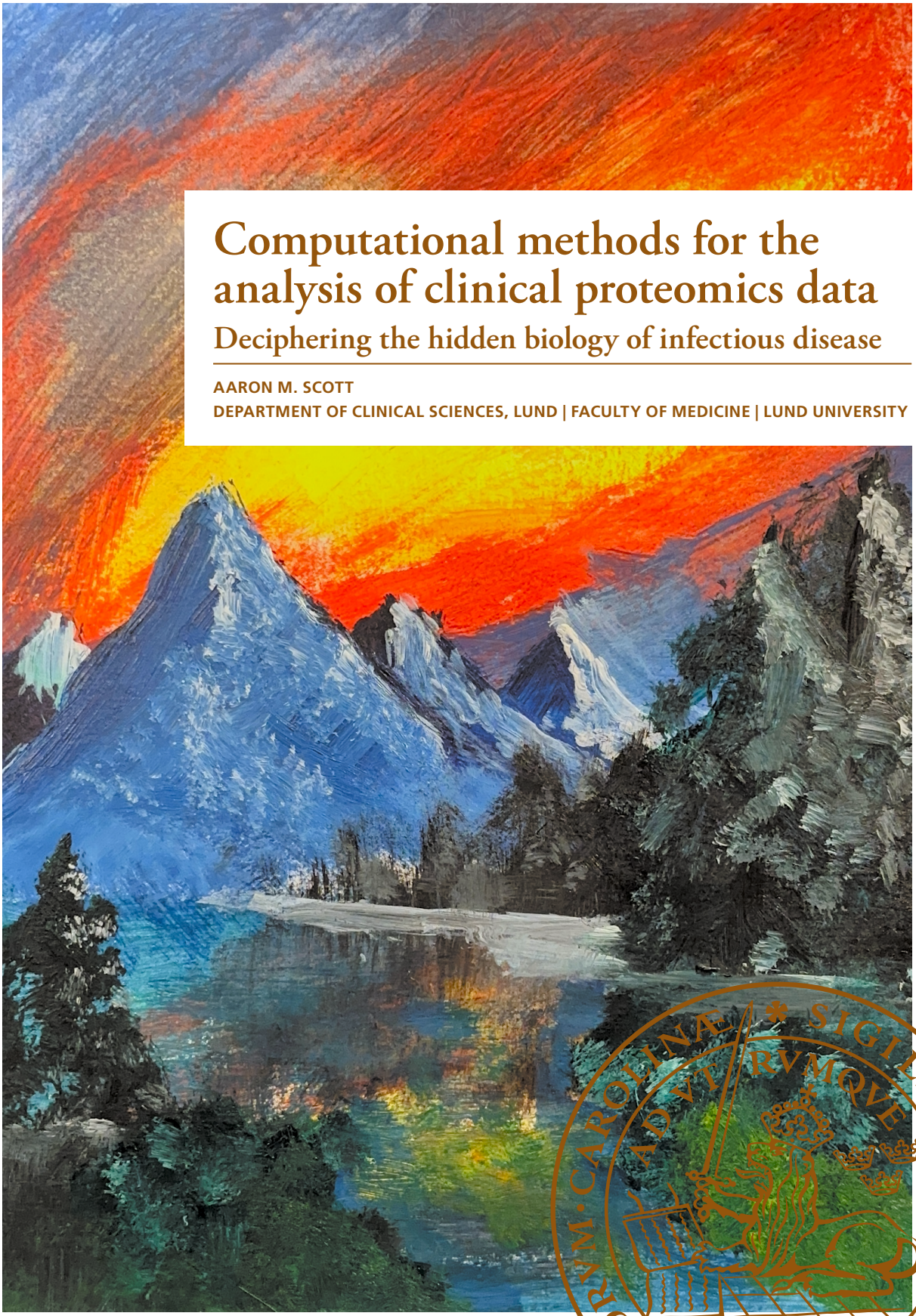
LUND UNIVERSITY

Computational methods for the analysis of clinical proteomics data - Deciphering the hidden biology of infectious disease

Scott, Aaron

2025

Document Version:
Publisher's PDF, also known as Version of record



Computational methods for the analysis of clinical proteomics data

Deciphering the hidden biology of infectious disease

AARON M. SCOTT
DEPARTMENT OF CLINICAL SCIENCES, LUND | FACULTY OF MEDICINE | LUND UNIVERSITY

Printed by Media-Tryck, Lund 2025 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

FACULTY OF MEDICINE

Department of Clinical Sciences, Lund

Lund University, Faculty of Medicine
Doctoral Dissertation Series 2025:3
ISBN 978-91-8021-656-2
ISSN 1652-8220

Computational methods for the analysis of clinical proteomics data

Deciphering the hidden biology of infectious disease

Aaron M. Scott

LUND
UNIVERSITY

DOCTORAL DISSERTATION

Doctoral dissertation for the degree of Doctor of Philosophy (PhD) at the Faculty of Medicine at Lund University to be publicly defended on the 10th of January at 09.00 in Belfragesalen, Biomedical Center, Lund, Sweden

Faculty opponent
Professor Lukas Käll
Science for Life Laboratory, KTH – Royal Institute of Technology
Stockholm, Sweden

Organization: LUND UNIVERSITY

Document name: Doctoral dissertation **Date of issue:** 2025-01-10

Author(s): Aaron M. Scott **Sponsoring organization:**

Title and subtitle: Computation methods for the analysis of clinical proteomics data: Deciphering the hidden biology of infectious disease

Abstract:

Infectious diseases are one of the leading causes of mortality in the world. Severe infections can manifest in many ways, creating a heterogeneous clinical and molecular disease landscape that renders these diseases difficult to research, diagnose, and treat. To investigate the molecular mechanisms of infectious disease, we apply mass spectrometry-based proteomics to analyze blood plasma samples for the dynamic stratification of infectious disease and sepsis patients. In this thesis, we focus on the development of computational methods that facilitate the interrogation of these complex proteomes towards the goal of translational medicine and personalized care.

The overall goal of this thesis was to enable the in-depth analysis of large-scale clinical proteomic cohorts. As a first step, we leveraged computational methods to facilitate discovery data-independent acquisition (DIA) mass spectrometry (MS) and maximize the number of identified proteins in plasma samples. Using large-scale machine learning methods, we optimize the search space using a multi-pass prediction-based filtration step that allows for robust control of the false discovery rate (FDR) while optimizing the number of quantified proteins. From here, we introduce explainable machine learning methods to select the most important proteins involved in predicting severe disease. We substantially expand these explainable machine learning methods, formalizing them into easy-to-use software packages that support reproducible research and in-depth proteomic analysis. Finally, we combine our novel computational methods to analyze 1400 clinical plasma samples from patients suspected of sepsis. Using samples taken at the time-of-admission to the hospital, we developed an inherently interpretable architecture to match new patients to similar groups of existing patients from a database to create digital families. These digital families could accurately stratify patients suspected of sepsis, predict disease trajectories, predict mortality, and identify hidden cohorts within the data.

In combination, the results contained within this thesis provide a strong basis for further studies and movement towards personalized health care for infectious diseases.

Key words: Proteomics, infectious disease, sepsis, machine learning, bioinformatics, software engineering, mass spectrometry

Language: English **Number of pages**: 85

ISSN and key title: 1652-8220

Lund University, Faculty of Medicine Doctoral Dissertation Series 2025:3

ISBN: 978-91-8021-656-2

Signature Date 2024-12-02

Computational methods for the analysis of clinical proteomics data

Deciphering the hidden biology of infectious disease

Aaron M. Scott

LUND
UNIVERSITY

Coverphoto by Aaron Scott

Faculty of Medicine
Department of Clinical Sciences, Lund

Lund University, Faculty of Medicine Doctoral Dissertation Series 2025:3
ISBN 978-91-8021-656-2
ISSN 1652-8220

Printed in Sweden by Media-Tryck, Lund University
Lund 2025

To Jossan

“There is nothing like looking,
if you want to find something.
You certainly usually find something,
if you look, but it is not always quite the
something you were after.”

- *J.R.R Tolkien, The Hobbit*

Table of Contents

Abbreviations

MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
LC	Liquid chromatography
DIA	Data independent acquisition
DDA	Data dependent acquisition
ESI	Electron spray ionization
MS1	Mass spectrum of intact analyte
MS2	Mass spectrum of fragmented analyte
RT	Retention time
IM	Ion mobility
LFQ	Label-free quantification
ML	Machine learning
XML	Explainable machine learning (Explainable artificial intelligence)
AI	Artificial intelligence
m/z	Mass-to-charge ratio
FDR	False discovery rate
TIMS	Trapped ion mobility
CCS	Collisional cross section
PASEF	Parallel acquisition serial fragmentation
CID	Collision induced dissociation
HCD	High-energy collisional dissociation
PSM	Peptide spectrum match
[M]	Monoisotopic peak
[M+1,2]	Isotopic peaks
MBR	Match between runs
LOWESS	Locally weighted scatterplot smoothing
MAR	Missing at random
MNAR	Missing not at random

KNN	K-nearest neighbor
LOD	Limit of detection
ANOVA	Analysis of variance
GSEA	Gene set enrichment analysis
Low-n	Low sample number
CV	Cross validation
TOA	Time-of-admission
SOFA	Sequential organ failure assessment
SPD	Samples per day
BINN	Biologically informed neural network
DPKS	Data processing kitchen sink
AKI	Acute kidney injury
COVID	Coronavirus disease
GPS	Generalized precursor scoring
API	Application programming interface
ILS	Interpretable latent space
CNS	Central nervous system
FAIR	Findability, accessibility, interoperability, and reusability

Included Publications

Paper I

"Generalized precursor prediction boosts identification rates and accuracy in mass spectrometry-based proteomics"

Aaron M. Scott, Christofer Karlsson, Tirthankar Mohanty, Erik Hartman, Suvi T. Vaara, Adam Linder, Johan Malmström & Lars Malmström

Communications Biology

Paper II

"Interpreting biologically informed neural networks for enhanced proteomic biomarker discovery and pathway analysis"

Erik Hartman*, **Aaron M. Scott***, Christofer Karlsson, Tirthankar Mohanty, Suvi T. Vaara, Adam Linder, Lars Malmström & Johan Malmström

*Authors contributed equally

Nature Communications

Paper III

"Explainable machine learning for the identification of proteome states via the data processing kitchen sink"

Aaron M. Scott, Erik Hartman, Johan Malmström, and Lars Malmström

bioRxiv

Paper IV

"Population scale proteomics enables adaptive digital twin modelling in sepsis"

Aaron M. Scott*, Lisa Mellhammar*, Erik Malmström*, Axel Goch Gustafsson, Anahita Bakochi, Marc Isaksson, Tirthankar Mohanty, Louise Thelaus, Fredrik Kahn, Lars Malmström1, Johan Malmström**, Adam Linder**

*Authors contributed equally

medRxiv

Excluded Publications

Paper

"Peptide clustering enhances large-scale analyses and reveals proteolytic signatures in mass spectrometry data"

Erik Hartman, Fredrik Forsberg, Sven Kjellström, Jitka Petrlova, Congyu Luo, **Aaron M. Scott**, Manoj Puthia, Johan Malmström, Artur Schmidtchen

Nature Communications

Paper

"Inferring the composition of the blood plasma proteome by a human proteome distribution atlas"

Erik Malmström, Simon Hauri, Tirthankar Mohanty, **Aaron M. Scott**, Christofer Karlsson, Carlos Gueto-Tettay, Emma Åhrman, Shahab Nozohoor, Bobby Tingstedt, Sara Regner, Peter Elfving, Leif Bjermer, Andreas Forsvall, Alexander Doyle, Mattias Magnusson, Ingrid Hedenfalk, Päivi Kannisto, Christian Brandt, Emma Nilsson, Lars B Dahlin, Johan Malm, Adam Linder, Lars Malmström, Emma Nimeus, Johan Malmström

bioRxiv (in review)

Abstract

Infectious diseases are one of the leading causes of mortality in the world. Severe infections can manifest in many ways, creating a heterogeneous clinical and molecular disease landscape that renders these diseases difficult to research, diagnose, and treat. To investigate the molecular mechanisms of infectious disease, we apply mass spectrometry-based proteomics to analyze blood plasma samples for the dynamic stratification of infectious disease and sepsis patients. In this thesis, we focus on the development of computational methods that facilitate the interrogation of these complex proteomes towards the goal of translational medicine and personalized care.

The overall goal of this thesis was to enable the in-depth analysis of large-scale clinical proteomic cohorts. As a first step, we leveraged computational methods to facilitate discovery data-independent acquisition (DIA) mass spectrometry (MS) and maximize the number of identified proteins in plasma samples. Using large-scale machine learning methods, we optimize the search space using a multi-pass prediction-based filtration step that allows for robust control of the false discovery rate (FDR) while optimizing the number of quantified proteins. From here, we introduce explainable machine learning methods to select the most important proteins involved in predicting severe disease. We substantially expand these explainable machine learning methods, formalizing them into easy-to-use software packages that support reproducible research and in-depth proteomic analysis. Finally, we combine our novel computational methods to analyze 1400 clinical plasma samples from patients suspected of sepsis. Using samples taken at the time-of-admission to the hospital, we developed an inherently interpretable architecture to match new patients to similar groups of existing patients from a database to create digital families. These digital families could accurately stratify patients suspected of sepsis, predict disease trajectories, predict mortality, and identify hidden cohorts within the data.

In combination, the results contained within this thesis provide a strong basis for further studies and movement towards personalized health care for infectious diseases.

Popular Science Summary

The human body is a complex hierarchical machine. At the smallest level, genes provide the blueprint for the body to build tiny mechanisms, such as proteins, that drive the bodily functions and physical characteristics of a human. These little proteins are essential to biological processes in everyday life, from regulating metabolism, providing energy to cells, allowing physical movement, and establishing the defense mechanisms against disease. Just like in a machine where a faulty cog can break the entire system, small perturbations in the levels of these proteins can drive the development of disease and death in humans. For this reason, it is important to study the types and quantities of proteins under different biological conditions to understand disease. If we can associate certain proteins with certain types of diseases, it is possible to provide personalized treatment by administering medication that target those specific dysregulated proteins. Just like a mechanic can identify which part of a machine is broken and provide a specific fix, if researchers can identify the specific proteins that are faulty then clinicians can provide the precise treatment that can fix the patient. However, this is only possible if we can identify and quantify proteins in a human during disease in a robust and reproducible manner.

Infectious diseases are one of the leading causes of death in the world, affecting millions of patients and putting substantial strain on the health care system. In this thesis, we attempt to identify and quantify proteins from the blood of patients with infectious diseases and sepsis using a machine called the mass spectrometer. The mass spectrometer can provide extremely precise measurements for theoretically all proteins in a sample, being able to distinguish molecules that differ by only a millionth of a mass unit. Using these mass measurements, we can identify the proteins in a sample and provide quantitative measurements based on the intensity of those molecules. From here we can investigate the composition of the proteins of a patient that is affected by a particular disease. However, the measurements obtained by a mass spectrometer are not simple to analyze, and advanced data-analysis techniques need to be employed to get to some sort of biological conclusion from the raw mass measurements.

To study infectious diseases in a practical manner, biological material to analyze the proteins of infected individuals should be collected from some sort of minimally invasive media. In our case, we analyze proteins from the liquid component of blood, or the blood plasma, to infer health and disease. Plasma can contain signaling

proteins, proteins leaking from damaged organs, functional plasma proteins, and many other protein types, making it a perfect media to analyze to get a snapshot picture of the molecular health of the human body. However, due to limitations of the mass spectrometer, and the range of protein concentrations in plasma, it is difficult to analyze enough proteins in a sample to provide an accurate picture of the health of that individual.

This thesis aims to address the computational issues associated with maximizing the number of proteins that can be quantified from blood plasma and making biological sense of those complex mass spectrometry-based results. We apply novel computational methods, including advanced machine learning models and algorithms, to investigate and elucidate the proteins in blood that are associated with specific manifestations of infectious diseases. With a focus on explainability and interpretability, we were able to develop novel methods for the stratification of patients suspected of sepsis on admission to the hospital using only a few clinical parameters and proteins quantified from the blood. Our unique approach leverages explainable machine learning to match new patients with groups of patients from a database to create digital families. These digital families can be used to model outcomes, such as mortality and the development of sepsis, and could potentially be used to drive the treatment of patients in the clinic. Overall, we hope that the contents of this thesis provide evidence that the analysis of proteins from the blood of patients could be used to more effectively treat infectious disease and sepsis, improving the lives of the millions affected by these severe diseases.

Populärvetenskaplig Sammanfattning

Människokroppen är en komplex hierarkisk maskin. På den lägsta nivån tillhandahåller gener ritningar för kroppen att bygga små mekanismer, såsom proteiner, som driver kroppens funktioner och fysiska egenskaper. Dessa små proteiner är avgörande för biologiska processer i vardagen, från att reglera ämnesomsättningen, ge energi till celler, möjliggöra fysisk rörelse och etablera försvarsmekanismer mot sjukdomar. Precis som i en maskin där ett defekt kugghjul kan förstöra hela systemet, kan små störningar i nivåerna av dessa proteiner driva utvecklingen av sjukdomar och död hos människor. Av denna anledning är det viktigt att studera typer och mängder av proteiner under olika biologiska förhållanden för att förstå sjukdom. Om vi kan koppla vissa proteiner till specifika typer av sjukdomar är det möjligt att erbjuda personanpassad behandling genom att ge läkemedel som riktar in sig på just de proteiner som är i obalans. Precis som en mekaniker kan identifiera vilken del av en maskin som är trasig och erbjuda en specifik lösning, kan kliniker ge precis behandling om forskare identifierar de specifika proteiner som är felaktiga. Detta är dock bara möjligt om vi kan identifiera och kvantifiera proteiner hos en sjuk människa på ett robust och reproducerbart sätt.

Infektionssjukdomar är en av de främsta dödsorsakerna i världen och påverkar miljontals patienter samtidigt som de innebär en stor belastning för hälso- och sjukvårdssystemet. I denna avhandling försöker vi identifiera och kvantifiera proteiner från blodet hos patienter med infektionssjukdomar och sepsis med hjälp av en maskin som kallas masspektrometer. Masspektrometern kan ge extremt precisa mätningar för teoretiskt sett alla proteiner i ett prov och har förmågan att särskilja molekyler som endast skiljer sig med en miljondel av en massenhet. Med hjälp av dessa massmätningar kan vi identifiera proteiner i ett prov och ge kvantitativa mätningar baserade på intensiteten hos dessa molekyler. Därifrån kan vi undersöka sammansättningen av proteiner hos en patient som påverkas av en specifik sjukdom. Mätningarna från en masspektrometer är dock inte enkla att analysera, och avancerade dataanalystekniker behöver tillämpas för att nå någon form av biologisk slutsats från de råa massmätningarna.

För att studera infektionssjukdomar på ett praktiskt sätt, bör biologiskt material för att analysera proteiner hos infekterade individer samlas in från ett minimalt invasivt medium. I vårt fall analyserar vi proteiner från den flytande komponenten i blodet, eller blodplasman, för att dra slutsatser om hälsa och sjukdom. Plasma kan innehålla signaleringsproteiner, proteiner som läcker från skadade organ, funktionella

plasmaproteiner och många andra proteintyper, vilket gör det till ett perfekt medium att analysera för att få en ögonblicksbild av människokroppens molekylära hälsa. På grund av masspektrometerns begränsningar och det breda koncentrationsspannet av proteiner i plasma är det dock svårt att analysera tillräckligt många proteiner i ett prov för att ge en korrekt bild av individens hälsa.

Denna avhandling syftar till adressera beräkningsmässiga utmaningar som är kopplade till att maximera antalet proteiner som kan kvantifieras från blodplasma och ge biologisk förståelse för dessa komplexa resultat från masspektrometri. Vi tillämpar nya beräkningsmetoder, inklusive avancerade maskininlärningsmodeller och algoritmer, för att undersöka och klargöra proteiner i blodet som är associerade med specifika manifestationer av infektionssjukdomar. Med fokus på förklarbarhet och tolkbarhet utvecklade vi nya metoder för att stratifiera patienter som misstänks ha sepsis vid ankomsten till sjukhuset, med endast några få kliniska parametrar och proteiner kvantifierade från blodet. Vår unika metod använder förklarbar maskininlärning för att matcha nya patienter med grupper av patienter från en databas och skapa digitala familjer. Dessa digitala familjer kan användas för att modellera utfall, såsom dödlighet och utveckling av sepsis, och kan potentiellt användas för att vägleda behandling av patienter i kliniken. Sammantaget hoppas vi att innehållet i denna avhandling ger bevis för att analys av proteiner från patienters blod kan användas för att mer effektivt behandla infektionssjukdomar och sepsis, och därigenom förbättra livet för de miljontals som drabbas av dessa allvarliga sjukdomar.

Introduction

Proteomics

If genes provide the blueprint of a cell, the proteins provide the machinery. Proteins are responsible for many of the physiological actions of an organism. Among many other functions, proteins can regulate metabolism, respond to external stimuli, defend against external pathogens, provide structure, and replicate genes. Based on the central dogma of molecular biology, genes are transcribed and translated to proteins, which can be further modified post-translationally (**Figure 1**). Therefore, studying the proteins of a particular biological system allows us to investigate closer to the level of physical expression rather than genetic blueprint. The large-scale study of proteins in a biological system, or proteomics[1], can help us understand the molecular mechanisms involved in a physiological response. In the past, proteomics may have been considered the little brother of genomics, but in the last 10 years, advances, including drafts of the human proteome[2,3], have allowed researchers to cement proteomics as a mature and powerful platform for the analysis of biological systems.

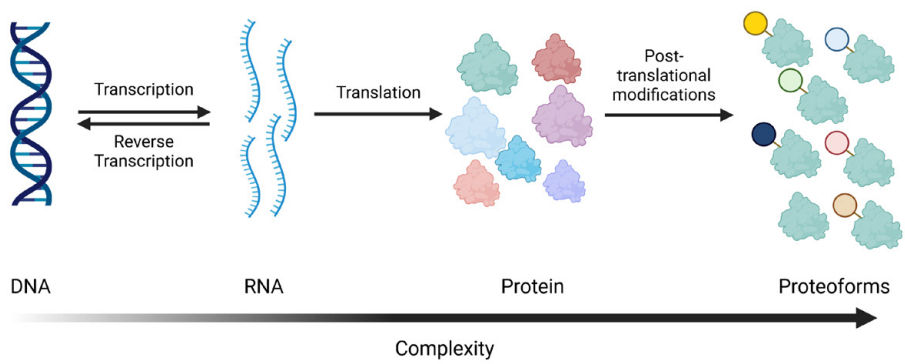


Figure 1 The Central Dogma of Molecular Biology
A schematic representing the flow of genes, encoded in DNA, to RNA, proteins, and modified proteins, or proteoforms. The diversity of the proteome and proteoforms can arise from post-translationally modified proteins, alternative splicing, or sequence variants. The complexity and number of uniquely expressed entities increases from left to right. (Created using Biorender)

In order to effectively study the proteome, a robust method to identify and quantify proteins in a sample is needed. In the past, the potential of proteomics was hindered by low throughput and shallow technologies that could not adequately analyze the proteome at a global scale. To mitigate the issues of throughput and analytical depth, mass spectrometry coupled with liquid chromatography (LC-MS/MS) provided an efficient mechanism for the comprehensive analysis of the proteome [4,5]. Mass spectrometry-based proteomics has been successfully applied to identify biomarkers and subtypes in disease[6–17], elucidate potential drug targets [18–22], and quantify and characterize immune response [23–26]. Among many other applications, it can also be used to analyze protein structure[27–29], including protein-protein interactions and binding sites[30,31], and network biology [32]. Recently, LC-MS/MS has even been applied to analyze systems at the single-cell level, reaching unprecedented depth for the technology[33–40]. Although there are different platforms available to identify and quantify proteins in a sample, mass spectrometry remains one of the most common, and is the method focused on in this thesis.

Mass Spectrometry Proteomics

Although it is possible to analyze an intact protein using mass spectrometry, termed top-down proteomics[41], the analysis of peptides with a mass spectrometer is technically and computationally more efficient[42]. In a process known as bottom-up proteomics, proteins are first denatured and then digested into peptides using an enzyme such as trypsin[43], and the resulting peptides are identified and quantified (**Figure 2**). In a process known as label-free quantification (LFQ), these annotated and quantified peptides are then mapped back to their parent protein to estimate the quantity of that protein in a sample. However, the digestion of proteins into peptides creates a complex sample with hundreds of thousands of analytes, resulting in highly multiplexed and convoluted signals from MS analysis. To mitigate this, in a standard bottom-up proteomics workflow, peptides are first separated using reversed-phase liquid chromatography (LC). In this step, peptides are separated based on their hydrophobicity as a gradient of buffers is passed through the LC column. Peptides are bound to the inside of the column and elute at different times depending on the composition of the buffer. The amount of time that it takes a peptide to elute from the column is recorded as the retention time (RT) and used to deconvolute and annotate peptides correctly[44] (**Figure 2**).

To enter the mass spectrometer, peptides are elevated to the gas-phase as charged ions using electrospray ionization (ESI), which allows them to travel from the liquid-phase of the LC to the vacuum of the MS[45]. These charged ions, known as precursors, are the entities that are analyzed by the mass spectrometer. After elevation to the gas phase, depending on the type of MS instrument that is being used, precursors can be further separated based on their ion mobility[46–49]. In the case of trapped ion mobility (TIMS), precursor ions are accumulated and separated based

on their collisional cross section (CCS) and then released serially into the mass spectrometer using a ramped electric field in a process known as parallel accumulation serial fragmentation (PASEF)[46,47].

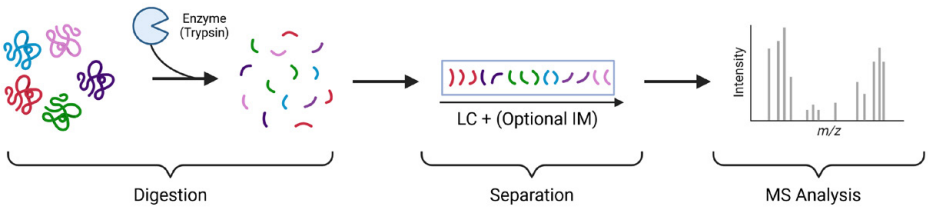


Figure 2 Simplified Schematic of a typical LC-MS workflow
This schematic depicts a classic bottom-up proteomics workflow, where proteins are first digested into peptides, separated using liquid chromatography and sometimes ion mobility (IM) spectrometry, and then finally analyzed using mass spectrometry. (Created using Biorender)

In a general sense, the mass-to-charge ratio (m/z) is recorded for each precursor ion at an elution point and intensities for each precursor are recorded in an MS1 spectra by the mass analyzer. Depending on the acquisition method used, precursors are then selected for fragmentation in the collision cell. To fragment the precursors, collision with an inert gas is used to induce dissociation (CID) of the peptide into fragments[50] in the collision cell. A common subtype of CID is high-energy collisional dissociation (HCD)[51], which is used throughout the methods in this thesis. The m/z values and intensities of the resulting fragments are recorded in an MS2 spectrum by another mass analyzer known as the detector. A simplified depiction of this workflow is visualized in **Figure 3**. As different precursors can have very similar m/z values and retention times, MS2 spectra enable precursors to be annotated based on the matching of multiple fragment ions to the theoretical spectra of a peptide rather than a single matching at the MS1 level. Although MS1 spectra give a comprehensive overview of all analytes available in a sample, MS2 spectra can increase the accuracy and precision of annotating a precursor with the correct peptide sequence. Depending on the type of experiment, or the biological question at hand, different data acquisition methods on the mass spectrometer acquire MS2 or MS/MS spectra in different ways.

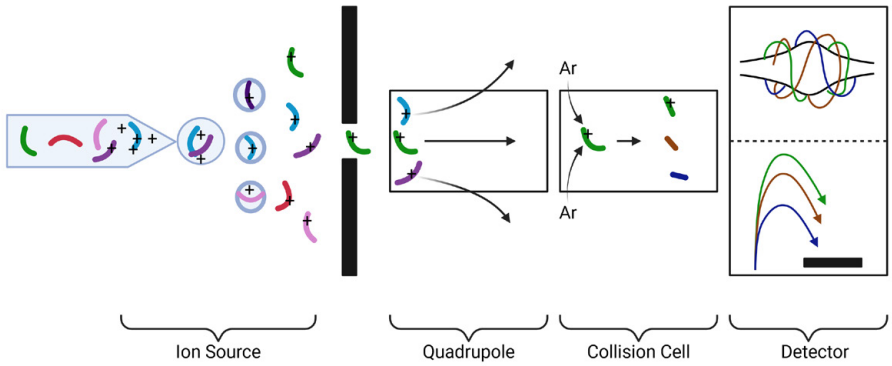


Figure 3 A cartoon representation of a generalized mass spectrometer
This schematic represents a simplified generalization of a mass spectrometer, including the ion source, depicted as using electrospray ionization (ESI), a quadrupole mass analyzer for precursor selection, a generalized collision cell for fragmentation, and a final mass analyzer, or detector. (Created using Biorender)

Data Dependent Acquisition

Data Dependent Acquisition (DDA) is a precursor isolation method commonly used in bottom-up proteomics where the most intense precursors at a particular elution event are selected for fragmentation. The advantage of this method is that the fragment ions can be directly linked to the precursor ion that they originate from, making the process of identifying and quantifying a particular precursor relatively straightforward. However, the selection of precursors for fragmentation can be stochastic, as only a few of the most intense precursors from an MS1 spectra are selected for fragmentation[5]. This leads to missing values for precursors that are present in a sample but were not selected for fragmentation. Instead of selecting the same most intense precursors for fragmentation after every MS1 scan, modern mass spectrometers can temporarily exclude precursors that were recently fragmented using dynamic exclusion lists to expand the number of precursors that are selected for fragmentation[52].

Data Independent Acquisition

Data Independent Acquisition (DIA) is an alternative precursor isolation method that attempts to mitigate the missing values and stochasticity of DDA. Instead of selecting individual precursors for fragmentation, DIA selects precursor m/z windows in a serialized manner to theoretically provide fragment ion series for all precursors in an isolation window[53]. Although this method theoretically provides a complete map of the proteome in a sample, the direct link of fragment ion series to precursor ion is lost, making data interpretation and analysis more complex. Apart from when directly specified, DIA is the method used in all studies covered in this

thesis. As large populations and individual diseases can be extremely heterogenous, the complexity in the data needs to be carefully handled.

Mass Spectrometry Data Analysis

The interpretation and analysis of mass spectrometry-based proteomics data can be complex. As multiple analytes elute from the LC column and are analyzed simultaneously by the mass spectrometer, the raw signal is highly multiplexed and convoluted. These complex signals need to be annotated and quantified before they can be used for downstream analysis to answer actual biological questions. In this section we will outline the main analytical steps for the analysis of DDA and DIA data and how they are applied for the quantification of the proteome.

DDA Analysis

Identification

In DDA, a singular MS2 spectra should be theoretically linked to a singular precursor ion. Exploiting this known link, the MS2 spectra are searched against a database of potential theoretical protein sequences that are believed to be contained within the sample being analyzed. These proteins are computationally digested into peptides and MS2 spectra are matched with the theoretical peptides within the specified mass range surrounding the precursor mass. This type of mass tolerance threshold can filter out many peptides in a database and only include those that are the closest matches. Theoretical fragment ion spectra of the potential peptides from the database are compared to the MS2 spectra and potential matches are ranked using a similarity function. The top matches are returned for further processing and are known as peptide spectrum matches (PSMs)[54–60]. Different software calculate different types of similarity scores, and it has been shown empirically that combining multiple search engines can improve the number of confident PSMs in a database search[61–63]. Once a list of highest scoring hits is aggregated for each MS2 spectra, it is important to employ measures to ensure that the PSMs that are passed through to downstream analysis are true positives and not false positives.

Validation

One of the most critical steps in the analysis of mass spectrometry-based proteomics data is the validation of the extracted precursors to ensure that they are correctly annotated. Without this step, there is no confidence in the downstream statistical analysis and biological interpretation will suffer if the false discovery rate (FDR) is not controlled correctly. The most prominent and accepted method in the field to provide confident control of the FDR, is the target-decoy approach[64]. In the target-

decoy approach, decoy protein sequences are generated by reversing or shuffling the proteins contained in the database being searched. These reversed protein sequences are appended to the database and compete with the targets to be matched with MS2 spectra. Any MS2 spectra that are annotated decoy sequences can be assumed to be false PSMs, since the protein sequence of origin is randomized and not contained in the sample. Using the scores assigned from the database search engines, distributions for decoy PSMs and target PSMs can be modelled, and confidence can be assigned to each PSM. A particular score cutoff, the FDR, or q-value, at that score can be calculated as follows:

$$FDR = \frac{Decoy\ PSMs}{Target\ PSMs}$$

The above equation gives the ratio of decoy PSMs that are greater than the given threshold to the number of target PSMs that are greater than a given threshold. To avoid situations where q-values are 0, the above equation can be expressed as:

$$FDR = \frac{Decoy\ PSMs + 1}{Target\ PSMs}$$

An alternative approach is to measure the FDR as a percentage:

$$FDR = \frac{Decoy\ PSMs}{(Decoy\ PSMs + Target\ PSMs)}$$

This equation gives the percentage of PSMs that are false compared to all PSMs past a given score threshold. In this case, each q-value represents the expected percent of false positives at a given score threshold if the feature is considered significant[65]. If all PSMs are filtered for a specific q-value, then that is the expected FDR of the resulting list of PSMs.

Although the raw scores from the database search engines can be used to calculate q-values directly, it has been shown that the inclusion of multiple features, scores, and [66–75]properties predicted using machine learning can boost the number of true target PSMs that pass a given FDR threshold. However, since many of the PSMs may still be false, semi-supervised methods machine learning algorithms, such as Percolator[66], have been developed to iteratively remove false targets from the training data, resulting in classifiers that can more accurately differentiate between true targets and decoys and provide confident FDR control. Another advantage of using machine learning algorithms to control the FDR is that multiple features can be considered as input and combined into a consensus PSM score.

Quantification

In typical experiments, annotated spectra that pass FDR control are quantified using either a targeted or untargeted approach. The targeted approach extracts intensity values for a peptide precursor along the retention time axis, or chromatograms, from consecutive MS1 spectra to assemble peaks related directly to an identified peptide at a specified retention time[76,77]. Additionally, as peptides contain heavier carbon isotopes, there are additional isotopic peaks associated with each peptide signal in a sample. The monoisotopic peak ([M]) is associated with the lowest mass of the precursor, and additional isotopic peaks ([M+1], [M+2], etc.) arise from different isotopic compositions at predicted m/z intervals to form an isotopic envelope. Typically, targeted quantification approaches extract chromatograms for the [M], [M+1], and [M+2] isotopes for each identification. If IM was used to further separate the sample, mobilograms, or intensity values along the IM axis, are also extracted for each isotope[76]. The area under each of these chromatograms is then integrated and summed to give an estimate of the abundance for a precursor in a sample.

Similar to the targeted approach, the untargeted approach extracts chromatograms, also called hills, from consecutive scans by linking peaks that are within a narrow m/z range. New hills are started when a peak is encountered that cannot be linked to any of the existing hills being traced. These extracted hills are then clustered into isotopic envelopes based on estimated isotopic patterns and resolved by traversing a graph data-structure. The charge state is assigned based on the measured interval between linked isotopic peaks. Once the hills have been clustered, these MS1 features are quantified by integrating the area under each hill and summing[78–81]. The downside of the untargeted approach is that the assembled features need to be matched to confident peptide identifications from the database search step instead of directly quantifying identified peptides in a targeted manner. This approach is also more sensitive to the hyperparameters set during hill extraction, particularly the m/z error threshold allowed to link peaks in consecutive scans together. Alternatively, the untargeted approach makes it possible to theoretically quantify every peptide in a sample within the dynamic range of the MS. Even if most of these peptide features may remain unannotated, the untargeted approach results in a comprehensive map of all peptide-like MS1 features in a sample. The most common steps in the analysis of DDA data are summarized in **Figure 4**.

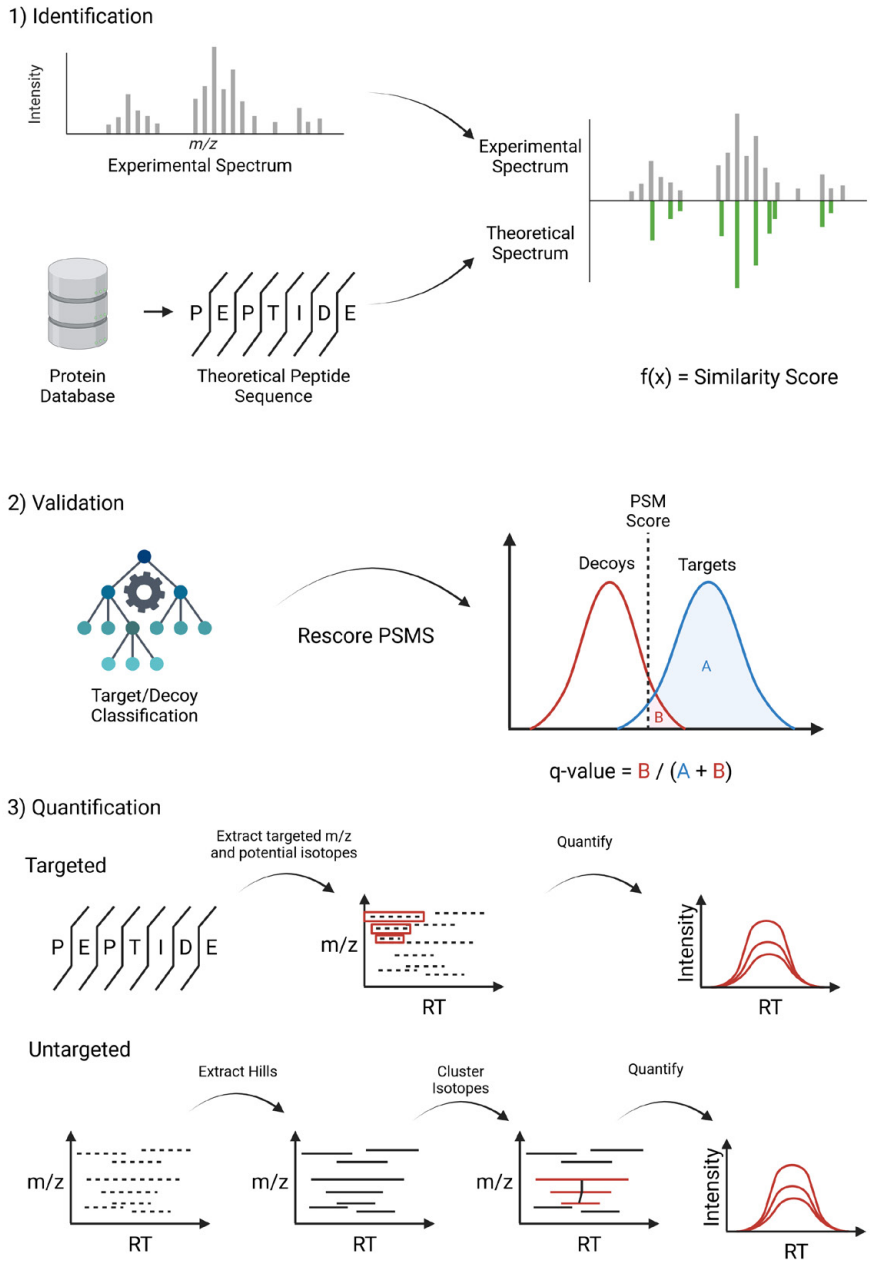


Figure 4 The 3 main steps for analyzing DDA data

This figure depicts the main steps involved in identifying and quantifying peptides from DDA data. In step 1, spectra are first searched against a protein database to identify peptides that may match the spectrum (PSMs). In step 2, those PSMs are scored using the target-decoy approach and the false

discovery rate is controlled. In step 3, these identified peptides are quantified using either a targeted or untargeted approach to match quantitative values to identification. (Created using Biorender)

Match-between-runs

It is possible to minimize the issue of missing values in DDA data through the application of data-driven alignment algorithms that attempt to match unidentified precursors to confident identifications from other runs in an experiment. These alignment, or match-between-runs (MBR), algorithms aggregate annotated peptides across runs, correct retention time between the runs, and then assign MS1 features to the global list of peptide identifications to fill in missing values[76,82–86].

Chimeric Spectra Correction

Occasionally, when a precursor is selected for fragmentation in DDA-MS, multiple other precursors can be selected in the same narrow isolation window. In these cases, it is possible to utilize untargeted MS1 feature extraction methods to identify what additional precursors could potentially fall in the DDA isolation window for the selected precursor. Different algorithms handle this scenario differently, but the main concept is that if multiple MS1 features can potentially be associated with an MS2 spectra, then that spectra will be searched multiple times with each MS1 feature considered the parent of the MS2 spectra and providing the quantitative information at the MS1 level[55,78,81].

DIA Analysis

Spectral Libraries

As all precursors in DIA are fragmented and measured for a particular isolation window, the resulting MS2 spectra are complex and consist of all fragments from multiple precursors. To deconvolute this highly multiplexed signal, it is possible to use a library of previously identified fragment ions in the form of a spectral library that are used to guide targeted signal extraction from the sample being analyzed. Spectral libraries typically contain information about the retention time, m/z, and IM of a precursor, as well as m/z values and intensities for fragment ions. Spectral libraries are typically created by first analyzing a group of samples from the experiment using DDA. Confident PSMs passing FDR control from the DDA samples are aggregated and fragment ions for these PSMs are extracted to build an experiment specific spectral library[87] (**Figure 5**). To increase the number of identifications included in the spectral library, additional offline chromatographic fractionation steps can be used to split each sample into multiple parts that are analyzed separately with the resulting data being combined downstream. The increased MS time per fraction of the sample allows for a deeper analysis of the proteome. Although effective, these methods can be extremely time consuming and expensive, as this library creation process may need to be repeated for each new

experiment. As these libraries are based on previously identified precursors for a particular sample type, another downside of experiment specific library guided DIA analysis is that only previously found compounds in a sample can be identified. This can bottleneck discovery analysis of large cohorts, as the number of analytes is restricted to a set list before downstream analysis can identify what is important. As a result, proteins important to a biological system may not be quantified.

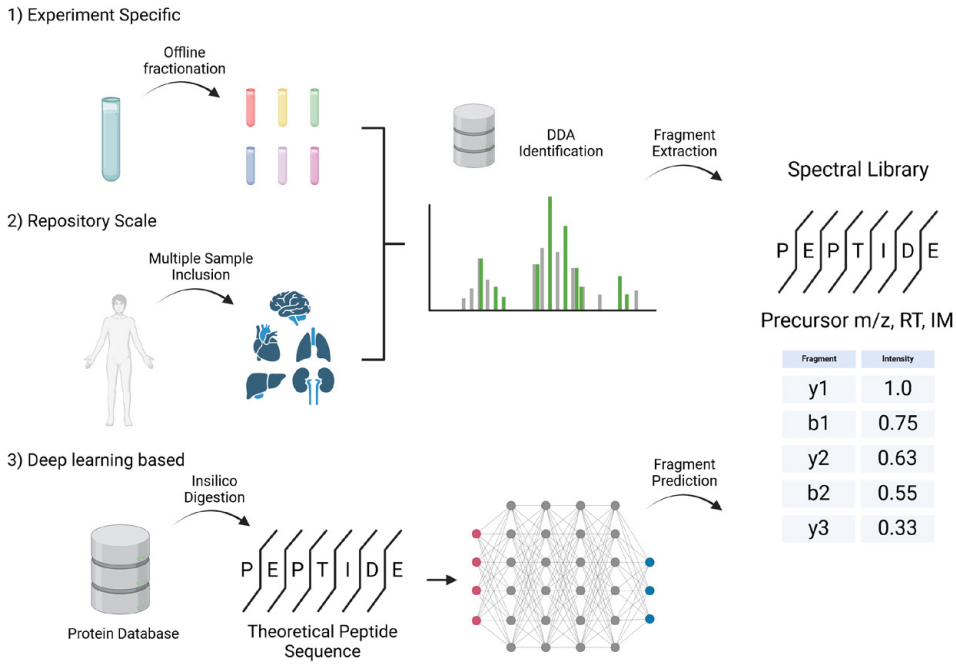


Figure 5: Spectral library creation workflows
3 of the most common methods for creating spectral libraries for DIA analysis. Experiment specific libraries usually consist of identifications from fractionated samples that are analyzed separately with the results combined downstream. Repository scale libraries consist of full proteome measurements from multiple tissue and sample types that are then aggregated into a single spectral library. Deep learning based spectral libraries predict precursor and fragment properties from peptide sequences using deep neural networks. (Created using Biorender)

Library-free Methods

One way to circumvent this problem is to search the complex MS2 spectra as DDA data[88–90]. Although this process can help annotate previously unidentified compounds, they suffer from a significant increase in computational complexity and do not provide the same level of sensitivity for quantification. Another option is to apply deep learning to predict spectral libraries for the entire proteome of the

biological system in question[71,72,75,91–95] (**Figure 5**). Although this may provide the potential to identify compounds not previously identified in a system, predicted spectral libraries introduce a non-trivial amount of computational complexity for library generation, signal extraction and annotation, and validation of the identified precursors. For large-scale experiments, the run-time cost and computational resources needed for using predicted spectral libraries for analysis is a significant deterrent and can lead to downstream issues. As a compromise between using full proteome predicted spectral libraries and small scale experimental spectral libraries, it is also possible to use repository scale spectral libraries. Repository scale libraries are generally created once using many samples and diverse sample types with offline fractionation to create an extensive generalized library that can be repeatedly used for different experiments and sample types (**Figure 5**). Although more computational resources are needed than with sample or experiment specific spectral libraries, it is possible to mitigate these issues with intelligent algorithms in the signal extraction and validation stages of analysis.

Identification and Quantification

The time resolved nature of LC-MS/MS makes it possible to extract intensities, or chromatograms, for each precursor and their corresponding fragments along the retention time axis. These extracted chromatograms for each entry in the spectral library are assembled into peak groups and passed to the validation step in the pipeline (**Figure 6**). This type of targeted analysis is termed peptide-centric, since the peptides in the spectral library lead the extraction of chromatograms from the data. That is, the presence of a peptide is queried in the spectra, rather than matching spectra to potential theoretical peptide sequences, as is common in DDA. Common analysis tools such as OpenSwath[96], EncyclopeDIA[97], DIA-NN[98], and Spectronaut all perform signal extraction and annotation using this targeted peptide-centric method. For typical analysis, this method performs very well, especially if all precursors in the spectral library are possibly contained in the sample that is being analyzed. However, this method can run into issues when the majority of precursors in the library are not actually contained in the sample, which can be common in plasma proteomics experiments or analysis with full proteome predicted spectral libraries. In these cases, empty chromatograms will be extracted for the missing precursors, which will unnecessarily increase run-time and file sizes, and propagate errors to the validation of the extracted and annotated precursors. There remains a need for methods which are designed specifically for these scenarios.

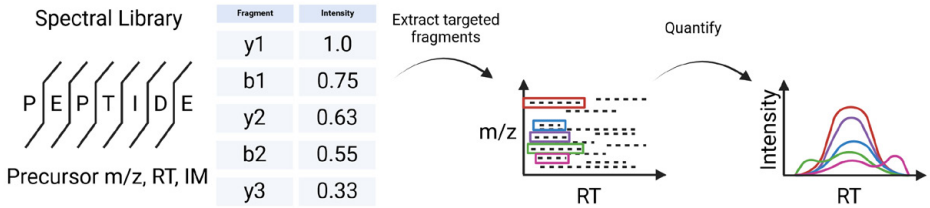


Figure 6: Targeted DIA Analysis Workflow
In standard DIA analysis, fragment ion chromatograms are extracted for each entry in the spectral library to quantify the associated precursor. Each spectral library entry consists of a precursor ion and related information such as the m/z, CCS, and RT values so that the precursor can be isolated algorithmically in the data. Information about the fragment ion m/z values are then used to extract fragment peaks at those coordinates. These extracted fragment peaks are then assembled into a peak group that provides fragment level specificity for annotation and precise quantification. (Created using Biorender)

Validation

In DIA, the target-decoy approach creates false randomized entries, or decoys, for all entries in the spectral library. It is assumed that the decoy precursors are not contained in the sample, so any annotations to decoy library entries must be erroneous. Using these decoy peak groups, machine learning models can be trained to differentiate between the false decoys and the true target peak groups. Similar to DDA validation, semi-supervised algorithms to distinguish true peak groups from decoy peak groups for have emerged as the prominent method FDR control in DIA analysis[99,100]. In cases such as plasma proteomics where predicted or repository scale libraries are used, the number of true targets in a spectral library may be extremely low, and these methods may struggle to train classifiers to validate new experimental data. These issues are caused by a severe class imbalance[101] and the presence of noisy annotations through mislabeled data[102] due to the high proportion of false targets to true targets in the data (**Figure 7**). Although there are possibly ways to configure tools such as PyProphet[100] to mitigate these problems, the issue may be solved by training generalizable validation models using one reliable dataset and then applying these models to new experiments to perform validation. This has been done previously for DDA experiments using Percolator[103], but is understudied in DIA, particularly for plasma proteomics experiments that utilize large spectral libraries. For plasma DIA experiments with large libraries, even if a generalizable model is applied for validation, the number of true targets is exceedingly small and the FDR may not be accurately controlled, so a method to minimize the issues caused by this search space imbalance would be useful.

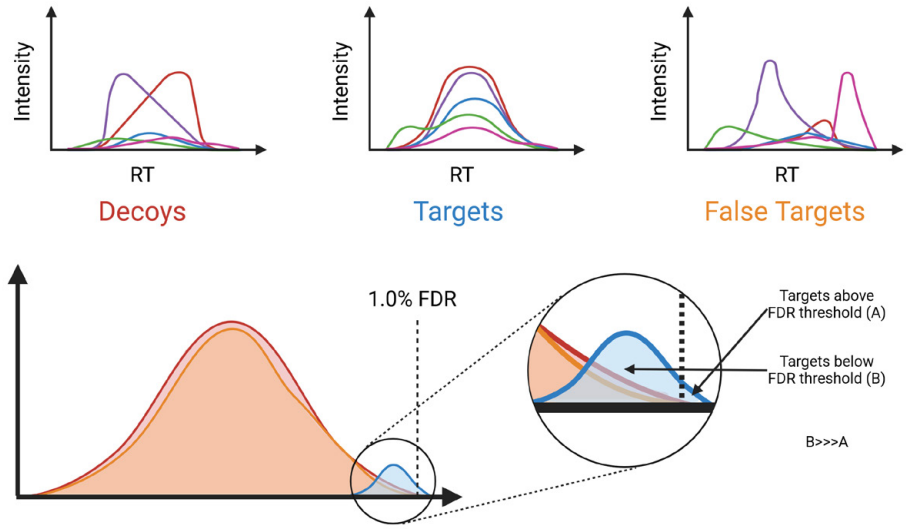


Figure 7: Search space imbalance and validation in DIA proteomics
Depending on the sample type and spectral library, it is possible that many of the extracted peak groups in DIA analysis are false targets that more closely resemble decoys. If many of the entries in a spectral library are not contained within the sample being searched, then many of these extracted peak groups will be false targets, that more closely resemble peak groups for shuffled decoy precursors. As these false targets are incorrectly passed to downstream validation algorithms as correct, true, targets, this can lead to issues in accurately controlling the FDR. (Created using Biorender)

Statistical and Biological Analysis

Once the precursor signals in a sample have been successfully extracted, annotated, and validated, the MS data is then ready for downstream analysis and biological interpretation. In bottom-up up proteomics, since proteins are digested into peptides to be analyzed using mass spectrometry, we first need to quantify the proteins that the peptides originate from by combining the extracted peptide signals in an appropriate manner. Once proteins are accurately quantified, statistical tests that evaluate the differences in protein abundance between experimental groups can be performed and the results can be interpreted (**Figure 8**).

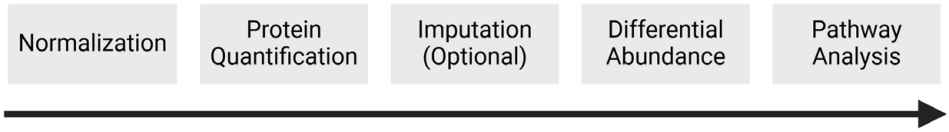


Figure 8: Recommended order of operations
A flow chart showing the recommended order of operations for the statistical analysis of proteomics data. Normalization should occur first, followed by protein quantification with the optional imputation of missing values, then differential abundance testing, and finally pathway analysis.

Normalization

The mass spectrometer is an extremely sensitive instrument. Slight changes in temperature, pressure, and variations in sample preparation can cause small, but sometimes detectable, fluctuations in the intensities that are measured for the same precursor across different samples. Because of this, the first necessary step when analyzing MS data is to minimize technical variation across samples, placing them on the same scale so that they may be compared. This process, termed normalization, is essential to ensure that the biological findings of an experiment are rooted in biology instead of technical artifacts of the mass spectrometer. There are several different methods used for normalizing data in MS proteomics. The main idea is to use some sort of aggregation function that minimizes the variation across all samples. This can be the mean, median, or the sum of all the signal for a sample. This process can be expressed as generalized mathematical framework following the following algorithm:

For an input matrix X:

$$X = \begin{bmatrix} X_{11} & \cdots & X_{1j} \\ \vdots & \ddots & \vdots \\ X_{i1} & \cdots & X_{ij} \end{bmatrix}$$

Where X_{ij} represents the abundance for the *i*-th precursor of the *j*-th sample. An aggregation function S (mean, median, or sum, for example) is used to calculate a column (sample) wise statistic. For each sample *j*, the statistic is calculated:

$$S_j = S(X_{1j}, X_{2j}, \ldots, X_{ij})$$

Giving a vector of statistics for each column we calculate the mean for this sample-wise statistic:

$$S = [S_1 \quad \ldots \quad S_j]$$

$$S_{mean} = \frac{1}{n}\sum_{j=1}^{n} S_j$$

Finally, the normalized signal can be expressed using vectorized operations by dividing each sample measurement by the sample statistic and multiplying by the mean of the sample-wise statistic:

$$X_{normalized} = \frac{X}{S} S_{mean}$$

Additionally, it has been shown empirically that performing normalization along a sliding-window of the RT axis can substantially improve the accuracy of normalization by removing technical bias in a non-linear manner while preserving biological signal[104]. After these transformations are complete, it is also typical to log2 transform the sample normalized data to obtain the final quantitative matrix. There are a number of different normalization algorithms that exist, and certain considerations should be accounted for when choosing a normalization algorithm for analysis. Methods such as quantile normalization can amplify certain effects in the data and change the abundance distributions dramatically[105]. Other methods such as locally weighted scatterplot smoothing (LOWESS), require extensive and non-trivial parameter optimization[106]. In these cases, it is necessary to consider how much will these more complex and computationally expensive methods improve the interpretable results of an experiment. If a method successfully removes technical bias from a dataset, then it is a successful normalization method, even if it is a simple algorithm.

Protein Quantification

In bottom-up proteomics the process of combining multiple peptide signals into a single accurate protein abundance is an area of research that is constantly evolving. The process can be exceedingly complex, as it is possible for peptides to originate from multiple proteins, and the problem of how much of the peptide signal contributes to each of the parent protein abundances must be answered. For DIA analysis, one very simple method of mitigating this issue is to only include peptides that are unique to a single protein in the spectral library. The inclusion of only these unique, or proteotypic, peptides allow for the simple assumption that the full abundance profile for that peptide belongs to the unique parent protein. This may restrict the quantification of some high-quality peptides, but the trade-off will lead to more interpretable protein quantities as a result. Combining proteotypic peptides into protein quantities can be as simple as choosing the appropriate aggregation function for a particular use case. Quantification algorithms can be broadly split into 2 classes, total quantification algorithms and relative quantification algorithms. Total quantification algorithms use an aggregation function, such as mean, median, or sum, to combine the top-N selected peptide quantities for a given protein. This method ensures the abundance rank of a protein in relation to other proteins in a sample to be preserved, allowing for the inter-sample comparison of protein abundances. Relative quantification algorithms aim to minimize the variance

between samples while still preserving biological signal. These algorithms, such as MaxLFQ[107], iq[108], and DirectLFQ[109], focus on solving systems of equations to minimize the variance between samples. Due to this, the absolute rank of a protein is not always preserved, as the optimal median signal may be underestimated for proteins that have more quantified peptides. Relative quantification algorithms provide a means to very accurately perform intra-sample comparisons, i.e. perform differential statistical tests between experimental sample groups, but they do not provide accurate within sample comparisons of protein groups. A combination of these methods could be practically beneficial, so multiple different types of quantification algorithms do not need to be run depending on the biological question.

Imputation

Unfortunately, in MS proteomics there can be a substantial number of missing values in each dataset. It can be appealing to impute new values to fill in these missing values, but this must be done carefully to avoid introducing artifacts in the data that are misinterpreted as biological results. Imputation can increase the number of proteins that are found statistically significant but may mask other biological findings or introduce false positives. If possible, it is best to avoid missing value imputation, but in certain cases, such as with certain machine learning algorithms, missing values must be filled in. Depending on the MS method used, DDA or DIA, the assumptions driving the cause of missing values, and thus the choice of imputation algorithm differ. Due to the stochasticity of DDA, missing values can broadly be defined as missing at random (MAR), as the missing values in a DDA dataset and the present values could have the same distributions but are selected based on their abundance[110]. In these cases, methods such as k-nearest neighbor (KNN) imputation could be effective. In KNN imputation, a group of specified nearest neighbors are calculated using a distance metric, such as the Euclidian distance, from values that are not missing for a sample. Missing values are then interpolated by aggregating values from the group of k-nearest neighbors[111]. In DIA, since all precursors in a window are selected for fragmentation, the data can be considered missing not at random (MNAR). In this case the MNAR missing values from DIA data are generally assumed to be caused by peptides that are below the limit-of-detection (LOD), that is low-abundance peptides[112]. In this case, an acceptable method of imputation DIA data could be to randomly select values from a distribution centered around the low percentage (1-5%) of observed abundance per feature. This way missing values are interpolated with values simulating low abundance features near the limit of detection. If performed, imputation should almost always occur after protein quantification. Protein quantification can help fill in missing values already, and if imputed values are used to infer protein quantities, then the resulting quantitative values may be drastically skewed.

Statistical Analysis

One common method to determine proteins that are related to a particular phenotype state, or proteome state, is to perform statistical tests between experimental groups to determine proteins that are significantly differentially abundant. Often times methods used for determining differentially expressed genes are applied to proteomics data[113,114], and many methods have been fine tuned to work specifically with proteomics data[115–117]. This is another subfield of research within proteomics that is constantly evolving, and complex methods utilizing individual peptide quantities[118] or Bayesian techniques[119,120], and other advanced methods are constantly being developed. Depending on the type of mass spectrometer used, the type of data acquired, and the type of biological question that is being asked, easily interpretable statistical tests, such as t-tests, analysis of variance (ANOVA), or linear regression, can provide an efficient alternative. As thousands of these statistical tests can be performed in a single proteomics experiment, it is imperative that the resulting p-values are corrected to control the FDR. Many times, a cutoff is applied to the resultant corrected p-values to aggregate all proteins that are significantly differentially abundant. These cutoffs can be arbitrary, and if the background data is altered, the p-values can change. Some argue that p-values should not be used to strictly infer or discount biological findings[121]. Occasionally, proteins with the most significant p-values or highest fold-change may not be the most interesting proteins for a particular biological question, leading to difficulty in sifting through all statistically significant proteins in an experiment to find what is important. There remains room for improvement in this area, especially for the data driven selection of proteins that are associated with a particular proteome state.

Pathway Analysis

As proteins represent the biological machinery of a system, they are generally linked together in sequence with other proteins in pathways to carry out different biological processes. Once statistically significant proteins are gathered from differential abundance analysis, it is often more relevant to determine which processes and pathways these proteins are associated with. If certain groups of statistically significant and high fold change proteins are associated with the same biological pathways, then it is plausible to say that those pathways are relevant to the biological question that is being asked. As pathways and processes are even closer to the expressed phenotype than proteins, this type of analysis will provide substantial biological context and facilitate interpretation of results. To perform pathway analysis, a list of proteins can be compared against biological pathways and statistical tests are performed to determine if those mappings occur more often than at random compared to a background set of proteins. The pathways that are detected depend on the pathway database being used and there are a variety to choose from[122–128]. Meta-analysis tools, such as Metascape[129], exist that combine the results

from these databases with a singular searchable interface. Gene set enrichment analysis (GSEA) is an alternative method that can be used to analyze the expression patterns of proteins across experiments rather than just lists of significant proteins[130]. Although these methods can be extremely useful to narrow down which pathways in a system are the most important based on the abundance of proteins in a dataset, the reliance on p-values and the commonality of shared pathways can make the results difficult to interpret.

Choice Paralysis

As there are so many different options for each of the analytical steps involved in the statistical analysis of proteomics data, it can a paralyzing and daunting task to choose what tool to apply for each particular use case. Each of these tools might require different dependencies and computational environments, which can be non-trivial to string together in a useful pipeline. Additionally, many of these tools are created for publication without any documentation and then abandoned to wither away without any maintenance. There are some comprehensive tools available to perform end-to-end analysis[115,131–133], but they can be inflexible and opinionated in their available functionality. To that end, a comprehensive software suite that incorporates all the above statistical methods and facilitates biological inference for an analysis would help ease choice paralysis during analysis and facilitate reproducible research.

Applied Machine Learning in Biology

Machine learning can be described as a subfield of math and statistics that encompasses algorithms that learn from data. Machine learning algorithms are able to ingest training data, learn patterns within that data, and generalize to new unseen data depending on the task that the algorithm was trained for. In recent years it has become increasingly more common to apply machine learning algorithms to answer biological questions. For example, in the medical field, machine learning excels in tasks such as the early prediction of disease to provide better and more personalized treatment, the identification and stratification of subgroups within a particular disease, and other tasks associated towards precision medicine[14,15,134–152]. Although machine learning can provide high performance predictive models for many different tasks, it is not often completely transparent how different models arrive at different predictions. To that end, it is important to design and utilize algorithms in machine learning that can be interpreted and explained, especially if the end goal of a predictive model is to apply it in a clinic to drive individualized patient treatment.

Machine Learning

Most machine learning algorithms can be broadly grouped into 2 main different types of algorithms, supervised and unsupervised machine learning. Supervised machine learning consists of predictive tasks where known sample annotations, or labels, are passed in with data and the model learns from the data how to predict the given labels. Supervised learning can further be grouped into 2 main types of problems, classification and regression. Classification problems involve predicting a label from data, while regression problems involve predicting a continuous value from data. Supervised learning has become increasingly common in biology, where binary classification problems are often presented in the form of predicting disease from healthy controls, or predicting a more severe subtype of disease from another. This type of binary classification problem is the easiest to understand and interpret, and perhaps the most useful in practice, as most biological questions involve determining the differences between 2 groups of samples in an experiment. Unsupervised learning typically involves algorithms that learn from the data without the guidance of labels. Clustering algorithms that attempt to group similar samples close together based on the features in the dataset are one of the most common applications of unsupervised learning methods. Unsupervised learning can be leveraged in biology to identify subgroups within a particular disease and cluster samples together if explicitly defined experimental groups are not defined beforehand. Although these unsupervised methods can be extremely powerful at identifying new patterns in data, they require rigorous validation to ensure that the identified and annotated clusters are accurate.

Deep Learning

In recent years, there has been an explosion of products and tools that utilize so-called "Artificial Intelligence" or AI. These types of machine learning models are based around deep learning algorithms, particularly deep neural networks. Deep neural networks are machine learning models that consist of densely connected nodes, resembling biological neurons with adjustable weights that are tuned during training. These deep neural networks excel at learning and engineering features from the data without an explicit feature engineering step[153]. This process, known as representation learning, allows deep neural networks to learn complex non-linear relationships between input features and project these features into latent representations for classification. The predictive power of deep neural networks in many cases is unparalleled compared to classical machine learning methods, providing state-of-the-art performance for classification and regression in a variety of fields[134,154–162]. However, deep neural networks suffer from a lack of interpretability, as the inner workings of the dense networks resemble a black block of computation, and it is not explicitly known how some algorithms reach their predictions. Additionally, these algorithms can be excessively data hungry. It

requires a lot of data to reliably train deep learning algorithms that will generalize to samples outside the training set. For that reason, the application of deep learning in typical proteomics experiments is not utilized as often as classic machine learning algorithms for sample classification.

Explainable Machine Learning

Explainable machine learning (XML) refers to a group of methods that are designed to elucidate how particular machine learning and deep learning algorithms reach their conclusions. In a broad sense, XML methods can be divided into 2 different categories, by-design methods and post-hoc methods[163].

By-design methods refer to cases where the design of the machine learning algorithm provides inherent explainability. The simplest case of by-design explainability is the architecture of the machine learning algorithm itself. Linear models, such as linear or logistic regression, learn weights for each feature that can be analyzed as a proxy for importance. In tree-based models, the paths of trained decision trees can be extracted for each sample and analyzed to infer importance. In more complex cases, such as neural networks, nodal connections in the hidden layers can based on some prior knowledge that gives the algorithm an interpretable design[164]. By-design methods provide a good start towards providing an explainable base for machine learning algorithms. In the context of proteomics, the interpretability of by-design methods can be complemented by using algorithms to determine which proteins are most important in predicting a certain biological state, or proteome state.

Post-hoc methods provide a means to calculate feature importance towards prediction after the training of a machine learning algorithm. In the context of proteomics, this can be explained as the importance of each protein in a dataset towards predicting a particular proteome state. For example, proteins with higher importance values can be considered more important in predicting disease. There are many ways to predict these values, from first calculating local feature importances using SHAP[165] or LIME[166], to looking at global feature perturbation importance (**Figure 9**). Conceptually, perturbation importance can be described as the impact that a shuffled feature has on the predictive output of a class. It can be expressed as follows:

$$Feature\ Importance = \left(\frac{1}{n}\right)\sum_{i=1}^{i=n}(y_i - \hat{y}_i)$$

Where y is the baseline prediction without any shuffled features and $\hat{y}$ is the prediction with the feature shuffled. This delta is calculated for all samples in the

dataset to get local explanations and the mean is taken to get the global explanation for a feature (**Figure 9**).

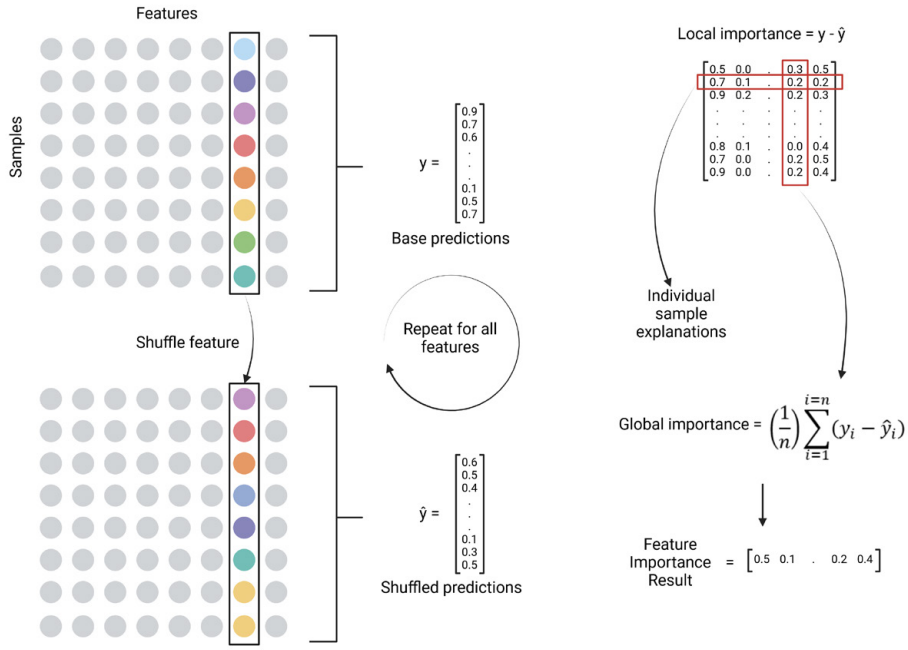


Figure 9: Simplified workflow for calculating feature importance
Data for each sample is shuffled and new predictions are made. The difference in model output can be inferred as the feature importance. This should be repeated with multiple rounds of shuffling. (Created using Biorender)

These post-hoc methods are particularly useful in identifying proteins that are most highly associated with a proteome state based on a binary question, such as predicting a particular disease. Occasionally, these proteins may not always be the proteins with the lowest corrected p-values or highest fold-change between groups. In these cases, a combination of classic statistical methods and explainable machine learning could accelerate the process of identifying the most biologically important proteins for a particular proteome state. These feature importance methods can be further used in conjunction with various feature selection methods to identify panels of proteins that are the most important for driving a prediction in a certain direction. However, both feature importance and feature selection methods are notoriously sensitive to perturbations in background data or model hyperparameters and care should be used when applying these tools.

Overall, by using both by-design and post-hoc XML methods, it is possible to drive the identification of potentially biologically relevant proteins using data-driven methods rather than using pre-defined p-value cutoffs and basic differential abundance analysis. However useful they might be, these types of methods in the context of proteomics remain underdeveloped and require extensive computational expertise to apply them.

Limitations

Although machine learning can be extremely powerful, there are certain limitations that need to be considered when applying these methods. In supervised learning, models can be overfit by learning the idiosyncrasies and noise of the training data instead of the underlying distribution of the data[167]. Overfit models may appear to make correct predictions, but they are biased to the training set and cannot generalize to new data. As the goal of supervised learning is to produce highly predictive models that can generalize effectively to new data, this is a serious problem. Overfitting can be a result of inappropriate model complexity, a lack of training samples (low-n), high dimensionality, or a lack of rigorous testing[168]. In experimental biology low-n is a common problem, which means that careful consideration and testing needs to be performed to ensure that any trained classification models are not overfitting. If there is not enough data for a split test set, this can be done using cross-validation (CV), where models are trained in a loop using a subset of the data and evaluated on a held-out fold of the data. This estimates the accuracy of a model by testing on unseen data and maximizing the availability of the training data. The sampling of the data in a CV loop may also help provide a better idea of the estimated error than a held-out test set if the data set is too small. In supervised learning, noise within the labels of the dataset can also be detrimental to the accuracy of the model[102]. If the positive labels that are being used to train a model contain a large number of false-positives, bias is introduced into the model, and it will not learn the correct underlying distribution. Ideally, labels can be corrected by experts, but that is not feasible for most datasets, so automated methods can be used, such as label-ranking mechanisms[169,170], or loss based methods[171], to try to correct or remove noisy labels. This can also lead to class imbalance issues, where one of the classes in the labels constitutes the majority of the samples. When the majority class greatly outnumbers the minority class, the training of algorithms can be destabilized as there are not enough minority instances for the model to learn[101]. In these cases, it is possible to resample the training data in an attempt to balance the class ratios but causes changes to the underlying distributions of the data. It is also possible to apply weights to certain samples based on their class label, so the training algorithm can account for the class imbalance. Noisy labels and class imbalances are very common in real world data, especially biological data, and methods to mitigate the issues caused by these problems should always be considered when appropriate.

Applications in Infectious Disease

Infectious diseases are one of the leading causes of global death. They can progress rapidly, and seemingly healthy patients may deteriorate quickly if the infection is not quickly identified and treated. Both patients and medical doctors could benefit substantially from accurate time-of-admission (TOA) tests that predict the outcome of patients. As proteomics has been previously successful in other clinical studies at identifying early detection biomarkers for disease, this strategy can be applied directly to the study of infectious disease[6,8,10,143,148,150,172–193]. If proteomics can help improve the molecular definition of infectious disease and identify predictive molecular markers for early detection and patient stratification, the lofty goal of true translational medicine, where proteomics can be used directly in clinical human applications[194] may become more commonplace.

Sepsis

As a result of infectious disease, sepsis is one the leading causes of mortality in the world. Sepsis affects almost 50 million people a year with around 11 million deaths[195], and has been most recently defined as life-threatening organ dysfunction caused by a severe host response to infection[196]. Despite the mortality of sepsis, the definition has changed many times, and there is still no strict definition for what constitutes an infection[197]. Infections in sepsis may originate from many different pathogens at different sites of infection, causing different host responses and clinical presentations. Organ dysfunction is measured using the Sequential Organ Failure Assessment (SOFA) system, which uses a combination of different clinical parameters to define SOFA scores for individual organ systems and a combined overall score. In most cases, patients present multiple types of organ dysfunction together, leading to further heterogeneity in the clinical manifestation of sepsis. Additionally, sepsis can be extremely difficult to detect early based on just the clinical parameters available at TOA, so there is a niche for predictive tooling that can stratify patients fast and effectively. Additionally, although there have been numerous studies that look to define potential subphenotypes of sepsis[148,198–203], the results do not agree upon a concrete definition of what subtypes of sepsis actually are. To reinforce the fact that sepsis is a complex and heterogeneous disease, there have been hundreds of proposed treatments that have failed clinical trials[197]. It is clear that a method to study sepsis while taking into account the heterogenous nature of the syndrome is desperately needed. Ideally, proteomics and translational medicine would be able to elucidate personalized treatment plans for new individual patients, predict outcomes, and stratify patients, enabling true precision medicine in the clinic. However, for this to be possible the issues caused by the severe heterogeneity of sepsis still need to be addressed.

Plasma Proteomics

To enable translational studies in sepsis, molecular signatures need to be detected from a quickly accessible and minimally invasive biological material that is representative of the biological state of the host. Plasma, or the liquid component of blood that remains after the blood cells are removed, is perfectly suited for this task. Plasma is rich in protein content, containing functional plasma proteins, receptor ligands, immunoglobulins, tissue leakage products, disease markers, pathogenic proteins, and other transport proteins[204]. This creates a perfect environment for the proteomic analysis of infectious disease, where perturbations in metabolism, immune activation, hemostasis, and protein leakage can all be monitored in the plasma[25,205]. Unfortunately, the relative composition of proteins in the plasma is substantially skewed. Albumins make up around 55% of the protein content[204], while globulins and fibrinogen can make up 30-40%, meaning those 3 protein groups alone can account for up to 95% of the protein content in plasma. Based on concentrations from The Human Protein Atlas (proteinatlas.org), the top 20 most abundant proteins constitute 92.1% of the protein content[206]. More than 10 orders of magnitude alone can separate albumin from the low abundant proteins in plasma[204] (**Figure 10**). This high dynamic range in biological samples can make it difficult to identify low abundance proteins in the presence of high abundance proteins. As both high and low-abundance precursors are often coeluted and analyzed simultaneously by the MS, the presence of high-abundance precursors can make it difficult to differentiate fragments from low-abundance precursors[207]. Coeluting precursors from more abundant proteins may also block less abundant precursors from being ionized in a process called ion suppression. Ion suppression reduces the detection capability of the mass spectrometer by decreasing the amount of suppressed analyte that enters the instrument, therefor causing a decrease in the dynamic range[208]. Advances in MS instrumentation, and the addition of IM separation after the LC has improved the dynamic range of MS analysis, but many important low abundance proteins are still missed in standard experiments. In plasma proteomics, the dynamic range will cause substantial issues for MS-based analysis, as significantly more MS-time would be spent analyzing the proteins that constitute more than 90% of the protein content. The identification and quantification of tissue leakage proteins and disease markers will suffer as they orders of magnitude less abundant than the most prominent proteins in plasma[193,209]. As these proteins are much more interesting for the detection and molecular understanding of infectious diseases and sepsis, new methods to increase the dynamic range and depth of MS analysis in plasma are crucial. There are existing methods available that can remove high-abundance proteins from plasma or enrich low-abundance proteins to increase dynamic range[210–219], but these methods can complicate sample preparation, be very expensive, and may have unforeseen consequences on quantification. An ideal situation would be if neat plasma, or non-depleted plasma, could be analyzed deeply enough to provide insight into the proteome state of sepsis and other infectious diseases.

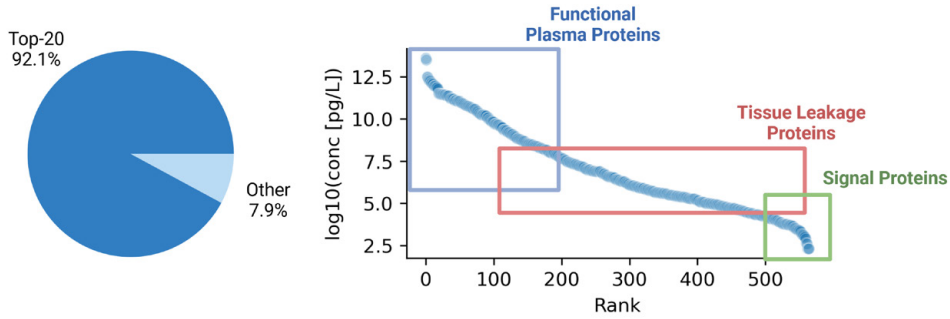


Figure 10: Protein content and dynamic range of blood plasma
A pie chart representing the abundance distribution of proteins in plasma and a recreation of the dynamic range of plasma proteins from Geyer et al. 2017[209], based on concentrations measured for The Human Protein Atlas (proteinatlas.org) .

Population Scale Proteomics

In recent years, advancements in LC and MS instrumentation have increased the throughput of mass spectrometry-based proteomics[46,47,220,221]. Samples can now be processed at a rate of up to 500 samples per day (SPD), allowing the size of proteomics experiments to increase from 100s of samples to 1000s of samples in an efficient manner. The addition of trapped ion mobility spectrometry as an extra dimension to the data has also enabled enhanced sensitivity, allowing for peptides with similar m/z and retention time to be resolved and uniquely identified. This increase in throughput is extremely promising from a translational perspective as it is easier to distinguish true biological signal in larger cohorts as they more closely represent the molecular population of a given geographic location where the samples originate from. This type of population scale proteomics also unlocks the potential of applied machine learning algorithms to identify complex relationships in the data, and we have already seen this applied to study sepsis[148]. However, such large sample sizes introduce complexities in data analysis, as large populations can be extremely heterogenous in reality, and the noise in the data needs to be carefully handled.

Aim of the thesis

Problem Statement

Although blood plasma is one of the more easily accessible and minimally invasive biological media available to obtain, it is exceedingly difficult to analyze and quantify the proteins contained in these samples, and therefor translate meaningful proteome measurements into biological insight. Unlocking the plasma proteome for discovery analysis and clinical use is a crucial step to enable translational research and guide personalized care. This is especially true regarding infectious diseases, including sepsis, where small proteomic differences and convoluted signals can reflect a rapid progression to a poor outcome. It is thus critical that we can analyze the molecular profiles of patients quickly while providing interpretable and clinically actionable results.

To that end, each project included in this thesis was carried out with 2 overarching ideas in mind:

1. How can we computationally extract more protein identifications from plasma?
2. How can we simplify biological and clinical interpretation?

The more proteins that we quantify, the more detailed and nuanced patterns we can identify in the biology. We can leverage these improvements to better stratify patients on admission and provide treatment for new patients suspected of sepsis.

Aim

The aim of this thesis was to develop computational methods, algorithms, and machine learning models that push forward the potential of using mass spectrometry-based proteomics in translational research and precision medicine. Although infectious diseases and sepsis were the main application of the methods developed during this thesis, all algorithms were designed to generalize to other diseases.

Results

Overview

The results presented in **Papers I-IV** represent the chronological steps taken to enable the effective large-scale analysis of clinical proteomics data. Using novel computational methods, we were able to facilitate the analysis of infectious diseases from neat blood plasma samples and identify patterns within the biology to help stratify heterogenous patient groups and predict outcomes. Each paper builds on concepts from the previous paper and will enable the comprehensive analysis of population scale proteomics datasets in the future, for infectious diseases and potentially other pathologies.

In **Paper I,** we wanted to demonstrate that discovery analysis of neat plasma could be improved and confirm that a biological signal in infectious disease could be identified. If we can identify biological patterns and predict clinical outcomes from patient plasma samples, we are one step closer towards data driven patient care. **Paper I** shows how large-scale spectral libraries can be leveraged to increase the analytical depth of plasma proteomics experiments if the FDR is handled appropriately. Additionally, we provide an initial glimpse into how XML can be used to identify groups of proteins that are specific to a particular proteome state that could be missed using classic differential abundance analysis.

In **Paper II**, we wanted to extend the capabilities of XML introduced in **Paper I** to leverage the predictive power of deep neural networks. Additionally, we hypothesized that XML methods could provide a data-driven approach for the analysis and identification of important biological pathways. Using a combination of by-design, and post-hoc interpretations, we formalized a framework for creating and interpreting biologically informed neural networks (BINNs). We were able to demonstrate how BINNs can be generalized to different diseases, and how they can provide an intelligent mechanism for the identification of important pathways in a system. Perhaps the most interesting finding regarding BINNs, that provided a basis for the explainable machine learning algorithms used in **Paper III** and **Paper IV**, was that the importance of nodes in the neural network and input features changed significantly depending on the background data. To calculate robust importance values, it is critical to run the interpretations many times, with different training batches or bootstrapped data, as the values will vary for each iteration. These findings will allow us to identify which proteins are most important to a particular

proteome state in a robust manner and build models that can generalize effectively to new data.

During our ongoing development of the different analytical methods in **Papers I-II**, we noticed a lack of standardization and availability of open-source software to perform basic statistical analysis and explainable machine learning for proteomics data in Python. In **Paper III**, we began to tie together and formalize the different analytical methods we have developed over the last few years in a common and easy-to-use Python package called the Data Processing Kitchen Sink (DPKS). In this paper we showcase and provide proof of performance for algorithms used in every step of the analysis of proteomics data. We provide easy access to methods for filtering, normalization, imputation, protein quantification, differential abundance testing, explainable machine learning, and pathway analysis. Many of these methods were used previously in **Papers I-II**, and many other papers not included in this thesis, and many of the more advanced methods are used extensively in **Paper IV**. We continually update and add new functionality to DPKS and provide detailed documentation to facilitate easy use.

All developed computational algorithms from **Papers I-III** were applied in **Paper IV** and demonstrate how we can use these findings to effectively analyze a population scale cohort and provide clinically relevant results. For **Paper IV**, we apply previously developed methods, substantially expanding many of them, and develop novel computational ideas for the interpretable analysis of a population scale cohort of sepsis patients. To efficiently stratify patients suspected of sepsis for personalized treatment, we analyze time-of-admission plasma samples and apply digital twin modelling to identify hidden cohorts within the data and stratify new patients in an adaptive manner. This project provides a unique approach to digital twin modelling where we aggregate multiple patients into a digital family and use this digital family to model outcome and make predictions. To our knowledge, **Paper IV** represents the largest DIA study of sepsis and the largest application of digital twin modelling in sepsis that has been performed.

Paper I

Title: Generalized precursor prediction boosts identification rates and accuracy in mass spectrometry-based proteomics

Authors: **Aaron M. Scott**, Christofer Karlsson, Tirthankar Mohanty, Erik Hartman, Suvi T. Vaara, Adam Linder, Johan Malmström & Lars Malmström

Journal: Communications Biology

Background

As a relatively non-invasive and biological comprehensive media, blood plasma can be extremely useful to identify the molecular profiles associated with different types of disease. However, the proteomic analysis of plasma remains non-trivial. One major issue with plasma proteomics, and in particular discovery plasma proteomics, is that the number of identifiable peptides in a sample is relatively small compared to tissue samples for example, which can complicate downstream analysis. One computational reason for this in DIA plasma proteomics is that the number of precursors in a spectral library may substantially outnumber the precursors in a plasma sample. Since the established method for controlling the FDR rely on training experiment or sample specific classifiers to score the true targets, this leads to a massive class imbalance which detrimentally effects the ability of experiment specific classifiers to accurately control the FDR. Due to the quantitative accuracy and depth available via DIA-MS and the non-invasive availability of plasma as a biological media, it is important to have computational tools available that can effectively analyze the plasma proteome and provide biologically interpretable results. The main aim of this study was to develop a generalizable machine learning model that facilitates accurate FDR control for discovery plasma proteomics with large spectral libraries. We then demonstrate how our novel computational tools can be used to effectively analyze a cohort of patients with Acute Kidney Injury (AKI) from neat plasma samples and provide biologically interpretable results.

Result

Using 2,988,116 peak groups we first trained models to predict true target peak groups from decoys. Using a novel label denoising algorithm, we removed false target labels from the training set to ensure the models were trained on correctly labeled data (**Figure 11**). We then demonstrated how this model can generalize to 3 different biologically diverse sets of data from 4 different organisms.

Once the ability to generalize was demonstrated on multiple datasets, we show that the predictive power of our generalized precursor scoring (GPS) model can be used

to subset large-scale spectral libraries to more effectively control the FDR and enhance the rate of identification in an experiment. Using GPS, we were able to provide a 50.57% increase in the number of precursor identifications from 31 mouse-kidney samples compared to PyProphet. We also show that precursor prediction can eliminate false positives from a dataset using entrapment yeast proteins in the mouse-kidney samples.

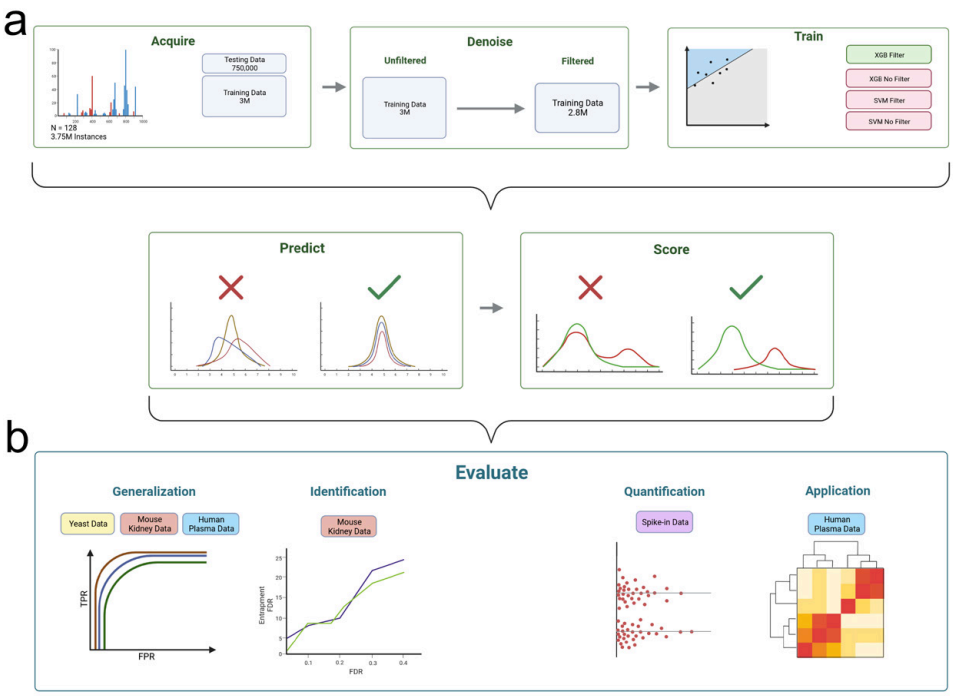


Figure 11: GPS workflow overview
Reproduced from *Scott et.al., (2023) "Generalized precursor prediction boosts identification rates and accuracy in mass spectrometry based proteomics", Communications Biology, **6**, 628*. An overview of the workflow proposed in **Paper I**. The workflow demonstrates how GPS models are trained on a curated dataset, denoised to remove false target labels, and trained. These models are used to predict precursor groups for several new datasets to control the search space and provide accurate FDR control.

In addition to providing more identifications, we also show that GPS can improve quantitative accuracy on mouse-kidney samples with known ratios of yeast peptides spiked-in to the samples. Compared to PyProphet, GPS identified 18.97% more precursors, 17.96% more peptides, and 5.28% more proteins in ratio validated

regions. GPS also decreased the number of missing values by 60.51% compared to PyProphet.

Finally, we applied GPS to analyze a cohort of 141 patients with Acute Kidney Injury (AKI). Of these samples 60 patients had less severe AKI while 81 had more severe AKI. We analyzed the dataset using large-scale spectral libraries, precursor prediction, and FDR control with GPS. GPS identified 53.81% more proteins than PyProphet (1312 proteins), 24.35% more proteins in at least 10 replicates compared to PyProphet (771 vs. 620), and 22.91% more differentially abundant proteins. Using this increased depth, we applied classic statistics and explainable machine learning to identify a panel of proteins that is highly accurate in stratifying and predicting severe AKI.

Conclusion

Overall, this study demonstrated that discovery DIA-MS can be effectively applied to analyze plasma samples and identify biologically interpretable results. We show that FDR control in DIA experiments has issues when the sample does not match closely with the proteins and peptides contained in the spectral library, and we provide a computational approach to mitigate those issues. Additionally, we demonstrate that explainable machine learning is a potentially powerful tool for the identification of proteins of interest for a particular proteome state.

One additional conclusion that is not stated in the manuscript, is that significant changes are needed in the open-source community of mass spectrometry proteomics software when it comes to the analysis of DIA data. There are relatively few open source options to analyze DIA data[96,97] and the associate workflows can be prohibitively difficult to customize, as we have done in this paper. When it comes to analyzing diaPASEF data acquired on a timsTOF[220], there is only one open source option, OpenSWATH, and unfortunately it is difficult to analyze large cohorts of data with the current implementation. We are left with closed source or commercial software to analyze DIA data, leaving a niche to be filled by open-source options. Certain tools have begun to emerge to fill this niche[222], but further work is needed.

Paper II

Title: Interpreting biologically informed neural networks for enhanced proteomic biomarker discovery and pathway analysis

Authors: Erik Hartman*, **Aaron M. Scott***, Christofer Karlsson, Tirthankar Mohanty, Suvi T. Vaara, Adam Linder, Lars Malmström & Johan Malmström

*Authors contributed equally

Journal: Nature Communications

Background

As demonstrated in **Paper I**, the incorporation of machine learning methods into omics workflows can help improve the identification of features related to particular proteome states. This is particularly important in infectious disease, as patient outlook can deteriorate rapidly if the patients are not stratified and treated quickly in the hospital. Although these machine learning methods may prove useful in quickly predicting a patient phenotype, if they are not interpretable and explainable, it will be difficult to implement these models in a clinical setting. Deep neural networks are a subset of machine learning algorithms with state-of-the-art predictive power. However, they suffer from a lack of interpretability, creating a computational black box where certain features are plugged in and certain answers are shot out, without knowing how those answers are obtained. The idea for this project started out with a desire to leverage the predictive power of deep neural networks while providing transparent interpretations, allowing for each prediction from the network to be explained computationally and biologically. Utilizing the previously published biologically informed neural networks for pancreatic cancer[164], we generalized the concept allowing for the creation of any hierarchical neural network where the input and nodes in the hidden layers are connected in an ontology, such as biological pathways. We then applied these biologically informed neural networks (BINNs) to 3 different datasets regarding infectious disease (**Figure 12**). Additionally, we also hypothesized that we could use the by-design interpretable framework to identify the most important biological pathways in a system using feature attribution methods. Since the connections in the neural network are based on biological pathways, we can determine which nodes in the network are the most important in each layer of the hierarchy, essentially allowing data-driven optimization to determine which pathways are the most important based on a set of input proteins and the proteome state that the BINN is predicting.

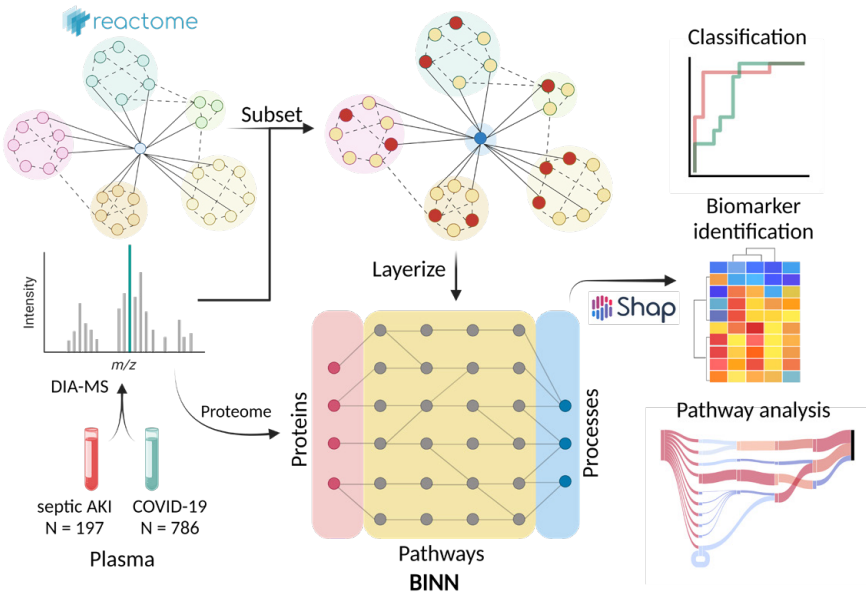


Figure 12: An Overview of the BINN package and analysis
Reproduced from *Hartman and Scott et al. (2023) "Interpreting biologically informed neural networks for enhanced proteomic biomarker discovery and pathway analysis", Nature Communications, **14**, 5359*. This figure provides an overview of the analysis in **Paper II** and the overall workflow for creating and interpreting BINNs.

Result

To first demonstrate that BINNs can provide state-of-the-art performance compared to other models, we compared the performance of BINN to 5 other types of machine learning models in predicting severe AKI and severe COVID. BINNs outperformed all other models based on the area under the receiver operating characteristic curve and the area under the precision-recall curve. Additionally, BINNs provided the highest precision and recall rates compared to other models.

We additionally demonstrated that we could rank proteins by feature importance as input to a BINN and select protein panels that are able to stratify patients from both the AKI and COVID datasets with high accuracy. Simultaneously, we can provide customized pathway analysis that allows you investigate which pathways are important in a biologically intuitive manner through the ranking of nodal connections in the BINN. These results reconfirm what we identified in **Paper I**, that explainable machine learning can be used to identify proteins and biology that are highly important to a particular proteome state.

To provide additional confirmation that BINNs can generalize to a variety of biological questions and across platforms, we analyzed a previously generated O-link dataset consisting of patients with different sources of ARDS[223].

Conclusion

The overarching conclusion of this study is that advanced machine learning algorithms can be effectively used to analyze the proteome in infectious disease, particularly if feature attribution methods and explainable machine learning algorithms are used properly in conjunction with these methods. Machine learning can assist in identifying biomarkers, providing highly predictive models for disease, and assist in identifying biologically relevant pathways for further downstream analysis.

Not mentioned in the article, but an additional conclusion from this paper that serves as framework for **Paper IV**, is that by-design (the embedding of biological pathways into the architecture of the neural network) and post-hoc (the calculation of the most important proteins and pathways in the BINN) methods of interpretation used in combination provide the greatest benefit towards explainability of an algorithm. If it is possible to interpret the architecture of a machine learning algorithm in a biologically relevant manner, while also providing quantitative interpretations for the input features within the algorithm, you can both maximize the extracted biological information and provide powerful predictive models simultaneously. These findings help cement explainable machine learning as a powerful tool for the analysis of the proteome, potentially moving both machine learning and proteomics closer towards translational medicine and personalized treatment in the clinic.

Paper III

Title: Explainable machine learning for the identification of proteome states via the data processing kitchen sink

Authors: **Aaron M. Scott**, Erik Hartman, Johan Malmström, and Lars Malmström

Journal: bioRxiv

Background

The beginnings of **Paper III** started around the same time as **Paper I**, but constantly developed over the course of the thesis. The aim of this project was to develop an easy-to-use and powerful analytical software package for the general analysis of mass spectrometry-based proteomics data. Although there are many different tools published in many different journals that go into detail about different steps in the analytical process, they are all implemented separately, sometimes in different programming languages, and are not developed with longevity and good software practices in mind. The goal of this project was to develop a data processing kitchen sink (DPKS) that can take quantified precursors from a variety of tools as input and provide a modular and accessible application programming interface (API) for the application of a wide range of tools. DPKS provides access to advanced normalization techniques, missing data imputation, protein quantification, differential abundance analysis, and pathway analysis. Additionally, we have formalized many of the explainable machine learning methods developed in **Papers I-II** and incorporated them into the easy-to-use API, allowing practitioners with minimal experience in machine learning to apply advanced methods of feature attribution to their datasets. All features are described in the documentation (https://infectionmedicineproteomics.github.io/DPKS/) and the code is available on Github (https://github.com/InfectionMedicineProteomics/DPKS). In this paper, we wanted to provide additional benchmarks for the available statistical methods as well as provide examples for the possible applications of explainable machine learning.

Result

In general, this paper showcases the functionality and capabilities of DPKS. We first demonstrate that DPKS provides easy access to common statistical methods for preprocessing proteomics data for downstream analysis. We provide a number of different normalization algorithms to minimize technical bias between samples, and include a retention time sliding window algorithm to accommodate the selected liquid chromatography gradient[104]. We provide multiple options for protein quantification based on relative comparisons[108], and a topN method that can

summarize the most intense peptides per protein using the specified function. We also implement a novel combined method that provides the signal smoothing benefits of relative quantification and the rank preserving capabilities of the topN method. Finally, we provide a number of options for differential abundance analysis between experimental groups, including paired sample statistical tests.

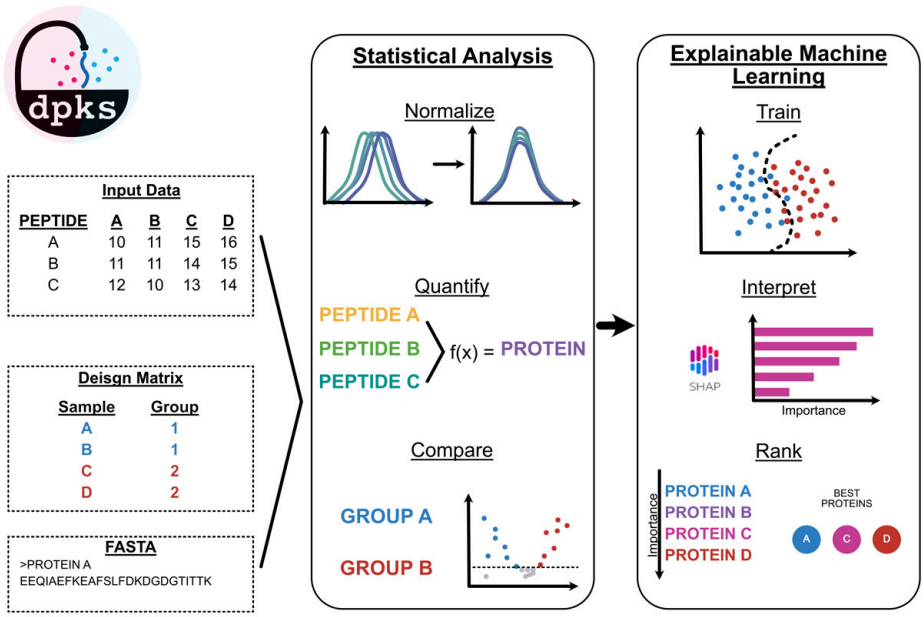


Figure 13: Overview of the functionality in DPKS
This figure provides an overview of the functionality in the DPKS software package. Input data in the form of a quantitative matrix and a design matrix are first parsed into an internal data structure that allows for the modular application of multiple statistical analysis and explainable machine learning algorithms.

We also demonstrate and describe in detail the types of explainable machine learning methods that are available in DPKS and how they can be applied to investigate a proteome state. In **Paper III** we reanalyze the COVID-19 data used in **Paper II** and apply feature selection using explainable machine learning to identify a panel of proteins that are highly accurate in stratifying severe COVID from less severe COVID. We reemphasize that, occasionally, the most important proteins associated with predicting a particular proteome state are not always the proteins with the lowest p-values. The functionality of DPKS is summarized in **Figure 13**.

Conclusion

The important conclusion in this paper is that we provide the described functionality in an easily accessible Python package that utilizes industry standard coding practices, test coverage, and in-depth documentation. We provide additional benchmarks for many of the methods available in DPKS as well as further showcase how explainable machine learning can be used to investigate biological questions. One of the most important aspects of research is reproducibility, and by providing tested code in a packaged environment, we hope to contribute towards the overall goal of reproducible research, particularly in a computational setting. Often, code is written for publication and abandoned immediately after. From the onset of implementation, this is something we strive to avoid with DPKS.

Paper IV

Title: Population scale proteomics enables adaptive digital twin modelling in sepsis

Authors: **Aaron M. Scott***, Lisa Mellhammar*, Erik Malmström*, Axel Goch Gustafsson, Anahita Bakochi, Marc Isaksson, Tirthankar Mohanty, Louise Thelaus, Fredrik Kahn, Lars Malmström1, Johan Malmström**, Adam Linder**

*Authors contributed equally

**Corresponding authors: Adam Linder, Johan Malmström for proteomics and mass spectrometry related aspects of the study

Journal: medRxiv

Background

Sepsis is a severe and potentially lethal reaction to an infection that effects nearly 50 million people year, causing 11 million deaths annually[195]. In Sweden alone, around 50,000 people are affected per year. Even though sepsis is one of the leading causes of mortality in the world, the definition is vague[197], it is difficult to diagnose in the clinic, and there remains a lack of personalized treatments for a massively heterogeneous syndrome. In Sweden, there is a triage program known as Sepsis Alert[224–226], where patients that are suspected of sepsis on admission to the hospital are placed under specific care and time of admission blood samples are taken for further analysis. From this program, we have analyzed 1364 patients suspected of sepsis to investigate the molecular mechanisms of the syndrome and to develop computational methods that can stratify patients on admission to the clinic. Since sepsis patients present with a variety of different organ dysfunctions, different localizations of infections, and different pathogens, the molecular landscape of

sepsis is extremely diverse. Due to this, we hypothesize that breaking patients into rule-based subphenotypes may be a suboptimal method for stratifying sepsis patients for treatment. Instead, we propose to leverage a means of neighborhood analysis akin to the digital twin paradigm proposed in mechanical engineering[227]. Using a rich compendium of sepsis patients as a search space, we can identify groups of patients that are most similar to new patients and use this digital family to model and predict outcome. As an alternative to classic stratification methods, digital twin models provide an unprecedented flexibility, allowing for hidden cohorts within the data to be identified and predictions made without explicitly training individual models.

Result

The first significant result we present in this study is that sepsis is a massively heterogeneous syndrome, and even using the proteome it is almost impossible to stratify sepsis patients from sepsis mimics. We also found that grouping patients into subphenotypes may not be the most effective way to stratify patients as the number of distinct and accurate phenotypes can be arbitrarily high. Due to the heterogeneity observed in sepsis, it is necessary to take the full proteome and project it down into a lower dimensional latent space to see if novel patterns within the data can be identified.

Utilizing a similar approach to BINNs from **Paper II**, we developed a by-design interpretable latent space (ILS) where we train individual models to predict organ dysfunction (respiratory, renal, cardiovascular, coagulation, liver, and central nervous system (CNS)), infection, and sepsis individually and then use these models to predict probabilities for a subset of the data that was not used to train the models. From here, we can used post-hoc interpretation methods to identify which proteins are the most important to predict each individual proteome state, allowing us to learn about the biology of each clinical outcome predicted and create a latent space that was optimized based on supervised explainable machine learning. Each of these classifiers identifies molecular features that are specific to that particular outcome and can be linked to biological pathways that are associated with the given organ dysfunction or infection type.

Using this ILS, we adapted the digital twin approach to aggregate groups of patients from the training set that most closely resemble new patients based on a distance metric calculated using the ILS as features. With this approach, we can predict and model patient trajectories and outcome without explicitly training individual models to do so. We found that we can group and stratify patients based on outcome effectively without the need for grouping patients into rule-based phenotypes (**Figure 14**).

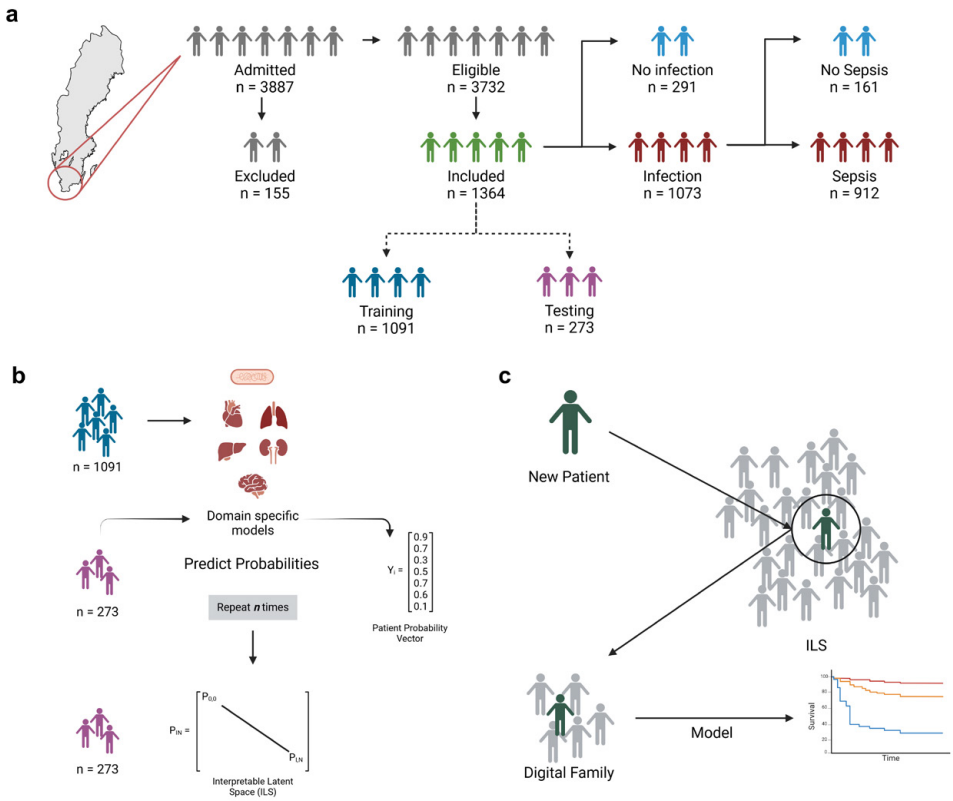


Figure 14: Overview of digital twin modelling in the Sepsis Alert cohort
Produced using data from **Paper IV**. This figure provides an overview for the inclusion of patients from the Sepsis Alert triage program. The proteome of these patients is then used to train domain specific models for predicting organ dysfunction, infection, and sepsis, and then used to create an interpretable latent space (ILS) to facilitate digital twin modelling for patients suspected of sepsis. (Created using Biorender)

Conclusion

In **Paper IV**, we demonstrate how stratifying patients into rule-based subphenotypes can be problematic as sepsis is a highly heterogenous syndrome. Although, subphenotypes can be useful when there is no strict definition of the syndrome (i.e. they are better than not having guided treatment at all), we propose that digital twin models can provide a more dynamic approach for stratifying patients on admission to the hospital.

One main conclusion that is not specified explicitly in the text, is that this study represents one of largest DIA-MS studies in existence and that population scale proteomics can be used to identify molecular markers for disease in complex syndromes such as sepsis.

“Real world” data can be noisy and messy, containing missing values and overlapping distributions that make it difficult to identify the signal within the noise. Precisely designed cohorts will allow specific effects to be identified, but it is possible that these findings may not translate to a larger scale representation of the population. Our study allows for these types of phenomena to be investigated in-depth. In **Paper IV**, we showcase the predictive capabilities of digital families to predict mortality, organ dysfunction, and identify hidden patterns in the data, but these applications can be greatly expanded in the future.

Discussion

The aim of this thesis was to develop computational methods that push forward the potential of using mass spectrometry-based proteomics in translational research and precision medicine. To that end, each project was carried out with two main overall goals. First, can we algorithmically increase the depth of plasma proteomics by extracting more identifications from the MS data, and second, can we simplify biological or clinical interpretation of complex results? The results contained in this thesis provide a promising outlook for the use of plasma proteomics towards the understanding of infectious diseases and sepsis. By unlocking the plasma proteome, it will be possible to comprehensively study complex diseases using relatively non-invasive procedures to interrogate the biology of a patient. Ideally, continued advances in this field would lead to the adoption of proteomics in the clinic, where blood samples from patients in the hospital can be used to rapidly stratify patients and guide treatment. Although much needs to be done to enable real-time clinical proteomics, the ability to comprehensively analyze population scale cohorts of clinical time-of-admission samples is a step in the right direction.

Through strategies implemented in **Paper I**, we have demonstrated that we can increase the number and quality of quantified proteins by manipulating the search space using generalized precursor prediction and robustly control the FDR using generalized machine learning models. This is a post-processing step in the signal extraction pipeline of DIA-MS data that can substantially increase the number of precursors, peptides, and proteins quantified in an experiment. However, more than 90% of potential peptide features in a plasma sample remain unannotated and thus unused for downstream analysis in a typical DIA-MS experiment. A successful DIA-MS experiment using neat plasma may identify around 10,000 precursors in a sample, while if you extract potential MS1 precursors from the same sample there will be around 200,000. Even if many of the MS1 features extracted are not identifiable peptides, either because they contain some sort of post-translational modification or are false positive features, this still represents an underutilization of the available precursor ions in a sample. If 10,000 precursors can quantify 700-800 proteins per sample, fully utilizing hundreds of thousands of precursor ions could provide a massive boost to the depth of plasma proteomics, and proteomics in general.

When most DIA-MS analysis software was designed, it was with the idea that each fragment and precursor in a spectral library is probably contained in the sample

being analyzed. Due to this assumption, chromatograms are extracted for every single entry in a spectral library. For a tissue sample that is expected to have close to the full proteome being analyzed, this does not cause any problems. However, in plasma samples for example, the number of detectable proteins may be considerably less. If a spectral library consisting of the full human proteome is used to analyze a plasma sample using standard DIA-MS software, false positive chromatograms will be extracted across the entire gradient for hundreds of thousands to millions of ions that are not contained in the sample. This will cause errors in accurately controlling the FDR, as many of the "true" target peak groups used to train the classifiers for validation are empty chromatograms. This issue is partly addressed in **Paper I**, where we established that precursor prediction can remove false target peak groups in a first pass filtering step to allow for accurate FDR control down the line. However, this can go a step further. Similar to library-free DIA analysis approaches[88–90], detectable MS1 features could be used to guide the extraction of peak groups in a data-driven approach. This would represent a hybrid approach to DIA-MS analysis, combining the peptide-centric spectral library chromatogram extraction with library-free feature extraction. This work was started throughout the thesis but will need to be continued.

The results from **Paper I** already suggest that we can successfully leverage DIA-MS in plasma proteomics to study infectious disease at a relatively large scale. This allows us to utilize the technology for the proteomic investigation of infectious disease. However, to increase the depth of experiments from hundreds to thousands of samples, changes to the hardware as well as the software need to be addressed. The utilization of Bruker timsTOF Pro 2 and EvoSep One allowed the throughput of samples to immensely increase, facilitating the analysis of 1400 samples in weeks for **Paper IV**. However, the original pipeline designed in GPS could not be applied to this data. OpenSWATH was expanded to analyze timsTOF data[220], however this implementation did not adequately scale to utilizing the 4th ion mobility dimension provided by the timsTOF in a fast or user-friendly manner. The signal extraction software was switched to DIA-NN[98], which is closed source and not easily customizable. However, DIA-NN can be configured to use a similar algorithm, that they refer to as match-between-runs (MBR). The MBR algorithm in DIA-NN extracts peak groups in a first pass analysis and builds a second spectral library with all confidently extracted peak groups across samples for a second-pass analysis. This is very similar to precursor prediction, except the confidently extracted precursors are still based on the FDR computed using a large search space and the issues described in **Paper I**, when many extracted targets are false targets, may still occur. DIA-NN could potentially benefit from the inclusion of a GPS-like algorithm, but unfortunately this is not easy to implement as the source code of the tool is closed. There remains a niche for robust, scalable, open-source tools for signal extraction for timsTOF data. These tools must scale reliably to thousands of samples and process them at close to real-time speed to continue to push proteomics closer to clinic.

In this thesis, we have focused on algorithms and computational methods to increase the number of proteins identified in a sample. However, sample preparation methods, such as antibody-based enrichment, depletion columns that remove the most abundant proteins in plasma, or equalizer beads, can be used to increase the number of measured proteins from hundreds to thousands before any signal processing is performed[210,211,213,215,216,218,219,228,229]. Combined with the algorithmic improvements suggested in this thesis, the increased depth could help identify subtle proteome changes regarding organ dysfunction and infection, which are critical clinical outcomes related to the diagnosis and treatment of sepsis. Organ dysfunction can be detected using proteins in the plasma that have leaked in from the damaged organs, but these proteins are general present in low abundance[205]. Due to issues with dynamic range in neat plasma, they are difficult to detect and quantify. The same is true when trying to identify the pathogen causing infection from the host plasma proteome. In **Paper IV**, we explored the idea of training classifiers to predict the infecting pathogen from plasma, but there was no detectible signal for differentiating specific pathogens. Instead, we had to group pathogens into broad groups, gram-positive and gram-negative, to investigate proteome changes in the plasma between the groups. With improved depth, we may be able to better understand questions such as: What proteins represent this specific type of organ dysfunction? Or, What proteins indicate an infection caused by a particular pathogen? Or, where does a particular infection originate? Many of these questions are addressed in **Paper IV**, but improving the depth of analysis could be invaluable. With increased depth comes additional analytical challenges, and since there are more proteins detectable in a sample, signal extraction algorithms could potentially take longer. Improvements to analytical software have significantly improved processing time, but with large scale population cohorts, additional work may be needed to ensure that signal extraction does not become a bottleneck in the pipeline. It will also be necessary to have tools to allow the increased dimensionality of the data to be interpreted. **Papers I-IV** all focus on providing biologically interpretable results within an efficient timeframe and would be well suited for this particular task.

In addition to enabling increased analytical depth, **Paper I** demonstrated that explainable machine learning can be useful in identifying important proteins in infectious disease. In infectious disease, many times the most differentially abundant proteins are non-specific inflammatory proteins that do not provide specific biological relevance to a proteome state, particularly in heterogenous syndromes such as sepsis (**Paper IV**), or in stratifying severity in AKI or COVID (**Papers I-III**). In **Paper I**, we showcased that explainable machine learning can identify proteins specific to a proteome state that are not necessarily the most differentially abundant or have the highest statistical significance (lowest p-value). This concept is revisited in **Paper II**, formalized in **Paper III**, and massively expanded in **Paper IV.** Although infectious diseases and sepsis were the main

application of these methods developed during this thesis, all algorithms were designed to generalize to other diseases.

In **Paper II**, we further investigated the concept of feature attribution methods and explainable machine learning towards proteomic biomarker discovery and pathway analysis. The appeal of deep neural networks is justified by their state-of-the-art predictive performance, however the use of deep learning for biomarker discovery and clinical applications may be hindered by their lack of interpretability. BINNs allow for a fully interpretable structure with explainable predictions, which could aid in the adoption of deep learning closer to a clinical setting if the model and the predictions can be explained. In **Paper II**, we restricted the pathway database to the Reactome[122], but theoretically, BINNs can be used with any sort of hierarchical structure, where the input features can be mapped to a directed graph that can flow to an output. In fact, we have used BINN in unpublished projects using a customized hierarchy of data, not actually related to biological pathways. This flexibility is something that is made possible due to our focus on creating an extensible software package that can generalize to many different use cases. Many times, the paradigm of reproducible and reusable software is lost in academic software development, but as research becomes more computationally heavy, software design should become a more important aspect of the development cycle. Some researchers have begun to expand the principles of findable, accessible, interoperable, and reusable (FAIR) research to software development[230], which is a step in the right direction but far from widely adopted. This is a focus throughout the thesis, particularly in **Paper III**, where all tools developed should have a structure and design that makes them easily adaptable for future use.

During the development of both the BINN and DPKS software packages, we noticed that if the background data is slightly altered, the importance of certain proteins and pathways in a model can change. This may seem obvious, however in practice it is not consistently accounted for. In many cases where explainable machine learning is used in biology, a model is trained, the most important features are extracted, and they are validated using a test set. This may lead to a classifier that generalizes well to the test data, but if explainable machine learning is being used to identify proteins that are most consistently important to a proteome state, this may lead to false positive results. In **Paper I**, we attempted to mitigate these issues using recursive feature elimination with a cross-validation loop, and we further develop a customized bootstrapping feature attribution method available in DPKS that is used extensively in **Paper IV**. Although the applications in this thesis have been proteomic and clinical data, the bootstrapping feature importance algorithm could be applied to any sort of classification problem, meaning that studies could easily integrate additional modalities, such as metabolomics or transcriptomics data, or the software could be used for applications outside of biology.

The combination of results from **Papers I-III** allowed for a comprehensive investigation of one the largest cohorts of sepsis patients using DIA-MS to date. As

sepsis is a complex and molecularly heterogenous syndrome, we found that the classic methods of stratifying patients by endotypes or subtypes[198,200–203,231–233] may not be the most optimal approach to guide treatment, as the heterogeneity of sepsis patients provides near-endless combinations and sizes of clusters that can be molecularly and clinically explained. Digital twin models provide flexibility and adaptability that allows for patients to be stratified in a dynamic manner and group new patients with those most similar to them without predefining subgroup definitions. In **Paper IV**, we demonstrated that this can be used to predict patient outcomes and identify hidden subgroups within a cohort. Although these types of predictions may already be of use to the clinic to guide treatment, we believe that the applications of this technology may be far greater than what we used as a proof-of-concept. However, to expand the possible applications of digital twin modelling in sepsis, there are several possible improvements that should be considered. As stated previously, if the depth of plasma proteomics is increased, then we can predict much more subtle changes in the proteome. This would increase the performance of the organ dysfunction classifiers, infection classifiers, and lead to the creation of a more representative interpretable latent space (ILS). Additional clinical outcomes could also be included in the latent space to improve the distance calculations between neighbors to select more accurate digital families. For example, outcomes such as drug response and infection localizations could be included to group similar patients closer together. As with organ dysfunction and infection, treatment response can be predicted by labelling patients in the training set as responsive to a particular treatment and classifiers can be trained to predict which new patients will also be responsive to a particular treatment. If much of the digital family of a new patient was responsive to a treatment, this treatment may prove effective for that patient. In the case of treatment response, an expansion to the time dimension would allow for the trajectory of patients to be tracked as the treatment is applied. This same logic can be applied to predict infection localization, or any other desired clinical outcome. Although we need to be careful to not over expand the latent space, as distance calculations can become less meaningful if the dimensionality is too high[234], certain important clinical manifestations that were not detectable with the current analytical depth should be included. Extra modalities, such as metabolomics and lipidomics could improve the digital twin models in **Paper IV** as sepsis has a substantial impact on metabolism. These modalities could be included to improve predictions and help learn about the biology of sepsis, while in turn creating a more powerful latent space.

Finally, one of the most important mechanisms to improve digital twin modelling is to include more patients. As we have shown in **Paper IV**, these models are adaptive, and the inclusion of more patients, even outside of sepsis, could boost the potential and clinical importance of these models without any technical changes. Including patients with non-septic infections could help differentiate the type of infecting pathogen and the localization of the infection while increasing the precision of digital families. In **Paper IV**, the type of infecting pathogen is skewed heavily

towards E. coli, so the inclusion of other infection types would greatly improve the sensitivity of digital family models. Additionally, patients with organ dysfunction unrelated to an infection would help differentiate and specify the type of organ dysfunction a patient may have. In **Paper IV**, most patients in the dataset have some degree of respiratory dysfunction which makes it difficult to train truly organ specific predictive models as most patients present with multiple organ dysfunction. These additions would allow patients to be grouped with more similar digital families and improve the overall predictive power of the model.

Conclusion and Future Perspectives

Conclusion

Overall, the work in this thesis has contributed to advancing the potential of plasma proteomics for the study of infectious disease in large scale clinical cohorts. With a focus on interpretability, we applied advanced machine learning techniques to find hidden biology in heterogenous data to better understand complex infectious disease. Concordant with the aims of this thesis, the methods developed in **Papers I-IV** can be applied to aid translational research and help bring proteomics closer to the clinic, for sepsis and for other diseases in the future.

Future Perspectives

Following the completion of this thesis, an immediate area of future development revolves around the expansion and development of signal processing algorithms to increase the depth of plasma proteomics. For the sake of discovery and plasma proteomics, changes to the quantification strategy that maximize ion annotation would be a straight-forward first step for the expansion of these tools. As previously mentioned in the discussion, sometimes only around 10,000 precursors from a potential 200,000 are annotated. During the thesis, we began to investigate the potential of utilizing MS1 features for the guided extraction of peak groups from DIA, but as is common with many PhD projects, it was not possible to complete this in time. MS1 features can be used to filter large search spaces or provide an additional data dimension enabling a hybrid approach to analysis that utilizes the flexibility of library-free analysis and the efficiency of targeted extraction. If we can rescue even a small percentage of the unannotated precursors in a sample, the number of quantified proteins could increase dramatically and could massively improve the effectiveness of neat plasma proteomics.

Additionally, the findings from **Paper I** should also be incorporated into modern DIA processing pipelines to ensure that the precursor prediction and search space minimization techniques can be used to accurately control the FDR when large libraries are being analyzed. Currently, this is not easily achievable as the prominent tools in the field for analyzing DIA data are closed source and do not provide access

to the intermediate data needed to train precursor prediction models. In this thesis, this became especially relevant in **Paper IV** as we exclusively utilized timsTOF machines to produce large clinically relevant cohorts. The few open-source tools available for DIA analysis either do not support timsTOF data or they do not adequately compete with the closed-source state-of-the-art tools available. As a solution, we could reach out to the groups that develop these tools and inquire about a collaboration. However, an open-source solution that is completely transparent would be preferrable, so an in-house solution may be needed. As tools to produce deep learning-based predicted spectral libraries keep evolving, algorithms that can maximize signal extraction from plasma samples will be even more relevant, helping us move away from a reliance on experimentally created spectral libraries to true discovery DIA proteomics.

Outside of the algorithmic next steps that are relevant to this thesis, there are some experimental aspects that would be extremely interesting to immediately follow-up on. In **Papers I-IV**, we rely on the plasma proteome to interrogate the molecular phenotypes of different disease states. However, plasma contains much more than just proteins. Expanding the modalities used to investigate infectious disease to include lipidomics and metabolomics could provide an instantaneous boost to our understanding of the molecular profiles associated with infectious disease and benefit our predictive models. Using the modular structure we developed in **Paper IV**, we could easily include predictive models that utilize lipids and metabolites in the ILS and digital family modelling. Hopefully, these modalities would help improve the prediction of individual organ dysfunctions, infection, infection localizations, and pathogen types, filling in where the proteome is lacking.

Another clear step past the addition of new modalities, especially in the digital twin model proposed in **Paper IV**, is the addition of multiple cohorts of diverse patients. Besides providing extensive validation sets, if we can include patients with other types of infections and injuries, we can improve the performance of models that predict specific pathogen types, specific types of organ dysfunction, infection localization, and more**.** In **Paper IV**, although the cohort was more than 1000 samples, some specific types of organ dysfunction and infection were rare, leading to imbalanced classes that can degrade the performance of machine learning algorithms. If we increase the number of samples included, we hope to more effectively detect these specific clinical phenotypes and more accurately stratify patients on admission to the hospital. With the addition of new modalities and more diverse sample sets, we can further explore the definition of the ILS and develop digital family models that can stratify patients more effectively.

So far, the future prospectives discussed are all focused on the pre-clinical level, meaning they are focused on maximizing discovery experiments and providing results to better understand the molecular definitions of infectious disease. In **Paper IV**, we describe a method that could be of actual potential use in the clinic, but a substantial amount of work needs to be done to evaluate if digital family models can

be used to guide treatment for sepsis, infectious diseases, or any other diseases. First, doctors would most likely be extremely hesitant to utilize this technology without rigorous benchmarking and proof that it can improve patient outcomes and shorten hospital stays. To that end, it would be extremely beneficial to perform some sort of prospective experiment in the clinic, where we can utilize our digital family model in a simulation to predict outcome and guide treatment. We would compare our predicted results to the actual outcome of each patient in the clinic to see if we can predict outcome in real time and stratify patients in a way that can guide treatment. If successful, doctors would hopefully gain confidence in the model, which would be a major barrier to overcome in the implementation of this technology in the clinic.

To truly move digital twin modelling to the clinic, there would have to be some way to measure the proteome, lipidome, and metabolome in the clinic for each patient at time-of-admission. Mass spectrometers are expensive and can be difficult to operate, so instead of providing a mass spectrometer to the clinic to perform global analysis, maybe a more effective route is to determine the minimal set of proteins, lipids, and metabolites that are needed to accurately stratify patients into digital families. This way we could provide targeted assays, either using MS, or some other technology to measure the minimum molecular signature needed for clinical stratification.

Although there are some significant obstacles that need to be overcome before the technology presented in this thesis can be used to guide treatment in the clinic, the methods developed in **Papers I-IV** provide a solid base of knowledge to build upon. Using these concepts, we can begin to explore the possibility of personalized medicine in the clinic for infectious disease, and far beyond that, which is an exciting prospect.

Acknowledgements

A PhD is a long journey, and there are countless people to thank that help you along the way. I have pointed out a few specific people here, but to anyone who helped me along the way, I sincerely thank you.

First, I would like to thank my supervisor, **Lars**. Your experience has been a huge boost during my time here, and being able to talk through my ideas always helps to solidify them into something useful.

To my co-supervisor **Christofer**. Thanks for always answering my randomly timed questions! You were always really supportive and happy to help, and it was much appreciated.

Johan, thank you so much for all your help over the course of the last few years. Your mentorship has been invaluable and your excitement with science is always inspiring.

Anita, everything runs so smoothly with your help. Thanks for making things so much easier.

To all the clinicians I have worked with, especially **Adam, Lisa, and Erik**, thank you for answering my repetitive questions and always being happy to explain the most basic parts of the clinic that I do not understand.

Thank you to all the **IMP** and **BioMS** group members, you have made the last 4(+) years a great experience overall:

Tommaso, it has been so fun working with you the last few years, and I really appreciate having someone around who is equally as loud as I am.

Tirthankar, your encyclopaedic knowledge of everything molecular always leads to interesting and informative conversations. Looking forward to swapping more guitars with you!

Carlos, your positivity and approach to problems helps me see things in a unique way, which is always a good thing.

Di, thanks for being there on this PhD journey to navigate all the mysterious and different admin things we need to do.

Erik, even during your master's thesis our discussions always seem to lead to cool ideas and things to investigate. Looking forward to what we can come up with next!

Sounak, I will always remember the shared office days. Thanks for all the discussions and ignoring me yelling at my computer occasionally.

Elizabeth, your energy is always fun to have around the lab, and I am glad you are part of the computational army we have accumulated around here.

Alejandro, talking with you always encourages me to put things in perspective and look at the big picture. You help remind me to ask why questions.

Joel, you were the first student I supervised! I'm glad I didn't scare you off, it's made the last few years much more fun. Thanks for helping me see things from a different angle.

Anahita, I am never quite sure what the story is going to be when you stop by my office, but I always know it's going to be good! I wouldn't have made it through without your organizational skills and magical spreadsheets.

Filip, we have known each other from back in the Bioinformatics masters' days. Thanks for helping me with all things data and always being down for an AW.

Lotta, thank you for being friendly right from the beginning and willing to listen whenever I needed to rant about something.

Lukas and **Simon**, I don't venture into your world of structural proteomics very often, but I'm thankful for your AW spontaneity.

Marc, even if it was for a short time, it was great being able to work with you and discuss various data-related and bioinformatics things.

Moritz, thanks for being my office mate for the first few years. The discussions we had were always interesting and I was lucky to learn so much from you, even during the pandemic.

To all the friends that I've made here. You made it so much easier to feel at home halfway around the world from where I came from.

Mandy, you were one of my first friends in Sweden and introduced me to the Sweden gang. I will always be thankful for this! To **Louise, Andreas, Sebastian, Erik, Henrik, Hannah, Alma, and Sibel**, you all have helped keep me sane outside of my PhD.

Martin, we made it through our master's program and then both jumped into the crazy world of computational proteomics. Happy you have been there along the way with me.

Jakob, thank you for initializing my interest in proteomics and for all the interesting discussions we had during my master's thesis.

Ryan, although we don't talk about science so often when we hangout, you were also one of the original Bioinformatics crew. Glad you decided to move back up to the cold northern part of the world.

Enrique, Derek. Part of my Californian home away from home! Even if we all live in different places, so glad that you both decided to stay out this way so we can complain about the winter together.

To all the friends from back in California, even though we are halfway around the world from each other, whenever we visit it feels like I never left. There are too many to name, but to those I see the most often nowadays and have convinced (most) to visit me up in the North, **Tigran, Bo, Max, Curt, Scott, Jordan, Greg, Mike, and Zach**.

Ali, thank you for always being there for me, even though we live on opposite sides of the world now. Myla & I miss you!

Mom & Dad, thank you for everything. You always have fostered and supported my curiosity, especially when I asked a million questions as a hyperactive little kid. If not for what you both taught me, I would not be where I am today.

Josefin, none of this would have been possible without you. You have been an unwavering pillar of support as I became a student again, and I know you always will be. With that said, I do promise not to be a perpetual student forever, though. I never imagined that when I moved to Sweden 7 years ago that I would still be here, married and buying houses with you, but here we are, and I wouldn't have it any other way.

References

1. Wilkins, M. R. *et al.* Progress with Proteome Projects: Why all Proteins Expressed by a Genome Should be Identified and How To Do It. *Biotechnol. Genet. Eng. Rev.* **13**, 19–50 (1996).
2. Kim, M.-S. *et al.* A draft map of the human proteome. *Nature* **509**, 575–581 (2014).
3. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).
4. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature 2003 422:6928* **422**, 198–207 (2003).
5. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature 2016 537:7620* **537**, 347–355 (2016).
6. Niu, L. *et al.* Noninvasive proteomic biomarkers for alcohol-related liver disease. *Nature Medicine 2022 28:6* **28**, 1277–1287 (2022).
7. Tijms, B. M. *et al.* Cerebrospinal fluid proteomics in patients with Alzheimer's disease reveals five molecular subtypes with distinct genetic risk profiles. *Nat. Aging* 1–15 (2024) doi:10.1038/s43587-023-00550-7.
8. Consortium, Co.-19 M. B. At. (COMBAT) *et al.* A blood atlas of COVID-19 defines hallmarks of disease severity and specificity. *Cell* **185**, 916-938.e58 (2022).
9. Latterich, M. & Schnitzer, J. E. Streamlining biomarker discovery. *Nat. Biotechnol.* **29**, 600–602 (2011).
10. Xu, J.-Y. *et al.* Integrative Proteomic Characterization of Human Lung Adenocarcinoma. *Cell* **182**, 245-261.e17 (2020).
11. Liu, W. *et al.* Large-scale and high-resolution mass spectrometry-based proteomics profiling defines molecular subtypes of esophageal cancer for therapeutic targeting. *Nat. Commun.* **12**, 4961 (2021).
12. Chen, F., Chandrashekar, D. S., Varambally, S. & Creighton, C. J. Pan-cancer molecular subtypes revealed by mass-spectrometry-based proteomic characterization of more than 500 human cancers. *Nat. Commun.* **10**, 5679 (2019).
13. Khoo, A. *et al.* Proteomic discovery of non-invasive biomarkers of localized prostate cancer using mass spectrometry. *Nat. Rev. Urol.* **18**, 707–724 (2021).
14. Casado, P. & Cutillas, P. R. Proteomic Characterization of Acute Myeloid Leukemia for Precision Medicine. *Mol. Cell. Proteom.* **22**, 100517 (2023).
15. Carnielli, C. M. *et al.* Combining discovery and targeted proteomics reveals a prognostic signature in oral cancer. *Nat. Commun.* **9**, 3598 (2018).

16. Skowronek, P. & Meier, F. Proteomics in Systems Biology, Methods and Protocols. *Methods Mol. Biol.* **2456**, 15–27 (2022).
17. Vitko, D. *et al.* timsTOF HT Improves Protein Identification and Quantitative Reproducibility for Deep Unbiased Plasma Protein Biomarker Discovery. *J. Proteome Res.* **23**, 929–938 (2024).
18. Batth, T. S. *et al.* Streamlined analysis of drug targets by proteome integral solubility alteration indicates organ-specific engagement. *Nat. Commun.* **15**, 8923 (2024).
19. Zhai, L., Chen, K., Hao, B. & Tan, M. Proteomic characterization of post-translational modifications in drug discovery. *Acta Pharmacol. Sin.* **43**, 3112–3129 (2022).
20. Eckert, S. *et al.* Decrypting the molecular basis of cellular drug phenotypes by dose-resolved expression proteomics. *Nat. Biotechnol.* 1–10 (2024) doi:10.1038/s41587-024-02218-y.
21. Mitchell, D. C. *et al.* A proteome-wide atlas of drug mechanism of action. *Nat. Biotechnol.* **41**, 845–857 (2023).
22. Ruprecht, B. *et al.* A mass spectrometry-based proteome map of drug action in lung cancer cell lines. *Nature Chemical Biology* **16**, (2020).
23. Mohanty, T. *et al.* A pharmacoproteomic landscape of organotypic intervention responses in Gram-negative sepsis. *Nat. Commun.* **14**, 3603 (2023).
24. Bakochi, A. *et al.* Cerebrospinal fluid proteome maps detect pathogen-specific host response patterns in meningitis. *eLife* **10**, (2021).
25. Karlsson, C. A. Q. *et al.* Streptococcus pyogenes Infection and the Human Proteome with a Special Focus on the Immunoglobulin G-cleaving Enzyme IdeS*. *Mol. Cell. Proteom.* **17**, 1097–1111 (2018).
26. Toledo, A. G. *et al.* Pathogen-driven degradation of endogenous and therapeutic antibodies during streptococcal infections. *Nat. Commun.* **14**, 6693 (2023).
27. Hauri, S. *et al.* Rapid determination of quaternary protein structures in complex biological samples. *Nature Communications* **10**, 1–10 (2019).
28. Chen, Z. A. & Rappsilber, J. Quantitative cross-linking/mass spectrometry to elucidate structural changes in proteins and their complexes. *Nat. Protoc.* **14**, 171–201 (2019).
29. Masson, G. R. *et al.* Recommendations for performing, interpreting and reporting hydrogen deuterium exchange mass spectrometry (HDX-MS) experiments. *Nat. Methods* **16**, 595–602 (2019).
30. Morris, J. H. *et al.* Affinity purification–mass spectrometry and network analysis to understand protein-protein interactions. *Nat. Protoc.* **9**, 2539–2554 (2014).
31. Liu, X., Salokas, K., Weldatsadik, R. G., Gawriyski, L. & Varjosalo, M. Combined proximity labeling and affinity purification−mass spectrometry workflow for mapping and visualizing protein interaction networks. *Nat. Protoc.* **15**, 3182–3211 (2020).
32. Santos, A. *et al.* A knowledge graph to interpret clinical proteomics data. *Nature Biotechnology 2022* 1–11 (2022) doi:10.1038/s41587-021-01145-6.
33. Kelly, R. T. Single-cell Proteomics: Progress and Prospects. *Mol Cell Proteomics* **19**, 1739–1748 (2020).

34. Bennett, H. M., Stephenson, W., Rose, C. M. & Darmanis, S. Single-cell proteomics enabled by next-generation sequencing or mass spectrometry. *Nat Methods* **20**, 363–374 (2023).

35. Tajik, M., Baharfar, M. & Donald, W. A. Single-cell mass spectrometry. *Trends Biotechnol.* **40**, 1374–1392 (2022).

36. Su, P. *et al.* Single Cell Analysis of Proteoforms. *J. Proteome Res.* **23**, 1883–1893 (2024).

37. Li, W. *et al.* scPROTEIN: a versatile deep graph contrastive learning framework for single-cell proteomics embedding. *Nat. Methods* **21**, 623–634 (2024).

38. MacCoss, M. J. *et al.* Sampling the proteome by emerging single-molecule and mass spectrometry methods. *Nat Methods* **20**, 339–346 (2023).

39. Huffman, R. G. *et al.* Prioritized mass spectrometry increases the depth, sensitivity and data completeness of single-cell proteomics. *Nat Methods* **20**, 714–722 (2023).

40. Gatto, L. *et al.* Initial recommendations for performing, benchmarking and reporting single-cell proteomics experiments. *Nat Methods* **20**, 375–386 (2023).

41. Tran, J. C. *et al.* Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **480**, 254–258 (2011).

42. Duncan, M. W., Aebersold, R. & Caprioli, R. M. The pros and cons of peptide-centric proteomics. *Nat. Biotechnol.* **28**, 659–664 (2010).

43. Vandermarliere, E., Mueller, M. & Martens, L. Getting intimate with trypsin, the leading protease in proteomics. *Mass Spectrom. Rev.* **32**, 453–465 (2013).

44. Jiang, Y. *et al.* Comprehensive Overview of Bottom-Up Proteomics Using Mass Spectrometry. *ACS Meas. Sci. Au* **4**, 338–417 (2024).

45. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science* **246**, 64–71 (1989).

46. Meier, F. *et al.* Parallel Accumulation–Serial Fragmentation (PASEF): Multiplying Sequencing Speed and Sensitivity by Synchronized Scans in a Trapped Ion Mobility Device. *J. Proteome Res.* **14**, 5378–5387 (2015).

47. Meier, F. *et al.* Online parallel accumulation–serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer. *Molecular and Cellular Proteomics* **17**, 2534–2545 (2018).

48. Guevremont, R. High-field asymmetric waveform ion mobility spectrometry: A new tool for mass spectrometry. *J. Chromatogr. A* **1058**, 3–19 (2004).

49. Fernandez-Lima, F., Kaplan, D. A., Suetering, J. & Park, M. A. Gas-phase separation using a trapped ion mobility spectrometer. *Int. J. Ion Mobil. Spectrom.* **14**, 93–98 (2011).

50. Wells, J. M. & McLuckey, S. A. Collision-Induced Dissociation (CID) of Peptides and Proteins. *Methods Enzymol* **402**, 148–185 (2005).

51. Olsen, J. V. *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nature Methods* (2007) doi:10.1038/nmeth1060.

52. Old, W. M. *et al.* Comparison of Label-free Methods for Quantifying Human Proteins by Shotgun Proteomics* S. *Mol. Cell. Proteom.* **4**, 1487–1502 (2005).

53. Gillet, L. C. *et al.* 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* **11**, O111.016717 (2012).

54. Eng, J. K., Searle, B. C., Clauser, K. R. & Tabb, D. L. A Face in the Crowd: Recognizing Peptides Through Database Search. *Molecular & Cellular Proteomics* **10**, R111.009522 (2011).

55. Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of proteome research* **10**, 1794–1805 (2011).

56. Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: An open-source MS/MS sequence database search tool. *Proteomics* **13**, 22–24 (2013).

57. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications* **5**, 1–10 (2014).

58. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nature Methods* **14**, (2017).

59. Tabb, D. L., Fernando, C. G. & Chambers, M. C. MyriMatch: Highly Accurate Tandem Mass Spectral Peptide Identification by Multivariate Hypergeometric Analysis. *J Proteome Res* **6**, 654–661 (2007).

60. Lazear, M. R. Sage: An Open-Source Tool for Fast Proteomics Searching and Quantification at Scale. *J. Proteome Res.* (2023) doi:10.1021/acs.jproteome.3c00486.

61. Farag, Y. M., Horro, C., Vaudel, M. & Barsnes, H. PeptideShaker Online: A User-Friendly Web-Based Framework for the Identification of Mass Spectrometry-Based Proteomics Data. *J. Proteome Res.* **20**, 5419–5423 (2021).

62. Barsnes, H. & Vaudel, M. SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines. *J. Proteome Res.* **17**, 2552–2555 (2018).

63. Kremer, L. P. M., Leufken, J., Oyunchimeg, P., Schulze, S. & Fufezan, C. Ursgal, Universal Python Module Combining Common Bottom-Up Proteomics Tools for Large-Scale Analysis. *J. Proteome Res.* **15**, 788–794 (2016).

64. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).

65. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* **100**, 9440–9445 (2003).

66. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925 (2007).

67. Yang, K. L. *et al.* MSBooster: improving peptide identification rates using deep learning-based features. *Nat. Commun.* **14**, 4539 (2023).

68. Li, K., Jain, A., Malovannaya, A., Wen, B. & Zhang, B. DeepRescore: Leveraging Deep Learning to Improve Peptide Identification in Immunopeptidomics. *Proteomics* **20**, (2020).

69. Adams, C. *et al.* Fragment ion intensity prediction improves the identification rate of non-tryptic peptides in timsTOF. *Nat. Commun.* **15**, 3956 (2024).

70. Zolg, D. P. *et al.* INFERYS rescoring: Boosting peptide identifications and scoring confidence of database search results. *Rapid Commun. Mass Spectrom.* (2021) doi:10.1002/rcm.9128.

71. Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods* **16**, 509–518 (2019).

72. Picciani, M. *et al.* Oktoberfest: Open-source spectral library generation and rescoring pipeline based on Prosit. *PROTEOMICS* e2300112 (2023) doi:10.1002/pmic.202300112.

73. Buur, L. M. *et al.* MS2Rescore 3.0 Is a Modular, Flexible, and User-Friendly Platform to Boost Peptide Identifications, as Showcased with MS Amanda 3.0. *J. Proteome Res.* (2024) doi:10.1021/acs.jproteome.3c00785.

74. Kalhor, M., Lapin, J., Picciani, M. & Wilhelm, M. Rescoring Peptide Spectrum Matches: Boosting Proteomics Performance by Integrating Peptide Property Predictors Into Peptide Identification. *Mol. Cell. Proteom.* **23**, 100798 (2024).

75. Cox, J. Prediction of peptide mass spectral libraries with machine learning. *Nature Biotechnology 2022* 1–11 (2022) doi:10.1038/s41587-022-01424-w.

76. Yu, F., Haynes, S. E. & Nesvizhskii, A. I. IonQuant Enables Accurate and Sensitive Label-Free Quantification With FDR-Controlled Match-Between-Runs. *Mol. Cell. Proteom. : MCP* **20**, 100077 (2021).

77. Weisser, H. & Choudhary, J. S. Targeted Feature Detection for Data-Dependent Shotgun Proteomics. *Journal of Proteome Research* **16**, 2964–2974 (2017).

78. Teleman, J., Chawade, A., Sandin, M., Levander, F. & Malmström, J. Dinosaur: A Refined Open-Source Peptide MS Feature Detector. *J. Proteome Res.* **15**, 2143–2151 (2016).

79. Abdrakhimov, D. A. *et al.* Biosaur: An open-source Python software for liquid chromatography–mass spectrometry peptide feature detection with ion mobility support. *Rapid Commun. Mass Spectrom.* e9045 (2021) doi:10.1002/rcm.9045.

80. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).

81. Argentini, A. *et al.* MoFF: A robust and automated approach to extract peptide ion intensities. *Nature Methods* **13**, 964–966 (2016).

82. Lim, M. Y., Paulo, J. A. & Gygi, S. P. Evaluating False Transfer Rates from the Match-between-Runs Algorithm with a Two-Proteome Model. *Journal of Proteome Research* **18**, 4020–4026 (2019).

83. Prianichnikov, N. *et al.* MaxQuant Software for Ion Mobility Enhanced Shotgun Proteomics*. *Mol. Cell. Proteom. : MCP* **19**, 1058–1069 (2020).

84. Sandin, M. *et al.* An adaptive alignment algorithm for quality-controlled label-free LC-MS. *Molecular and Cellular Proteomics* **12**, 1407–1420 (2013).

85. Wang, J. & Lam, H. Graph-based peak alignment algorithms for multiple liquid chromatography-mass spectrometry datasets. *Bioinformatics* **29**, 2469–2476 (2013).

86. Smith, R., Ventura, D. & Prince, J. T. LC-MS alignment in theory and practice: A comprehensive algorithmic review. *Briefings in Bioinformatics* (2013) doi:10.1093/bib/bbt080.

87. Schubert, O. T. *et al.* Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nature Protocols 2014 10:3* **10**, 426–441 (2015).

88. Tsou, C. C. *et al.* DIA-Umpire: Comprehensive computational framework for data-independent acquisition proteomics. *Nature Methods* **12**, 258–264 (2015).

89. Yu, F. *et al.* Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. *Nat. Commun.* **14**, 4154 (2023).

90. Lu, Y. Y., Bilmes, J., Rodriguez-Mias, R. A., Villén, J. & Noble, W. S. DIAmeter: matching peptides to data-independent acquisition mass spectrometry data. *Bioinformatics* **37**, i434–i442 (2021).

91. Yang, Y. *et al.* In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nature Communications* **11**, 1–11 (2020).

92. Song, J. & Yu, C. Alpha-Tri: a deep neural network for scoring the similarity between predicted and measured spectra improves peptide identification of DIA data. *Bioinformatics* **38**, 1525–1531 (2022).

93. Zeng, W.-F. *et al.* AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics. *Nat Commun* **13**, 7238 (2022).

94. Tiwary, S. *et al.* High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nature Methods* **16**, 519–525 (2019).

95. Searle, B. C. *et al.* Generating high quality libraries for DIA MS with empirically corrected peptide predictions. *Nature Communications* (2020) doi:10.1038/s41467-020-15346-1.

96. Röst, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature Biotechnology* **32**, 219–223 (2014).

97. Searle, B. C. *et al.* Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nature Communications* **9**, (2018).

98. 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. *Nature Methods* **17**, (2020).

99. Reiter, L. *et al.* MProphet: Automated data processing and statistical validation for large-scale SRM experiments. *Nature Methods* **8**, 430–435 (2011).

100. Rosenberger, G. *et al.* Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nature Methods* **14**, 921–927 (2017).

101. Megahed, F. M. *et al.* The class imbalance problem. *Nat. Methods* **18**, 1270–1272 (2021).

102. Northcutt, C. G., Athalye, A. & Mueller, J. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. *arXiv* (2021) doi:10.48550/arxiv.2103.14749.

103. Fondrie, W. E. & Noble, W. S. Machine Learning Strategy That Leverages Large Data sets to Boost Statistical Power in Small-Scale Experiments. *Journal of Proteome Research* **19**, 1267–1274 (2020).

104. Willforss, J., Chawade, A. & Levander, F. NormalyzerDE: Online Tool for Improved Normalization of Omics Expression Data and High-Sensitivity Differential Expression Analysis. *Journal of Proteome Research* (2019) doi:10.1021/acs.jproteome.8b00523.

105. Zhao, Y., Wong, L. & Goh, W. W. B. How to do quantile normalization correctly for gene expression data analyses. *Sci. Rep.* **10**, 15534 (2020).

106. Berger, J. A. *et al.* Optimized LOWESS normalization parameter selection for DNA microarray data. *BMC Bioinform.* **5**, 194 (2004).

107. Cox, J. *et al.* Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Molecular & Cellular Proteomics* **13**, 2513–2526 (2014).

108. Pham, T. V., Henneman, A. A. & Jimenez, C. R. iq: an R package to estimate relative protein abundances from ion quantification in DIA-MS-based proteomics. *Bioinformatics* **36**, 2611–2613 (2020).

109. Ammar, C., Schessner, J. P., Willems, S., Michaelis, A. C. & Mann, M. Accurate Label-Free Quantification by directLFQ to Compare Unlimited Numbers of Proteomes. *Mol. Cell. Proteom.* **22**, 100581 (2023).

110. Bhaskaran, K. & Smeeth, L. What is the difference between missing completely at random and missing at random? *Int. J. Epidemiology* **43**, 1336–1339 (2014).

111. Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525 (2001).

112. Harris, L., Fondrie, W. E., Oh, S. & Noble, W. S. Evaluating Proteomics Imputation Methods with Improved Criteria. *J. Proteome Res.* **22**, 3427–3438 (2023).

113. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47–e47 (2015).

114. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

115. Kohler, D. *et al.* MSstats Version 4.0: Statistical Analyses of Quantitative Mass Spectrometry-Based Proteomic Experiments with Chromatography-Based Quantification at Scale. *J Proteome Res* (2023) doi:10.1021/acs.jproteome.2c00834.

116. Goeminne, L. J. E., Sticker, A., Martens, L., Gevaert, K. & Clement, L. MSqRob Takes the Missing Hurdle: Uniting Intensity- and Count-Based Proteomics. *Anal Chem* **92**, 6278–6287 (2020).

117. Zhu, Y. *et al.* DEqMS: A Method for Accurate Variance Estimation in Differential Protein Expression Analysis. *Molecular & cellular proteomics : MCP* **19**, 1047–1057 (2020).

118. Suomi, T. & Elo, L. L. Enhanced differential expression statistics for data-independent acquisition proteomics. *Sci Rep-uk* **7**, 5869 (2017).

119. Truong, P., The, M. & Käll, L. Triqler for Protein Summarization of Data from Data-Independent Acquisition Mass Spectrometry. *J Proteome Res* **22**, 1359–1366 (2023).

120. The, M. & Käll, L. Integrated identification and quantification error probabilities for shotgun proteomics. *Molecular and Cellular Proteomics* **18**, 561–570 (2019).

121. Wasserstein, R. L., Schirm, A. L. & Lazar, N. A. Moving to a World Beyond "p < 0.05." *Am. Stat.* **73**, 1–19 (2019).

122. Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Res* **50**, D687–D692 (2021).

123. Aleksander, S. A. *et al.* The Gene Ontology knowledgebase in 2023. *GENETICS* **224**, iyad031 (2023).

124. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).

125. Szklarczyk, D. *et al.* The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2022).

126. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

127. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–D592 (2022).

128. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **28**, 1947–1951 (2019).

129. Zhou, Y. *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communications 2019 10:1* **10**, 1–10 (2019).

130. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).

131. Krismer, E., Bludau, I., Strauss, M. T. & Mann, M. AlphaPeptStats: an open-source Python package for automated and scalable statistical analysis of mass spectrometry-based proteomics. *Bioinformatics* **39**, btad461 (2023).

132. Kohler, D., Staniak, M., Yu, F., Nesvizhskii, A. I. & Vitek, O. An MSstats workflow for detecting differentially abundant proteins in large-scale data-independent acquisition mass spectrometry experiments with FragPipe processing. *Nat. Protoc.* 1–24 (2024) doi:10.1038/s41596-024-01000-3.

133. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods 2016 13:9* **13**, 731–740 (2016).

134. Arasteh, S. T. *et al.* Large language models streamline automated machine learning for clinical studies. *Nat. Commun.* **15**, 1603 (2024).

135. Foglierini, M. *et al.* RAIN: machine learning-based identification for HIV-1 bNAbs. *Nat. Commun.* **15**, 5339 (2024).

136. Yun, T. *et al.* Unsupervised representation learning on high-dimensional clinical data improves genomic discovery and prediction. *Nat. Genet.* 1–10 (2024) doi:10.1038/s41588-024-01831-6.

137. Hällqvist, J. *et al.* Plasma proteomics identify biomarkers predicting Parkinson's disease up to 7 years before symptom onset. *Nat. Commun.* **15**, 4759 (2024).

138. Williams, S. A. *et al.* A proteomic surrogate for cardiovascular outcomes that is sensitive to multiple mechanisms of change in risk. *Sci. Transl. Med.* **14**, eabj9625 (2022).

139. Isom, M., Go, E. P. & Desaire, H. Enabling Lipidomic Biomarker Studies for Protected Populations by Combining Noninvasive Fingerprint Sampling with MS Analysis and Machine Learning. *J. Proteome Res.* (2024) doi:10.1021/acs.jproteome.3c00368.

140. Pavlović, M. *et al.* Improving generalization of machine learning-identified biomarkers using causal modelling with examples from immune receptor diagnostics. *Nat. Mach. Intell.* **6**, 15–24 (2024).

141. Qumsiyeh, E., Showe, L. & Yousef, M. GediNET for discovering gene associations across diseases using knowledge based machine learning approach. *Sci. Rep.* **12**, 19955 (2022).

142. Dadu, A. *et al.* Identification and prediction of Parkinson's disease subtypes and progression using machine learning in two cohorts. *npj Park.'s Dis.* **8**, 172 (2022).

143. Kotol, D. *et al.* Absolute Quantification of Pan-Cancer Plasma Proteomes Reveals Unique Signature in Multiple Myeloma. *Cancers* **15**, 4764 (2023).

144. Medina, L. M. P. *et al.* Targeted plasma proteomics reveals signatures discriminating COVID-19 from sepsis with pneumonia. *Respir. Res.* **24**, 62 (2023).

145. Placido, D. *et al.* A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nat Med* 1–10 (2023) doi:10.1038/s41591-023-02332-5.

146. Cosentino, J. *et al.* Inference of chronic obstructive pulmonary disease with deep learning on raw spirograms identifies new genetic loci and improves risk models. *Nat Genet* 1–9 (2023) doi:10.1038/s41588-023-01372-4.

147. Demichev, V. *et al.* A time-resolved proteomic and prognostic map of COVID-19. *Cell Syst* **12**, 780-794.e7 (2021).

148. Mi, Y. *et al.* High-throughput mass spectrometry maps the sepsis plasma proteome and differences in patient response. *Sci. Transl. Med.* **16**, eadh0185 (2024).

149. Cho, N. Molecular subtypes and imaging phenotypes of breast cancer. *Ultrasonography* **35**, 281–288 (2016).

150. Mann, M., Kumar, C., Zeng, W. F. & Strauss, M. T. Artificial intelligence for proteomics and biomarker discovery. *Cell Systems* **12**, 759–770 (2021).

151. Ciaramella, A. *et al.* A new biomarker panel of ultraconserved long non-coding RNAs for bladder cancer prognosis by a machine learning based methodology. *Bmc Bioinformatics* **23**, 569 (2022).

152. Gao, F. *et al.* DeepCC: a novel deep learning-based framework for cancer molecular subtype classification. *Oncogenesis* **8**, 44 (2019).

153. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

154. Shen, D., Wu, G. & Suk, H.-I. Deep Learning in Medical Image Analysis. *Annu. Rev. Biomed. Eng.* **19**, 1–28 (2016).

155. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

156. Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).

157. Angelis, L. D. *et al.* ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Front. Public Heal.* **11**, 1166120 (2023).

158. Clusmann, J. *et al.* The future landscape of large language models in medicine. *Commun. Med.* **3**, 141 (2023).

159. Christin, S., Hervet, É. & Lecomte, N. Applications for deep learning in ecology. *Methods Ecol. Evol.* **10**, 1632–1644 (2019).

160. Gupta, A., Anpalagan, A., Guan, L. & Khwaja, A. S. Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. *Array* **10**, 100057 (2021).

161. Vaswani, A. *et al.* Attention Is All You Need. *arXiv* (2017).

162. Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).

163. Chen, V. *et al.* Applying interpretable machine learning in computational biology—pitfalls, recommendations and opportunities for new developments. *Nat. Methods* **21**, 1454–1461 (2024).

164. Elmarakeby, H. A. *et al.* Biologically informed deep neural network for prostate cancer discovery. *Nature 2021 598:7880* **598**, 348–352 (2021).

165. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Arxiv* (2017) doi:10.48550/arxiv.1705.07874.

166. Ribeiro, M. T., Singh, S. & Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv* (2016).

167. Webb, G. I. Encyclopedia of Machine Learning. 744–744 (2011) doi:10.1007/978-0-387-30164-8_623.

168. Lever, J., Krzywinski, M. & Altman, N. Model selection and overfitting. *Nat. Methods* **13**, 703–704 (2016).

169. Northcutt, C. G., Wu, T. & Chuang, I. L. Learning with Confident Examples: Rank Pruning for Robust Classification with Noisy Labels. *arXiv* (2017).

170. Bernhardt, M. *et al.* Active label cleaning for improved dataset quality under resource constraints. *Nat. Commun.* **13**, 1161 (2022).

171. Natarajan, N., Dhillon, L. with N. L. N. N. I. S., Ravikumar, P. & Tewari, A. Learning with Noisy Labels. *Conference on Neural Information Processing Systems (NeurIPS)* 1196–1204 (2013).

172. Shen, B. *et al.* Proteomic and Metabolomic Characterization of COVID-19 Patient Sera. *Cell* **182**, 59-72.e15 (2020).

173. Birhanu, A. G. Mass spectrometry-based proteomics as an emerging tool in clinical laboratories. *Clin. Proteom.* **20**, 32 (2023).

174. Whiteaker, J. R. *et al.* A targeted proteomics–based pipeline for verification of biomarkers in plasma. *Nat. Biotechnol.* **29**, 625–634 (2011).

175. Puyvelde, B. V. *et al.* Acoustic ejection mass spectrometry empowers ultra-fast protein biomarker quantification. *Nat. Commun.* **15**, 5114 (2024).

176. Bruderer, R. *et al.* Analysis of 1508 plasma samples by capillary-flow data-independent acquisition profiles proteomics of weight loss and maintenance. *Molecular and Cellular Proteomics* **18**, (2019).

177. Paramasivan, S. *et al.* Automated Proteomics Workflows for High-Throughput Library Generation and Biomarker Detection Using Data-Independent Acquisition. *J Proteome Res* (2023) doi:10.1021/acs.jproteome.3c00074.

178. Maciel, I. de S. *et al.* Plasma proteomics discovery of mental health risk biomarkers in adolescents. *Nat. Ment. Heal.* **1**, 596–605 (2023).

179. Dammer, E. B. *et al.* Multi-platform proteomic analysis of Alzheimer's disease cerebrospinal fluid and plasma reveals network biomarkers associated with proteostasis and the matrisome. *Alzheimer's Res Ther* **14**, 174 (2022).

180. Kim, H. *et al.* Development of a Fit-For-Purpose Multi-Marker Panel for Early Diagnosis of Pancreatic Ductal Adenocarcinoma. *Mol. Cell. Proteom.* **23**, 100824 (2024).

181. Garrett, M. E., Foster, M. W., Telen, M. J. & Ashley-Koch, A. E. Nontargeted Plasma Proteomic Analysis of Renal Disease and Pulmonary Hypertension in Patients with Sickle Cell Disease. *J. Proteome Res.* **23**, 1039–1048 (2024).

182. Sun, B. B. *et al.* Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).

183. Tanaka, T. *et al.* Plasma proteomic signature of age in healthy humans. *Aging Cell* **17**, 1–13 (2018).

184. Schmidt, I. M. *et al.* Plasma proteomics of acute tubular injury. *Nat. Commun.* **15**, 7368 (2024).

185. Cheishvili, D. *et al.* A high-throughput test enables specific detection of hepatocellular carcinoma. *Nat. Commun.* **14**, 3306 (2023).

186. Ward, B. *et al.* Deep Plasma Proteomics with Data-Independent Acquisition: Clinical Study Protocol Optimization with a COVID-19 Cohort. *J. Proteome Res.* **23**, 3806–3822 (2024).

187. Wahle, M. *et al.* IMBAS-MS discovers organ-specific HLA peptide patterns in plasma. *Mol. Cell. Proteom.* 100689 (2023) doi:10.1016/j.mcpro.2023.100689.

188. Ferkingstad, E. *et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet.* **53**, 1712–1721 (2021).

189. Lazar, J. *et al.* Large-Scale Plasma Proteome Epitome Profiling is an Efficient Tool for the Discovery of Cancer Biomarkers. *Mol. Cell. Proteom.* **22**, 100580 (2023).

190. Duijvelaar, E., Gisby, J., Peters, J. E., Bogaard, H. J. & Aman, J. Longitudinal plasma proteomics reveals biomarkers of alveolar-capillary barrier disruption in critically ill COVID-19 patients. *Nat. Commun.* **15**, 744 (2024).

191. Metatla, I. *et al.* Neat plasma proteomics: getting the best out of the worst. *Clin. Proteom.* **21**, 22 (2024).

192. Oh, H. S.-H. *et al.* Organ aging signatures in the plasma proteome track health and disease. *Nature* **624**, 164–172 (2023).

193. Geyer, P. E. *et al.* Plasma Proteome Profiling to Assess Human Health and Disease. *Cell Syst* **2**, 185–195 (2016).

194. Wehling, M. Principles of Translational Science in Medicine (Second Edition). 1–12 (2015) doi:10.1016/b978-0-12-800687-0.00001-3.

195. Rudd, K. E. *et al.* Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *Lancet* **395**, 200–211 (2020).

196. Singer, M. *et al.* The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* **315**, 801–810 (2016).

197. Marshall, J. C. Why have clinical trials in sepsis failed? *Trends Mol. Med.* **20**, 195–203 (2014).

198. Lodge, S. *et al.* Stratification of Sepsis Patients on Admission into the Intensive Care Unit According to Differential Plasma Metabolic Phenotypes. *J. Proteome Res.* **23**, 1328–1340 (2024).

199. Bhavani, S. V. *et al.* Identifying Novel Sepsis Subphenotypes Using Temperature Trajectories. *Am. J. Respir. Crit. Care Med.* **200**, 327–335 (2019).

200. Scicluna, B. P. *et al.* Classification of patients with sepsis according to blood genomic endotype: a prospective cohort study. *Lancet Respir. Med.* **5**, 816–826 (2017).

201. Antcliffe, D. B. *et al.* Transcriptomic Signatures in Sepsis and a Differential Response to Steroids. From the VANISH Randomized Trial. *Am. J. Respir. Crit. Care Med.* **199**, 980–986 (2018).

202. Bhavani, S. V. *et al.* Development and validation of novel sepsis subphenotypes using trajectories of vital signs. *Intensive Care Med* **48**, 1582–1592 (2022).

203. Xu, Z. *et al.* Sepsis subphenotyping based on organ dysfunction trajectory. *Crit Care* **26**, 197 (2022).

204. Anderson, N. L. & Anderson, N. G. The Human Plasma Proteome History, Character, and Diagnostic Prospects*. *Mol. Cell. Proteom.* **1**, 845–867 (2002).

205. Malmström, E. *et al.* Large-scale inference of protein tissue origin in gram-positive sepsis plasma using quantitative targeted proteomics. *Nature Communications* **7**, 1–10 (2016).

206. Uhlén, M. *et al.* The human secretome. *Sci. Signal.* **12**, (2019).

207. Mann, M. & Kelleher, N. L. Precision proteomics: the case for high resolution and high mass accuracy. *Proc Natl Acad Sci U S A* **105**, 18132–18138 (2008).

208. Furey, A., Moriarty, M., Bane, V., Kinsella, B. & Lehane, M. Ion suppression; A critical review on causes, evaluation, prevention and applications. *Talanta* **115**, 104–122 (2013).

209. Geyer, P. E., Holdt, L. M., Teupser, D. & Mann, M. Revisiting biomarker discovery by plasma proteomics. *Molecular Systems Biology* **13**, 942 (2017).

210. Viode, A. *et al.* A simple, time- and cost-effective, high-throughput depletion strategy for deep plasma proteomics. *Sci. Adv.* **9**, eadf9717 (2023).

211. Tu, C. *et al.* Depletion of Abundant Plasma Proteins and Limitations of Plasma Proteomics. *J Proteome Res* **9**, 4982–4991 (2010).

212. Gharibi, H. *et al.* A uniform data processing pipeline enables harmonized nanoparticle protein corona analysis across proteomics core facilities. *Nat. Commun.* **15**, 342 (2024).

213. Qian, W.-J., Jacobs, J. M., Liu, T., Camp, D. G. & Smith, R. D. Advances and Challenges in Liquid Chromatography-Mass Spectrometry-based Proteomics Profiling for Clinical Applications*. *Mol. Cell. Proteom.* **5**, 1727–1744 (2006).

214. Deutsch, E. W. *et al.* Advances and Utility of the Human Plasma Proteome. *J. Proteome Res.* **20**, 5241–5263 (2021).

215. Reymond, S., Gruaz, L. & Sanchez, J.-C. Depletion of abundant plasma proteins for extracellular vesicle proteome characterization: benefits and pitfalls. *Anal. Bioanal. Chem.* **415**, 3177–3187 (2023).

216. Liu, Z. *et al.* Enhanced Detection of Low-Abundance Human Plasma Proteins by Integrating Polyethylene Glycol Fractionation and Immunoaffinity Depletion. *PLoS ONE* **11**, e0166306 (2016).

217. Suhre, K. *et al.* Nanoparticle enrichment mass-spectrometry proteomics identifies protein-altering variants for precise pQTL mapping. *Nat. Commun.* **15**, 989 (2024).

218. Zhou, Y. *et al.* Optimizing and integrating depletion and precipitation methods for plasma proteomics through data-independent acquisition-mass spectrometry. *J. Chromatogr. B* **1235**, 124046 (2024).

219. Keshishian, H. *et al.* Quantitative, multiplexed workflow for deep analysis of human blood plasma and biomarker discovery by mass spectrometry. *Nat. Protoc.* **12**, 1683–1701 (2017).

220. Meier, F. *et al.* diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nature Methods* **17**, 1229–1236 (2020).

221. Kverneland, A. H. *et al.* Fully Automated Workflow for Integrated Sample Digestion and Evotip Loading Enabling High-Throughput Clinical Proteomics. *Mol. Cell. Proteom.* **23**, 100790 (2024).

222. Wallmann, G. *et al.* AlphaDIA enables End-to-End Transfer Learning for Feature-Free Proteomics. *bioRxiv* 2024.05.28.596182 (2024) doi:10.1101/2024.05.28.596182.

223. Batra, R. *et al.* Urine-based multi-omic comparative analysis of COVID-19 and bacterial sepsis-induced ARDS. *Mol. Med.* **29**, 13 (2023).

224. Rosenqvist, M. *et al.* Improved Outcomes After Regional Implementation of Sepsis Alert: A Novel Triage Model*. *Crit. Care Med.* **48**, 484–490 (2020).

225. Strålin, K. *et al.* Design of a national patient-centred clinical pathway for sepsis in Sweden. *Infect. Dis.* **55**, 716–724 (2023).

226. Rosenqvist, M., Fagerstrand, E., Lanbeck, P., Melander, O. & Åkesson, P. Sepsis Alert – a triage model that reduces time to antibiotics and length of hospital stay. *Infect. Dis.* **49**, 507–513 (2017).

227. Grieves, M. & Vickers, J. Transdisciplinary Perspectives on Complex Systems, New Findings and Approaches. 85–113 (2016) doi:10.1007/978-3-319-38756-7_4.

228. Blume, J. E. *et al.* Rapid, deep and precise profiling of the plasma proteome with multi-nanoparticle protein corona. *Nat. Commun.* **11**, 3662 (2020).

229. Wu, C. C. *et al.* Mag-Net: Rapid enrichment of membrane-bound particles enables high coverage quantitative analysis of the plasma proteome. *bioRxiv* 2023.06.10.544439 (2024) doi:10.1101/2023.06.10.544439.

230. Barker, M. *et al.* Introducing the FAIR Principles for research software. *Sci. Data* **9**, 622 (2022).

231. Poll, T. van der, Shankar-Hari, M. & Wiersinga, W. J. The immunology of sepsis. *Immunity* **54**, 2450–2464 (2021).

232. Calfee, C. S. *et al.* Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials. *Lancet Respir. Med.* **2**, 611–620 (2014).

233. Seymour, C. W. *et al.* Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis. *Jama* **321**, 2003–2017 (2019).

234. Aggarwal, C. C., Hinneburg, A. & Keim, D. A. Database Theory — ICDT 2001, 8th International Conference London, UK, January 4–6, 2001 Proceedings. *Lect. Notes Comput. Sci.* 420–434 (2001) doi:10.1007/3-540-44503-x_27.