

DISSERTATION FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (PhD)

Bridging Experimental and Computational Barriers in Needle-Biopsy
Proteomics

by Gontse Mabuse Moagi

Supervisor: Éva Csősz, Ph.D., D.Sc.



UNIVERSITY OF DEBRECEN
DOCTORAL SCHOOL OF MOLECULAR CELLULAR AND IMMUNE BIOLOGY

DEBRECEN, 2026

Table of Contents

ABBREVIATIONS	4
1. INTRODUCTION	6
2. LITERATURE REVIEW	8
2.1. Proteomics investigation of tissue samples	8
2.1.1. FFPE dissection strategies	9
2.1.2. Mass Spectrometry-Based Proteomics of FFPE samples	10
2.1.3. Deparaffinization	11
2.1.4. Reversal of formalin-induced crosslinks	12
2.1.5. Enzymatic digestion	13
2.2. Data acquisition, analysis and interpretation	14
2.2.1. Data acquisition	14
2.2.2. Data analysis	15
2.2.2.1. Data analysis of DIA data using DIA-NN software	16
2.2.2.2. Low-code application platforms	17
3. AIMS AND OBJECTIVES	19
4. MATERIALS AND METHODS	20
4.1. Reagents	20
4.2. FFPE Tissue Processing and Protocol Development	20
4.2.1. Approach 1: Chaotropic-Based in-solution digestion (App_1)	21
4.2.2. Approach 2: Chaotropic-, Reductant-, and Surfactant-Free in-solution digestion (App_2)	21
4.2.3. Approach 2M: Chaotropic- and Surfactant-Free Buffer with Reduction/Alkylation (App_2M)	22
4.2.4. Approach 3: Surfactant-Based in-solution digestion (App_3)	22
4.2.5. Approach 4: On-Surface Surfactant-Based digestion (App_4)	23
4.3. Samples used for mass spectrometry analyses	23
4.3.1. Examination of needle biopsy samples obtained from patients with liver disease	23
4.3.2. Examination of proteins involved in adipocyte differentiation	24

4.4. Chromatographic and Mass Spectrometric Conditions	24
4.5. DEF workflow development	26
4.5.1. Benchmark Dataset Case Studies	28
4.6. Data Analysis	28
4.6.1. Data analysis of the FFPE liver needle biopsy sample set	28
4.6.2. Data analysis of the adipose tissue differentiation sample set	30
4.6.3. Data analysis of the benchmark datasets	31
4.6.3.1. Data analysis of the HEK Benchmark dataset	31
4.6.3.2. Data analysis of the Mouse–Yeast Spike-In Benchmark dataset	31
4.6.3.3. Data analysis of the HYE124 Tri-Species Benchmark dataset	31
4.7. Functional analysis	32
5. RESULTS	33
5.1. DIA-NN Easy Filter (DEF) Workflow Development	33
5.1.1. Evaluation of DEF workflow using external, published datasets	34
5.1.2. Evaluation of DEF workflow using dataset produced in our laboratory	40
5.1.3. Examination of FFPE samples	49
5.2. Xylene-Free FFPE Sample Preparation for Proteomics Analysis	49
5.2.1. Method Refinement and Comparative Evaluation	50
5.3. Examination of needle biopsy samples from patients with liver disease	57
6. DISCUSSION	66
7. SUMMARY	73
8. NEW FINDINGS	74
9. REFERENCES	75
10. KEY WORDS	87
11. ACKNOWLEDGEMENTS	88
12. APPENDIX	89

ABBREVIATIONS

ABC	Ammonium bicarbonate	LCM	Laser capture microdissection
ACN	Acetonitrile	LC-MS/MS	Liquid chromatography-tandem mass spectrometry
AFA	Adaptive focused acoustics	MMD	Manual macrodissection
AIF	all-ion fragmentation	MMi	Manual microdissection
APIs	Application programming interfaces	NNs	Neural network
App	Approach	NSI	Nanoelectrospray ion source
CALD	Computer-aided laser dissection	PA	Pre-adipocytes
CV	Coefficients of variation	PACIFIC	Precursor Acquisition Independent From Ion Count
DAP	Differentially abundant proteins	PCA	Principal component analysis
DDA	Data dependent acquisition	PG	Protein group
DEF	DIA-NN_easyfilter	PHAD	Projected hot air deparaffinization
DIA	Data-independent acquisition	PPI	Protein-protein interaction
dll	Dynamic-link library	PTM	Post-translational modifications
DTT	Dithiothreitol	QuantUMS	Quantification method using an Uncertainty Minimising Solution
FA	Formic acid	RSD	Relative standard deviation
FDR	False discovery rate	SDC	Sodium deoxycholate
FFPE	Formaldehyde-fixed paraffin-embedded	SDS	Sodium dodecyl sulfate
GPF	Gas phase fractionation	SGBS	Simpson-Golabi-Behmel syndrome

HEK	Human embryonic kidney	SL	Sodium laurate
HIAR	Heat-induced antigen retrieval	TEAB	Triethylammonium bicarbonate
HMQ	Hot Milli-Q water	TFA	Trifluoroacetic acid
IPA	Ingenuity pathway analysis	WA	Differentiated white adipocytes
IQR	Interquartile range	WP	White adipocytes differentiated in the presence of palmitate
iRT	Indexed retention time	XICs	Extracted ion chromatogram
KNIME	Konstanz information miner	xMD	Expression microdissection
LCAPs	Low-code/no-code data science and integration platforms		

1. INTRODUCTION

Liquid chromatography-tandem mass spectrometry (LC-MS/MS)-based proteomics has become a central analytical platform for the quantitative and qualitative characterization of proteins in complex biological systems, underpinning major advances in both life sciences and clinical research. Enabled by developments in protein ionization technologies and comprehensive sequence databases, this approach allows high-confidence identification and reproducible quantification of thousands of proteins in a single experiment. Importantly, MS-based proteomics enables the simultaneous assessment of protein abundance and a wide spectrum of post-translational modifications (PTMs), including phosphorylation and acetylation, without reliance on target-specific antibodies. Its versatility further allows application across diverse biological materials, ranging from cultured cells and fresh-frozen tissues to formalin-fixed, paraffin-embedded (FFPE) specimens.

FFPE tissues represent the most extensive and accessible archive of human biological material, comprising billions of specimens stored in pathology departments and biobanks worldwide. These samples are frequently linked to long-term clinical follow-up data, positioning them as a uniquely powerful resource for retrospective and translational studies. While genomics and transcriptomics define cellular potential and metabolomics reflects downstream biochemical states, proteins constitute the principal functional layer of biological systems, acting as enzymes, structural elements, signaling mediators, and therapeutic targets. Despite this, protein-level analysis in clinical practice remains largely dependent on immunohistochemistry (IHC). Although well established, IHC is inherently limited by its dependence on highly specific antibodies, restricting analyses to predefined targets. Moreover, its relatively narrow dynamic range, sensitivity to variable and often limited sample input, and inability to resolve protein isoforms constrain its utility for comprehensive and discovery-driven investigations.

Discovery-based MS proteomics, particularly through data-independent acquisition (DIA), overcomes many of these limitations by enabling unbiased, large-scale, and highly reproducible proteome profiling. However, the application of DIA-based proteomics to FFPE

tissues remains far from routine. Significant challenges persist, including the absence of standardized and broadly applicable sample preparation protocols, time-intensive and technically demanding workflows, and the reliance on advanced bioinformatics tools for data analysis and interpretation. These limitations represent critical bottlenecks that hinder clinical implementation and, to date, have precluded the establishment of FDA-approved FFPE-based proteomic diagnostics in routine clinical practice.

2. LITERATURE REVIEW

Proteins, rather than genes, are the principal functional molecules within cells, acting as enzymes, structural components, signaling mediators, and therapeutic targets. Consequently, comprehensive characterization of the proteome is essential for understanding biological processes and disease mechanisms [1]. In clinical pathology, IHC remains the gold standard for protein-level analysis because of its ability to localize proteins within tissue architecture [2]. However, IHC is inherently limited by its dependence on highly specific and validated antibodies, restricting analyses to predefined targets and reducing its suitability for discovery-driven investigations. Additional limitations include a relatively narrow dynamic range, sensitivity to variable and often limited sample input, and limited capacity to distinguish between protein isoforms with distinct biological and clinical significance [3,4].

2.1. Proteomics investigation of tissue samples

Proteomic investigations can be performed using both biofluids and tissue specimens. While biofluid-based proteomics is widely used for biomarker discovery and disease monitoring [5–7], tissue-based proteomics provides important spatial and cellular context that cannot be obtained from biofluids alone [8–12]. Tissue specimens are commonly preserved either as fresh-frozen (FrFr) or FFPE samples. FrFr tissues are generally regarded as the gold standard for proteomic analyses because of their superior preservation of protein integrity. However, their widespread use is limited by the logistical and financial demands associated with long-term cold-chain storage. In contrast, FFPE tissues constitute the largest repository of preserved human biological material worldwide, encompassing billions of archived specimens stored in pathology departments and biobanks [13,14]. These collections are frequently linked to extensive longitudinal clinical data, making them an invaluable resource for retrospective and translational research.

2.1.1. FFPE dissection strategies

FFPE proteomic workflows can be performed using either tissue scrolls collected directly from FFPE blocks or tissue sections mounted on microscope slides. Slide-based sampling strategies include bulk tissue scraping, manual macrodissection (MMD), manual microdissection (MMi), laser capture microdissection (LCM), and expression microdissection (xMD) [15]. Among these approaches, MMD is performed without the use of a microscope or specialized instrumentation, whereas MMi and LCM enable more precise isolation of defined tissue regions.

For most slide-based dissection methods, a pathologist or trained histologist is required to identify regions of interest (ROIs) prior to sample collection. This additional step can increase both costs and turnaround times. Expression microdissection differs from other approaches by employing molecular tags, typically antibodies, to selectively isolate target cell populations; however, its application remains constrained by the availability and specificity of suitable antibodies.

Recent advances in machine learning and digital pathology have facilitated the development of automated ROI identification tools aimed at reducing subjectivity and increasing throughput. For example, computer-aided laser dissection (CALD) workflows utilizing image analysis platforms such as ImageJ have shown promise for improving the accuracy and efficiency of tissue selection [16]. Nevertheless, LCM still requires specialized instrumentation and dedicated consumables, while xMD remains limited by antibody-dependent targeting strategies.

Consequently, despite its lower spatial resolution compared with microdissection-based methods, MMD remains one of the most widely adopted tissue sampling strategies in clinical proteomics and oncology research due to its simplicity, low cost, and compatibility with routine pathology workflows [17,18].

2.1.2. Mass Spectrometry-Based Proteomics of FFPE samples

Mass spectrometry-based proteomics has emerged as a powerful and unbiased alternative for large-scale protein characterization, with broad applications across life sciences and clinical research [19,20]. Among these approaches, liquid chromatography-tandem mass spectrometry (LC-MS/MS)-based proteomics has become a fundamental analytical platform for the comprehensive analysis of proteins in complex biological systems [21].

Most LC-MS/MS proteomic studies employ a bottom-up strategy, in which proteins are enzymatically digested into peptides, which are subsequently identified and quantified to infer the original protein composition of a sample. Because peptide sequence coverage is typically incomplete, proteins are assembled into protein groups (PGs) based on shared peptide evidence [22]. Conventionally, a protein group is considered confidently identified when supported by at least two or more unique peptides. Although this criterion improves identification confidence, it may also reduce proteome depth by excluding proteins represented by fewer unique peptides, an approach that utilizes the so-called proteotypic peptides, that are specific to the proteins [23].

Despite the considerable value of FFPE tissue archives, their application in LC-MS/MS-based proteomics remains technically challenging because formalin fixation induces extensive protein crosslinking reactions [24]. Formaldehyde primarily reacts with ϵ -amino groups of lysine residues and N-terminal amino groups to form methylol adducts (**Figure 1**). These intermediates can subsequently undergo dehydration to generate Schiff bases [25], which may further participate in reactions involving amino acid residues such as arginine, asparagine, glutamine, histidine, tryptophan, and tyrosine [26]. However, the predominant crosslinking pathway proceeds through methylol intermediates, ultimately resulting in the formation of stable methylene bridges [27].

These covalent modifications reduce protein solubility, hinder protease accessibility, and impair enzymatic digestion efficiency, thereby limiting peptide recovery and protein identification. In addition, paraffin embedding introduces the need for effective deparaffinization prior to protein extraction. Consequently, successful FFPE LC-MS/MS-based

proteomics workflows require optimized strategies for deparaffinization, reversal of formalin-induced crosslinks, protein extraction, and enzymatic digestion to enable efficient downstream analysis.

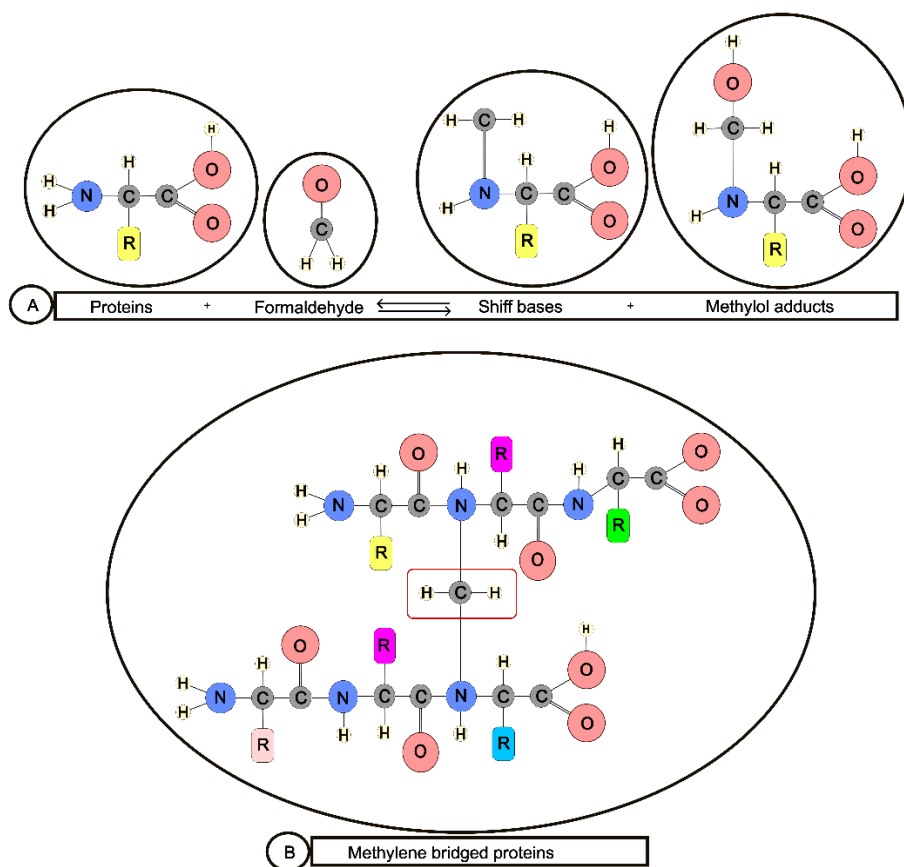


Figure 1. Formalin-induced fixation. (A) Reaction of formaldehyde with proteins and (B) formation of methylene bridged proteins.

2.1.3. Deparaffinization

Traditional deparaffinization methods rely on organic solvents, most commonly xylene, followed by graded ethanol washes for tissue rehydration [28,29]. Xylene remains the most widely used deparaffinization agent due to its rapid action and cost-effectiveness. However, it is a harsh solvent, and exposure is associated with adverse health effects, including headaches,

dizziness, nausea, and irritation of the eyes and respiratory tract, and requires constant monitoring of the workers [30].

In response to these limitations, solvent-free or “green chemistry” deparaffinization strategies have been proposed to simplify workflows and reduce chemical exposure. These approaches include hot water-based methods, projected hot air deparaffinization [31–34], oven-based techniques [35], as well as the use of alternative reagents such as coconut and turpentine oil [36]. Such strategies aim to provide safer and more environmentally sustainable alternatives while maintaining effective paraffin removal.

2.1.4. Reversal of formalin-induced crosslinks

Crosslink reversal and protein extraction from FFPE tissues can be achieved through a chemical de-crosslinker such as hydroxylamine under mild conditions [37], or more commonly through the application of thermal energy (typically 95-120 °C) in combination with optimized buffer systems, a process known as heat-induced antigen retrieval (HIAR), originally developed for immunohistochemistry [3,14]. Efficient protein solubilization during this process requires the use of detergents and chaotropic agents, which play a critical role in disrupting protein-protein interactions, unfold proteins, and enhancing extraction efficiency. Among chaotropic agents, urea is widely used because it effectively denatures proteins while remaining compatible with trypsin digestion at sufficiently low concentrations. Furthermore, urea can be readily removed during peptide desalting prior to LC-MS/MS analysis [9]. Commonly employed strong detergents include sodium dodecyl sulfate (SDS), sodium deoxycholate (SDC) and sodium laurate (SL). SDC and SL offer the advantage of being compatible with proteolytic digestion and can be easily removed by acid precipitation prior to LC-MS/MS analysis. In contrast, although SDS is inexpensive, it is incompatible with downstream LC-MS/MS analysis and necessitate additional cleanup procedures, such as protein precipitation or filter-based buffer exchange, which can introduce sample loss, particularly in low-input settings [38–40].

The need for low-input proteomic workflows is driven by the intrinsic limitations of clinically relevant and biologically informative sample types. Rare cell populations, including

specific immune subsets, stem cells, and spatially defined microenvironments, often constitute only a small fraction of total tissue and cannot be effectively studied using bulk approaches without loss of biological specificity [41–43]. In parallel, contemporary clinical practice increasingly relies on minimally invasive sampling techniques, such as core needle biopsies and fine-needle aspirates, which yield limited amounts of material but are essential for reducing patient risk and enabling longitudinal sampling [32,44]. As a result, the available tissue frequently falls below the input requirements of conventional proteomic workflows [45].

These challenges can be partially mitigated by non-detergent alternatives [8], or through the use of mass spectrometry-compatible surfactants, such as RapiGest and PPS Silent Surfactant, which are designed to degrade under mild acidic conditions prior to LC-MS/MS analysis [46]. This property enables efficient protein solubilization while eliminating the need for additional cleanup steps, thereby reducing sample loss and improving workflow suitability for low-input applications.

2.1.5. Enzymatic digestion

LC-MS/MS-based proteomics relies on the analysis of peptides typically ranging from 7 to 30 amino acids in length, which are generated through proteolytic digestion of proteins. Trypsin is the most commonly used protease due to its high specificity, cleaving at the C-terminus of lysine and arginine residues, except when followed by proline [14,47]. In FFPE tissues, however, formalin-induced crosslinking and chemical modifications can adversely affect proteolytic efficiency. Although many crosslinks are reversible, certain amino acid residues, particularly lysine, may remain chemically modified, for example through methylation [26,48]. These modifications can hinder trypsin cleavage, resulting in incomplete digestion and the generation of peptides outside the optimal size range for LC-MS/MS analysis.

Consequently, the reduced availability of detectable peptide precursors negatively impacts protein identification, scoring, and overall analytical reliability. The use of alternative or complementary proteases with different cleavage specificities may partially alleviate this limitation; however, such approaches would increase sample and data complexity and can

introduce additional challenges in data analysis. As a result, extended digestion times, typically overnight or longer, are often required to improve digestion efficiency [49]. Alternatively, accelerated digestion strategies, such as those employing adaptive focused acoustics (AFA) [50], have been developed, although these approaches may involve increased cost and specialized instrumentation.

Despite ongoing methodological developments, the absence of standardized workflows, extended processing times, and variability in protocol performance remain major barriers to reproducibility and clinical implementation. These limitations contribute to the current lack of FFPE-based proteomic diagnostic assays in routine clinical practice.

2.2. Data acquisition, analysis and interpretation

2.2.1. Data acquisition

Data acquisition strategies in mass spectrometry-based proteomics differ in the sequence of precursor and fragment ion scans, as well as in the manner by which analytes are selected for fragmentation. These strategies are broadly classified into two categories: data-dependent acquisition (DDA) and DIA. In DDA, the instrument selectively fragments the most abundant precursor ions detected during the MS1 scan. Because ions are selected as they are observed, this strategy is inherently biased toward high-abundance species and can result in stochastic precursor selection, reducing reproducibility between analyses. In contrast, DIA systematically fragments all precursor ions within predefined mass-to-charge windows regardless of their observed abundance. This comprehensive sampling strategy improves reproducibility and quantitative completeness [51–53] (**Figure 2**).

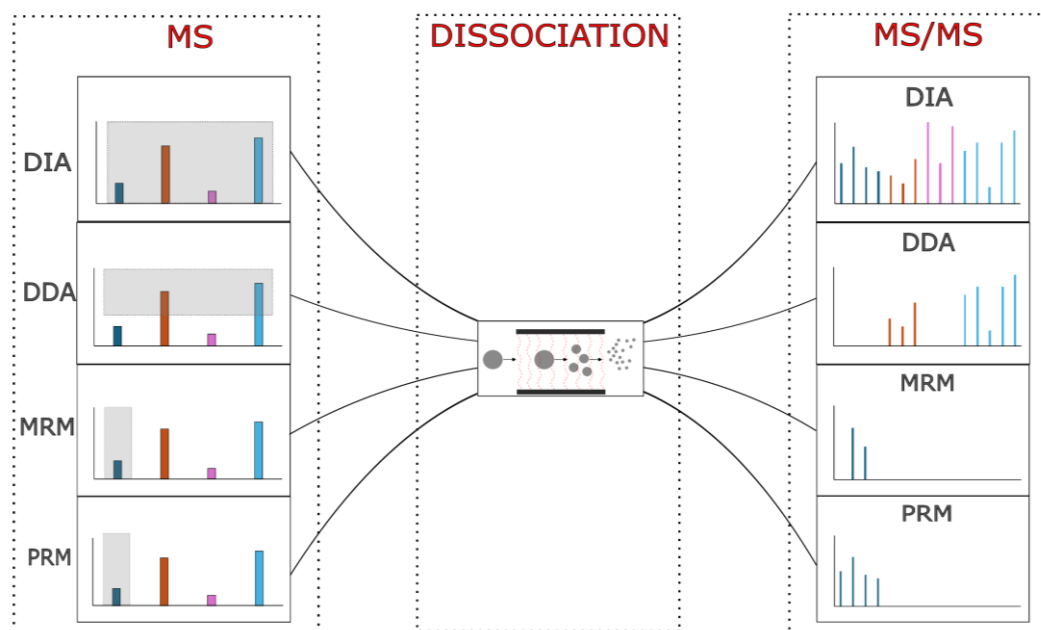


Figure 2. Conceptual diagrams of DIA, DDA, MRM, and PRM strategies at MS and MS/MS level.

A widely adopted DIA implementation is SWATH (Sequential Window Acquisition of All Theoretical Mass Spectra), in which consecutive precursor isolation windows are fragmented across the full mass range. Other DIA variants include MSE, Precursor Acquisition Independent From Ion Count (PACIFIC), and all-ion fragmentation (AIF) [51,54,55].

Targeted acquisition strategies can also be implemented independently of DDA and DIA. The most widely used targeted methods are selected/multiple reaction monitoring (SRM/MRM), and parallel reaction monitoring (PRM) [56,57]. These approaches focus on predefined analytes and are commonly employed for the sensitive and reproducible quantification of specific proteins or peptides [58,59].

2.2.2. Data analysis

A wide range of bioinformatics tools has been developed for the analysis of DDA data and the assessment of peptide and protein identification quality. Commonly used database search engines include Andromeda, MSGF+, Comet, MSFragger, and SAGE [60–63]. The

analysis of DIA data presents additional computational challenges due to the highly multiplexed nature of the acquired MS/MS spectra, in which fragment ions originating from multiple co-isolated precursor ions are recorded simultaneously [51,52]. Consequently, specialized computational approaches have been developed to interpret DIA datasets. These approaches can be broadly categorized as peptide-centric and spectrum-centric. Peptide-centric methods directly interrogate DIA data using experimentally acquired or computationally predicted spectral libraries. By leveraging prior knowledge of expected fragment ion patterns, retention times, and other peptide characteristics, peptide-centric approaches achieve high sensitivity and quantitative precision but limited to peptides represented in the library. In contrast, spectrum-centric approaches attempt to deconvolute co-eluting precursor and fragment ions from the acquired DIA spectra first, generating pseudo-MS/MS spectra that can then be searched using traditional database-search engines developed for DDA analysis. Spectrum-centric methods enable library-free identification of novel peptides but is computationally intensive and can be less sensitive for low-abundance peptides due to mixed-precursor interference at fragment level [20,39,53].

Recent advances in DIA data analysis have increasingly combined peptide-centric and spectrum-centric strategies, integrating spectral library searching with library-free database searching and machine learning-based prediction of peptide properties to improve identification coverage and quantitative performance [22,64].

Consequently, the widespread adoption of DIA has shifted key analytical challenges from data acquisition toward data processing and interpretation. To address this complexity, a range of computational tools, including Spectronaut [65], ScaffoldDIA (Proteome Software), and EncyclopeDIA [66], have been developed with DIA-NN [22,65,67] emerging as one of the most widely used and high-performing solutions.

2.2.2.1. Data analysis of DIA data using DIA-NN software

DIA-NN employs machine learning and deep neural networks to enhance peptide identification and quantification accuracy by integrating multiple scoring features derived from precursor and fragment ion signals and applying target-decoy strategies for false discovery rate control. Recent developments, including DIA-NN version 2.0, have further improved

performance, particularly through the introduction of the QuantUMS module, which enables machine learning-based relative quantification of precursors and proteins, as well as enhanced handling of post-translational modification (PTM)-focused analyses [68,69].

Despite these advances, filtering low-confidence identifications and quantitative values remains a critical step when analyzing complex proteomic datasets. DIA-NN generates its primary output in a compact, compressed .parquet format, which requires specialized tools, typically in R or Python with packages such as PyArrow modules for downstream processing. R-based solutions can represent powerful and, in some cases, more streamlined alternatives. Script-based workflows provide full analytical traceability, version control, and high statistical flexibility, particularly when custom models or advanced statistical frameworks are required. However, their effective implementation generally requires proficiency in R, and fully script-based environments may increase the risk of misconfiguration or unintended analytical errors for non-programming users. R-based solutions such as MSDAP [70] and DIAgiu [71] primarily focus on quantitative statistics and provide limited support for integrated evaluation of identification performance. Consequently, some users continue to prefer earlier versions of DIA-NN that provided .tsv outputs compatible with standard spreadsheet software. Manual filtering approaches, however, are labor-intensive, time-consuming, and prone to user bias and error. Therefore, there is a clear need for solutions that retain the advanced capabilities of modern DIA-NN versions while enabling intuitive, reproducible data analysis for users without coding expertise.

2.2.2.2. Low-code application platforms

Low-code application platforms (LCAPs) have emerged as a promising solution to this challenge. These environments enable users to develop applications and automate data analysis workflows through graphical interfaces, without requiring advanced programming skills. LCAPs have been shown to significantly accelerate development compared to traditional coding approaches [72,73], contributing to their increasing adoption in biomedical research. Platforms such as Galaxy and Konstanz Information Miner (KNIME) are widely used in this context [74–76].

While Galaxy provides a web-based environment, KNIME is a self-contained, open-source, model-driven LCAP with a broad range of applications, including business data analysis research [77] and cheminformatics [78]. KNIME was built around three main principles: i) visually interactive data processing and analysis framework, ii) modularity, and iii) easy extensibility. A workflow in KNIME is a visually interactive direct acyclic graph with a modular setup built from nodes as basic processing units. KNIME facilitates easy extensibility allowing access to external data sources using a variety of application programming interfaces (APIs) and nodes that allow conventional coding [79].

To support collaborative development and address limitations commonly associated with open-source software, KNIME provides a community-driven Hub for sharing workflows, components, and extensions. Through this platform, users can access and reuse proteomics-related workflows developed by the OpenMS initiative [80] and the broader KNIME community, thereby promoting reproducibility and accelerating method development. However, no dedicated workflow currently supports comprehensive filtering and evaluation of DIA-NN main reports.

3. AIMS AND OBJECTIVES

The overarching aim of this thesis was to reduce key technical and analytical barriers limiting the adoption of FFPE tissue proteomics, thereby facilitating its broader application in clinical and translational research.

To achieve this aim, the following objectives were pursued:

- i. To optimize FFPE protein extraction and digestion workflows with an emphasis on simplicity, reproducibility, and compatibility with low-input clinical specimens.
 - Evaluate alternative deparaffinization strategies and mass spectrometry-compatible sample preparation methods.
 - Assess the impact of workflow modifications on protein recovery, proteome coverage, and analytical reproducibility.
 - Identify a practical and robust workflow that minimizes sample loss and experimental complexity.
 - Apply the optimized protocol to FFPE liver needle biopsy specimens as a proof-of-concept.
- ii. To develop an accessible data analysis workflow for DIA proteomics.
 - Design DIA-NN EasyFilter (DEF), a user-friendly data processing and quality assessment workflow implemented using a low-code/no-code platform.
 - Evaluate its performance using both publicly available and in-house generated DIA proteomics datasets across diverse data types.

4. MATERIALS AND METHODS

4.1. Reagents

All reagents were purchased from Sigma Aldrich (St. Louis, MI, USA) if not stated otherwise. Pierce Trypsin/Lys-C was purchased from ThermoFisher Scientific (San Jose, CA, USA) and MS-grade trypsin purchased from AB Sciex (Framingham, MA, USA).

4.2. FFPE Tissue Processing and Protocol Development

FFPE tissue material used for protocol development consisted of archived patient samples obtained from routine diagnostic procedures. Serial FFPE tissue sections (10 μ m thickness) were mounted onto glass slides and baked at 60 °C for 2 h to improve tissue adherence and prevent detachment during processing. Deparaffinization was subsequently performed using either projected hot air deparaffinization (PHAD) or a hot water-based method (HMQ). For PHAD, paraffin was melted using a conventional hair dryer operated at maximum power for 20 min. Slides were immediately immersed in distilled water for 5 min to promote tissue rehydration and solidification of residual paraffin. For the HMQ approach, 1 mL of distilled water preheated to 80 °C was applied dropwise onto the tissue sections to facilitate paraffin removal.

Following deparaffinization, the tissue area was delineated using a ruler to ensure consistent sampling. Approximately 0.025mm³ of tissue was collected using a sterile scalpel in five technical replicates and processed according to the workflows described below. Following all workflows, peptides were purified using Pierce™ C18 Tips (Thermo Fisher Scientific), dried, and stored at -20 °C until LC-MS/MS analysis.

4.2.1. Approach 1: Chaotropic-Based in-solution digestion (App_1)

This protocol was adapted with minor modifications from a previously described method [81]. Heat-induced antigen retrieval (HIAR) was performed prior to tissue collection by incubating slides in citrate buffer (95 mM trisodium citrate, 21 mM citric acid, pH 6.0) at 75 °C for 60 min. Tissue material was subsequently scraped from the slides and resuspended in 75 µL of 8 M urea. Protein solubilization was achieved by sonication in a chilled (~1.5 °C) water bath using 15 cycles of 1 min sonication followed by 30 s pauses.

Samples were reduced with 10 mM dithiothreitol (DTT) for 30 min at 37 °C and alkylated with 15 mM iodoacetamide (IAA) for 20 min at room temperature in the dark. The solution was diluted to 4 M urea using 50 mM Tris-HCl (pH 8.0), followed by incubation with 40 ng Pierce™ Trypsin/Lys-C for 4 h at 37 °C. Prior to overnight digestion, the urea concentration was further reduced to 0.8 M and an additional 60 ng trypsin was added. Digestion was terminated by acidification with 6% trifluoroacetic acid (TFA) to pH < 3.

4.2.2. Approach 2: Chaotropic-, Reductant-, and Surfactant-Free in-solution digestion (App_2)

This protocol was adapted from a previously published method [8]. Collected tissue was washed with 50 µL of 100% acetonitrile (ACN) and dried using a SpeedVac system (Thermo Fisher Scientific). Samples were resuspended in 40 µL of 60 mM triethylammonium bicarbonate (TEAB) and incubated at 95 °C for 60 min. Subsequently, 10 µL of 60% ACN in 60 mM TEAB was added, followed by incubation at 75 °C for an additional 60 min. Pre-digestion was performed using 40 ng Trypsin/Lys-C for 4 h at 37 °C, followed by overnight digestion with 60 ng trypsin at 37 °C. The reaction was quenched by acidification with 6% TFA to pH < 3.

4.2.3. Approach 2M: Chaotropic- and Surfactant-Free Buffer with Reduction/Alkylation (App_2M)

This modified workflow incorporated reduction and alkylation steps into the App_2 protocol. Following incubation at 95 °C in 60 mM TEAB, samples were reduced using 10 µL of 60% ACN containing 80 mM DTT and incubated at 75 °C for 60 min. Alkylation was then performed with 15 mM IAA for 30 min at room temperature in the dark. Subsequent pre-digestion and overnight digestion steps were performed as described for App_2. The digestions were terminated by acidification with 6% TFA to pH < 3.

4.2.4. Approach 3: Surfactant-Based in-solution digestion (App_3)

This protocol was adapted with minor modifications from a previously described method [9]. Tissue samples were solubilized in 100 µL of freshly prepared buffer containing 80 mM HEPES (pH 8.0), 80 mM DTT, and 4% SDS. Samples underwent two rounds of sonication in a chilled (~1.5 °C) water bath, each consisting of 15 cycles of 1 min sonication and 30 s pauses, followed by incubation at 99 °C for 1 h. Following centrifugation (15 min, 13,000 × g, room temperature), the supernatant was collected and alkylated with IAA to a final concentration of 15 mM for 30 min in the dark. Proteins were precipitated by adding four volumes of ice-cold acetone and incubating samples overnight at -20 °C. Samples were then centrifuged (40 min, 16,000 × g, 4 °C), and the resulting pellets were washed twice with 80% ice-cold acetone.

Protein pellets were resuspended in 25 µL digestion buffer containing 1.5 M urea and 100 mM HEPES (pH 8.0) using brief sonication (3 cycles). Trypsin/Lys-C digestion (1:30, w/w) was performed for 4 h at 37 °C with shaking (600 rpm), followed by dilution to 0.75 M urea and overnight trypsin digestion (1:20, w/w). Digestion was terminated by acidification with TFA to pH < 3.

4.2.5. Approach 4: On-Surface Surfactant-Based digestion (App_4)

This method was adapted with minor modifications from a previously reported workflow [82]. HIAR was performed as described for App_1. Reduction was carried out directly on the tissue surface using 3 μ L droplets containing 20% glycerol, 0.1% SDS, and 5 mM DTT, followed by incubation for 40 min at 37 °C. Alkylation was subsequently performed using 3 μ L droplets containing 10 mM IAA in 25 mM ammonium bicarbonate (ABC) for 20 min at room temperature in the dark. Pre-digestion was conducted in two cycles using 3 μ L droplets containing 20% glycerol and 40 ng Trypsin/Lys-C in 50 mM ABC for 40 min at 37 °C in a humidified chamber. Final digestion was performed in three cycles using 200 ng trypsin under identical conditions. Peptides were extracted by repeated pipetting (5 $\times$) with 3 μ L of 10% acetic acid and collected into LoBind tubes.

4.3. Samples used for mass spectrometry analyses

4.3.1. Examination of needle biopsy samples obtained from patients with liver disease

This cohort included patients with histologically normal liver tissue (n = 3), fibrosis (n = 3), or hepatic steatosis (n = 3), aged between 32 and 72 years, with an approximately balanced female-to-male ratio. Needle biopsy specimens were fixed in 4% paraformaldehyde at 4 °C for 6 h and sequentially dehydrated in 75%, 80%, 90%, 95%, and 100% ethanol. FFPE blocks were sectioned at 10 μ m thickness using an HM34E rotary microtome (Eppendorf Hungary, Budapest, Hungary) according to standard histological procedures and mounted onto Superfrost™ glass slides. These steps were performed at the Pathology Department of the University of Debrecen. The study was approved by the Scientific and Research Ethics Committee of the Scientific Council of Health of the Hungarian Government (approval number: IV/8465-3/2021/EKU).

4.3.2. Examination of proteins involved in adipocyte differentiation

This case study involved analysis of three biological replicates from three experimental groups: pre-adipocytes (PA), differentiated white adipocytes (WA), and white adipocytes differentiated in the presence of palmitate (WP) to mimic high-fat conditions. All cells were derived from subcutaneous adipose tissue obtained from an infant diagnosed with Simpson-Golabi-Behmel syndrome (SGBS) [83]. Differentiation was induced for 7 days using a standard adipogenic differentiation medium as previously described by [84]. WP cells were additionally treated with 100 μ M sodium palmitate during differentiation. Following cell harvesting, samples were lysed in 1% SDS and protein concentrations were determined using the Pierce™ BCA Protein Assay Kit (Thermo Fisher Scientific). A total of 110 μ g protein per sample was digested using micro S-Trap™ columns (ProtiFi, Fairport, NY, USA).

4.4. Chromatographic and Mass Spectrometric Conditions

The chromatographic and mass spectrometric conditions used throughout this study are presented for FFPE needle biopsy protocol development, followed by the liver disease sample set and adipose differentiation samples.

Digested peptides from FFPE needle biopsy protocol development samples were reconstituted in 10 μ L of 1% formic acid (FA) and analyzed using single-shot DIA LC-MS/MS. Chromatographic separation was performed using an EASY-nLC 1200 system coupled to an Orbitrap Fusion mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA). A total of 5 μ L sample was injected onto a nanoACQUITY™ trap column (180 μ m $\times$ 20 mm, 5 μ m particle size, Symmetry C18; Waters, Milford, MA, USA) at a constant flow rate of 1 μ L/min using solvent A. Peptide separation was subsequently achieved on an ACQUITY UPLC™ M-Class Peptide BEH C18 analytical column (75 μ m $\times$ 150 mm, 1.7 μ m particle size; Waters) operated at a flow rate of 300 nL/min. Peptides were eluted using a multistep gradient consisting of 0-2% solvent B over 2 min, followed by an increase to 25% solvent B over 52 min and to 38% solvent B over an additional 11 min. Solvent B was then increased to 90% over 4 min and maintained for 3

min before re-equilibration to 2% solvent B over 1 min and a final 2 min hold period. Mass spectrometric analysis was performed using a Nanospray Flex™ nanoelectrospray ion source (Thermo Fisher Scientific) operated in positive ion mode at 2.3 kV, with the ion-transfer tube maintained at 275 °C. MS1 spectra were acquired in the Orbitrap analyzer at 120,000 resolution across an m/z range of 350-1200. DIA fragmentation was performed using fixed 12 m/z isolation windows with 55 scan events and higher-energy collisional dissociation (HCD) at 30% normalized collision energy. MS2 spectra were acquired in the Orbitrap analyzer at 30,000 resolution over an m/z range of 145-1450, using an automatic gain control (AGC) target of 2.0×10^5 and a maximum injection time of 60 ms.

LC-MS/MS analysis for the liver disease sample set was performed using a Waters ACQUITY UPLC M-Class system (Waters, Milford, MA, USA) coupled to an Orbitrap Exploris™ 240 mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA). A total of 2 µL sample was injected onto a Symmetry® C18 trap column (180 µm × 20 mm, 5 µm particle size, 100 Å pore size; Waters) at a flow rate of 7 µL/min using solvent A. Peptide separation was performed on an ACQUITY UPLC® M-Class Peptide BEH C18 analytical column (75 µm × 250 mm, 1.7 µm particle size, 130 Å pore size; Waters) operated at 200 nL/min. The chromatographic gradient consisted of 3-8% solvent B over 1 min, followed by an increase to 32% solvent B over 81 min and subsequently to 40% solvent B over 3 min. Solvent B was then increased to 90% over 1 min and maintained for 2 min prior to re-equilibration to 3% solvent B over 1 min, followed by a 15 min equilibration period. The nanoelectrospray ion source was operated at 1.8 kV in positive ion mode with an ion-transfer tube temperature of 280 °C. MS1 spectra were acquired at 30,000 resolution across an m/z range of 380-985. DIA fragmentation was performed using fixed 10 m/z isolation windows with 60 scan events and HCD fragmentation at 28% normalized collision energy. MS2 spectra were acquired in the Orbitrap analyzer at 11,250 resolution using an AGC target of 2.0×10^6 and a maximum injection time of 40 ms. DIA scan ranges were automatically assigned by the instrument software.

LC-MS/MS analysis for adipose differentiation sample set was performed using an EASY-nLC 1200 system (Thermo Fisher Scientific, San Jose, CA, USA). Peptide samples were reconstituted in 10 µL of 1% FA, and 5 µL was injected onto an ACQUITY UPLC M-Class

Symmetry C18 trap column (180 μm $\times$ 20 mm, 5 μm particle size, 100 Å pore size; Waters Corporation, Milford, MA, USA) at a constant flow rate of 1 $\mu\text{L}/\text{min}$. Analytical separation was achieved using a Biozen XB-C18 column (150 $\times$ 0.075 mm, 2.6 μm particle size; Phenomenex, Torrance, CA, USA) operated at a flow rate of 700 nL/min. Peptides were eluted over 50 min using a segmented gradient consisting of 5-6% solvent B over 3 min, followed by sequential increases to 9%, 14%, 19%, 24%, 30%, and 51% solvent B over 2, 6, 6, 6, 7, and 11 min, respectively. Solvent B was subsequently increased to 99% over 0.5 min and maintained for 2 min before re-equilibration to 1% solvent B over 0.5 min and a final 6 min hold period. Mass spectrometric analysis was performed using a Nanospray Flex™ nanoelectrospray ion source (Thermo Fisher Scientific) operated in positive ion mode at 2.3 kV, with the ion-transfer tube maintained at 275 °C. MS1 spectra were acquired in the Orbitrap analyzer at 120,000 resolution across an m/z range of 350-1200. DIA fragmentation was performed using fixed 12 m/z isolation windows with 55 scan events and higher-energy collisional dissociation (HCD) at 30% normalized collision energy. MS2 spectra were acquired in the Orbitrap analyzer at 30,000 resolution over an m/z range of 145-1450, using an automatic gain control (AGC) target of 2.0×10^5 and a maximum injection time of 60 ms.

4.5. DEF workflow development

The DEF workflow was developed using the open-source KNIME Analytics Platform 5.4.3 (www.knime.com). Installation and usage instructions are provided in the DEF user guide available through the DIA-NN EasyFilters GitHub repository (https://github.com/MabuseM/DIA-NN_EasyFilters/blob/main/DEF_User_guide.pdf). All required extensions were installed directly within KNIME, and no external libraries beyond those available in the platform were used. The workflow accepts DIA-NN output files, including PARQUET or TSV reports, pg.matrix files, and optional extracted ion chromatogram (XIC) reports. Data import was performed using PARQUET and File Reader nodes, after which downstream analyses were organized into modular sub-workflows, including “XICs filter,” “Library choice filter,” “Task selector,” and “Quantity filter.” These modules were designed to

be user-configurable through graphical dialog interfaces without requiring detailed knowledge of node structure or data flow.

Within the “XICs filter” module, fragment ion features corresponding to b- and y-ions were extracted from DIA-NN XIC outputs and concatenated into unique feature sets. Only peptides supported by at least four consecutive fragment ions were retained and used for filtering the main DIA-NN report.

The “Library choice filter” incorporated curated contaminant databases compiled from MaxQuant [85], cRAP (<https://www.thegpm.org/crap/>), and Hao-Group [86] contaminant datasets. Contaminant entries and Biognosys iRT peptides were flagged and excluded from downstream protein group (PG) identification. Users were additionally able to upload custom contaminant FASTA files.

Protein inference was implemented within the “Task selector” module using either a two-unique-peptide rule or a proteotypic peptide-based strategy. Under the two-peptide rule, duplicate protein entries were removed and only proteins supported by more than one unique peptide sequence were retained. In the proteotypic approach, non-proteotypic peptides were filtered analogously, whereas proteins supported by at least one proteotypic peptide were retained.

Quality thresholds were applied at both the run and library levels using $PG.Q.Value \leq 0.05$ and $Lib.PG.Q.Value$ and $Lib.Q.Value \leq 0.01$. $PG.MaxLFQ$ values not meeting these thresholds were replaced with an arbitrary value of “50” to indicate identified but non-quantified proteins. Quantitative information was derived directly from the DIA-NN output using the quantification algorithms implemented in the corresponding DIA-NN version, including MaxLFQ and QuantUMS.

Visualization modules were implemented as KNIME composite components to summarize qualitative and quantitative results. These included bar plots displaying the number of identified PGs, peptide summary tables reporting retained and excluded peptides following XIC filtering, stacked bar plots summarizing run-level quantification performance, and binary parallel coordinate plots representing protein identification patterns across runs. In addition,

pie charts were generated to visualize the abundance distribution of all identified PGs, including contaminants.

Group-level visualization was implemented within an interactive component supporting comparisons across two to five experimental groups. User-defined parameters included missing value thresholds and relative standard deviation (RSD) cut-offs. Finally, all output tables were reformatted to ensure compatibility with downstream statistical analysis in MetaboAnalyst [87].

4.5.1. Benchmark Dataset Case Studies

Three publicly available benchmark datasets were used to evaluate analytical performance. These included: (i) the HEK cell dataset (PXD029738), containing triplicate LC-MS/MS measurements of 100 ng peptide loads derived from Human Embryonic Kidney cells [67]; (ii) the mouse-yeast spike-in dataset (PXD034709), comprising five technical replicates from six samples containing varying proportions (5-40%) of mouse membrane proteins spiked into a yeast protein background [88]; and (iii) the HYE124 tri-species dataset (PXD002952), consisting of human, *Escherichia coli*, and yeast protein digests [89].

4.6. Data Analysis

Different data analysis strategies were applied in case of each sample set to meet the specific criteria and needs required for the sample sets.

4.6.1. Data analysis of the FFPE liver needle biopsy sample set

Raw DIA files were processed using DIA-NN version 2.2 in library-free search mode [22,69]. As DIA-NN supports endogenous peptide-based retention time alignment, indexed retention time (iRT) peptides were not added to the samples. Protein identification was performed against a reviewed human protein FASTA database (20,420 entries) downloaded

from UniProtKB on 2025-07-26. *In silico* spectral libraries were generated using the “FASTA digest” and “Deep-learning spectra” options. Search parameters included Trypsin/P specificity with a maximum of two missed cleavages and up to three variable modifications per peptide. Protein N-terminal methionine excision was enabled. Carbamidomethylation of cysteine residues was specified as a fixed modification for all workflows except App_2, whereas methionine oxidation and protein N-terminal acetylation were included as variable modifications in all cases. XIC output generation was enabled, and all remaining DIA-NN parameters were maintained at default settings. DIA-NN PARQUET output files, together with the corresponding XIC and report.pg_matrix files, were imported into the DEF workflow. Quantification settings were retained at default parameters, and the HaoGroup contaminant library [86] was applied. Protein quantification was based on DIA-NN PG.MaxLFQ intensities. To improve identification and quantification reliability, proteins were required to be supported by at least two peptides, thereby excluding single-peptide identifications. Peptides were retained only when supported by a minimum of four consecutive b- or y-fragment ions. Furthermore, proteins were required to be identified in at least two out of three biological replicates.

Statistical analyses were performed using GraphPad Prism version 8.0.1 (GraphPad Software, San Diego, CA, USA). Group comparisons were conducted using unpaired t-tests with Welch’s correction, and p-values < 0.05 were considered statistically significant. Principal component analysis (PCA) and protein identification overlap calculations were performed in Python version 3.12.7 using the sklearn.decomposition and pandas libraries, respectively. Venn diagrams were generated using the IntersectMe web application. Group-level comparisons were defined within the DEF “Group-Visuals” module, enabling direct comparison between control, fibrosis, and hepatic steatosis groups. A maximum of one missing value per replicate was permitted. Exported datasets were subsequently analyzed using MetaboAnalyst version 5 [87].

Because low-input FFPE biopsy samples inherently contain missing quantitative values, missing data were imputed using a left-censored approach in which values were replaced with one-fifth of the minimum positive intensity value. Prior to statistical testing, data were filtered

using a low-variance filter, median-normalized, and log₁₀-transformed. Differential abundance analysis was performed using unpaired two-tailed t-tests with Benjamini-Krieger false discovery rate (FDR) correction (1%). Proteins with q-values < 0.01 and absolute log₂ fold changes > 1 were considered differentially abundant proteins (DAPs).

4.6.2. Data analysis of the adipose tissue differentiation sample set

Raw files were processed as a single batch using DIA-NN version 2.2.0 in library-free search mode. Resulting PARQUET outputs, together with the corresponding XIC and report.pg_matrix files, were imported into the DEF workflow. Quantification settings were maintained at default values, and the HaoGroup contaminant library [86] was applied because it contains an expanded set of experimentally observed contaminants while excluding unassigned UniProt entries and human protein standards unique to MaxQuant and cRAP contaminant databases.

Group-level comparisons were defined within the DEF “Group_Visuals” component to enable direct comparisons between WA vs PA, WP vs PA, and WP vs WA conditions. A maximum of one missing value per replicate was permitted, and resulting datasets were exported using the “MetaboAnalyst Ready” module. Statistical analyses were subsequently performed using MetaboAnalyst version 6.0 [87]. Data were filtered using the low-variance filter, and missing values were imputed using a left-censored approach by replacing missing entries with one-fifth of the minimum positive intensity value [90]. Data were then median-normalized and log₁₀-transformed. Differential abundance analysis was performed using two-sample t-tests. Proteins with adjusted p-values < 0.05 and fold changes greater than 1.5 or lower than -1.5 were considered significantly altered. PCA and volcano plots were generated within MetaboAnalyst, with selected non-significant proteins omitted from volcano plots to improve visualization clarity.

4.6.3. Data analysis of the benchmark datasets

4.6.3.1. Data analysis of the HEK Benchmark dataset

Raw files were analyzed using DIA-NN version 2.2 against either the authors' provided human FASTA database, a DDA-derived spectral library, or a gas-phase fractionation (GPF)-generated spectral library produced from GPF-DIA runs searched against the FASTA database.

Batch .csv files were used to determine the number of identified proteins per replicate. Quantitative reproducibility between technical replicates was assessed using median coefficients of variation (CVs) calculated from PG.MaxLFQ intensities.

4.6.3.2. Data analysis of the Mouse–Yeast Spike-In Benchmark dataset

Original DIA datasets were analyzed using either a universal spectral library generated from DDA data processed with FragPipe or a library-free workflow based on *in silico* spectral library generation. Downloaded DIA-NN (version 1.8.1) TSV outputs were processed using the DEF workflow. Proteins were annotated as mouse- or yeast-specific using UniProt annotations, after which mouse-specific proteins were extracted for downstream evaluation. Proteins containing more than two missing values among five technical replicates were excluded from quantitative analyses.

4.6.3.3. Data analysis of the HYE124 Tri-Species Benchmark dataset

AB Sciex dynamic-link library (.dll) files generated using MSConvert were copied into the DIA-NN installation directory to enable direct import of WIFF files. WIFF datasets corresponding to TripleTOF 5600 and TripleTOF 6600 instruments were processed as independent batches and searched against a combined three-species spectral library. Proteins quantified in at least two out of three replicates per sample were retained for downstream analysis. Protein identification performance and quantitative reproducibility, expressed as CV values, were compared across instrument platforms and SWATH acquisition modes.

4.7. Functional analysis

DAPs from the liver disease sample set were further analyzed using Ingenuity Pathway Analysis (IPA) (QIAGEN Inc., Redwood City, CA, USA) to identify enriched canonical pathways, biological processes, interaction networks, upstream regulators, and predicted downstream effects. Pathway enrichment significance was evaluated using right-tailed Fisher's exact tests, whereas predicted activation or inhibition states were inferred from IPA z-scores under default parameters.

In case of adipose differentiation sample set, the protein-protein interaction (PPI) networks and enrichment analyses for biological processes, molecular functions, and KEGG pathways were generated using STRING database v12.0 [91]. Protein interactions were predicted using all available evidence channels, applying a high-confidence interaction score threshold of 0.7.

5. RESULTS

This chapter describes the results generated to address the two central objectives of this thesis. First, we present the development of an accessible workflow for the analysis of DIA proteomics data, evaluated across a range of benchmark and biological datasets. Second, the FFPE protein extraction and digestion workflows were systematically optimized with the aim of improving simplicity, reproducibility, and suitability for low-input clinical specimens.

5.1. DIA-NN Easy Filter (DEF) Workflow Development

DEF workflow paths are separated into functional container termed components: there are mandatory input components, mandatory yet once-off functions such as “saving location” definer, components for optional and visualization functions. A “Task selector” component is used to split user defined parameters into eight branches: (1) and (2) save the filtered output as .csv per sample; (3) and (4) visualize filtered PGs; (5) and (6) visualize filtered protein quantities; and (7) and (8) save the filtered output as .csv of batch samples (**Figure 3**).

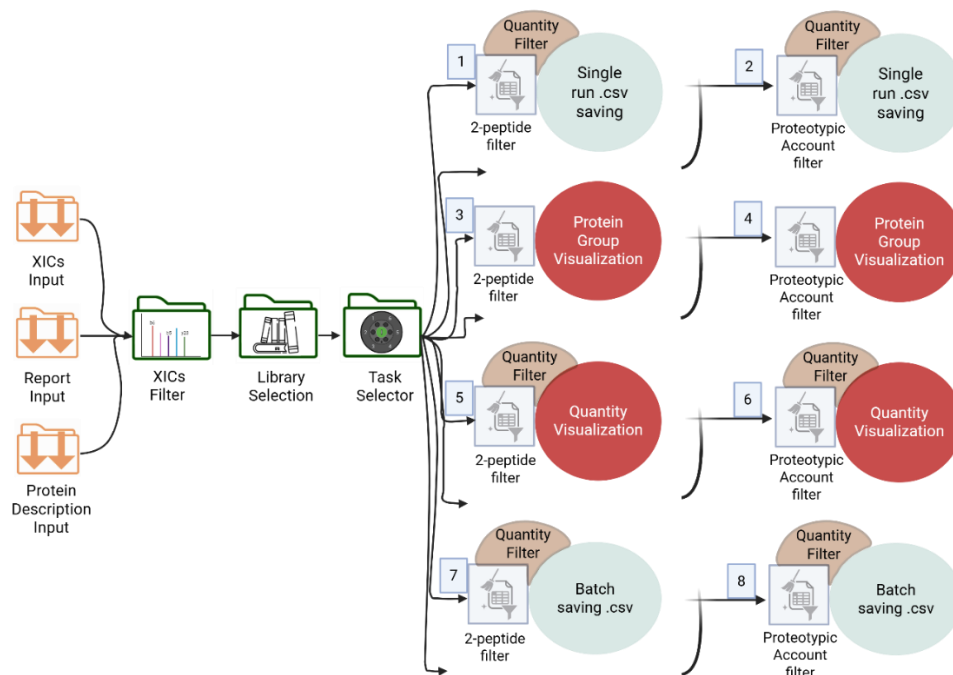


Figure 3. Workflow illustration showing input (orange), configuration components (green), optional filters (brown), and visual (red) components.

5.1.1. Evaluation of DEF workflow using external, published datasets

To evaluate the performance of the DEF workflow on different spectral library strategies, a benchmark DIA proteomics dataset from [67] was re-analyzed using DEF. Processing of the dataset resulted in the identification of more than 4200 protein groups (PGs), consistent with the values reported in the original publication (**Table 1**).

Table 1. Comparison of the number of protein identifications in single injections from the DDA-, FASTA-, and GPF-based strategies.

	GPF		FASTA		DDA	
	Jiang		Jiang		Jiang	
	et al.	DEF	et al.	DEF	et al.	DEF
	[67]		[67]		[67]	
HEK293_100ng_replicate_1	4752	4796	4691	4647	4270	4433
HEK293_100ng_replicate_2	4809	4863	4666	4689	4253	4506
HEK293_100ng_replicate_3	4959	4936	4766	4722	4467	4612

For quantitative assessment, only PGs identified across all three technical replicates without missing values were retained. Quantification reproducibility was subsequently evaluated using the distribution and median coefficients of variation (CVs) derived from PG.MaxLFQ values. Although the absolute number of reported proteins differed slightly because of differences in filtering strategies, the overall trends observed using DEF were comparable to those described in the original study (**Figure 4**).

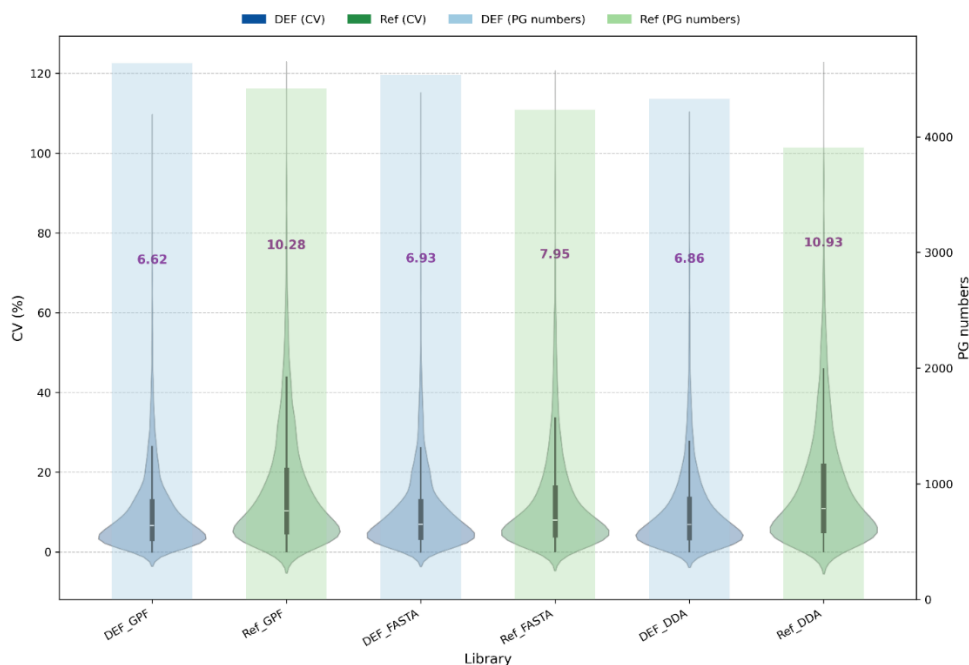


Figure 4. Quantification accuracies from a GPF-, FASTA-, and DDA-based library search. ‘DEF’ highlights workflow filtered results with an additional XIC filtration requiring consecutive b and y ions, while ‘Ref’ represents results from the original publication. Bars (right y-axis) indicate the number of quantified PGs per library, while the overlaid violins (left y-axis) illustrate the distribution of their coefficients of variation (CV, %). Median CV values are annotated above each violin.

Among the evaluated library generation strategies, DDA-based libraries resulted in slightly fewer protein identifications relative to FASTA- and GPF-based libraries. Nevertheless, all approaches demonstrated highly similar quantitative reproducibility, with median CV values of 6.62%, 6.93%, and 6.86% for GPF-, FASTA-, and DDA-based analyses, respectively.

In order to assess the compatibility of DEF with DIA-NN version 1.8.1 TSV outputs and to evaluate workflow performance across different instrument platforms and library generation strategies a mouse-yeast spike in benchmark dataset [88] was used. The dataset consisted of five technical replicates analyzed on Thermo Q Exactive HF and Bruker timsTOF Pro instruments using both universal spectral libraries and *in silico*-generated libraries. Protein group identifications generated by DEF using the proteotypic peptide inference strategy were comparable to those reported using the razor peptide-based inference approach in the original

publication. Similar numbers of mouse-specific protein groups were identified across both mass spectrometry platforms and library generation methods (**Table 2**).

Table 2. The number of identified mouse PGs from both Thermo QExactive HF (HF) and timsTOF Pro (TIMS) instruments when searched against a universal library and DDA-independent library (Library free) from both the original publication and DEF.

	Universal library		Library free	
	Lou et al. [88]	DEF	Lou et al. [88]	DEF
HF	4923	4903	5186	5166
TIMS	7128	7176	6616	6664

Mouse-specific PGs containing fewer than two missing values among five replicates were retained for downstream quantification analyses. Quantitative reproducibility metrics generated by DEF closely matched those reported in the original publication (**Figure 5**).

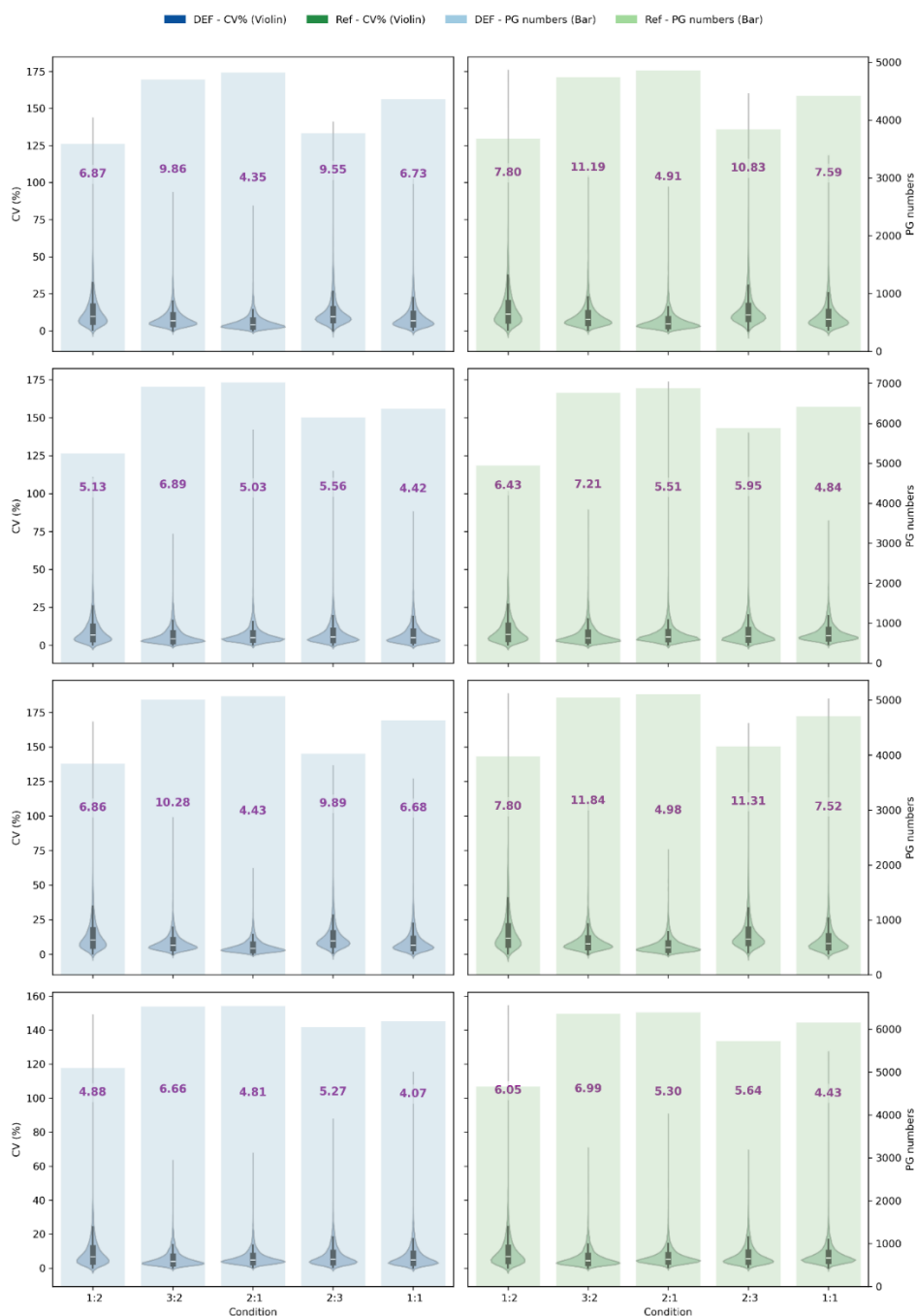


Figure 5. Oritrap HF and timsTOF Pro instruments comparative evaluation of quantification performance showing DEF (blue) and those reported in the reference publication (green) for various library generation methods (universal and library free). Bar plots (right axis) shows the number of quantified PGs, while overlaid violin and box plots (left axis) show the distribution of their coefficients of variation (CV, %). The median CV value for each group is indicated above the violins (purple).

Median CV values ranged from 4.35% to 10.28%, compared with previously reported values ranging from 4.43% to 11.84%.

For a cross-platform and acquisition mode DEF performance evaluation, a HYE124 tri-species benchmark dataset [89] was used. The dataset included acquisitions generated using TripleTOF 5600 and TripleTOF 6600 instruments with both 32 fixed-window and 64 variable-window SWATH acquisition methods. DEF analysis produced slightly higher numbers of protein identifications compared with those reported in the original benchmark publication (**Table 3**).

Table 3. Numbers of reported proteins, quantified in at least 2 replicates out of 3, identified by different software tools and DEF in HYE124 cell lines analysed by two types of Sciex QTOF instruments and number and types of SWATH windows.

		Number of proteins in HYE124 [89]		
		32 fixed	64 variable	Difference
OpenSWATH [89]	QTOF 5600	3,339	4,086	+22%
	QTOF 6600	4,035	4,636	+15%
	Difference	+21%	+13%	
SWATH 2.0 [89]	QTOF 5600	2,125	3,119	+47%
	QTOF 6600	2,656	3,946	+49%
	Difference	+25%	+27%	
Skyline [89]	QTOF 5600	2,878	3,629	+26%
	QTOF 6600	4,046	4,692	+16%
	Difference	41%	29%	
Spectronaut [89]	QTOF 5600	3,332	3,669	+10%
	QTOF 6600	4,226	4,675	+11%
	Difference	+27%	+27%	
DIA-Umpire [89]	QTOF 5600	1,566	1,781	+14%
	QTOF 6600	2,729	3,673	+35%
	Difference	+74%	+106%	
DEF	QTOF 5600	4873	5386	+11%
	QTOF 6600	5658	6707	+19%
	Difference	+16%	+25%	

Analysis of DEF outputs demonstrated that the 64-variable-window acquisition strategy consistently generated more protein identifications than the 32-fixed-window approach. In addition, instrument platform exerted a pronounced effect on proteome coverage, with substantially higher protein identifications observed for TripleTOF 6600 datasets relative to

TripleTOF 5600 acquisitions. These observations were consistent with the trends reported in the original publication, where transition from 32-fixed to 64-variable SWATH windows increased protein identifications by approximately 9-37%, while use of the TripleTOF 6600 system resulted in 14-102% higher protein identifications relative to TripleTOF 5600 datasets.

High technical reproducibility was observed across all replicate analyses, with CV distributions and quantitative trends closely matching those reported in the original study (**Table 4**).

Table 4. Quantification reproducibility expressed as a median coefficient of variation (CV%).

	TripleTOF 5600 32fix	TripleTOF 5600 64var	TripleTOF 6600 32fix	TripleTOF 6600 64var
OpenSWATH [89]	6%	5%	8%	6%
SWATH2.0 [89]	5%	4%	8%	6%
Skyline [89]	6%	6%	9%	5%
Spectronaut [89]	4%	4%	3%	3%
DIA-Umpire [89]	10%	10%	12%	12%
DEF	7%	9%	6%	7%

5.1.2. Evaluation of DEF workflow using dataset produced in our laboratory

In this data examination we wanted to apply the DEF workflow to analyze proteomic changes associated with adipocyte differentiation and palmitate treatment. In this case the data were generated in our laboratory and the comparisons were performed between preadipocytes (PA), differentiated white adipocytes (WA), and palmitate-treated white adipocytes (WP).

The total numbers of identified protein groups are summarized in **Table 5**. Application of the two-peptide inference criterion reduced the number of identified PGs relative to proteotypic peptide-based inference; however, more than 300 PGs per replicate remained

confidently identified under the two-peptide rule, while proteotypic peptide filtering retained more than 500 PGs.

Table 5. Number of identified PGs in Case Study 5. Number of PGs, with a 2-peptide and proteotypic inference rule, are shown for pre-adipocytes (PA), differentiated white adipocytes (WA), and white adipocytes differentiated in the presence of palmitate (WP) with respective replicates (rep).

	Number of PGs	
	2-Peptide rule	Proteotypic rule
PA_rep1	566	754
PA_rep 2	399	535
PA_rep 3	416	554
WA_rep 1	588	787
WA_rep 2	706	907
WA_rep 3	628	830
WP_rep 1	627	813
WP_rep 2	696	904
WP_rep 3	684	871

The influence of extracted ion chromatogram (XIC)-based filtering on protein group identifications is presented in **Figure 6**. DEF-generated peptide tables additionally included detailed modification annotations for each peptide sequence, including modification type and localization.

List of XICs Excluded Peptides Rows: 405 Columns: 3			Table: Number of filtered hits Rows: 9 Columns: 2		
☐ Run String	☐ Ion_series String	☐ Excluded_By_XICs String	☐ RowID String	☐ Runs String	☐ Number of PGs Number (Integer)
☐ PA_rep1	y14, y19, y3, y5, y7, b4, y6	MMCGAPSATQPATAETQHIAI	☐ Row0	PA_rep1	566
☐ WA_rep1	y14, y19, y3, y5, y7, b4, y6	MMCGAPSATQPATAETQHIAI	☐ Row1	PA_rep2	399
☐ WA_rep2	y14, y19, y3, y5, y7, b4, y6	MMCGAPSATQPATAETQHIAI			

Excluded Peptide Summary Rows: 9 Columns: 5			Peptide Summary Rows: 1-10 of 5310 Columns: 27		
☐ Run String	☐ Excluded... String	☐ Modifi... String	☐ Seq... Num...	☐ Modifica Number (n	☐ WP2_ID String
☐ PA_rep2	AASDIAMTELPPTHf	UniMod:4(3)	28	1	☐ WP2_-Mod String
☐ PA_rep3	AASDIAMTELPPTHf	UniMod:4(4)	31	1	☐ WP2_-Seq String
☐ WA_rep3	(UniMod:1)MMC(Un	UniMod:1(1),	53	2	☐ WP3... String

☐ P07437	M(UniMod:35)	AILVDLEPGTMD:	☐ P22234	C(UniMod:4),C(Uni	ACGNFGIPCELR
☐ P35558	C(UniMod:4),C(Uni	CLHSVGCPLPLQ			

Figure 6. A qualitative evaluation of tables as can be seen in the “Visualize PGs” composite view. The number of filtered hits displays a total of all identified PGs. Peptide tables include a peptide summary, and XICs included/excluded peptides.

The workflow also generated several interactive visualization modules to support qualitative and quantitative data assessment. The “Quantity Visuals” component enabled visualization of protein abundance distributions using interactive pie charts that included both retained proteins and filtered contaminants prior to export.

To assess quantification consistency across samples, DEF generated binary parallel coordinate plots in which detected PGs were assigned a value of 1 and missing values a value of 0 (**Figures 7A, 7B**). Additional parallel coordinate plots based on min-max-scaled quantitative intensities enabled visualization of relative abundance trends independent of absolute signal intensity (**Figures 7C, 7D**). Quantification completeness across runs was further summarized using bar plots displaying the percentage of quantified PGs per sample (**Figures 7E, 7F**).

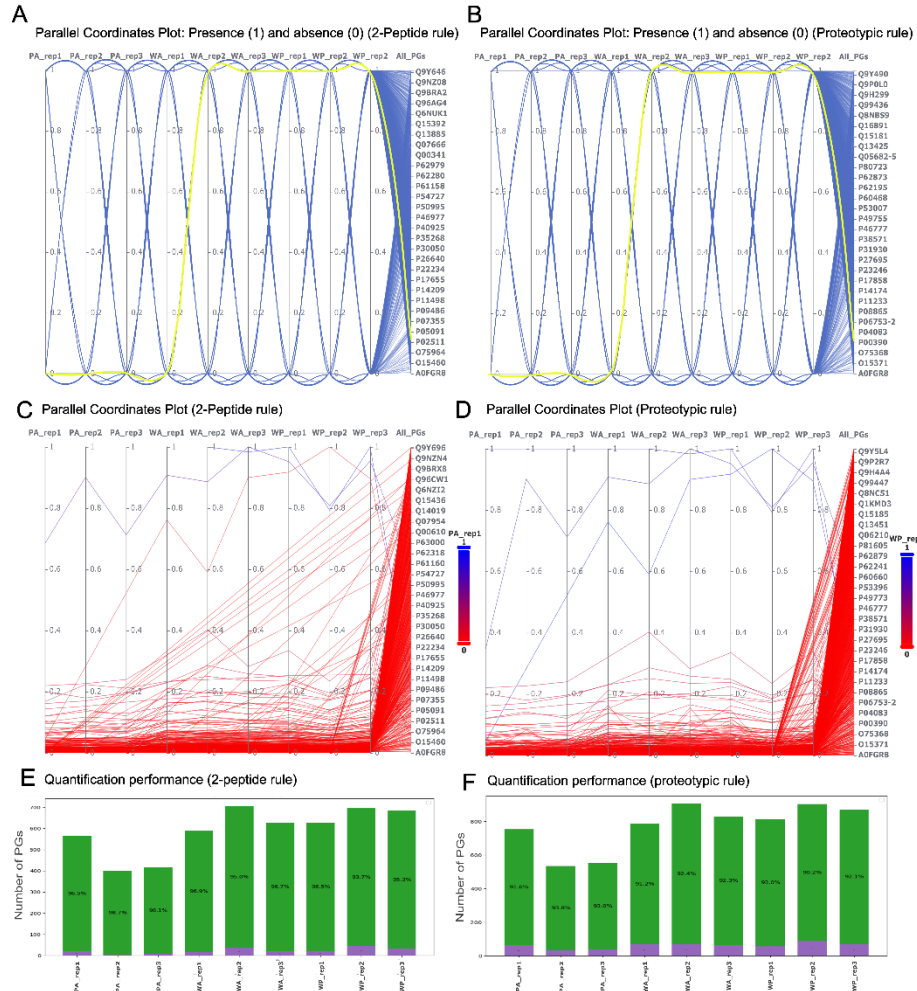


Figure 7. DEF workflow parallel coordinate plots showing a run-level comparison of protein identifications using 2-peptide rule (A) and proteotypic rule (B) with the arbitrarily selected O60240 protein (yellow line). DEF workflow parallel coordinate plot showing a run-level comparison of scaled protein quantities for a 2-peptide rule (C) and proteotypic rule (D) where the intensity dimension was color-coded according to user defined runs (in this case WP_rep2 for panel C and PA_rep1 for panel D). Panels E and F show the ratio of sample specific quantifiable PGs obtained using the two-peptide rule (E) and the proteotypic rule (F), with quantified PGs shown in green and non-quantified PGs in purple.

The workflow also enabled visualization and export of summarized protein abundance tables across all runs. Dynamic interaction between summary tables and visualization panels

allowed direct highlighting of selected protein groups within the corresponding parallel coordinate plots.

For downstream statistical analyses, DEF outputs were processed in MetaboAnalyst [87]. Low-variance protein groups were filtered based on interquartile range (IQR) thresholds, followed by median normalization and log₁₀ transformation of protein intensities. Principal component analysis (PCA) demonstrated separation between experimental conditions (**Figure 8**).

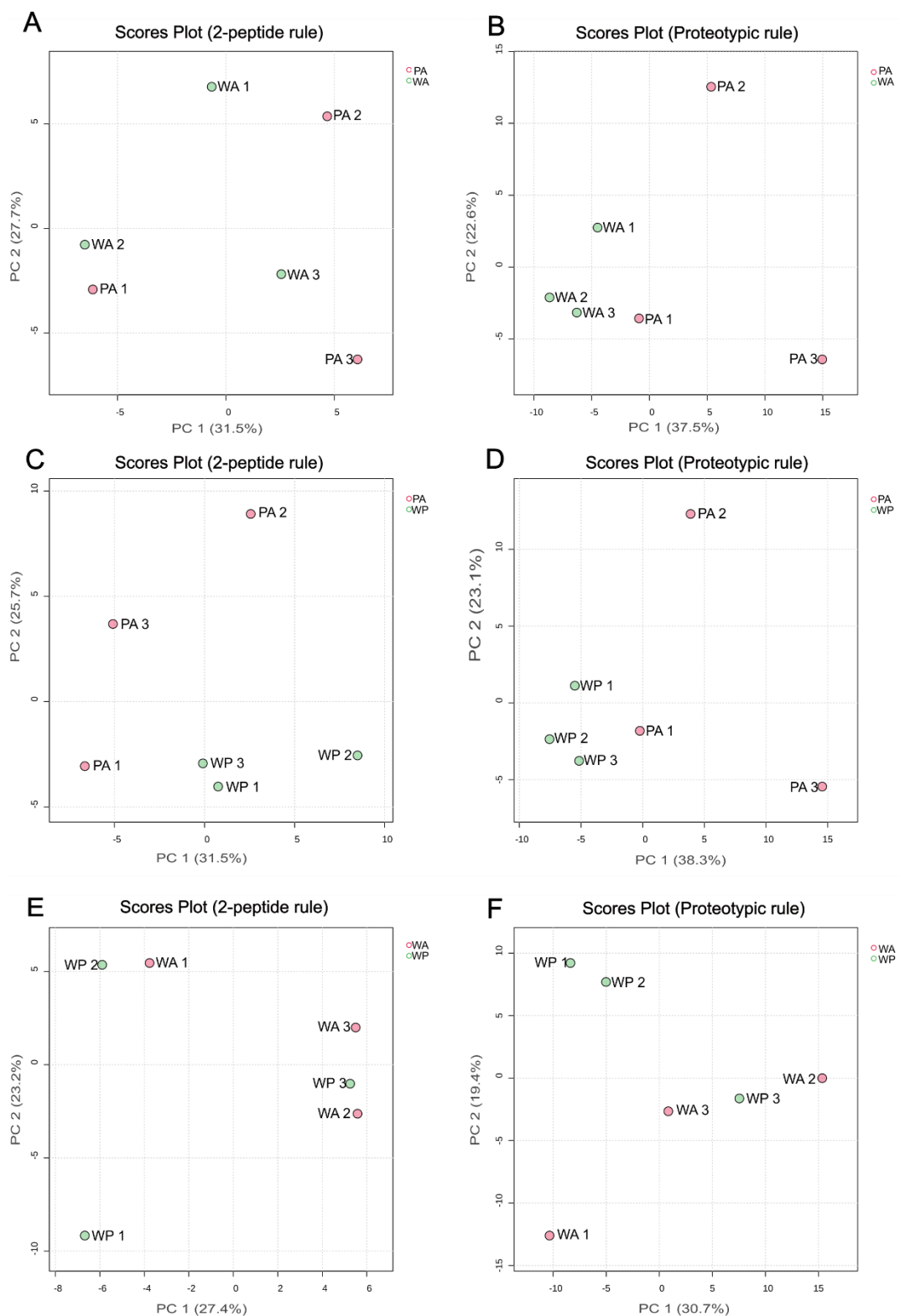


Figure 8. Principal component analysis (PCA) plots where (A) and (B) show WA vs PA; panels (C) and (D) show WP vs PA; and panels (E) and (F) show WP vs WA, for both inference rules.

Differential abundance analysis identified proteins exhibiting statistically significant abundance changes between sample groups using a log₂ fold-change threshold of 1.5 and Benjamini-Hochberg FDR-adjusted q-values < 0.05 (**Table 6**). Volcano plots were generated to visualize the magnitude and statistical significance of these protein abundance changes across experimental comparisons (**Figure 9**). Between the WA and WP groups 3 proteins showed statistically significant change. Cytochrome c oxidase subunit 2 (MT-CO2, P00403) was found to be less abundant, while serotransferrin (TF, P02787) and 60S ribosomal protein L10 (RPL10, P27635) were more abundant in WP.

Table 6. Group-level DEF output summary and statistics as performed in MetaboAnalyst showing the number of identified proteins, differentially abundant proteins (DAP) and significantly increased and decreased proteins.

DEF			Statistics (MetaboAnalyst)		
Inference rule	Group	Number of proteins	DAP	FC↑	FC↓
2-peptide	WA x PA	448	115	43	72
	WP x PA	448	132	49	83
	WP x WA	640	3	2	1
Proteotypic	WA x PA	601	126	57	69
	WP x PA	601	161	66	95
	WP x WA	860	2	1	1

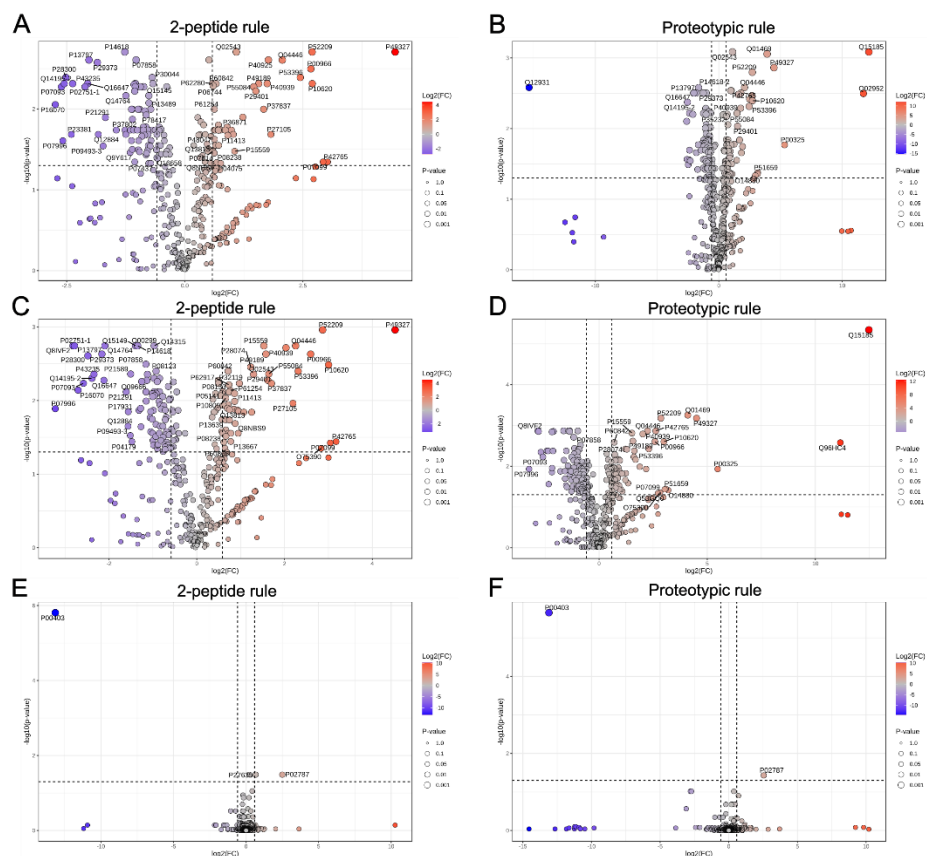


Figure 9. Volcano plots, constructed with MetaboAnalyst, showing differentially abundant proteins identified using a q -value cutoff of 0.05 and a fold-change threshold of 1.5. Panels (A) and (B) depict WA vs PA; panels (C) and (D) depict WP vs PA; and panels (E) and (F) depict WP vs WA.

Functional characterization of differentially abundant proteins was subsequently performed using enrichment analyses for Biological Process, Molecular Function, and KEGG pathways (Figure 10).

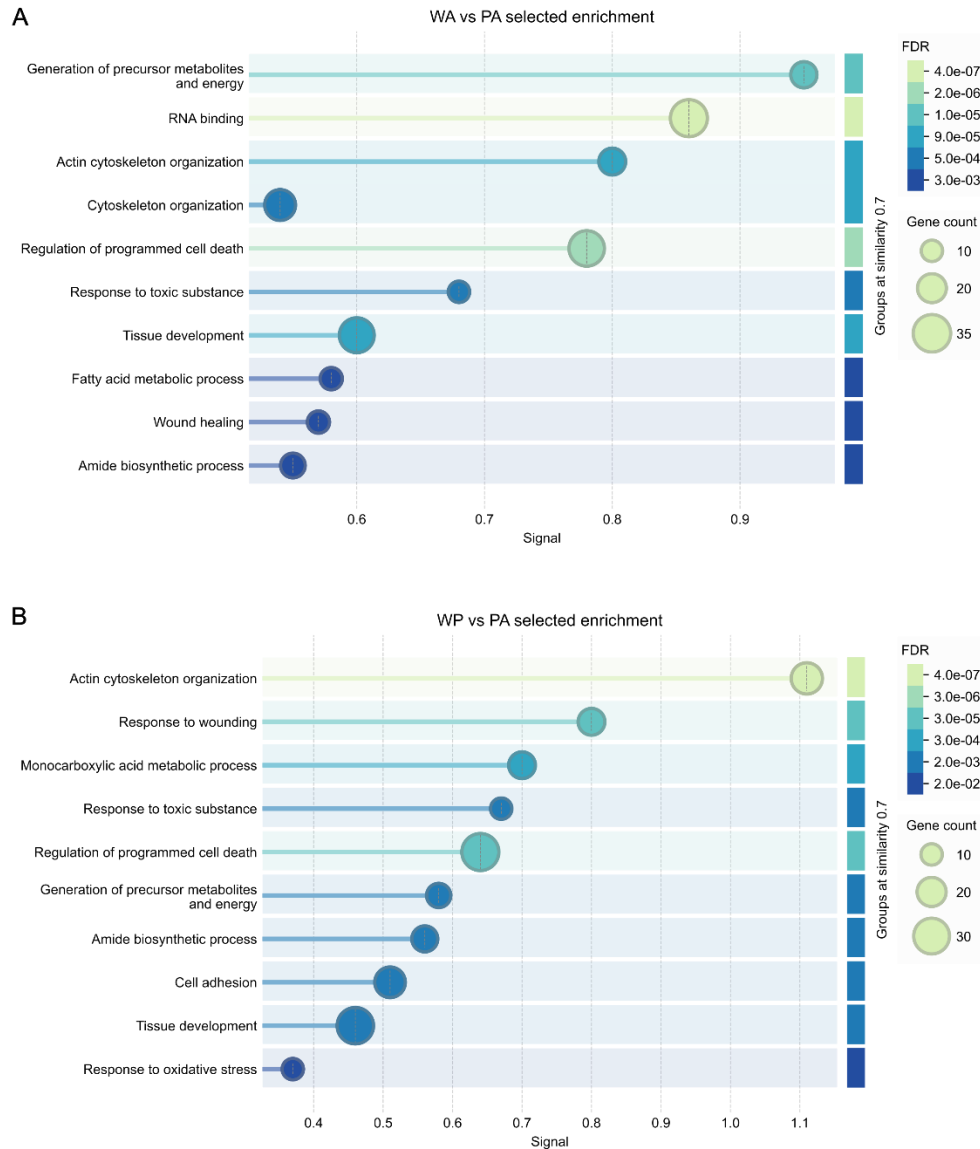


Figure 10. Gene Ontology (GO) biological process and molecular function terms, along with KEGG pathway enrichment results of the differentially abundant proteins for (A) WA vs PA, and (B) WP vs PA. Circle diameter indicates gene counts while the color scale represents false discovery rate (FDR).

Between the WA vs PA and WP vs PA comparisons, processes including cytoskeleton organization, metabolite and energy generation, cell death regulation, and tissue development were commonly enriched in both comparisons.

5.1.3. Examination of FFPE samples

Optimization experiments were conducted in three sequential stages. Initially, four FFPE sample preparation workflows (App_1-App_4) were evaluated using a limited tissue area ($\sim 0.025 \text{ mm}^3$) to assess protein extraction and digestion efficiency. Based on the initial screening results, the two best-performing approaches (App_2 and App_3) were selected for further qualitative and quantitative comparison. Based on the results App_2 was further optimized and in the final step, the modified form of App_2, the App_2M was used in a case study to examine biological differences characteristic to different forms of liver disease.

5.2. Xylene-Free FFPE Sample Preparation for Proteomics Analysis

To establish a simplified, safer, and cost-effective FFPE proteomics workflow, two xylene-free deparaffinization strategies were evaluated, including hot water-mediated deparaffinization and PHAD. Four non-commercial FFPE processing workflows (App_1-App_4), representing different approaches for reversing formalin-induced protein crosslinking, were subsequently assessed using low-input tissue material ($\sim 0.025 \text{ mm}^3$).

The evaluated protocols incorporated distinct extraction and digestion chemistries, including heat-induced antigen retrieval (HIAR) in citrate buffer, triethylammonium bicarbonate (TEAB)-based extraction, and SDS-containing surfactant buffers (**Figure 11**).

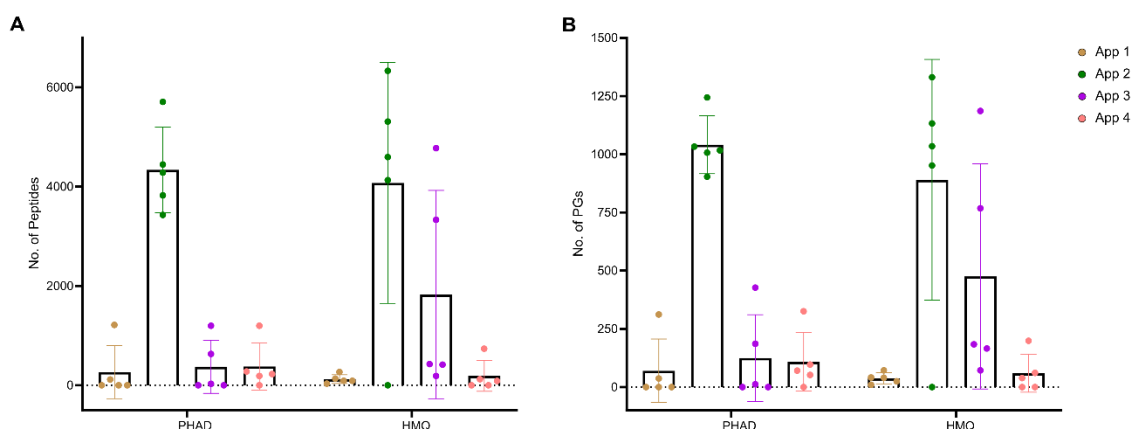


Figure 11. Evaluation of the results obtained by the application of approaches App_1-4. (A) The number of identified peptides, (B) protein groups. The y axis shows the number of peptides and protein groups, respectively, while on the x axis the type of deparaffinization is shown. PG: protein groups, PHAD: projected hot air deparaffinization, HMQ: deparaffinization with hot MilliQ water

Comparative analysis demonstrated that App_2 and App_3 yielded the highest numbers of identified peptides and protein groups (PGs), this is why they were selected for further optimization.

5.2.1. Method Refinement and Comparative Evaluation

To further compare App_2 and App_3, three randomly selected tissue regions from two consecutive needle biopsy sections were processed using each workflow, generating three technical replicates per condition. Proteomics data were filtered to allow peptide and protein lists evaluation. App_2 resulted in substantially higher protein group identifications compared with App_3, with mean values of 939 and 492 PGs, respectively (**Figure 12A**). In both workflows, a large proportion of proteins was consistently detected across all technical replicates, whereas a smaller subset of proteins was identified only in individual samples (**Figure 12B-12D**)

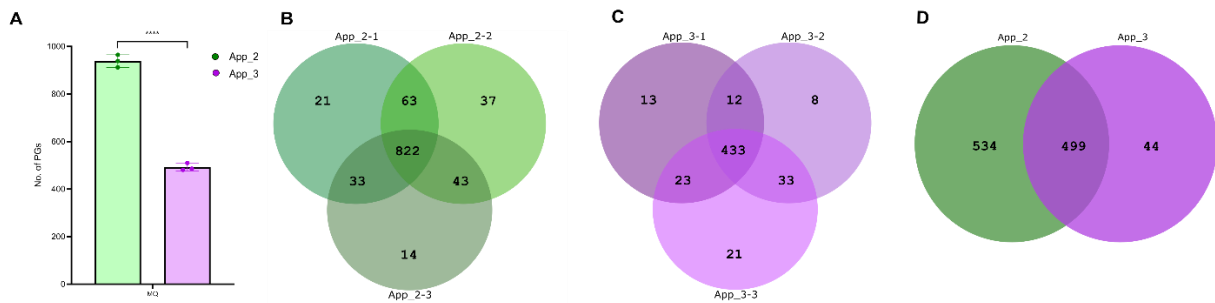


*Figure 12. Comparative evaluation of Approach 2 and Approach 3 for FFPE proteomic analysis in three technical replicates. (A) Comparison of the number of protein groups (PGs) identified using Approach 2 (App_2) and Approach 3 (App_3). Statistical significance was assessed using an unpaired t-test with Welch's correction (**** $p < 0.0001$). (B, C) Venn diagrams of PG overlap across three technical replicates for (B) App_2 and (C) App_3. (D) Venn diagrams of PG overlap across App_2 and App_3.*

Missed cleavage analysis demonstrated that more than 88% of peptides identified in both workflows contained no missed cleavages. However, App_3 exhibited an approximately 6% higher proportion of fully cleaved peptides compared with App_2 (**Figure 13A**). Consistent with this observation, App_3 generated a greater proportion of shorter peptides within the 7-12 amino acid range (**Figure 13B**).

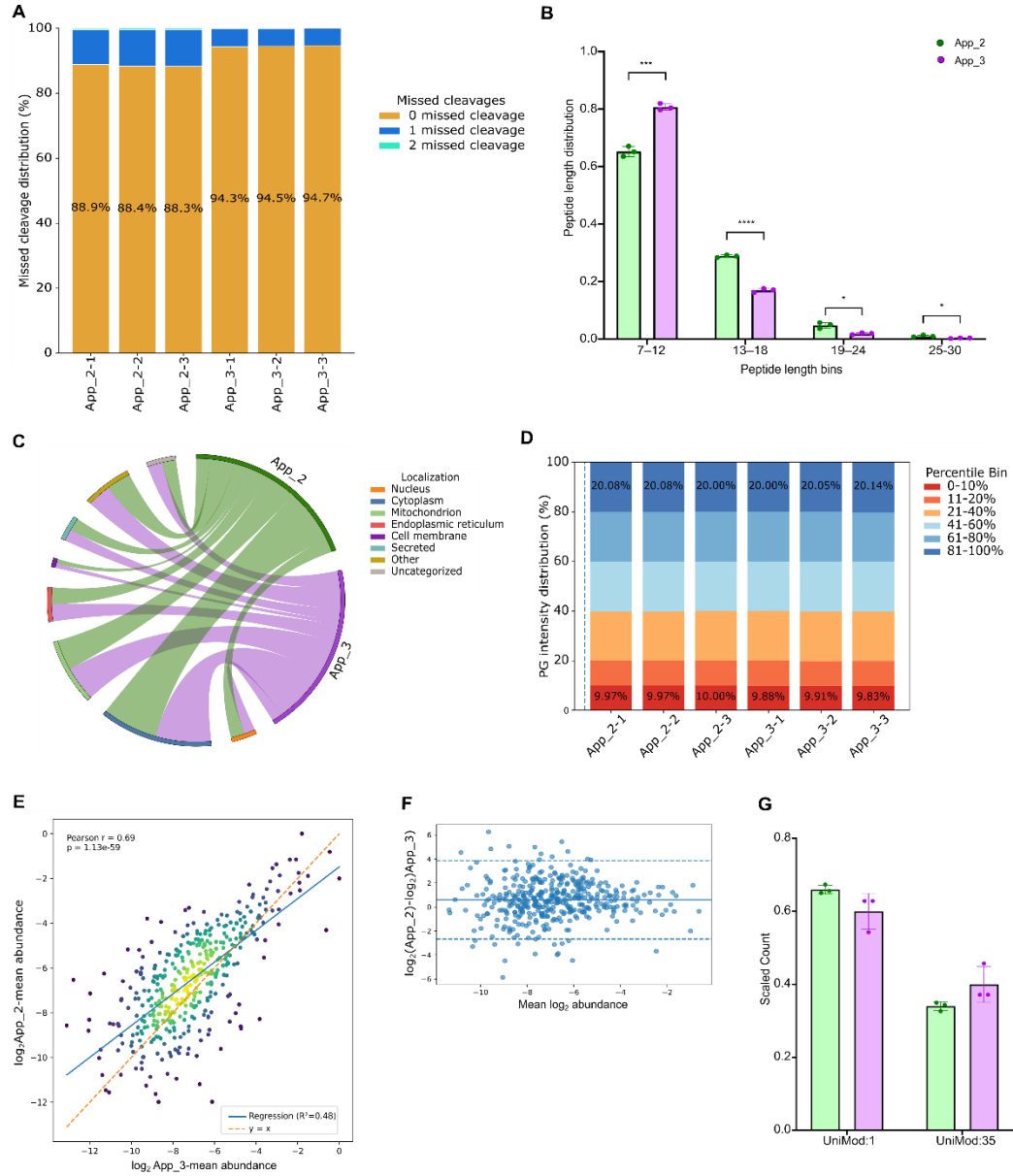


Figure 13. Comparative evaluation of Approach 2 and Approach 3. (A) Trypsin cleavage efficiency; (B) Peptide length distribution presented as the number of identified peptides within defined amino acid length bins (7-12, 13-18, 19-24, and 25-30 residues); (C) Distribution of identified proteins across subcellular compartments for App_2 and App_3; (D) Distribution of PG intensities across percentile bins (0-10, 11-20, 21-40, 41-60, 61-80, and 81-100); (E) Regression comparing mean abundances of the share 499 PG in App_2 and App_3. (F) Abundance differences between the share 499 PG App_2 and App_3; (G) The number of N-

terminus acetylation (UniMod:1) and Methionine oxidation (UniMod:35) modified peptides for App_2 and App_3 given as a ratio of the total modified peptides.

Subcellular localization analysis revealed highly similar distributions of identified proteins between the two workflows, with no apparent enrichment bias toward specific cellular compartments (**Figure 13C**). Likewise, quantitative intensity distributions across protein abundance percentiles were highly comparable between App_2 and App_3 (**Figure 13D**). Despite a moderate correlation between the quantitative profiles generated by App_2 and App_3 (Pearson $r = 0.69$; **Figure 13E**), considerable abundance discrepancies were observed. Among proteins quantified by both workflows, App_2 yielded approximately 1.5-fold higher abundances on average, and more than 50% of shared proteins differed by greater than twofold (**Figure 13F**). Together, these findings suggest that while both approaches capture similar overall proteomic landscapes, they vary substantially in protein extraction and/or recovery efficiency. Furthermore, analysis of peptides with abundant post-translational modifications (acetylation and methionine oxidation) showed no statistically significant differences between the two workflows (**Figure 13G**).

Because App_2 provided substantially higher protein group identifications while requiring fewer processing steps and reduced workflow complexity, this approach was selected for further optimization. Based on the improved cleavage efficiency observed with App_3, reduction and alkylation steps were incorporated into App_2 to generate the modified App_2M workflow.

Application of App_2M resulted in a modest increase in identified protein groups and peptides relative to the original App_2 protocol. Mean protein group counts increased from 1108 to 1202 for PHAD-based deparaffinization and from 1066 to 1149 for HMQ-based deparaffinization (**Figure 14A**). Similar trends were observed for peptide identifications (**Figure 14B**), however, these differences did not reach statistical significance. Principal component analysis (PCA) of protein intensity values demonstrated no clear clustering between App_2 and App_2M workflows or between the two deparaffinization strategies (**Figure 14C**).

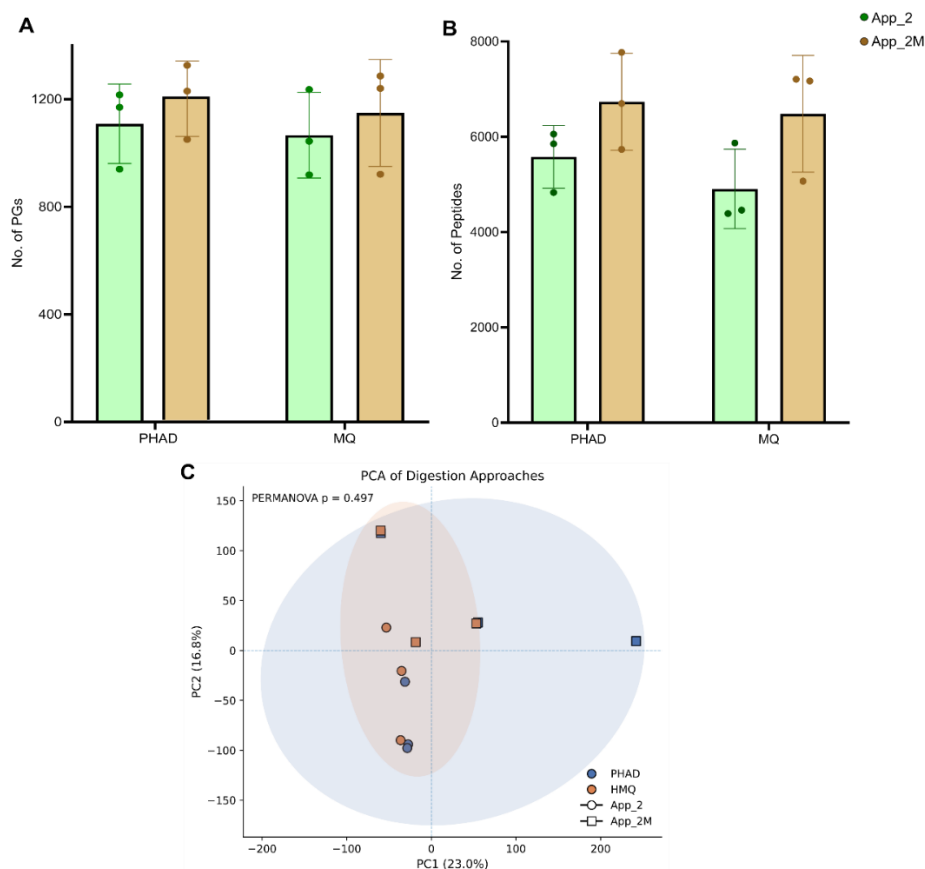


Figure 14. Comparative analysis of App_2 and App_2M using different deparaffinization methods in three technical replicates. (A) Mean number of identified peptides and (B) protein groups (PGs), presented with standard deviation. (C) Principal component analysis of protein intensities derived from three replicates per approach and deparaffinization condition.

Although the dependence of PG identification on sequence counts is comparable across methods (**Figure 15A, 15B**), App_2M showed an enrichment of protein groups supported by multiple unique peptides, particularly for proteins identified by 6-20 unique peptides. This change was significant particularly for HMQ-based deparaffinization, indicating improved peptide coverage per protein group (**Table 7, Figure 15C**).

Table 7. Percentage of proteins identified by unique peptides. The columns show the percentage of proteins identified by the number of unique peptides defined in the rows in case of different methods. Differences were calculated based on the comparison between App_2M and App_2. PHAD: projected hot air deparaffinization, HMQ: deparaffinization with hot MilliQ water.

Number of unique peptides per protein	PHAD- App_2	PHAD- App_2M	Change	HMQ- App_2	HMQ- App_2M	Change
1	40.88%	37.89%	-2.99%	42.48%	37.76%	-4.72%
2	17.57%	17.25%	-0.32%	17.80%	16.92%	-0.88%
3	10.08%	10.15%	0.07%	10.39%	10.88%	0.49%
4	6.59%	7.30%	0.71%	7.41%	6.72%	-0.69%
5	5.05%	5.59%	0.54%	4.47%	4.68%	0.21%
6-20	17.06%	18.73%	1.67%	15.45%	19.62%	4.17%
>20	2.77%	3.09%	0.32%	2.00%	3.41%	1.41%

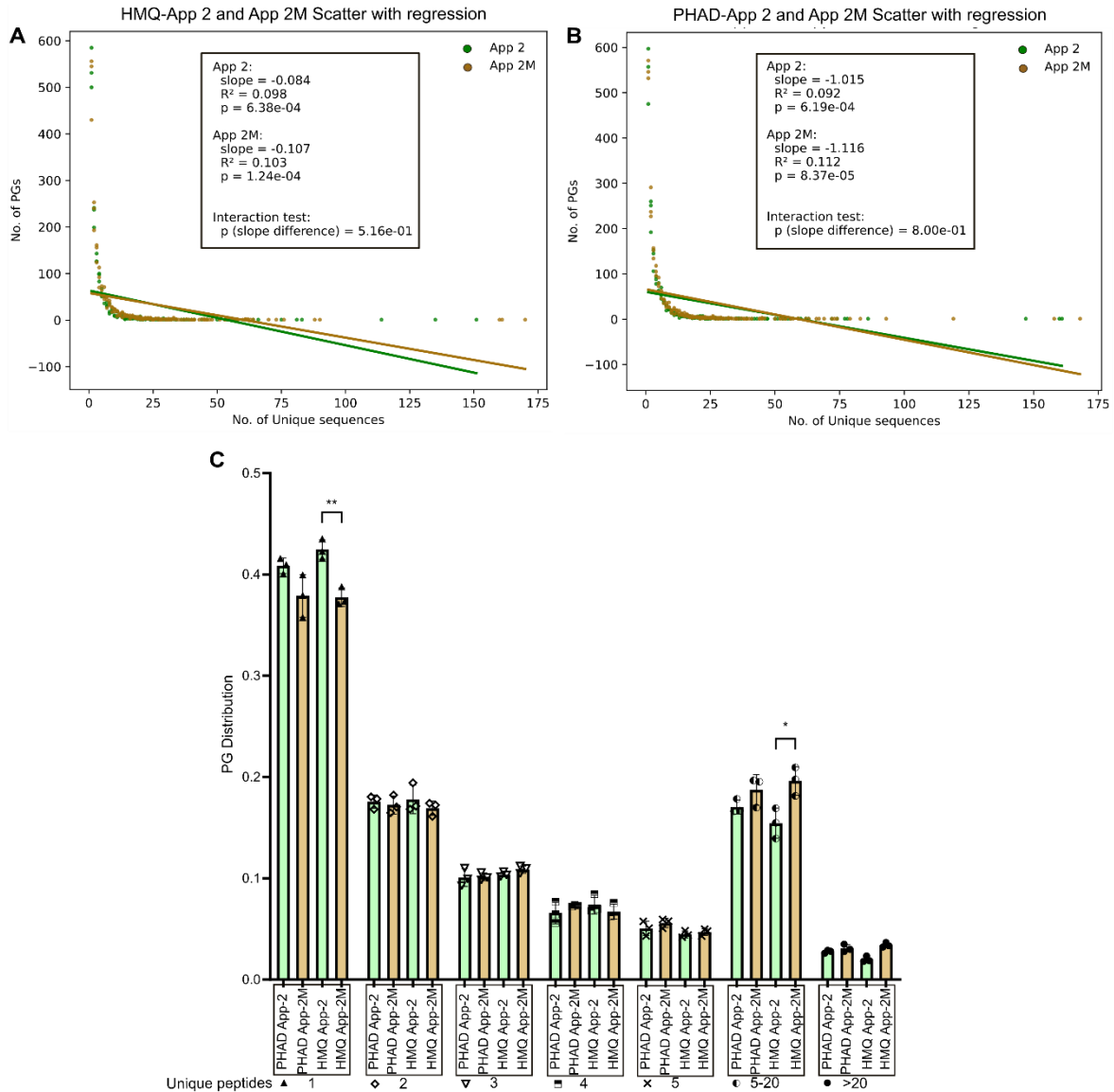


Figure 15: Proportion of proteins identified by unique peptides. Scatter plots showing the relationship between the number of identified proteins and the corresponding number of unique peptides, including fitted regression lines, for HMQ-App₂ and HMQ-App_{2M} (A), and PHAD -App₂ and PHAD-App_{2M} (B). PG distribution against the number of unique peptides bins (1, 2, 3, 4, 5, 6-20).

Our results show that modifying App₂ produced a protocol that delivers strong performance with fewer steps than the other approaches, resulting in App_{2M} that ultimately outperforms all previous protocols.

5.3. Examination of needle biopsy samples from patients with liver disease

The optimized App_2M workflow was subsequently applied to FFPE liver needle biopsy samples obtained from patients with hepatic steatosis and liver fibrosis, together with matched control samples. Approximately 0.03-0.05 mm³ tissue material was analyzed per sample using three biological replicates for each condition. The total number of identified proteins and peptides is summarized in **Table 8**. Variability in protein group identifications between control and disease samples was observed across the dataset.

Table 8. Summary of the number of peptide and proteins identifications in the liver fibrosis (LF), hepatic steatosis (HS), and control (C) groups.

Sample	No. of Peptides (1% FDR)	No. of PGs
LF-1	26416	2590
LF-2	2931	452
LF-3	26266	2583
HS-1	25593	2583
HS-2	26609	2589
HS-3	2470	422
C-1	11360	1161
C-2	7708	1113
C-3	571	83

Prior to downstream statistical analyses, one sample from each experimental group was excluded because of poor data quality. These samples exhibited markedly reduced peptide and protein identifications together with disproportionately high contributions of hemoglobin and keratin signals, exceeding 50% of total abundance (**Figure 16**). Correspondingly, these samples appeared as clear outliers in PCA and hierarchical clustering analyses (**Figure 17**).

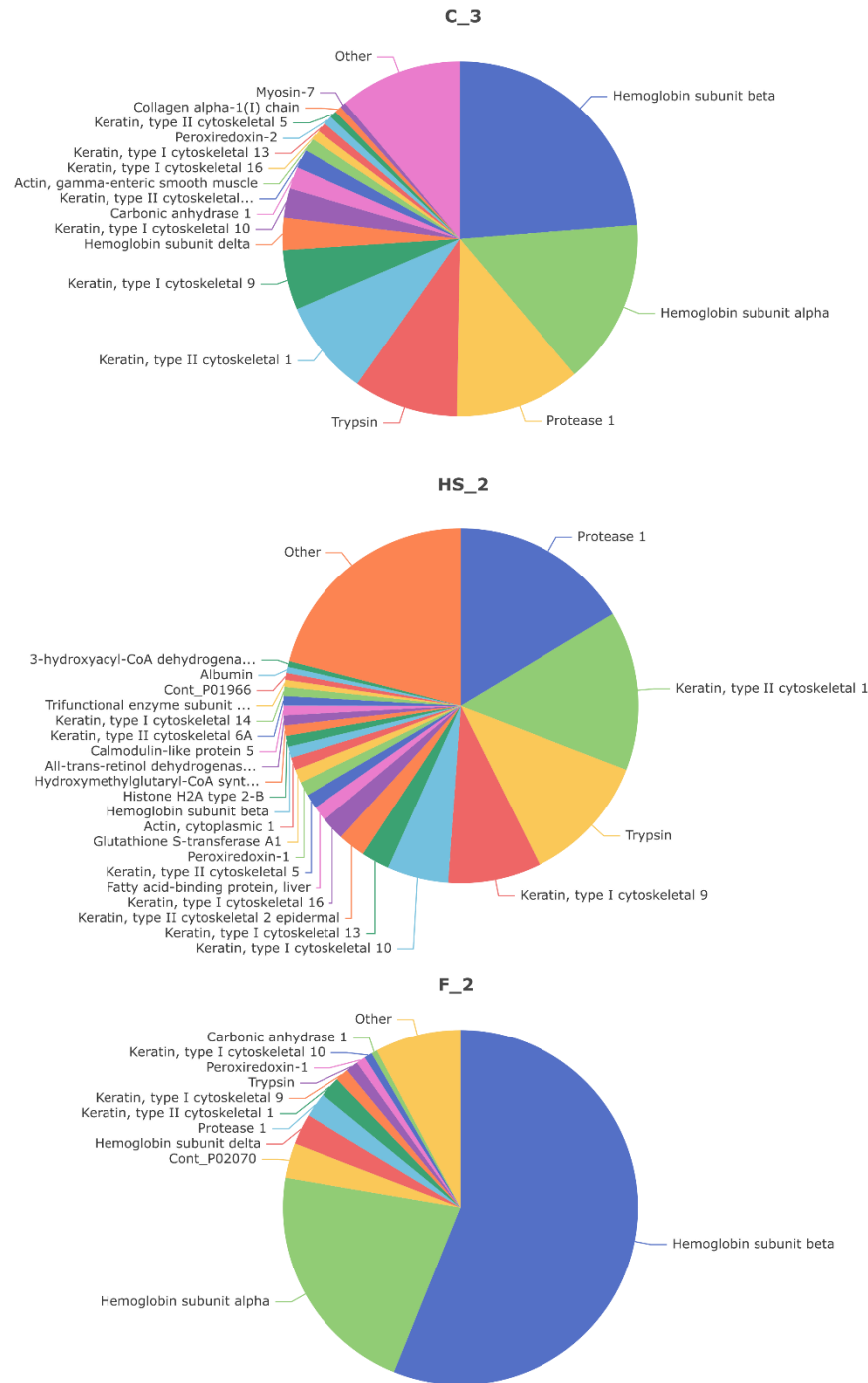


Figure 16: Pie chart showing intensity distribution of protein groups (PGs). Protein group intensities from excluded low-quality samples are displayed as scaled PG.MaxLFQ values and arranged in descending order. Labels C_3, HS_2, and F_2 are replicates of the control, hepatic steatosis, and fibrosis group, respectively.

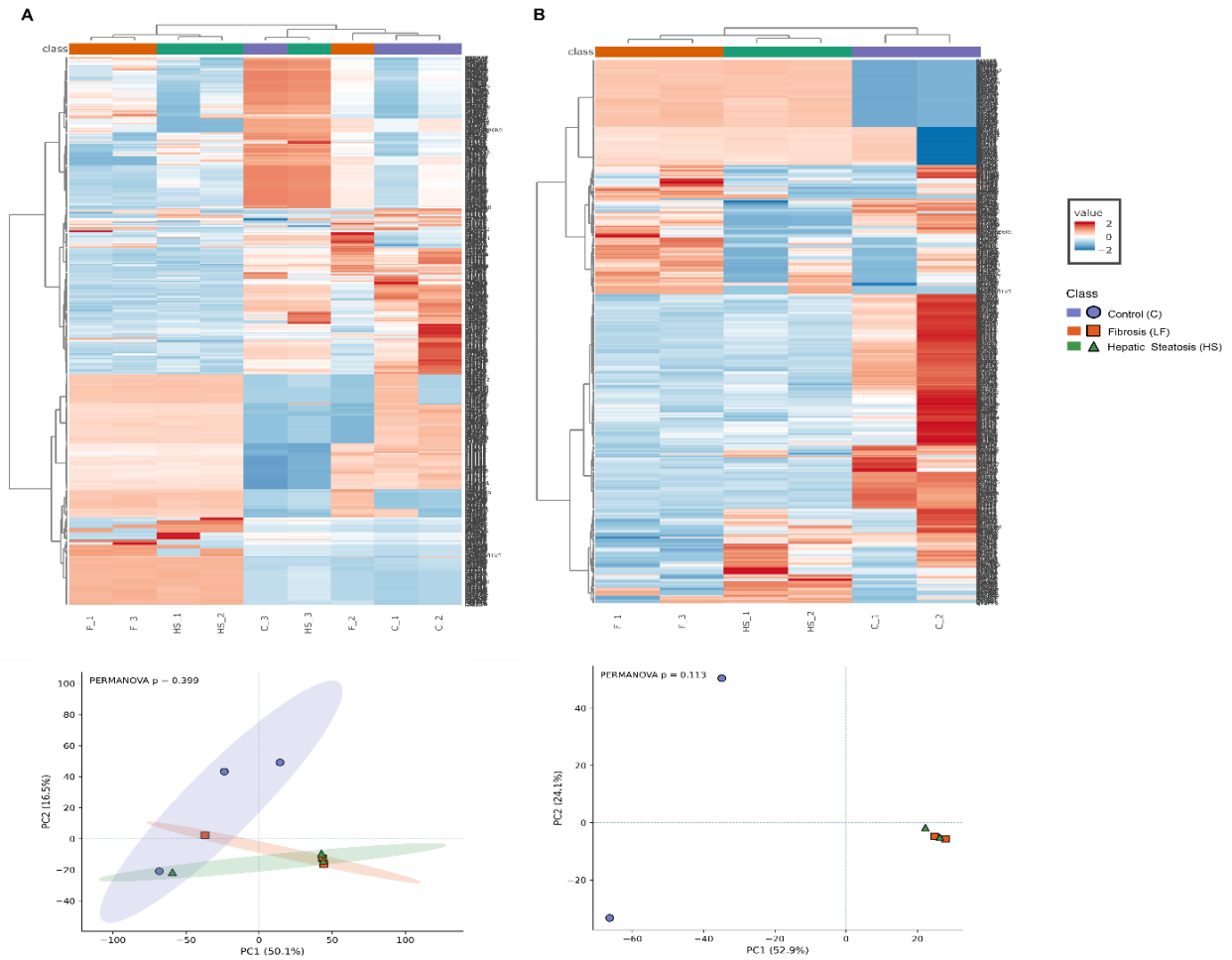
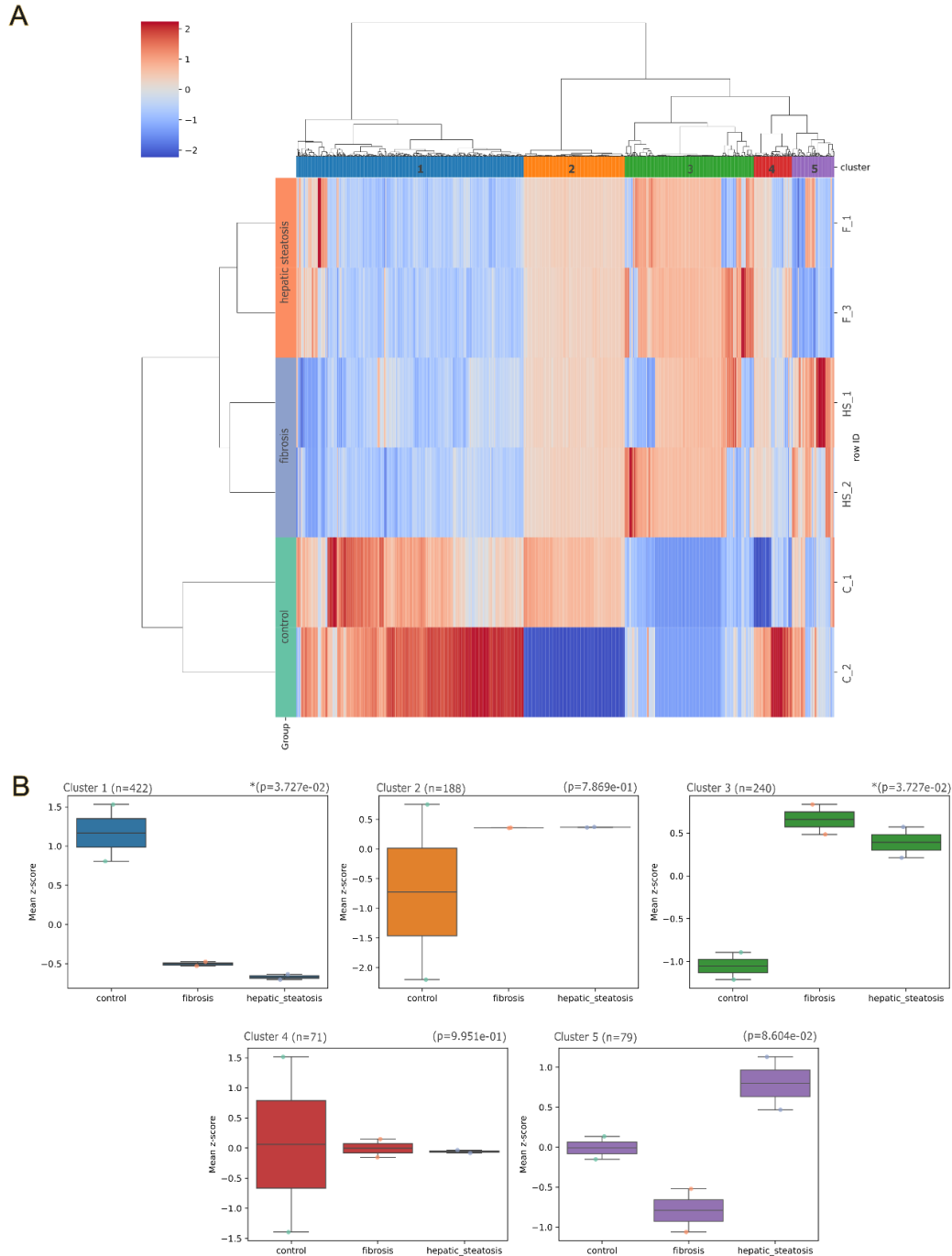


Figure 17: Heatmap and principal component analysis (PCA) of study samples. Clustered heatmap showing sample grouping across control (C_1, C_2, C_3), hepatic steatosis (HS_1, HS_2, HS_3), and fibrosis (F_1, F_3, F_3) groups before (A) and after exclusion (B) of low-quality samples, with the corresponding PCA plots presented below.

Clustered heat map visualization of normalized protein abundances demonstrated condition-dependent segregation between experimental groups (**Figure 18A**). Hierarchical clustering identified five major protein abundance clusters with distinct expression patterns across the analyzed conditions (**Figure 18B**).



Differential abundance analysis identified multiple proteins exhibiting statistically significant changes between groups. Using thresholds of absolute $\log_2$ fold change ≥ 1.5 and Benjamini-Hochberg FDR-adjusted q -values < 0.05 , hepatic steatosis samples displayed 147 proteins with increased abundance and 11 proteins with decreased abundance relative to controls. In fibrosis samples compared with controls, 23 proteins were increased whereas 151 proteins were decreased in abundance. Direct comparison between hepatic steatosis and fibrosis identified 13 proteins with increased abundance and 11 proteins with decreased abundance (Figure 19).

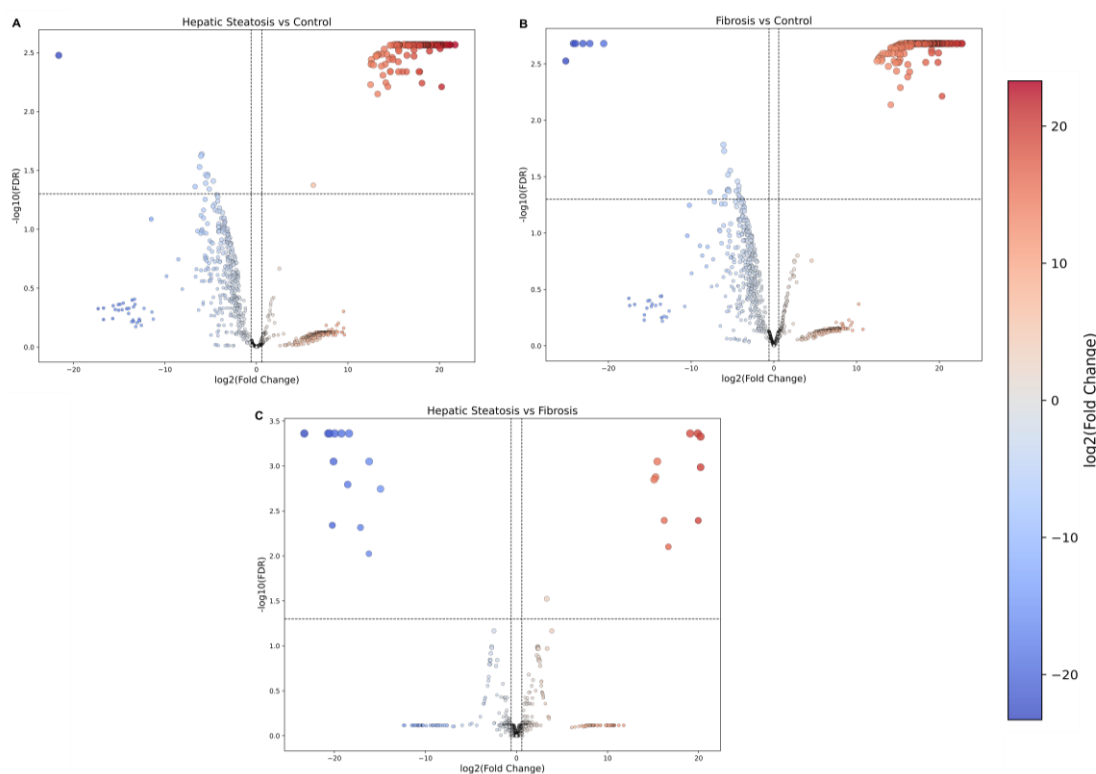


Figure 19. Differentially abundant proteins identified by limma analysis. Volcano plots showing differentially abundant proteins between (A) hepatic steatosis and control ($n=2$), (B) liver fibrosis and control ($n=2$), and (C) hepatic steatosis and liver fibrosis ($n=2$). Proteins with increased abundance are indicated in red, whereas proteins with decreased abundance are shown in blue.

Differentially abundant proteins (DAPs) were subsequently analyzed using Ingenuity Pathway Analysis (IPA). Regulator-effector network analyses were performed separately for hepatic steatosis versus control and fibrosis versus control comparisons (**Figures 20A, 20B**). Predicted upstream regulators, downstream biological functions, and associated causal relationships were inferred using the Ingenuity Knowledge Base. Detail lists of all molecules are provided in **Table 9**.

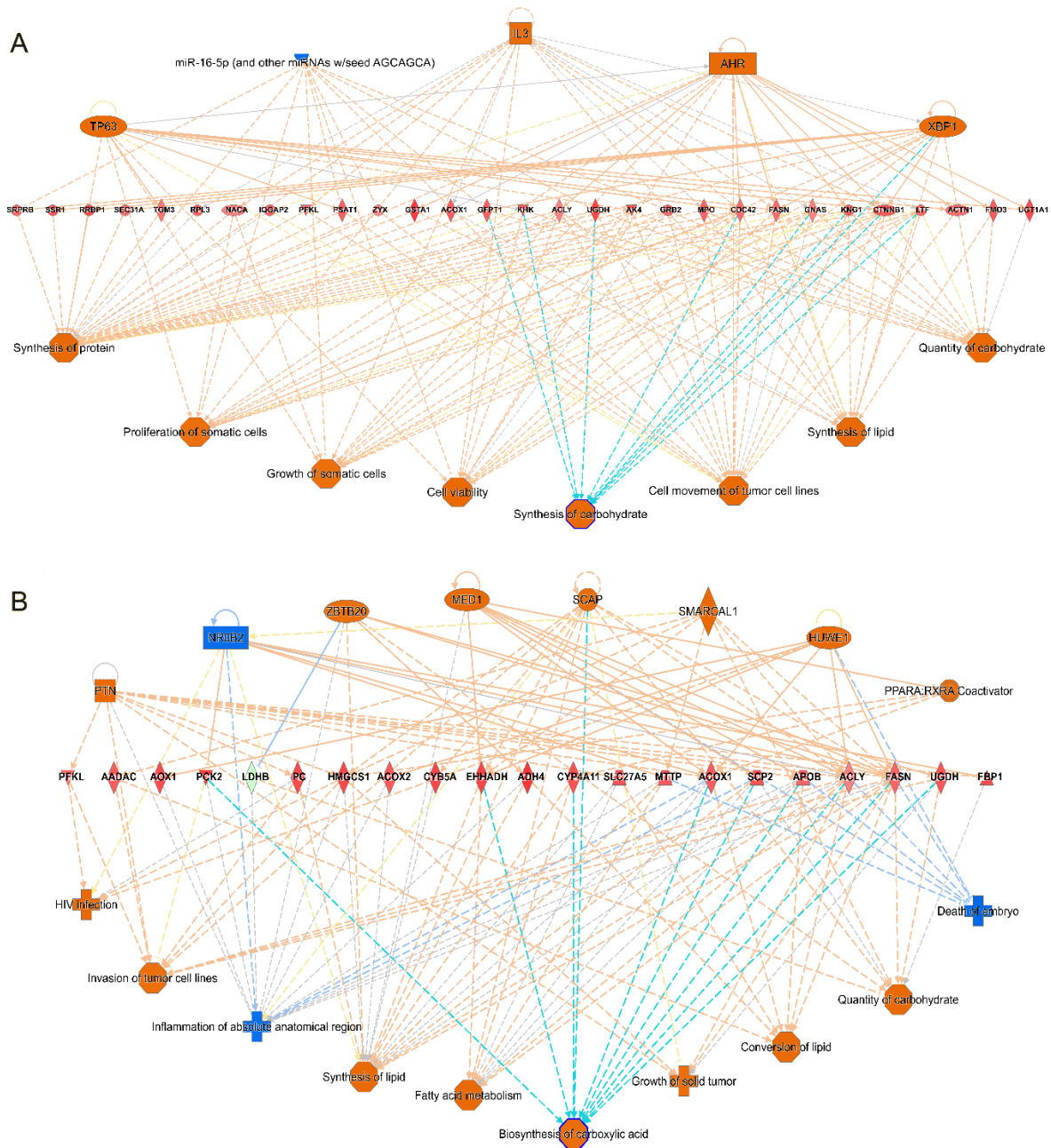


Figure 20. IPA regulator-effector networks of differentially abundant proteins. Regulator-effector networks generated by Ingenuity Pathway Analysis (IPA) for hepatic steatosis versus control (A) and fibrosis versus control (B). Node color reflects predicted activity state, with red indicating activation and blue indicating inhibition. Edges represent causal relationships, where orange lines denote predicted activation and blue lines denote predicted inhibition of downstream targets. Yellow edges indicate relationships that are inconsistent with the observed abundance patterns in the dataset.

Table 9. DAPs with sufficient pathway information from Ingenuity Pathway Analysis where at least four analysis molecules were mapped to the respective pathways.

Hepatic Steatosis - Control	Fibrosis - Control
ATP Citrate Lyase	Arylacetamide Deacetylase
Acyl-CoA Oxidase 1	ATP Citrate Lyase
Actinin Alpha 1	Acyl-CoA Oxidase 1
Adenylate Kinase 4	Acyl-CoA Oxidase 2
Cell Division Cycle 42	Alcohol Dehydrogenase 4 (Class II), Pi Polypeptide
Catenin Beta 1	Aldehyde Oxidase 1
Fatty Acid Synthase	Apolipoprotein B
Flavin Containing Dimethylaniline Monooxygenase 3	Cytochrome B5 Type A
Glutamine-Fructose-6-Phosphate Transaminase 1	Cytochrome P450 Family 4 Subfamily A Member 11
GNAS Complex Locus	Enoyl-CoA Hydratase And 3-Hydroxyacyl CoA Dehydrogenase
Growth Factor Receptor Bound Protein 2	Fatty Acid Synthase
Glutathione S-Transferase Alpha 1	Fructose-Bisphosphatase 1
IQ Motif Containing Gtpase Activating Protein 2	3-Hydroxy-3-Methylglutaryl-CoA Synthase 1
Ketohexokinase	Lactate Dehydrogenase B
Kininogen 1	Microsomal Triglyceride Transfer Protein
Lactotransferrin	Pyruvate Carboxylase
Myeloperoxidase	Phosphoenolpyruvate Carboxykinase 2, Mitochondrial

Nascent Polypeptide Associated Complex Subunit Alpha	Phosphofructokinase, Liver Type
Phosphofructokinase, Liver Type	Sterol Carrier Protein 2
Phosphoserine Aminotransferase 1	Solute Carrier Family 27-Member 5
Ribosomal Protein L3	UDP-Glucose 6-Dehydrogenase
Ribosome Binding Protein 1	
SEC31 Homolog A, COPII Component	
SRP Receptor Subunit Beta	
Signal Sequence Receptor Subunit 1	
Transglutaminase 3	
UDP-Glucose 6-Dehydrogenase	
UDP Glucuronosyltransferase Family 1 Member A1	
Zyxin	

6. DISCUSSION

The primary aim of this thesis was to overcome major technical and analytical challenges that hinder the widespread implementation of FFPE tissue proteomics, thereby supporting its broader use in clinical and translational research. To achieve this, two complementary strategies were pursued in parallel: the optimization of a data analysis workflow and the development of an improved sample preparation protocol for FFPE needle biopsy specimens.

The DEF workflow was developed to provide an accessible and user-friendly platform for DIA proteomics data post-processing and quality assessment. To improve usability and workflow organization, functionally related processing steps were grouped into modular KNIME components with integrated selection dialogs, thereby minimizing the need for programming expertise.

One of the key features implemented in DEF was extracted ion chromatogram (XIC)-based peptide quality filtering. When enabled, fragment ion series corresponding to b- and y-ions were concatenated for each peptide, after which an integrated Python script evaluated the presence of at least four consecutive fragment ions as an inclusion criterion for downstream analysis [92]. Although DIA-NN supports XIC visualization through DIA-NN Viewer and Skyline integration, practical application of these tools often requires additional software installation and manual import of peptide search results. DEF therefore provides a more streamlined approach for automated XIC-based filtering directly within the workflow environment. Contaminant management was implemented through a dedicated sub-workflow incorporating multiple commonly used contaminant databases, including MaxQuant, cRAP, and Hao Group contaminant lists. Because contaminant profiles vary substantially between laboratories and experimental workflows [93,94], users were given flexibility to apply individual or combined contaminant lists depending on their analytical requirements. Inclusion of updated contaminant resources is particularly important because several commonly used contaminant databases may contain obsolete entries or proteins that are not universally considered contaminants [86].

The workflow was additionally designed to maintain compatibility with recent DIA-NN quantification algorithms, including QuantUMS implemented in DIA-NN version 2.2 [95]. By default, DEF applies quantity-based quality thresholds to “Quantity” and “PG.MaxLFQ quality” metrics while preserving lower-confidence PG quantities (PG.MaxLFQ) as ‘50’ values rather than removing them entirely. This strategy enables users to perform additional downstream filtering according to experiment-specific requirements.

Evaluation of DEF across multiple benchmark datasets demonstrated broad compatibility with different mass spectrometry platforms, acquisition schemes, and DIA-NN output formats. Across all benchmark studies, DEF generated quantitative outputs comparable to those reported in the original publications. Slight improvements in quantitative reproducibility likely reflect methodological advances incorporated into recent DIA-NN versions, as DIA-NN has consistently demonstrated improved identification and quantification performance relative to several alternative DIA analysis platforms, including Spectronaut, OpenSWATH, Skyline, and DIA-Umpire [96,97]. Importantly, DEF enabled rapid post-processing of large DIA datasets, with a 35-sample batch processed in less than 14 minutes. These findings demonstrate that DEF provides efficient and reproducible DIA data filtering and quality assessment while remaining accessible to users with limited computational experience.

The application of the DEF workflow to study the adipocyte differentiation further demonstrated the utility of DEF for biological interpretation and interactive quality assessment of DIA proteomics datasets. The implemented “Quantity Visuals” component enabled visualization of protein abundance distributions prior to contaminant filtering, including high-abundance contaminants that may influence quantitative performance. Visualization of contaminant contribution is particularly relevant in mass spectrometry-based proteomics because highly abundant contaminant proteins can suppress detection of low-abundance peptides [86,98]. DEF-based differential abundance analysis identified substantial proteomic remodeling during adipocyte differentiation. Comparisons between white adipocytes and pre-adipocytes revealed numerous proteins exhibiting altered abundance under both the two-peptide and proteotypic peptide inference strategies. Similar trends were observed following palmitate treatment, but only a limited number of proteins differed significantly between

palmitate-treated and untreated adipocytes. Among these, cytochrome c oxidase subunit 2 (MT-CO2) displayed reduced abundance following palmitate exposure, whereas serotransferrin (TF) and ribosomal protein L10 (RPL10) exhibited increased abundance under the two-peptide inference strategy [91]. These proteins are associated with mitochondrial oxidative phosphorylation, iron transport, and translational activity, respectively, suggesting potential alterations in mitochondrial metabolism, cellular stress responses, and protein synthesis during fatty acid exposure.

Functional enrichment analyses demonstrated that adipocyte differentiation was accompanied by extensive cytoskeletal and structural remodeling. Biological processes associated with actin cytoskeleton organization, supramolecular fiber organization, apoptotic regulation, and energy metabolism were enriched across both differentiation comparisons. Correspondingly, molecular function analyses identified enrichment of actin-binding and cytoskeleton-associated proteins. These findings indicate that differentiation from precursor cells to mature adipocytes involves coordinated structural reorganization together with increased metabolic activity and oxidoreductase function. KEGG pathway enrichment analyses additionally demonstrated enrichment of carbon metabolism and broader metabolic pathways across differentiated adipocyte conditions, suggesting redirection of carbon flux toward anabolic and lipogenic processes required for triglyceride synthesis and energy storage. Because only a small number of proteins differed significantly between palmitate-treated and untreated adipocytes, pathway enrichment analyses for this comparison were limited. The reduced abundance of cytochrome c oxidase subunit 2, together with increased serotransferrin and ribosomal protein L10, suggests that differentiation in the presence of palmitate induces mitochondrial dysfunction while preserving biosynthetic capacity. The decreased mitochondrial respiratory component is consistent with impaired oxidative phosphorylation during lipotoxic stress, whereas the increased abundance of ribosomal machinery indicates continued protein synthesis required for cellular adaptation. Our data on palmitate-treated adipocytes may reflect extracellular matrix remodeling and stress-associated signaling responses induced by fatty acid exposure and provide further evidence of the protein profile

changes related to adipocyte differentiation supplying new information on the effect of palmitate, a damage associated molecular pattern, on the process.

Besides bulk tissue examinations it is critically important to be able to apply the highly performant proteomics techniques to the examination of FFPE samples, especially to the needle biopsies. We aimed to establish a simplified, safe, and cost-effective FFPE proteomics workflow that minimizes reliance on specialized instrumentation while maintaining compatibility with low-input clinical specimens. Although HMQ deparaffinization has previously been described in proteomics applications [31,32], PHAD has not yet been extensively explored in this context. Therefore, both xylene-free deparaffinization strategies were evaluated together with multiple extraction and digestion workflows to maximize proteome coverage from limited FFPE tissue material representative of needle biopsy specimens.

To improve proteome coverage and maintain methodological consistency across workflows, all digestion protocols incorporated a dual-protease strategy using Trypsin/Lys-C. Previous studies have demonstrated that sequential digestion with Lys-C and trypsin improves proteolytic efficiency and peptide recovery in complex proteomic samples [99].

Although App_3 exhibited improved cleavage efficiency relative to App_2, several factors likely contributed to this observation. The extensive reduction and homogenization conditions employed in App_3, including sonication in HEPES buffer containing DTT and SDS, may have enhanced protein unfolding and protease accessibility. This interpretation is supported by the increased proportion of shorter peptides and the reduced frequency of missed cleavages observed with App_3.

Detergent- and chaotrope-containing extraction buffers are known to influence trypsin activity [100]. Inadequate removal of these reagents may impair enzymatic digestion, resulting in incomplete cleavage and oversampling of abundant peptide variants. Conversely, surfactant-assisted extraction combined with appropriate cleanup procedures has been shown to improve digestion efficiency and peptide recovery [38,99,101]. The improved cleavage characteristics observed for App_3 are consistent with these previous findings. Despite the enhanced digestion efficiency associated with App_3, the workflow generated substantially fewer

peptide and protein identifications. This reduction may partially reflect peptide loss during acetone precipitation and cleanup procedures. Consequently, while App_3 demonstrated favorable proteolytic characteristics, its increased complexity and sample handling requirements may compromise overall peptide recovery, particularly in low-input workflows. These findings suggest that further optimization of surfactant-based extraction protocols may still improve performance.

In contrast, App_2 provided higher protein group identifications while requiring fewer processing steps and reduced handling complexity, making it more compatible with the practical objectives of this study. Evaluation of post-translational modifications (PTMs), particularly methionine oxidation and acetylation, revealed no apparent differences between App_2 and App_3 under the present experimental conditions. However, these analyses were performed without dedicated PTM enrichment strategies, limiting the depth and sensitivity of PTM characterization. Consequently, the suitability of the evaluated workflows for modification-specific proteomics applications requires further investigation using targeted enrichment approaches.

Overall, both App_2 and App_2M demonstrated good reproducibility, consistent quantitative performance between technical replicates, and relatively low frequencies of missed cleavages. The use of extended Lys-C predigestion prior to overnight trypsin digestion likely contributed to the observed digestion efficiency [49]. Incorporation of reduction and alkylation into App_2 resulted in the modified App_2M workflow, which demonstrated modest improvements in peptide and protein identifications together with slightly increased sequence coverage.

Using the optimized App_2M workflow, 1100-2500 proteins were identified from low-input FFPE needle biopsy material. These results are consistent with previously reported proteome depths obtained from biopsy-scale or LCM FFPE tissues analyzed using DIA-based proteomics workflows, where protein identifications typically range between approximately 800 and 2000 proteins [32,50,102]. Collectively, these findings support the suitability of the optimized workflow for low-input FFPE proteomics applications.

Analysis of the FFPE liver cohort demonstrated the applicability of the optimized App_2M workflow to clinically relevant low-input tissue specimens. However, several technical limitations characteristic of FFPE proteomics remained evident. Three samples were excluded prior to downstream analysis because of poor data quality associated with high keratin and hemoglobin contamination levels and reduced peptide identifications. Such variability is common in low-input FFPE proteomics workflows and may arise from differences in tissue composition, fixation quality, sample handling, and extraction efficiency [3].

The comparative analyses between hepatic steatosis, fibrosis, and control tissues revealed distinct proteomic alterations associated with disease progression. IPA regulator-effector network analyses indicated that both hepatic steatosis and fibrosis exhibited activation of metabolic and biosynthetic pathways; however, the upstream regulatory architecture differed between the two disease states.

In the hepatic steatosis vs control network, predicted activation of regulators including TP63, IL3, AHR, and XBP1 suggested involvement of inflammatory signaling, environmental stress sensing, unfolded protein response pathways, and transcriptional regulation of protein synthesis [103–105]. Simultaneously, predicted inhibition of miR-16-5p-related pathways further supports the presence of broad transcriptional and post-transcriptional regulatory alterations during early disease stages. At the protein level, coordinated abundance changes were observed among proteins associated with carbohydrate metabolism, lipid biosynthesis, and protein synthesis [106–109]. Functional predictions additionally suggested activation of pathways related to cellular growth, viability, and structural remodeling. Together, these observations indicate that hepatic steatosis is characterized by a metabolically adaptive and biosynthetically active phenotype rather than an exclusively lipid-centered state.

In contrast, the fibrosis vs control network displayed a more focused metabolic profile dominated by lipid metabolism and fatty acid-associated pathways. The fibrosis-associated regulator-effector network demonstrated stronger association with SCAP/SREBP-mediated lipogenic regulation [110], suggesting progression toward a more transcriptionally coordinated lipid metabolic phenotype.

Direct pathway comparison between fibrosis and hepatic steatosis was limited by the relatively small number of differentially abundant proteins identified between these conditions. Consequently, several pathways lacked sufficient molecule coverage to enable robust activity state prediction in IPA. The present findings are broadly consistent with emerging proteomic and transcriptomic evidence demonstrating that hepatic steatosis involves extensive metabolic, inflammatory, and stress-associated remodeling rather than solely dysregulated lipid accumulation [111,112]. Previous studies have similarly reported activation of unfolded protein response pathways, central carbon metabolism, and biosynthetic programs during early disease stages, including XBP1-mediated stress signaling.

Furthermore, inflammatory cytokine signaling and microRNA-associated regulatory mechanisms have been implicated in early steatotic progression [113,114], supporting the heterogeneous upstream regulatory landscape observed in the present study.

In contrast, fibrosis has been consistently associated with enhanced lipogenic signaling and SCAP/SREBP-driven metabolic reprogramming [115], suggesting transition from a broadly adaptive metabolic phenotype toward a more specialized lipid-centered regulatory state. Nevertheless, some studies indicate that lipogenic signaling may already be activated during steatosis, albeit at lower magnitude [110]. Likewise, the role of AHR signaling appears context-dependent and may contribute to both adaptive and pathogenic responses [116].

These observations support a progressive disease model in which hepatic steatosis represents an early stress-adaptive and metabolically flexible state that gradually converges toward a more coordinated lipid-centric phenotype during fibrosis progression.

Taken together, this work provides a straightforward, mass spectrometry-compatible workflow for FFPE proteomics together with a user-friendly computational platform for DIA data analysis, thereby supporting the wider adoption of FFPE-based proteomics in clinical and translational research.

7. SUMMARY

FFPE tissues represent one of the most valuable resources for clinical and translational research; however, their broader application in proteomics remains limited by the lack of standardized and accessible workflows. This thesis addressed these challenges through the optimization of xylene-free FFPE sample preparation methods and in parallel to this, the development of a user-friendly computational workflow for DIA-based proteomics data analysis.

Four FFPE extraction and digestion workflows were systematically evaluated. Among these, a chaotropic-, reductant-, and surfactant-free in-solution digestion approach (App_2) demonstrated the best balance between proteome depth, reproducibility, and workflow simplicity. A modified version incorporating reduction and alkylation (App_2M) further improved peptide recovery. Importantly, the optimized workflow avoids specialized instrumentation and extensive cleanup procedures, supporting its suitability for routine and low-input clinical applications. Application of the optimized workflow to FFPE liver needle biopsies enabled the identification of plausible proteomic differences associated with hepatic steatosis and liver fibrosis. The cohort size was not powered for biological discovery, and the observed disease-associated signatures should be interpreted with caution. Although not providing definitive insights into the molecular mechanisms underlying steatosis or fibrosis, this case study serves as a proof-of-concept demonstration of the workflow's technical capability to generate informative proteomic data from limited FFPE biopsy material.

In parallel, the developed DEF workflow provides a low-code/no-code solution for DIA proteomics post-processing. DEF integrates peptide and protein filtering, contaminant handling, quantitative assessment, and interactive visualization tools into a modular workflow. Benchmarking against multiple public datasets demonstrated identification and quantification performance comparable to previously published results across diverse instruments and acquisition strategies. In addition, the proteomic changes associated with adipocyte differentiation in the presence of palmitate suggest metabolic reprogramming characterized

by impaired mitochondrial respiration, maintained protein synthesis, and altered iron homeostasis.

Collectively, this work establishes a simple and mass spectrometry-compatible FFPE proteomics workflow together with an accessible computational platform for DIA data analysis, thereby contributing to the broader implementation of FFPE-based proteomics in clinical and translational research.

8. NEW FINDINGS

- Development of an optimized, xylene-free protein extraction and digestion protocol for FFPE samples;
- Development of DEF workflow, an accessible DIA-based data analysis workflow for comprehensive proteome profiling and data visualization;
- Proteomics examination of FFPE needle biopsy samples to study liver fibrosis;
- Generation of new insights into the proteomic changes accompanying the molecular effects of palmitic acid administration during adipocyte differentiation.

9. REFERENCES

1. Morris, R.; Black, K.A.; Stollar, E.J. Uncovering Protein Function: From Classification to Complexes. *Essays Biochem.* 2022, **66**, 255–285.
2. Yang, L.; Gregorich, Z.R.; Cai, W.; Zhang, P.; Young, B.; Gu, Y.; Zhang, J.; Ge, Y. Quantitative Proteomics and Immunohistochemistry Reveal Insights into Cellular and Molecular Processes in the Infarct Border Zone One Month after Myocardial Infarction. *J. Proteome Res.* **2017**, *16*, 2101–2112, doi:10.1021/acs.jproteome.7b00107.
3. Kim, S.W.; Roh, J.; Park, C.S. Immunohistochemistry for Pathologists: Protocols, Pitfalls, and Tips. *J. Pathol. Transl. Med.* 2016, *50*, 411–418.
4. Wakefield, C.; Hornick, J.L. Update on Immunohistochemistry in Bone and Soft Tissue Tumors: Cost-Effectively Replacing Molecular Testing with Immunohistochemistry. *Hum. Pathol.* **2024**, *147*, 58–71, doi:10.1016/j.humpath.2023.12.002.
5. Nokhoijav, E.; Káplár, M.; Aranyi, S.C.; Berzi, A.; Bergström, G.; Antonopoulos, K.; Edfors, F.; Emri, M.; Csősz, É. Machine Learning Uncovers Circulating Biomarkers and Molecular Heterogeneity in Obesity and Type 2 Diabetes. *bioRxiv* **2026**, doi:10.64898/2026.04.16.718836.
6. Carlyle, B.C.; Trombetta, B.A.; Arnold, S.E. Proteomic Approaches for the Discovery of Biofluid Biomarkers of Neurodegenerative Dementias. *Proteomes* **2018**, *6*, 32, doi:10.3390/proteomes6030032.
7. Teunissen, C.E.; Kimble, L.; Bayoumy, S.; Bolsewig, K.; Burtscher, F.; Coppens, S.; Das, S.; Gogishvili, D.; Fernandes Gomes, B.; Gómez de San José, N.; et al. Methods to Discover and Validate Biofluid-Based Biomarkers in Neurodegenerative Dementias. *Molecular & Cellular Proteomics* **2023**, *22*, 100629, doi:10.1016/j.mcpro.2023.100629.
8. Sinha, A.; Sachs, N.; Kratz, E.; Pauli, J.; Steigerwald, S.; Albrecht, V.; Nordmann, T.; Ugur, E.; Rodriguez, E.H.; Engl, M.-L.; et al. Spatially Resolved Proteomic Signatures of Atherosclerotic Carotid Artery Disease 2025.
9. Buczak, K.; Kirkpatrick, J.M.; Truckenmueller, F.; Santinha, D.; Ferreira, L.; Roessler, S.; Singer, S.; Beck, M.; Ori, A. Spatially Resolved Analysis of FFPE Tissue Proteomes by Quantitative Mass Spectrometry. *Nat. Protoc.* **2020**, *15*, 2956–2979, doi:10.1038/s41596-020-0356-y.
10. Dong, Z.; Jiang, W.; Wu, C.; Chen, T.; Chen, J.; Ding, X.; Zheng, S.; Piatkevich, K.D.; Zhu, Y.; Guo, T. Spatial Proteomics of Single Cells and Organelles on Tissue Slides Using

Filter-Aided Expansion Proteomics. *Nature Communications* **2024**, *15*, doi:10.1038/s41467-024-53683-7.

11. Nwosu, A.J.; Misal, S.A.; Truong, T.; Carson, R.H.; Webber, K.G.I.; Axtell, N.B.; Liang, Y.; Johnston, S.M.; Virgin, K.L.; Smith, E.G.; et al. In-Depth Mass Spectrometry-Based Proteomics of Formalin-Fixed, Paraffin-Embedded Tissues with a Spatial Resolution of 50-200 Mm. *J. Proteome Res.* **2022**, *21*, 2237–2245, doi:10.1021/acs.jproteome.2c00409.
12. Achter, J.S.; Jensen, T.H.L.; Pisano, P.; Bundgaard, J.S.; Raaschou-Oddershede, D.; Rossing, K.; Wierer, M.; Lundby, A. Quantitative Proteomics of Formalin-Fixed, Paraffin-Embedded Cardiac Specimens Uncovers Protein Signatures of Specialized Regions and Patient Groups. *Nature Cardiovascular Research* **2025**, *4*, 1409–1423, doi:10.1038/s44161-025-00721-2.
13. Liang, S.; Zheng, B.; Huang, D.; Zhang, J.; Li, H.; Zhao, Y.; Gao, X.; Xiao, T.; Ying, W. A Streamlined Technique for In-Depth Ubiquitomics Analysis of Formalin-Fixed Paraffin-Embedded Tissues. *Analytical Methods* **2026**, *18*, 138–147, doi:10.1039/d5ay01727k.
14. AlHammadi, S.A.; Nagshabandi, L.N.; Muhammad, H.; Sukkarieh, H.H.; Aljada, A. Mass Spectrometry-Based Proteomics of FFPE Tissues: Progress, Limitations, and Clinical Translation Barriers. *Clin. Proteomics* **2025**, *22*.
15. Friedrich, C.; Schallenberg, S.; Kirchner, M.; Ziehm, M.; Niquet, S.; Haji, M.; Beier, C.; Neudecker, J.; Klauschen, F.; Mertins, P. Comprehensive Micro-Scaled Proteome and Phosphoproteome Characterization of Archived Retrospective Cancer Repositories. *Nat. Commun.* **2021**, *12*, doi:10.1038/s41467-021-23855-w.
16. Hipp, J.D.; Johann, D.J.; Chen, Y.; Madabhushi, A.; Monaco, J.; Cheng, J.; Rodriguez-Canales, J.; Stumpe, M.C.; Riedlinger, G.; Rosenberg, A.Z.; et al. Computer-Aided Laser Dissection: A Microdissection Workflow Leveraging Image Analysis Tools. *J. Pathol. Inform.* **2018**, *9*, 45, doi:10.4103/jpi.jpi_60_18.
17. Heaton, K.J.; Master, S.R. Peptide Extraction from Formalin-Fixed Paraffin-Embedded Tissue. *Curr. Protoc. Protein Sci.* **2011**, *65*, doi:10.1002/0471140864.ps2305s65.
18. Walsh, E.M.; Halushka, M.K. A Comparison of Tissue Dissection Techniques for Diagnostic, Prognostic, and Theragnostic Analysis of Human Disease. *Pathobiology* **2023**, *90*, 199–208.
19. Domon, B.; Aebersold, R. *Mass Spectrometry and Protein Analysis*;

20. Guo, T.; Steen, J.A.; Mann, M. Mass-Spectrometry-Based Proteomics: From Single Cells to Clinical Applications. *Nature* **2025**, *638*, 901–911, doi:10.1038/s41586-025-08584-0.
21. Kalló, G.; Zaman, K.; Potor, L.; Hendrik, Z.; Méhes, G.; Tóth, C.; Gergely, P.; Tózsér, J.; Balla, G.; Balla, J.; et al. Identification of Protein Networks and Biological Pathways Driving the Progression of Atherosclerosis in Human Carotid Arteries Through Mass Spectrometry-Based Proteomics. *Int. J. Mol. Sci.* **2024**, *25*, doi:10.3390/ijms252413665.
22. Demichev, V.; Messner, C.B.; Vernardis, S.I.; Lilley, K.S.; Ralser, M. DIA-NN: Neural Networks and Interference Correction Enable Deep Proteome Coverage in High Throughput Europe PMC Funders Group. *Nat Methods* **2020**, *17*, 41–44, doi:10.17605/OSF.IO/6G3UX.
23. Mallick, P.; Schirle, M.; Chen, S.S.; Flory, M.R.; Lee, H.; Martin, D.; Ranish, J.; Raught, B.; Schmitt, R.; Werner, T.; et al. Computational Prediction of Proteotypic Peptides for Quantitative Proteomics. *Nat. Biotechnol.* **2007**, *25*, 125–131, doi:10.1038/nbt1275.
24. Piga, I.; Koenig, C.; Lechner, M.; Sabatier, P.; Olsen, J. V. Formaldehyde Fixation Helps Preserve the Proteome State during Single-Cell Proteomics Sample Processing and Analysis. *J. Proteome Res.* **2025**, *24*, 1624–1635, doi:10.1021/acs.jproteome.4c00656.
25. Tayri-Wilk, T.; Slavin, M.; Zamel, J.; Blass, A.; Cohen, S.; Motzik, A.; Sun, X.; Shalev, D.E.; Ram, O.; Kalisman, N. Mass Spectrometry Reveals the Chemistry of Formaldehyde Cross-Linking in Structured Proteins. *Nat. Commun.* **2020**, *11*, doi:10.1038/s41467-020-16935-w.
26. Metz, B.; Kersten, G.F.A.; Baart, G.J.E.; De Jong, A.; Meiring, H.; Ten Hove, J.; Van Steenberg, M.J.; Hennink, W.E.; Crommelin, D.J.A.; Jiskoot, W. Identification of Formaldehyde-Induced Modifications in Proteins: Reactions with Insulin. *Bioconjug. Chem.* **2006**, *17*, 815–822, doi:10.1021/bc050340f.
27. Kamps, J.J.A.G.; Hopkinson, R.J.; Schofield, C.J.; Claridge, T.D.W. How Formaldehyde Reacts with Amino Acids. *Commun. Chem.* **2019**, *2*, doi:10.1038/s42004-019-0224-2.
28. Sugár, S.N.; Molnár, B.A.; Bugyi, F.; Kecskeméti, G.; Szabó, Z.; Laczó, I.; Harkó, T.; Moldvay, J.; Turiák, L. Glycoproteomics Analysis of Triple Wild-Type Lung Adenocarcinoma Tissue Samples. *J. Proteome Res.* **2025**, *24*, 2419–2429, doi:10.1021/acs.jproteome.4c01063.
29. Eckert, S.; Chang, Y.C.; Bayer, F.P.; The, M.; Kuhn, P.H.; Weichert, W.; Kuster, B. Evaluation of Disposable Trap Column NanoLC-FAIMS-MS/MS for the Proteomic Analysis of FFPE Tissue. *J. Proteome Res.* **2021**, *20*, 5402–5411, doi:10.1021/acs.jproteome.1c00695.

30. Kandyala, R.; Raghavendra, S.P.; Rajasekharan, S. Xylene: An Overview of Its Health Hazards and Preventive Measures. *Journal of Oral and Maxillofacial Pathology* **2010**, *14*, 1, doi:10.4103/0973-029x.64299.
31. Mansour, A.G.; Abou-Khalil, P.; Bejjani, N.; Chatila, R.; Dagher-Hamalian, C.; Faour, W.H. An Optimized Xylene-Free Protein Extraction Method Adapted to Formalin-Fixed Paraffin Embedded Tissue Sections for Western Blot Analysis. *Histol. Histopathol.* **2017**, *32*, 307–313, doi:10.14670/HH-11-789.
32. Mitsa, G.; Guo, Q.; Goncalves, C.; Preston, S.E.J.; Lacasse, V.; Aguilar-Mahecha, A.; Benlimame, N.; Basik, M.; Spatz, A.; Batist, G.; et al. A Non-Hazardous Deparaffinization Protocol Enables Quantitative Proteomics of Core Needle Biopsy-Sized Formalin-Fixed and Paraffin-Embedded (FFPE) Tissue Specimens. *Int. J. Mol. Sci.* **2022**, *23*, doi:10.3390/ijms23084443.
33. Marinho, P.F.; Vieira, S.L.; Carvalho, T.G.; Peleteiro, M.C.; Hanscheid, T. A Novel and Simple Heat-Based Method Eliminates the Highly Detrimental Effect of Xylene Deparaffinization on Acid-Fast Stains. *Am. J. Clin. Pathol.* **2023**, *160*, 81–88, doi:10.1093/ajcp/aqad016.
34. Marinho, P.F.; Hanscheid, T. A Simple Heat-Based Alternative Method for Deparaffinization of Histological Sections Significantly Improves Acid-Fast Staining Results for Mycobacteria in Tissue. *MethodsX* **2023**, *10*, doi:10.1016/j.mex.2023.102079.
35. Scicchitano, M.S.; Dalmas, D.A.; Boyce, R.W.; Thomas, H.C.; Frazier, K.S. Protein Extraction of Formalin-Fixed, Paraffin-Embedded Tissue Enables Robust Proteomic Profiles by Mass Spectrometry. *Journal of Histochemistry and Cytochemistry* **2009**, *57*, 849–860, doi:10.1369/jhc.2009.953497.
36. Varshini, M.; Pushparaj, S.; Mangesh, S. Advancing Green Chemistry in Histopathology: A Comparative Study of Coconut Oil and Turpentine Oil as Deparaffinizing Agents. *Indian J. Pathol. Microbiol.* **2026**, doi:10.4103/ijpm.ijpm_836_24.
37. Lee, D.; Rubakhin, S.S.; Kusmartseva, I.; Wasserfall, C.; Atkinson, M.A.; Sweedler, J. V. Removing Formaldehyde-Induced Peptidyl Crosslinks Enables Mass Spectrometry Imaging of Peptide Hormone Distributions from Formalin-Fixed Paraffin-Embedded Tissues. *Angew. Chem. Int. Ed.* **2020**, *59*, 22584–22590, doi:10.1002/anie.202008847.
38. Varnavides, G.; Madern, M.; Anrather, D.; Hartl, N.; Reiter, W.; Hartl, M. In Search of a Universal Method: A Comparative Survey of Bottom-Up Proteomics Sample

Preparation Methods. *J. Proteome Res.* **2022**, *21*, 2397–2411, doi:10.1021/acs.jproteome.2c00265.

39. Jiang, Y.; Rex, D.A.B.; Schuster, D.; Neely, B.A.; Rosano, G.L.; Volkmar, N.; Momenzadeh, A.; Peters-Clarke, T.M.; Egbert, S.B.; Kreimer, S.; et al. Comprehensive Overview of Bottom-Up Proteomics Using Mass Spectrometry. *ACS Measurement Science Au* **2024**, *4*, 338–417, doi:10.1021/acsmeasuresciau.3c00068.
40. Darville, L.N.F.; Lockhart, J.H.; Putty Reddy, S.; Fang, B.; Izumi, V.; Boyle, T.A.; Haura, E.B.; Flores, E.R.; Koomen, J.M. A Fast-Tracking Sample Preparation Protocol for Proteomics of Formalin-Fixed Paraffin-Embedded Tumor Tissues. In *Methods in Molecular Biology*; Humana Press Inc., 2024; Vol. 2823, pp. 193–223.
41. Waas, M.; Kislinger, T. Addressing Cellular Heterogeneity in Cancer through Precision Proteomics. *J. Proteome Res.* **2020**, *19*, 3607–3619, doi:10.1021/acs.jproteome.0c00338.
42. Labib, M.; Kelley, S.O. Single-Cell Analysis Targeting the Proteome. *Nat. Rev. Chem.* **2020**, *4*, 143–158.
43. Ahmad, R.; Budnik, B. A Review of the Current State of Single-Cell Proteomics and Future Perspective. *Anal. Bioanal. Chem.* **2023**, *415*, 6889–6899.
44. Liotta, L.A.; Petricoin, E.F. Multiplex Proteomic Analysis of Cells in Fine Needle Aspiration Samples: How Soon Can We Expect to Individualize Therapy by Comparing Pre- and Posttreatment Biopsies? *Clin. Chem.* **2014**, *60*, 1143–1145, doi:10.1373/clinchem.2014.222828.
45. Csósz, É.; Kalló, G.; Márkus, B.; Deák, E.; Csutak, A.; Tózsér, J. Quantitative Body Fluid Proteomics in Medicine — A Focus on Minimal Invasiveness. *J. Proteomics* **2017**, *153*, 30–43, doi:10.1016/j.jprot.2016.08.009.
46. Wu, F.; Sun, D.; Wang, N.; Gong, Y.; Li, L. Comparison of Surfactant-Assisted Shotgun Methods Using Acid-Labile Surfactants and Sodium Dodecyl Sulfate for Membrane Proteome Analysis. *Anal. Chim. Acta* **2011**, *698*, 36–43, doi:10.1016/j.aca.2011.04.039.
47. Nesvizhskii, A.I. A Survey of Computational Methods and Error Rate Estimation Procedures for Peptide and Protein Identification in Shotgun Proteomics. *J. Proteomics* **2010**, *73*, 2092–2123, doi:10.1016/j.jprot.2010.08.009.
48. Coscia, F.; Doll, S.; Bech, J.M.; Schweizer, L.; Mund, A.; Lengyel, E.; Lindebjerg, J.; Madsen, G.I.; Moreira, J.M.; Mann, M. A Streamlined Mass Spectrometry–Based

Proteomics Workflow for Large-scale FFPE Tissue Analysis. *J. Pathol.* **2020**, *251*, 100–112, doi:10.1002/path.5420.

49. Sergei Saveliev, M.B.R.Z.M.S.H.B.& M.U. *Trypsin/Lys-C Protease Mix for Enhanced Protein Mass Spectrometry Analysis*; 2013;
50. Pujari, G.P.; Mangalaparthi, K.K.; Madden, B.J.; Bhat, F.A.; Charlesworth, M.C.; French, A.J.; Sachdeva, G.; Daviso, E.; Thomann, U.; McCarthy, P.; et al. A High-Throughput Workflow for FFPE Tissue Proteomics. *J. Am. Soc. Mass Spectrom.* **2023**, *34*, 1225–1229, doi:10.1021/jasms.3c00099.
51. Gillet, L.C.; Navarro, P.; Tate, S.; Röst, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. Targeted Data Extraction of the MS/MS Spectra Generated by Data-Independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Molecular and Cellular Proteomics* **2012**, *11*, doi:10.1074/mcp.O111.016717.
52. Egertson, J.D.; Kuehn, A.; Merrihew, G.E.; Bateman, N.W.; MacLean, B.X.; Ting, Y.S.; Canterbury, J.D.; Marsh, D.M.; Kellmann, M.; Zabrouskov, V.; et al. Multiplexed MS/MS for Improved Data-Independent Acquisition. *Nat. Methods* **2013**, *10*, 744–746, doi:10.1038/nmeth.2528.
53. Deutsch, E.W.; Lane, L.; Overall, C.M.; Bandeira, N.; Baker, M.S.; Pineau, C.; Moritz, R.L.; Corrales, F.; Orchard, S.; Van Eyk, J.E.; et al. Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 3.0. *J. Proteome Res.* 2019, *18*, 4108–4116.
54. Helm, S.; Baginsky, S. MSE for Label-Free Absolute Protein Quantification in Complex Proteomes. In; 2018; pp. 235–247.
55. Panchaud, A.; Jung, S.; Shaffer, S.A.; Aitchison, J.D.; Goodlett, D.R. Faster, Quantitative, and Accurate Precursor Acquisition Independent from Ion Count. *Anal. Chem.* **2011**, *83*, 2250–2257, doi:10.1021/ac103079q.
56. Colangelo, C.M.; Chung, L.; Bruce, C.; Cheung, K.-H. Review of Software Tools for Design and Analysis of Large Scale MRM Proteomic Datasets. *Methods* **2013**, *61*, 287–298, doi:10.1016/j.ymeth.2013.05.004.
57. Anderson, L.; Hunter, C.L. Quantitative Mass Spectrometric Multiple Reaction Monitoring Assays for Major Plasma Proteins. *Molecular and Cellular Proteomics* **2006**, *5*, 573–588, doi:10.1074/mcp.M500331-MCP200.

58. Kalló, G.; Emri, M.; Varga, Z.; Ujhelyi, B.; Tozsér, J.; Csutak, A.; Csosz, É. Changes in the Chemical Barrier Composition of Tears in Alzheimer's Disease Reveal Potential Tear Diagnostic Biomarkers. *PLoS One* **2016**, *11*, doi:10.1371/journal.pone.0158000.
59. Bons, J.; Husson, G.; Chion, M.; Bonnet, M.; Maumy-Bertrand, M.; Delalande, F.; Cianférani, S.; Bertrand, F.; Picard, B.; Carapito, C. Combining Label-free and Label-based Accurate Quantifications with SWATH-MS: Comparison with SRM and PRM for the Evaluation of Bovine Muscle Type Effects. *Proteomics* **2021**, *21*, doi:10.1002/pmic.202000214.
60. Eng, J.K.; Fischer, B.; Grossmann, J.; MacCoss, M.J. A Fast SEQUEST Cross Correlation Algorithm. *J. Proteome Res.* **2008**, *7*, 4598–4602, doi:10.1021/pr800420s.
61. Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R.A.; Olsen, J. V.; Mann, M. Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J. Proteome Res.* **2011**, *10*, 1794–1805, doi:10.1021/pr101065j.
62. Kim, S.; Pevzner, P.A. MS-GF+ Makes Progress towards a Universal Database Search Tool for Proteomics. *Nat. Commun.* **2014**, *5*, 5277, doi:10.1038/ncomms6277.
63. Yu, F.; Deng, Y.; Nesvizhskii, A.I. MSFragger-DDA+ Enhances Peptide Identification Sensitivity with Full Isolation Window Search. *Nature Communications* **2025**, *16*, doi:10.1038/s41467-025-58728-z.
64. Sinitcyn, P.; Hamzeiy, H.; Salinas Soto, F.; Itzhak, D.; McCarthy, F.; Wichmann, C.; Steger, M.; Ohmayer, U.; Distler, U.; Kaspar-Schoenefeld, S.; et al. MaxDIA Enables Library-Based and Library-Free Data-Independent Acquisition Proteomics. *Nat. Biotechnol.* **2021**, *39*, 1563–1573, doi:10.1038/s41587-021-00968-7.
65. Bruderer, R.; Bernhardt, O.M.; Gandhi, T.; Miladinović, S.M.; Cheng, L.Y.; Messner, S.; Ehrenberger, T.; Zanotelli, V.; Butscheid, Y.; Escher, C.; et al. Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues. *Molecular and Cellular Proteomics* **2015**, *14*, 1400–1410, doi:10.1074/mcp.M114.044305.
66. Searle, B.C.; Pino, L.K.; Egertson, J.D.; Ting, Y.S.; Lawrence, R.T.; MacLean, B.X.; Villén, J.; MacCoss, M.J. Chromatogram Libraries Improve Peptide Detection and Quantification by Data Independent Acquisition Mass Spectrometry. *Nat. Commun.* **2018**, *9*, doi:10.1038/s41467-018-07454-w.
67. Jiang, N.; Gao, Y.; Xu, J.; Luo, F.; Zhang, X.; Chen, R. A Data-Independent Acquisition (DIA)-Based Quantification Workflow for Proteome Analysis of 5000 Cells. *J. Pharm. Biomed. Anal.* **2022**, *216*, doi:10.1016/j.jpba.2022.114795.

68. Steger, M.; Demichev, V.; Backman, M.; Ohmayer, U.; Ihmor, P.; Müller, S.; Ralser, M.; Daub, H. Time-Resolved in Vivo Ubiquitinome Profiling by DIA-MS Reveals USP7 Targets on a Proteome-Wide Scale. *Nat. Commun.* **2021**, *12*, doi:10.1038/s41467-021-25454-1.
69. Kistner, F.; Grossmann, J.L.; Sinn, L.R.; Demichev, V. QuantUMS: Uncertainty Minimisation Enables Confident Quantification in Proteomics 2023.
70. Koopmans, F.; Li, K.W.; Klaassen, R. V.; Smit, A.B. MS-DAP Platform for Downstream Data Analysis of Label-Free Proteomics Uncovers Optimal Workflows in Benchmark Data Sets and Increased Sensitivity in Analysis of Alzheimer's Biomarker Data. *J. Proteome Res.* **2023**, *22*, 374–386, doi:10.1021/acs.jproteome.2c00513.
71. Gerault, M.A.; Camoin, L.; Granjeaud, S. DIAGui: A Shiny Application to Process the Output from DIA-NN. *Bioinformatics Advances* **2024**, *4*, doi:10.1093/bioadv/vbae001.
72. Khorram, F.; Mottu, J.M.; Sunyé, G. Challenges & Opportunities in Low-Code Testing. In Proceedings of the Proceedings - 23rd ACM/IEEE International Conference on Model Driven Engineering Languages and Systems, MODELS-C 2020 - Companion Proceedings; Association for Computing Machinery, Inc, October 16 2020; pp. 490–499.
73. Vincent, P.; Natis, Y.; Iijima, K.; Wong, J.; Ray, S.; Jain, A.; Leow, A. *Licensed for Distribution Magic Quadrant for Enterprise Low-Code Application Platforms*; 2020;
74. Berthold, M.R.; Cebron, N.; Dill, F.; Gabriel, T.R.; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B. *KNIME-The Konstanz Information Miner Version 2.0 and Beyond*;
75. Giardine, B.; Riemer, C.; Hardison, R.C.; Burhans, R.; Elnitski, L.; Shah, P.; Zhang, Y.; Blankenberg, D.; Albert, I.; Taylor, J.; et al. Galaxy: A Platform for Interactive Large-Scale Genome Analysis. *Genome Res.* **2005**, *15*, 1451–1455, doi:10.1101/gr.4086505.
76. Abueg, L.A.L.; Afgan, E.; Allart, O.; Awan, A.H.; Bacon, W.A.; Baker, D.; Bassetti, M.; Batut, B.; Bernt, M.; Blankenberg, D.; et al. The Galaxy Platform for Accessible, Reproducible, and Collaborative Data Analyses: 2024 Update. *Nucleic Acids Res.* **2024**, *52*, W83–W94, doi:10.1093/nar/gkae410.
77. Villarroel Ordenes, F.; Silipo, R. Machine Learning for Marketing on the KNIME Hub: The Development of a Live Repository for Marketing Applications. *J. Bus. Res.* **2021**, *137*, 393–410, doi:10.1016/j.jbusres.2021.08.036.
78. Moreira-Filho, J.T.; Ranganath, D.; Conway, M.; Schmitt, C.; Kleinstreuer, N.; Mansouri, K. Democratizing Cheminformatics: Interpretable Chemical Grouping Using an

Automated KNIME Workflow. *J. Cheminform.* **2024**, *16*, doi:10.1186/s13321-024-00894-1.

79. Fillbrunn, A.; Dietz, C.; Pfeuffer, J.; Rahn, R.; Landrum, G.A.; Berthold, M.R. KNIME for Reproducible Cross-Domain Analysis of Life Science Data. *J. Biotechnol.* 2017, *261*, 149–156.
80. Röst, H.L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H.C.; Gutenbrunner, P.; Kenar, E.; et al. OpenMS: A Flexible Open-Source Software Platform for Mass Spectrometry Data Analysis. *Nat. Methods* 2016, *13*, 741–748.
81. Bartolec, T.K.; Vázquez-Campos, X.; Norman, A.; Luong, C.; Payne, R.J.; Wilkins, M.R.; Mackay, J.P.; Low, J.K.K. Cross-Linking Mass Spectrometry Discovers, Evaluates, and Validates the Experimental and Predicted Structural Proteome. *bioRxiv* **2022**, 2022.11.16.516813, doi:10.1101/2022.11.16.516813.
82. Sugár, S.; Tóth, G.; Bugyi, F.; Vékey, K.; Karászi, K.; Drahos, L.; Turiák, L. Alterations in Protein Expression and Site-Specific N-Glycosylation of Prostate Cancer Tissues. *Sci. Rep.* **2021**, *11*, doi:10.1038/s41598-021-95417-5.
83. Wabitsch, M.; Melzner, I.; Braun, M.; Mo Èller, P.; Heinze, E.; Debatin, K.-M.; Hauner, H. *Characterization of a Human Preadipocyte Cell Strain with High Capacity for Adipose Differentiation*;
84. Vinnai, B.Á.; Arianti, R.; Fischer-Posovszky, P.; Wabitsch, M.; Fésüs, L.; Kristóf, E. The Importance of Thiamine Availability in the Thermogenic Competency of Human Adipocytes. *Mol. Cell. Endocrinol.* **2025**, *599*, doi:10.1016/j.mce.2025.112483.
85. Tyanova, S.; Temu, T.; Cox, J. The MaxQuant Computational Platform for Mass Spectrometry-Based Shotgun Proteomics. *Nat. Protoc.* **2016**, *11*, 2301–2319, doi:10.1038/nprot.2016.136.
86. Frankenfield, A.M.; Ni, J.; Ahmed, M.; Hao, L. Protein Contaminants Matter: Building Universal Protein Contaminant Libraries for DDA and DIA Proteomics. *J. Proteome Res.* **2022**, *21*, 2104–2113, doi:10.1021/acs.jproteome.2c00145.
87. Pang, Z.; Lu, Y.; Zhou, G.; Hui, F.; Xu, L.; Viau, C.; Spigelman, A.F.; Macdonald, P.E.; Wishart, D.S.; Li, S.; et al. MetaboAnalyst 6.0: Towards a Unified Platform for Metabolomics Data Processing, Analysis and Interpretation. *Nucleic Acids Res.* **2024**, *52*, W398–W406, doi:10.1093/nar/gkae253.

88. Lou, R.; Cao, Y.; Li, S.; Lang, X.; Li, Y.; Zhang, Y.; Shui, W. Benchmarking Commonly Used Software Suites and Analysis Workflows for DIA Proteomics and Phosphoproteomics. *Nat. Commun.* **2023**, *14*, doi:10.1038/s41467-022-35740-1.
89. Navarro, P.; Kuharev, J.; Gillet, L.C.; Bernhardt, O.M.; MacLean, B.; Röst, H.L.; Tate, S.A.; Tsou, C.C.; Reiter, L.; Distler, U.; et al. A Multicenter Study Benchmarks Software Tools for Label-Free Proteome Quantification. *Nat. Biotechnol.* **2016**, *34*, 1130–1136, doi:10.1038/nbt.3685.
90. Lazar, C.; Gatto, L.; Ferro, M.; Bruley, C.; Burger, T. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *J. Proteome Res.* **2016**, *15*, 1116–1125, doi:10.1021/acs.jproteome.5b00981.
91. Szklarczyk, D.; Kirsch, R.; Koutrouli, M.; Nastou, K.; Mehryary, F.; Hachilif, R.; Gable, A.L.; Fang, T.; Doncheva, N.T.; Pyysalo, S.; et al. The STRING Database in 2023: Protein-Protein Association Networks and Functional Enrichment Analyses for Any Sequenced Genome of Interest. *Nucleic Acids Res.* **2023**, *51*, D638–D646, doi:10.1093/nar/gkac1000.
92. Mann, M.; Wilm, M. *ErrorTolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags*; Academic Press, 1994; Vol. 66;.
93. Fox, K.; Castanha, E.; Fox, A.; Feigley, C.; Salzberg, D. Human K10 Epithelial Keratin Is the Most Abundant Protein in Airborne Dust of Both Occupied and Unoccupied School Rooms. *Journal of Environmental Monitoring* **2008**, *10*, 55–59, doi:10.1039/b714802j.
94. Keller, B.O.; Sui, J.; Young, A.B.; Whittall, R.M. Interferences and Contaminants Encountered in Modern Mass Spectrometry. *Anal. Chim. Acta* **2008**, *627*, 71–81.
95. Moschem, J. da C.; de Barros, B.C.S.C.; Serrano, S.M. de T.; Chaves, A.F.A. Decoding the Impact of Isolation Window Selection and QuantUMS Filtering in DIA-NN for DIA Quantification of Peptides and Proteins. *J. Proteome Res.* **2025**, doi:10.1021/acs.jproteome.5c00009.
96. Zhang, F.; Ge, W.; Huang, L.; Li, D.; Liu, L.; Dong, Z.; Xu, L.; Ding, X.; Zhang, C.; Sun, Y.; et al. A Comparative Analysis of Data Analysis Tools for Data-Independent Acquisition Mass Spectrometry. *Molecular and Cellular Proteomics* **2023**, *22*, doi:10.1016/j.mcpro.2023.100623.
97. Gotti, C.; Roux-Dalvai, F.; Joly-Beauparlant, C.; Mangnier, L.; Leclercq, M.; Droit, A. Extensive and Accurate Benchmarking of DIA Acquisition Methods and Software Tools

- Using a Complex Proteomic Standard. *J. Proteome Res.* **2021**, *20*, 4801–4814, doi:10.1021/acs.jproteome.1c00490.
98. Cox, J.; Mann, M. MaxQuant Enables High Peptide Identification Rates, Individualized p.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification. *Nat. Biotechnol.* **2008**, *26*, 1367–1372, doi:10.1038/nbt.1511.
 99. Uslan, T.; Quan, B.; Wang, T.Y.; Pang, M.; Qiu, Y.; Chou, T.F. In-Depth Comparison of Reagent-Based Digestion Methods and Two Commercially Available Kits for Bottom-Up Proteomics. *ACS Omega* **2025**, *10*, 10642–10652, doi:10.1021/acsomega.4c11585.
 100. Viswanatha, T.; Pallansch, M.J.; Liener, I.E. THE INHIBITION OF TRYPSIN. *Journal of Biological Chemistry* **1955**, *212*, 301–309, doi:10.1016/s0021-9258(18)71116-2.
 101. Masuda, T.; Tomita, M.; Ishihama, Y. Phase Transfer Surfactant-Aided Trypsin Digestion for Membrane Proteome Analysis. *J. Proteome Res.* **2008**, *7*, 731–740, doi:10.1021/pr700658q.
 102. Boys, E.L.; Liu, J.; Robinson, P.J.; Reddel, R.R. Clinical Applications of Mass Spectrometry-Based Proteomics in Cancer: Where Are We? *Proteomics* **2023**, *23*.
 103. Carambia, A.; Fenja, &; Schuran, A. The Aryl Hydrocarbon Receptor in Liver Inflammation., doi:10.1007/s00281-021-00867-8/Published.
 104. Tomankova, T.; Petrek, M.; Kriegova, E. Involvement of MicroRNAs in Physiological and Pathological Processes in the Lung. *Respir. Res.* **2010**, *11*.
 105. Bartoszewska, S.; Kochan, K.; Madanecki, P.; Piotrowski, A.; Ochocka, R.; Collawn, J.F.; Bartoszewski, R. Regulation of the Unfolded Protein Response by MicroRNAs. *Cell. Mol. Biol. Lett.* **2013**, *18*, 555–578.
 106. Shih, D.M.; Wang, Z.; Lee, R.; Meng, Y.; Che, N.; Charugundla, S.; Qi, H.; Wu, J.; Pan, C.; Mark Brown, J.; et al. Flavonoid Containing Monooxygenase 3 Exerts Broad Effects on Glucose and Lipid Metabolism and Atherosclerosis. *J. Lipid Res.* **2015**, *56*, 22–37, doi:10.1194/jlr.M051680.
 107. Bradbury, D.A.; Newton, R.; Zhu, Y.M.; El-Haroun, H.; Corbett, L.; Knox, A.J. Cyclooxygenase-2 Induction by Bradykinin in Human Pulmonary Artery Smooth Muscle Cells Is Mediated by the Cyclic AMP Response Element through a Novel Autocrine Loop Involving Endogenous Prostaglandin E₂, E-Prostanoid 2 (EP₂), and EP₄ Receptors. *Journal of Biological Chemistry* **2003**, *278*, 49954–49964, doi:10.1074/jbc.M307964200.

108. Ziegler, N.; Awwad, K.; Fisslthaler, B.; Reis, M.; Devraj, K.; Corada, M.; Minardi, S.P.; Dejana, E.; Plate, K.H.; Fleming, I.; et al. β -Catenin Is Required for Endothelial Cyp1b1 Regulation Influencing Metabolic Barrier Function. *Journal of Neuroscience* **2016**, *36*, 8921–8935, doi:10.1523/JNEUROSCI.0148-16.2016.
109. Fan, C.-Y.; Pan, J.; Chu, R.; Lee, D.; Kluckman, K.D.; Usuda, N.; Singh, I.; Yeldandi, A. V; Sambasiva Rao, M.; Maeda, N.; et al. *Hepatocellular and Hepatic Peroxisomal Alterations in Mice with a Disrupted Peroxisomal Fatty Acyl-Coenzyme A Oxidase Gene**; 1996;
110. Moon, Y.-A. The SCAP/SREBP Pathway: A Mediator of Hepatic Steatosis. *Endocrinology and Metabolism* **2017**, *32*, 6, doi:10.3803/EnM.2017.32.1.6.
111. De Nardo, W.; Lee, O.; Johari, Y.; Bayliss, J.; Pensa, M.; Miotto, P.M.; Keenan, S.N.; Ryan, A.; Rucinski, A.; Svinos, T.M.; et al. Integrated Liver-Secreted and Plasma Proteomics Identify a Predictive Model That Stratifies MASH. *Cell Rep. Med.* **2025**, *6*, doi:10.1016/j.xcrm.2025.102085.
112. Boel, F.; Akimov, V.; Teuchler, M.; Terkelsen, M.K.; Wernberg, C.W.; Larsen, F.T.; Hallenborg, P.; Lauridsen, M.M.; Krag, A.; Mandrup, S.; et al. Deep Proteome Profiling of Metabolic Dysfunction-Associated Steatotic Liver Disease. *Communications Medicine* **2025**, *5*, doi:10.1038/s43856-025-00780-3.
113. Endoplasmic Reticulum Stress and the Unfolded Protein.
114. MicroRNAs in Non-Alcoholic Fatty Liver Disease.
115. Chandrasekaran, P.; Weiskirchen, R. The Role of SCAP/SREBP as Central Regulators of Lipid Metabolism in Hepatic Steatosis. *Int. J. Mol. Sci.* **2024**, *25*.
116. Zheng, W.; Liu, M.; Lv, X.; He, C.; Yin, J.; Ma, J. AhR Governs Lipid Metabolism: The Role of Gut Microbiota. *Front. Microbiol.* **2025**, *16*.

10. KEY WORDS

DEF, DIA-NN, DIA, FFPE, HIAR, KNIME, mass spectrometry, needle biopsy, PARQUET, visualization, XICs, xylene-free,

11. ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor Prof. Éva Csősz for their unwavering guidance, encouragement, and support throughout the course of this research. Their expertise, patience, and willingness to provide assistance far beyond what could reasonably be expected have been invaluable to both my professional and personal development. This work would not have been possible without their mentorship.

I am also grateful to all my colleagues at the Proteomics Core Facility for their scientific support and collaborative spirit. Beyond their contributions to my research, their friendship has made my time in Hungary both productive and enjoyable.

I would like to extend my heartfelt thanks to Dr. Christine Carapito and Dr. Jeewan Babu Rijal at the University of Strasbourg (LSMBO), as well as Dr. Lilla Turiák at the HUN-REN Research Centre, for hosting my research visits. I am deeply appreciative of their scientific insights, technical expertise, generous support, and access to laboratory facilities, all of which significantly enriched this work and broadened my research experience.

I am grateful to the Tempus Public Foundation and DHET South Africa for providing me with the opportunity to pursue my studies in Hungary and for supporting my academic journey. Finally, and most importantly, I wish to thank my family and friends for their unconditional love, encouragement, and unwavering belief in me throughout the years. Their support has been a constant source of strength during both the challenges and achievements of this journey.

12. APPENDIX



**UNIVERSITY of
DEBRECEN**

**UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN**

H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

Registry number: DEENK/475/2026.PL
Subject: PhD Publication List

Candidate: Gontse Mabuse Moagi

Doctoral School: Doctoral School of Molecular Cellular and Immune Biology

List of publications related to the dissertation

1. **Moagi, G. M.**, Beke, L., Méhes, G., Kecskeméti, G., Szabó, Z., Turiák, L., Csósz, É.: Development of a Xylene-Free Sample Preparation Protocol for Quantitative Proteomics of Clinically Relevant Formaldehyde-Fixed Paraffin-Embedded Needle Biopsy Samples.
Proteomes. 14 (2), 1-20, 2026.
DOI: <http://dx.doi.org/10.3390/proteomes14020030>
IF: 4.3 (2025)
2. **Moagi, G. M.**, Ferreira, T. F., Kristóf, E., Arda, A. G., Arianti, R., Horvatovich, P., Csósz, É.: DIA-NN EasyFilter Workflow for the Fast and User-Friendly Critical Assessment and Visualization of DIA-NN Proteomics Analysis Outcome.
J. Proteome Res. 25 (6), 2927-2938, 2026.
DOI: <http://dx.doi.org/10.1021/acs.jproteome.5c01278>
IF: 3.8 (2025)

Total IF of journals (all publications): 8,1

Total IF of journals (publications related to the dissertation): 8,1

The Candidate's publication data submitted to the Tudóstér have been validated by DEENK on the basis of the Journal Citation Report (Impact Factor) database.

06 July, 2026

