbitrap Fusion	Proteome Discoverer	To assess the role of the K63-polyubiquitination site in the context of studying or learning-dependent synaptic plasticity. MS-based proteomics was used to analyse enriched K63-polyubiquitinated peptides	Rattus norvegicus/brain tissue slices	46,47

Table 2. Summary of PRIDE projects used as case studies for the capstone project.

The next steps of the workflow were optional in this course (Fig. 1d–g), but all teams enthusiastically carried them out except Team 6 (which had search results in.msf format). Besides identification results, students were encouraged to work with quantitative data. The QFeatures R-package[28] provides a suite of tools to manage and process proteomics quantitative data. Students were asked to create a QFeatures object in R using quantitative data and then to aggregate quantities from PSM to the peptide level. Three case projects did not include quantitative data (case 2, case 4, and case 5). Teams 2 and 4 performed their own search and quantification with MaxQuant software, setting it in a custom manner[29]. Team 5 was not able to reproduce the search and quantification with proprietary Progenesis QI software, as used by data creators; therefore, they contacted the corresponding authors to request the data. Case study 3 contained search results in Proteome Discoverer MSF format. Team 3 decided to perform their search in MaxQuant, rather than retrieving information from.msf, due to the unavailability of PD software. However, raw files were named in a generic manner. Team 3 contacted the authors of their chosen paper for the raw data-experimental conditions mapping. Finally, Team 1 is the only team that had results of the quantification in the mzTab standard. Nevertheless, this team tried to recreate the search with EncyclopeDIA software. The first challenge that the team faced was the absence of the spectral library that was utilized in the author's search. The team created their own spectral library. This step already diverged the analysis from the original one. Summarizing the observations, we encourage data creators to provide quantification data or the information on how to reproduce it (software parameters, FASTA database, spectral library, SDRF) since the identification by different search algorithms leads to divergent results. Preferably, the data should be provided in the proteomics standard format (.mzTab).

Next, teams performed normalization and imputation of the data. None of the PRIDE submissions contained detailed information (either missing or vague) to reproduce this step. Moreover, no data preparation/analysis scripts were provided. Teams 1, 2, and 4 used normalization and imputation tools from the QFeatures package. Teams 5 and 6 used the data they requested from data creators. The only exception was Team 3, which used one-fifth of the minimal value of abundance to impute missing values, reproducing the imputation method applied by the data creators. With prepared data, teams performed statistical analysis to identify DE proteins. The authors of the chosen papers provided limited information on the statistical analysis, which included the method used (limma, t-test) and cut-offs for the volcano plot [log2FC and -log10(p-value)], although not in all cases. For example, case study 2 did not provide a cut-off for the p-value, case study 3 did not provide any cut-offs,

case study 4 did not provide a cut-off for log2FC, and case study 5 did not provide information about the p-value cut-off and the method used to identify DE proteins. Like in the previous step, teams could not rely on the data analysis scripts, as they were not provided. Hence, teams performed the analysis with the method of their choice (msqrob2[30], limma[31], DEP[32], or OmicsQ[33]), inferring the experimental design from publications. Our conclusion here is that providing explicit experimental design (SDRF), data preparation and analysis steps (either as a script or a protocol), and naming data files in a meaningful way are crucial for both transparency and reproducibility.

At the final stage of the workflow, teams compared their results with those published by the authors of the selected paper (see the course GitHub for the complete team reports). Overall, it was complicated to compare both findings since different workflows were implemented, and some information was not provided by the authors. Team 1 tried to reproduce the author's results by searching files with EncyclopeDIA; however, they faced major issues like the absence of the spectral library, the used protein database, and a clear explanation of data preparation and data analysis. For example, in the original search, the authors of the paper reported 4,923 unique proteins (mzTab file from PRIDE). In the Excel table with protein results submitted by the authors as a supplementary file on the journal website, the number of unique proteins was lower than listed in the mzTab file. In contrast, the team's search identified 13,068 unique proteins after filtering. Additionally, the authors of their chosen paper focus on the TMEM237 protein, which is downregulated. However, the team could not find this protein in their own search. Additionally, Team 1 checked this protein in the reported identification mzTab file and found that this protein is reported only at the protein level. Team 2 assessed the original mzID file for the first three steps; however, to continue the analysis, they needed to perform a proteomics search in MaxQuant. The authors of the selected paper did not specify the number of identified proteins, making it impossible to directly compare the consistency of their findings. Instead, the team had to compare the number of peptides/proteins in the mzID file with the results from their MaxQuant search. The numbers seem to be comparable, although the quality of the replication is arguable. Lastly, the authors reported 108 DE proteins, while the team reported only 11. This discrepancy could have been explained by an explicit data analysis workflow. Team 3 analyzed the mzID file provided in PRIDE and compared this with the numbers in the publication. Their analysis and the original one found more proteins in males than in females. The team found good consistency between the file in PRIDE and the publication results (before filtration). The original study reported clear sex-related differences in locus coeruleus neuron activity and protein expression. While the team's analysis captures some of these patterns, such as partial sex separation in principal component analysis (PCA), there was no overlap in DE proteins. The lower number of identified proteins and inconsistencies in differential expression likely stem from differences in data processing, filtering, and imputation, particularly due to the use of MaxQuant. The analysis by Team 4 reported that statistical analysis was impossible based on the number of files the authors had provided to PRIDE. Since only one raw file was uploaded for each study group (i.e., wild type and knockout gene) and no replicates were available, their statistical analysis was based only on the difference of protein fold changes in the samples. Team 5 contacted authors for the data due to challenges in performing their own search. Thus, before starting downstream data analysis, the team had identical data to the authors. Their comparison revealed several key differences, highlighting the underlying challenge of reproducibility in proteomics. From the initial PSM data, they identified 5684 proteins, whereas the original study reported only 2855. Although both analyses used the same quantitative dataset, their pipeline, which employed OmicsQ for DE analysis, did not highlight APOE as significantly upregulated in both disease conditions, contrary to the original findings. While the number of peptides mapped to APOE exactly matched 64, likely due to the shared input files, their PCA results failed to replicate the trends observed in the study and appeared uninformative. This discrepancy, despite the use of identical data, emphasizes how variations in analytical tools and workflows can substantially influence results. It highlights the need for transparent and reproducible pipelines in proteomics research.

Conclusion

The main benefit of the proteomics database resource for researchers is to have an enriched database from many different fields to test or to have a new study based on the publicly available data. We think these valuable efforts would benefit the biomedical community in general to make sure about the reproducibility of results and the validity of biological outcomes.

With these examples, we wanted to show that even small data analyses during the university course can be challenging, let alone studies and research reusing this data to drive the decisions and findings in healthcare research. By applying a common, R-based workflow across six PRIDE projects, students showed how seemingly small gaps (unclear design, absent decoys, missing spectral libraries, and unavailable code/parameters) cascade into divergent identifications, quantification, and differential analyses. These observations underscore that reproducibility is not solely an instrument or algorithm problem; it is a metadata, provenance, and transparency problem across the entire pipeline. A common lesson was the systematic absence of SDRF files across the case studies, which obscured the experimental design, complicated statistical contrasts, and forced teams to infer group structure from manuscripts or file names. Routine provision of SDRF would have eliminated much of this ambiguity by explicitly linking samples, runs, factors, and covariates in a machine-readable way. In the same spirit, the label "Complete submission" in repositories does not necessarily equate to re-analysis readiness from a user's perspective; without SDRF, PSI-format identification files, QC summaries, and parameter/code archives, meaningful replication still stalls.

Our experience points to a minimum package that authors must provide to increase the credibility of their findings or the value of their data for others: (i) raw data plus open formats; (ii) identifications in PSI standards (e.g., mzIdentML/mzTab); (iii) SDRF for proteomics; (iv) basic QC summaries (e.g., mzQC when available); (v) spectral libraries for DIA where used; (vi) full parameter files and analysis code with pinned versions/containers[34]. Together, these elements operationalize transparency and the re-analysis routine. We encourage proteomics repositories to expand the list of mandatory information that data creators must provide in order to deposit their findings with a minimum data package. We think that it would not take a lot from data creators since some automatisation tools are available (e.g., the lesSDRF tool for SDRF creation)[35].

Proteomics as a discipline contains many intricate technical concepts that challenge students from diverse academic backgrounds. Similarly, the instructors were challenged by the balancing act of depth and accessibility of the course content. Nevertheless, the rewards were profound: witnessing students grasp cutting-edge scientific methods in proteomics analyses, fostering innovation in open and accessible science, and thereby contributing to advancements in biomedical research.

Finally, this educational model offers a scalable blueprint: students learn modern MS-proteomics by re-analyzing the literature, while the community gains sharper feedback on where data and metadata fall short. If authors deposit a minimum re-analysis package, reviewers assess it, and repositories nudge it at submission, we can collectively prevent valuable datasets from becoming data tombs and instead cultivate a culture where public proteomics is not only open but also reusable, auditable, and reproducible. At the same time, we recognize several limitations that may guide future courses of this kind. Time constraints restricted how comprehensively students could analyze multiple modules, and differences in computational resources or prior experience sometimes widened the learning curve. The inherent complexity and heterogeneity of real-world datasets, while educationally valuable, occasionally proved overwhelming for newcomers, especially without opportunities to clarify missing details with original authors. These limitations are not unique to our course, but acknowledging them may help educators design improved workflows, balance realism with pedagogy, and provide additional scaffolding for students approaching proteomics re-analysis for the first time.

Data availability

The proteomics search software outputs and parameter files generated by students are deposited on Zenodo (https://doi.org/10.5281/zenodo.17975984). Team 5 used quantification data that was obtained directly from data creators under verbal non-disclosure agreement.

Code availability

Students' R scripts and reports are available on the GitHub course page (https://github.com/jafarilab/MSProt2025).

Received: 16 October 2025; Accepted: 14 January 2026;
Published online: 22 January 2026

References

1. Deutsch, E. W. *et al*. The ProteomeXchange consortium at 10 years: 2023 update. *Nucleic Acids Res* **51**, D1539–D1548 (2023).
2. Perez-Riverol, Y. *et al*. The PRIDE database at 20 years: 2025 update. *Nucleic Acids Res* **53**, D543–D553 (2025).
3. Desiere, F. *et al*. The PeptideAtlas project. *Nucleic Acids Res* **34**, D655–8 (2006).
4. Moriya, Y. *et al*. The jPOST environment: an integrated proteomics data repository and database. *Nucleic Acids Res* **47**, D1218–D1224 (2019).
5. Ma, J. *et al*. iProX: an integrated proteome resource. *Nucleic Acids Res* **47**, D1211–D1217 (2019).
6. Sharma, V. *et al*. Panorama Public: A Public Repository for Quantitative Data Sets Processed in Skyline. *Mol Cell Proteomics* **17**, 1239–1244 (2018).
7. Wilkinson, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018 (2016).
8. Gharibi, H. *et al*. A uniform data processing pipeline enables harmonized nanoparticle protein corona analysis across proteomics core facilities. *Nat Commun* **15**, 342 (2024).
9. Tabb, D. L. *et al*. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J Proteome Res* **9**, 761–776 (2010).
10. Doncheva, N. T., Schwämmle, V. & Locard-Paulet, M. Understanding Data Analysis Steps in Mass-Spectrometry-Based Proteomics Is Key to Transparent Reporting. *J Proteome Res* https://doi.org/10.1021/acs.jproteome.5c00287 (2025).
11. Ashkarran, A. A. *et al*. Measurements of heterogeneity in proteomics analysis of the nanoparticle protein corona across core facilities. *Nat. Commun.* **13**, 6610 (2022).
12. Ashkarran, A. A., Gharibi, H., Modaresi, S. M., Saei, A. A. & Mahmoudi, M. Standardizing protein corona characterization in nanomedicine: A multicenter study to enhance reproducibility and data homogeneity. *Nano Lett.* **24**, 9874–9881 (2024).
13. Collins, B. C. *et al*. Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nat Commun* **8**, 291 (2017).
14. Varjosalo, M. *et al*. Interlaboratory reproducibility of large-scale human protein-complex analysis by standardized AP-MS. *Nat Methods* **10**, 307–314 (2013).
15. Taylor, C. F. *et al*. The minimum information about a proteomics experiment (MIAPE). *Nature Biotechnology* **25**, 887–893 (2007).
16. Palmer, L., Levett-Jones, T., Smith, R. & McMillan, M. Academic literacy diagnostic assessment in the first semester of first year at university. *Int. J. First Year High. Educ.* **5** (2014).
17. Nieminen, J. H. & Tuohilampi, L. 'Finally studying for myself' – examining student agency in summative and formative self-assessment models. *Assess. Eval. High. Educ.* **45**, 1031–1045 (2020).
18. Broadbent, J., Panadero, E. & Boud, D. Implementing summative assessment with a formative flavour: a case study in a large class. *Assess. Eval. High. Educ.* **43**, 307–322 (2018).
19. Lekholm, A. K. & Cliffordson, C. *Grades and Grade Assignment: Effects of Student and School Characteristics*. (2008).
20. Lok, B., McNaught, C. & Young, K. Criterion-referenced and norm-referenced assessments: compatibility and complementarity. *Assess. Eval. High. Educ.* **41**, 450–465 (2016).
21. Partanen, L. J., Myyry, L. & Asikainen, H. Physical chemistry students' learning profiles and their relation to study-related burnout and perceptions of peer and self-assessment. *Chem. Educ. Res. Pr.* **25**, 474–490 (2024).
22. Ragupathi, K. & Lee, A. Beyond fairness and consistency in grading: The role of rubrics in higher education. in *Diversity and Inclusion in Global Higher Education* 73–95 (Springer Singapore, Singapore, 2020).
23. Muukkonen, H. & Lakkala, M. Exploring metaskills of knowledge-creating inquiry in higher education. *Int. J. Comput. Support. Collab. Learn.* **4**, 187–211 (2009).
24. Tammeorg, P., Mykkänen, A., Rantamäki, T., Lakkala, M. & Muukkonen, H. Improving group work practices in teaching life sciences: Trialogical learning. *Res. Sci. Educ.* **49**, 809–828 (2019).
25. The mzTab Data Exchange Format: Communicating Mass-spectrometry-based Proteomics and Metabolomics Experimental Results to a Wider Audience. *Molecular & Cellular Proteomics* **13**, 2765–2775 (2014).
26. Combe, C. W. *et al*. mzIdentML 1.3.0 – Essential progress on the support of crosslinking and other identifications based on multiple spectra. *PROTEOMICS* **24**, 2300385 (2024).

27. Colaert, N. *et al*. thermo-msf-parser: An Open Source Java Library to Parse and Visualize Thermo Proteome Discoverer msf Files, https://doi.org/10.1021/pr2005154 (2011).
28. Laurent Gatto, C. V. *Qfeatures*, https://doi.org/10.18129/B9.BIOC.QFEATURES (Bioconductor, 2020).
29. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* **26**, 1367–1372 (2008).
30. Goeminne, L. J. E., Sticker, A., Martens, L., Gevaert, K. & Clement, L. MSqRob Takes the Missing Hurdle: Uniting Intensity- and Count-Based Proteomics. *Anal Chem* **92**, 6278–6287 (2020).
31. Ritchie, M. E. *et al*. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
32. Zhang, X. *et al*. Proteome-wide identification of ubiquitin interactions using UbIA-MS. *Nat Protoc* **13**, 530–550 (2018).
33. Trinh, X.-T., da Costa, A. A., Bouyssié, D., Rogowska-Wrzesinska, A. & Schwämmle, V. OmicsQ: A user-friendly platform for interactive quantitative omics data analysis. *arXiv [q-bio.QM]*, https://doi.org/10.48550/ARXIV.2504.19813 (2025).
34. Sandve, G. K., Nekrutenko, A., Taylor, J. & Hovig, E. Ten simple rules for reproducible computational research. *PLoS Comput. Biol.* **9**, e1003285 (2013).
35. Claeys, T. *et al*. lesSDRF is more: maximizing the value of proteomics data through streamlined metadata annotation. *Nature Communications* **14**, 1–4 (2023).
36. Patel, M. M. *et al*. SOX9-dependent fibrosis drives renal function in nephronophthisis. *EMBO Mol. Med.* **17**, 1238–1258 (2025).
37. PXD061542. https://doi.org/10.6019/PXD061542.
38. Chen, H.-C. *et al*. Proteomic landscape of pattern triggered immunity in the Arabidopsis leaf apoplast. *bioRxivorg*, https://doi.org/10.1101/2025.02.06.636724 (2025).
39. PXD060654, https://doi.org/10.6019/PXD060654.
40. Lee, J. *et al*. Sex differences in single neuron function and proteomics profiles examined by patch-clamp and mass spectrometry in the locus coeruleus of the adult mouse. *Acta Physiologica* **240**, e14123 (2024).
41. PXD045844, https://doi.org/10.6019/PXD045844.
42. Huang, H.-C. *et al*. The branched N-glycan of PD-L1 predicts immunotherapy responses in patients with recurrent/metastatic HNSCC. *Oncogenesis* **13**, 1–10 (2024).
43. PXD045506, https://doi.org/10.6019/PXD045506.
44. Farrell, C. *et al*. Apolipoprotein E abundance is elevated in the brains of individuals with Down syndrome–Alzheimer's disease. *Acta Neuropathologica* **149**, 1–33 (2025).
45. PXD058779, https://doi.org/10.6019/PXD058779.
46. Preveza, N. J. *et al*. Decreases in K63 Polyubiquitination in the Hippocampus Promote the Formation of Contextual Fear Memories in Both Males and Females. *Hippocampus* **35**, e23650 (2025).
47. PXD054726, https://doi.org/10.6019/PXD054726.

Acknowledgements

The course was supported by the DPBM Doctoral Program in Biomedicine at the University of Helsinki. We gratefully acknowledge Liisa Kauppi and Heidi Repo (DPBM) for their valuable guidance and support throughout the course. This study was partially supported by the Doctoral Programme in Biomedicine at the University of Helsinki (grant no. 924112). The open access funding was provided by the Helsinki University Library.

Author contributions

M.J. conceived the study, and M.J. and M.V. supervised and coordinated the course. U.V. and M.J. prepared the first draft of the manuscript. M.S., L.C., P.P., L.I., S.N.G., A.S., T.R., T.G.A., M.S., S.A., O.A., T.T., H.L., K.S., S.S., H.V., and J.R. contributed to the capstone project analysis. U.V. also served as a teaching assistant in the course. Guest speakers V.S. and A.A.S. provided external expertise, while internal speakers M.J., M.V., E.Z., R.S., and S.T.T. contributed to teaching and guidance. All authors participated in writing, editing, and revising the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Uladzislau Vadadokhau[1], Mai Soliman[2,3,4,22], Leticia Castillon[5,22], Paula Pastor Muñoz[6,22], Linda Id[6,22], Swethaa Natraj Gayathri[6,22], Ankita Srivastava[7,8,22], Tyko Runeberg[7,9,22], Tamara González-Armijos[6,22], Karina Šapovalovaitė[6,22], Milda Sakalauskaite[7,10,22],

Sadiksha Adhikari[7,11,22], Oluwatosin Abe[12,22], Tiialotta Tohmola[6,22], Hao Li[6,22], Srividhya Sundaresan[7,13,22], Hanna Vesikukka[6,22], Jannica Roininen[14,22], Ehsan Zangene[1], Rabah Soliymani[1,15,16], Sami T. Tuomivaara[1,15,16], Veit Schwämmle[17], Amir A. Saei[18], Markku Varjosalo[1,19] & Mohieddin Jafari[1,20,21] ✉

[1]Department of Biochemistry and Developmental Biology, Faculty of Medicine, University of Helsinki, Helsinki, Finland. [2]Division of Pharmaceutical Chemistry and Technology, Faculty of Pharmacy, University of Helsinki, Helsinki, Finland. [3]Institute of Biotechnology, Helsinki Institute of Life Sciences, University of Helsinki, Helsinki, Finland. [4]Department of Pharmaceutics, Faculty of Pharmacy, Alexandria University, Alexandria, Egypt. [5]Translational Cancer Medicine Program, Faculty of Medicine, University of Helsinki, Helsinki, Finland. [6]Faculty of Medicine, University of Helsinki, Helsinki, Finland. [7]iCAN Digital Precision Cancer Medicine Flagship, University of Helsinki, Helsinki, Finland. [8]Institute for Molecular Medicine Finland - FIMM, HiLIFE – Helsinki Institute of Life Science, University of Helsinki, Helsinki, Finland. [9]Neuroscience Center, Helsinki Institute of Life Sciences, University of Helsinki, Helsinki, Finland. [10]Faculty of Pharmacy, University of Helsinki, Helsinki, Finland. [11]Institute for Molecular Medicine Finland, Helsinki Institute of Life Science, University of Helsinki, Helsinki, Finland. [12]Turku Bioscience Centre, University of Turku and Åbo Akademi University, Turku, Finland. [13]Research Program in Systems Oncology, Research Programs Unit, Faculty of Medicine, University of Helsinki, Helsinki, Finland. [14]Faculty of Science and Engineering, Cell Biology, Åbo Akademi University, Turku, Finland. [15]Meilahti Proteomics Unit, University of Helsinki, Helsinki, Finland. [16]Helsinki Institute of Life Science, University of Helsinki, Helsinki, Finland. [17]Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense, Denmark. [18]Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden. [19]Molecular Systems Biology Group, Institute of Biotechnology, University of Helsinki, Helsinki, Finland. [20]Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. [21]Tampere Institute for Advanced Study, Tampere University, Tampere, Finland. [22]These authors contributed equally: Mai Soliman, Leticia Castillon, Paula Pastor Muñoz, Linda Id, Swethaa Natraj Gayathri, Ankita Srivastava, Tyko Runeberg, Tamara González-Armijos, Karina Šapovalovaitė, Milda Sakalauskaite, Sadiksha Adhikari, Oluwatosin Abe, Tiialotta Tohmola, Hao Li, Srividhya Sundaresan, Hanna Vesikukka, Jannica Roininen. ✉e-mail: mohieddin.jafari@helsinki.fi