Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

Next-Generation Mass Spectrometry for Clinical and Spatial Proteomics

Sophia Anna Victoria Steigerwald

aus
Mömbris, Deutschland

2024

Erklärung

Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Herrn Professor Dr. Matthias Mann betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, den 11.11.2024

Sophia Anna Victoria Steigerwald

Dissertation eingereicht am	11.11.2024
1. Gutachter:	Hon.-Prof. Dr. Matthias Mann
2. Gutachter:	Jun.-Prof. Dr. Florian Meier-Rosar
Mündliche Prüfung am	09.12.2024

Summary

While DNA provides the blueprint, proteins represent the functional and biologically active units of a cell. As such the proteome is our closest proxy to the phenotype, and can give important insights into cellular function and disease pathology. Although other approaches exist, mass spectrometry (MS) based proteomics remains the method of choice for fast, sensitive, quantitative, and high-throughput analysis of proteins. Over the years, MS-based proteomics has seen great advancements and now enables the routine analysis of thousands of clinical samples, near full proteomes and even single cells. A key factor in these advancements are innovations in MS technology that enable the instruments to push the boundaries of sensitivity, resolution, and acquisition speed. In this thesis I therefore first focus on evaluating MS technologies and optimizing MS acquisition strategies to expand the usability of MS instruments, and second to apply them to clinical and spatial proteomics.

Across the MS workflow, one can greatly improve performance by implementing novel technology, optimizing acquisition strategies and improving data analysis. In a first project, I evaluated the full mass range application of ΦSDM, a computational alternative to standard MS signal processing. By providing a two-fold increase in resolution or acquisition speed, as well as greatly improving signal-to-noise ratio, I showed that ΦSDM could be a useful addition to extend the potential of existing Orbitrap mass spectrometers for a wide range of proteomics applications. I then optimized a high-throughput acquisition strategy for plasma proteomics on a state-of-the-art LC/MS setup, which we applied to studying the effects of muscle loss in individuals undergoing bedrest in a study funded by the Italian Space Agency. While follow up is needed, the study identifies a potential biomarker candidate associated with muscle maintenance. To fully make use of the data obtained with state-of-the-art MS instruments and ever more complex data acquisition strategies, I contributed to benchmarking AlphaDIA, a modular, open-source framework for data independent acquisition data analysis developed in our lab.

I next contributed to applying novel MS technology for low input proteomics. The Orbitrap Astral, as well as other highly-sensitive TOF detector instruments have pushed the boundaries of sensitivity, acquisition speed and identification. This has shown to be particularly advantageous for low input applications such as Deep Visual Proteomics (DVP). Through a combination of these ultra-high sensitivity MS instruments, and tailored DIA acquisition strategies, we were able to decrease the required cell input

amount and broaden the range of application for DVP. Focusing first on tissues from a single patient with signet ring cell carcinoma, we showcased the potential of DVP for personalized medicine and were able to propose a treatment option that effectively halted tumor progression. We next evaluated the phenotypic shifts after xenotransplantation of organoid models. In a human mucosa model, we could show that xenotransplanted tissue was closer to human physiology and regained its functional profile in comparison to *in-vitro* organoid cultures and could provide valuable insights into human disease. Lastly, we extended the previously described single cell DVP workflow to formalin-fixed paraffin-embedded tissue, and applied it to study proteotoxic stress in alpha-1-antitrypsin deficiency. Using a tailored MS method with optimized variable DIA isolation windows, we were able to identify up to 3800 protein groups from a single hepatocyte shape, which is the equivalent to ~half of a full cell.

In summary, my thesis highlights how a combination of technical, methodological, and computational improvements can help to advance MS-based proteomics and bridge the gap to clinical applications and personalized medicine.

Table of contents

Abbreviations

AAT	Alpha-1-antitrypsin
AATD	Alpha-1-antitrypsin deficiency
API	Application programming interface
ART	Ataxia-telangiectasia mutated and Rad3-related
ASTRAL	Asymmetric track lossless analyzer
CCS	Collisional cross section
CEACAM	Carcinoembryonic antigen-related cell adhesion molecule
CID	Collision-induced dissociation
CV	Compensation voltage
DC	Direct current
DDA	Data-dependent acquisition
DDR	DNA damage response
DESI	Desorption electrospray ionization
DIA	Data-independent acquisition
DMS	Differential mobility spectrometry
DTIMS	Drift tube ion mobility spectrometry
DVP	Deep visual proteomics
EAD	Electron activated dissociation
eFT	Enhanced Fourier transformation
ESI	Electrospray ionization
ETD	Electron-transfer dissociation
FAIMS	Field asymmetric waveform ion mobility spectrometry
FDR	False discovery rate
FFPE	Formalin-fixed paraffin-embedded
FT	Fourier transform
HCD	Higher-energy collision dissociation
HDR	High dynamic range
HPLC	High performance liquid chromatography
HPR	Haptoglobin-related protein
HRAM	High-resolution accurate-mass
IEC	Intestinal epithelial cells
IM	Ion mobility
IMS	Ion mobility spectrometry
IRM	Ion routing multipole
IT	Ion trap
LC	Liquid chromatography
LFQ	Label-free quantification
LIT	Linear ion trap
LOD	Limit of detection
m/z	Mass-to-charge ratio
MALDI	Matrix-assisted laser desorption ionization
MMR	Miss-match repair

MR	Multi-reflector
MRM	Multiple reaction monitoring
MR-TOF	Multi-reflector time-of-flight analyzer
MS	Mass spectrometry
MUC	Mucin
OA	Orbitrap Astral
OE	Orbitrap Exploris
PASEF	Parallel accumulation – serial fragmentation
PD-L1	Programmed cell death ligand protein
PMT	Photomultiplier tube
ppm	Parts per million
PRM	Parallel reaction monitoring
PTCR	Proton-transfer charge reduction
PTM	Post-translational modification
Q	Quadrupole
QQQ	Triple quadrupole
R	Mass resolution
RF	Radio frequency
RT	Retention time
scDVP	Single cell deep visual proteomics
SNR	Signal-to-noise ratio
SPD	Samples per day
SRCC	Signet ring cell carcinoma
SRM	Selective reaction monitoring
TFS	Thermo Fisher Scientific
TIMS	Trapped ion mobility spectrometry
TOF	Time-of-flight
TWIMS	Traveling wave ion mobility spectrometry
UPR	Unfolded protein response
UVPD	Ultraviolet photodissociation
ΦSDM	Phase-constraint spectrum deconvolution method

« Le mieux est l'ennemi du bien »

To my family

1. Introduction

1.1 The human proteome

Nature's ingenuity is perhaps most evident in the intricate design of living cells, which form the foundation of all biological complexity. While all cells carry the same genetic information or genome, their individual functions and roles within a tissue or organism can greatly differ. In order to explain how genetic information is translated into functional diversity of living systems it is important to look at the different molecular components of a cell.

The genome represents the complete set of genetic information of an organism. Comprised of nucleotide sequences, the genome consists of only 1-2% protein coding genes, as well as non-protein coding genes. These can have regulatory, structural and other functional elements and seemingly non-functional elements, including “junk” DNA.[1–3] With the aim to use the genetic information to understand and potentially treat genetic or multifactorial diseases, the Human Genome Society established the Human Genome Project in 1990 to sequence the full human proteome.[4] After an initial draft in 2003, which was missing 8% of the genome, a first complete human reference genome was published in 2022, with additional information on the Y-chromosome following in 2023.[5–7] This complete reference genome, termed T2T-CHM13, encompassed more than 3 billion base pairs of nuclear DNA and the annotation lead to more than 63,000 genes of which close to 20,000 are predicted to be protein coding.

Through efforts of the Human Proteome Project, 18,397 or about 93% of these genome encoded or canonical proteins have been identified.[8] However the full human proteome is expected to consist of hundreds of thousands or even millions of protein species.[9–11] The portmanteau “proteome” was first coined at a conference in 1994 by scientist Marc Wilkins, who described it as “the protein complement expressed by a genome”, but now the term includes the set of all protein isoforms, modifications as well as protein-protein interactions and protein complex assemblies.[11–13] These discrepancies between canonical proteins and the total number of possible proteoforms arises from multiple regulatory mechanisms operating between the transcription of DNA and the translated protein. These biological processes include alternative splicing of mRNA transcripts, genetic variations such as single nucleotide polymorphisms, co-transcriptional mRNA

editing, and diverse post-translational modifications (PTMs). Each variation, or combination of such, yields a different proteoform, with potentially unique biological function. Highly adaptable to intrinsic and extrinsic stimuli and essentially the functional and biologically active unit of every cell, the proteome is the closest proxy to cellular phenotypes available.[14] Due to this close connection between the proteome and cellular function, diseases phenotypes often manifest at the protein level. This makes proteins ideal biomarker candidates for disease diagnosis, prognosis, treatment response, as well as therapeutic targets.[13,15–17] With more than 600 canonical proteins being target by FDA-approved drugs, and projected to represent half of the top ten selling drugs in 2023, protein-targeted therapies have revolutionized current treatment approaches.[18,19] While this includes important classes such as kinase and proteasome inhibitors used in the treatment of cancers, a notable and very recent example is the protein Semaglutide, a glucagon-like peptide-1 receptor agonist sold under the brand name Ozempic. Initially approved for the treatment of type 2 diabetes, it received much notice for its potential as an anti-obesity drug, with more promising treatment effects than other available medication.[20,21]

These clinical applications highlight the importance of investigating the proteome and shedding light on its dark side that lies beyond the canonical sequences.[22,23] While other approaches to study proteins exist, including gel electrophoresis and protein or antibody arrays, the clear advantages of mass spectrometry (MS) based proteomics have made it the method of choice for fast, sensitive, quantitative, and high-throughput analysis of proteins.[14,24,25]

1.2 Mass spectrometry-based proteomics

1.2.1 Principles of proteomic data acquisition

Mass spectrometry-based proteomics primarily uses two main approaches: bottom-up (shotgun) and top-down analysis. While bottom-up breaks proteins into small pieces and top-down analyzes whole proteins, a third approach called middle-down has emerged in recent years as an intermediate method (**Figure 1**).[14,26–29]

Top-down proteomics focusses on the analysis of intact proteins and omits any kind of proteolytic digest. Single protein or protein mixtures are directly injected and subjected to a full scan and subsequently fragmented for fragment ion scans (**Figure 1**, left). In comparison to other MS-based proteomics approaches, it provides complete protein sequence coverage and a holistic view of the proteoforms, including a high retention of

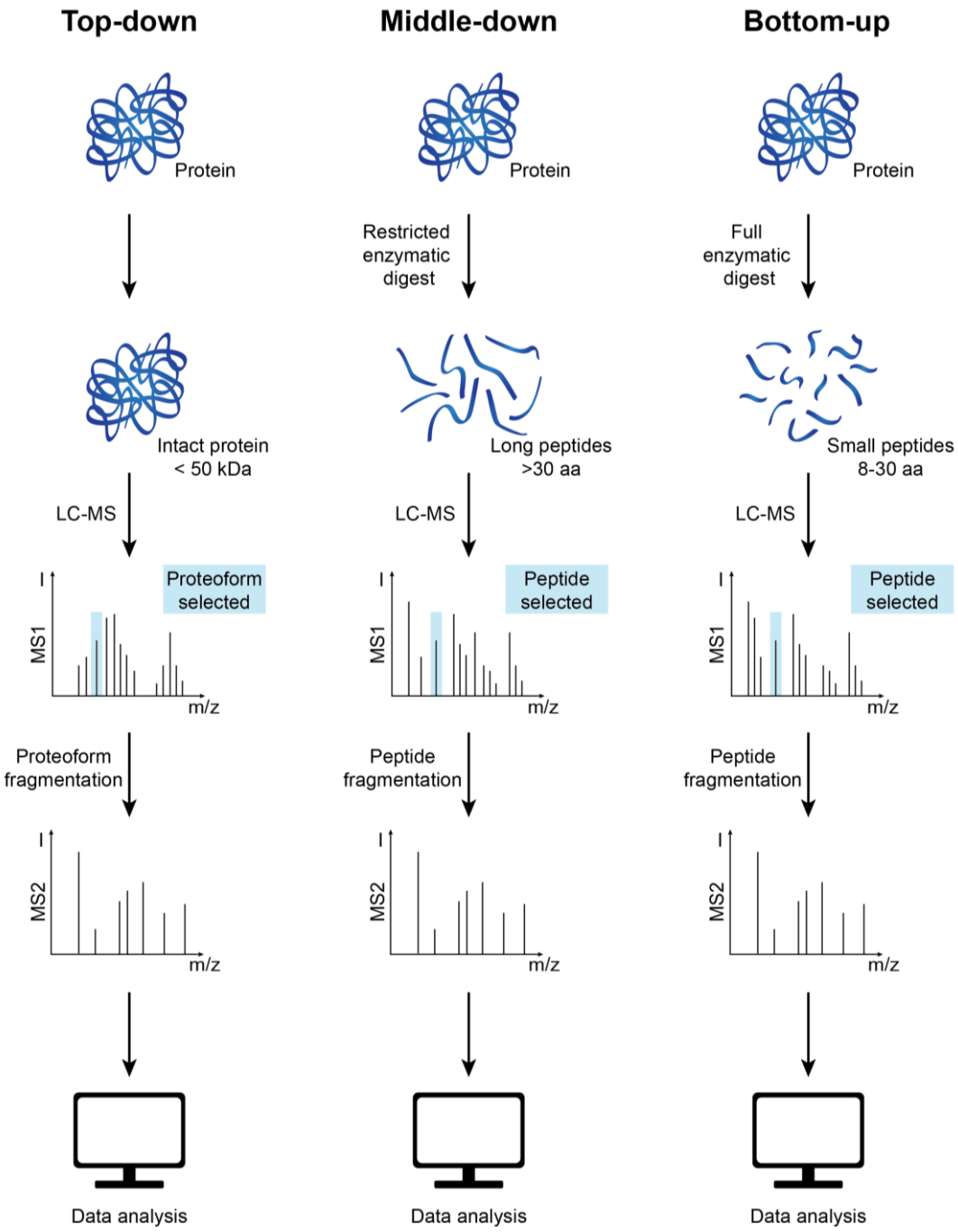


Figure 1 Schematic overview of mass spectrometry-based proteomic approaches. In both bottom-up (right) and middle down (middle) proteomics, proteins are proteolytically digested, resulting in small and large peptide fragments respectively. Digested peptides are separated by liquid chromatography and measured by mass spectrometry. Peptide and fragment levels are consequently used to infer protein information during data analysis. In top-down (left) proteomics intact proteins are injected, which allows for direct protein or proteoform-level information. Proteins are then fragmented prior to MS2 scans.

PTMs, which allows the analysis of co-occurring PTMs. Top-down proteomics, however,

is limited to the analysis of a few proteins at a time, with an additional upper limit on protein size, and data analysis is more complex.[26,30,31] In bottom-up and middle down proteomics, proteins are subjected to proteolytic digest (**Figure 1**, right and middle), however proteolysis in middle-down is restricted to achieve longer peptide fragments.[32–35] Using this strategy, middle-down proteomics can achieve higher sequence coverage than bottom-up proteomics and has an improved ability to characterize PTMs. As with top-down proteomics, however, throughput is limited and data analysis is more challenging.[28,34] While both top- and middle-down approaches have their benefits, the most common approach in mass spectrometry-based proteomics remains bottom-up or shotgun proteomics.[14,36] This can mainly be attributed to the high sensitivity and throughput this approach provides, as well as the ability to analyze complex samples, such as tissues or whole cell lysates. Additionally, more mature technology and data analysis tools make bottom-up more accessible and user-friendly than other MS-based proteomics approaches.[14,24,37–40] Bottom-up proteomic consists of three mayor steps, i) sample preparation, ii) liquid chromatography coupled to tandem MS (LC-MS/MS), and iii) data analysis (**Figure 2**).[41]

1.2.2 Sample preparation in bottom-up proteomics

Starting with sample preparation, proteins can be extracted from a plethora of biological material, including cell culture, body-fluids as well fresh-frozen or formalin-fixed paraffin embedded (FFPE) tissue samples. For effective protein extraction and improved enzymatic digest in tissue or cell culture samples, lysis buffers often contain protein denaturants, such as detergents. Commonly, sample lysis is followed by a reduction and alkylation step, where a reducing agent is used to disrupt disulfide bonds, followed by the alkylation of free cysteines.[42–44] Extracted proteins are then digested using sequence-specific proteases. Trypsin and LysC, the most commonly used proteases in bottom-up proteomics, cleave C-terminal to arginine and lysine residues, which results in peptides of 8-30 amino acid length and with a known proteolytic cleavage pattern. The cleavage pattern of trypsin and LysC leaves a positively charged amino acid on the C-terminal of the newly cleaved peptides, which increases subsequent ionization and fragmentation efficieny.[32,33,45,46] Prior to MS analysis, samples might require sample cleanup or can be subjected to various forms of offline fractionations for deeper proteomic depth.[47–52] Additionally, the analysis of PTMs often requires a separate and specific enrichment of modified peptides for optimal coverage.[53–55]

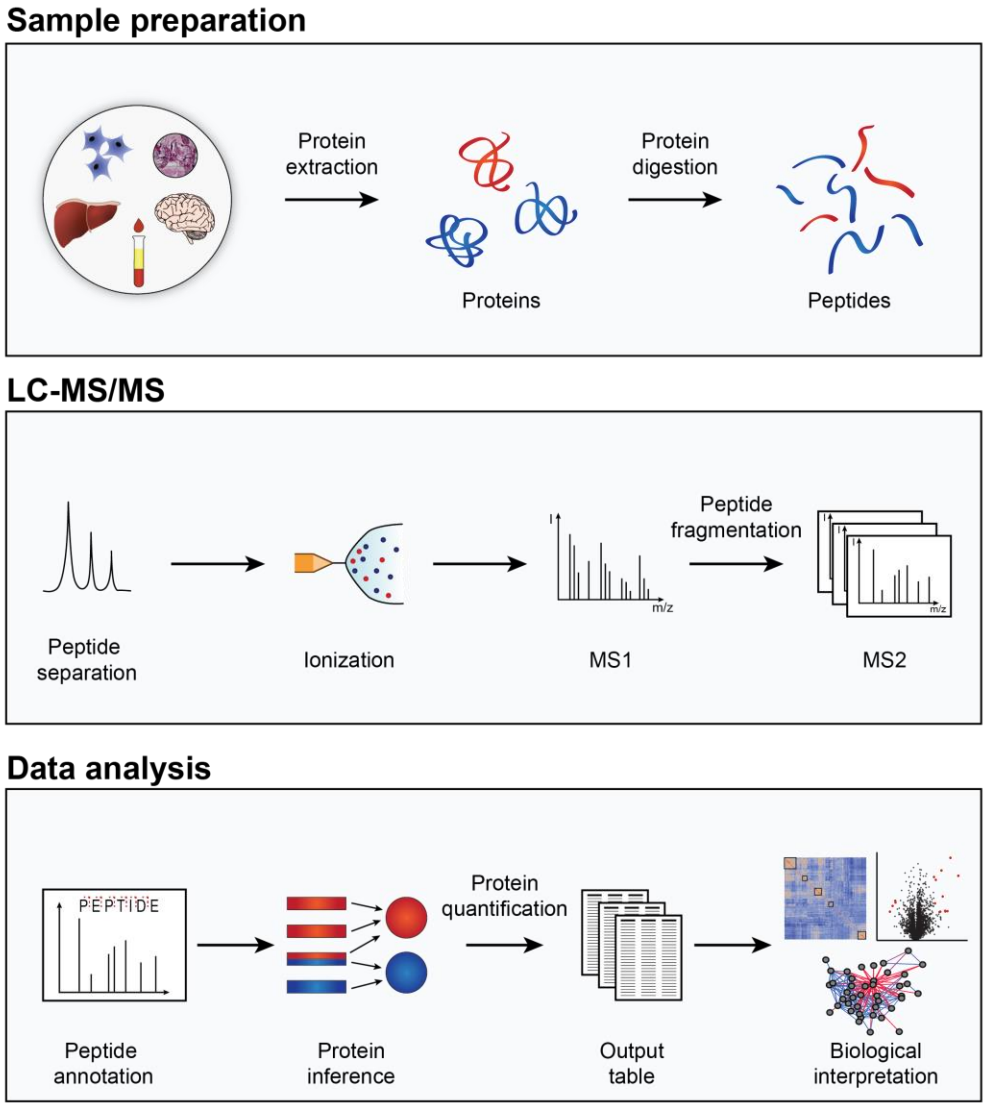


Figure 2 Schematic of the bottom-up proteomics workflow. The overall bottom-up workflow can be divided into three steps, i) sample preparation, ii) liquid chromatography coupled to mass spectrometry (LC-MS/MS) and iii) data analysis. In i) sample preparation, proteins are extracted from biological or clinical samples of interest, such as cell culture, body-fluids or tissues, including archived formalin-fixed paraffin embedded (FFPE) tissue. Extracted and solubilized proteins are enzymatically digested into peptides using trypsin, LysC or other proteases. In ii) LC/MS/MS, extracted peptides are separated using liquid chromatography and transferred to the mass spectrometer via electrospray ionization. First, the MS acquires a full mass spectrum (MS or MS1 spectra, before selected peptide precursor are fragmented for fragment ion scans (MS/MS or MS2 spectra). In iii) data analysis, the obtained MS1 and MS2 spectra are compared to a database to confidently identify peptides, infer protein sequences and quantify the identified proteins. These steps are commonly handled by bioinformatics tools. The resulting peptide or protein group output

tables are then used as the base for bioinformatic data analysis, statistics and data visualization. Adapted from ref.[41]

1.2.3 Liquid chromatography mass spectrometry

In a second step, the peptide mixture is separated based on their hydrophobicity using high performance liquid chromatography (HPLC). In reverse-phase LC, the complex peptide mixture is loaded onto an analytical column filled with porous silica beads that have C18 hydrocarbon chains attached to them - this forms the non-polar stationary phase commonly used in bottom-up proteomics. As peptides interact with these C18 chains through hydrophobic interactions, they can be gradually eluted from the column using increasing concentrations of a nonpolar solvent like acetonitrile.[56] This separation step helps reduce the complexity of the sample before it enters the mass spectrometer.

For optimal separation of peptides, analytical columns, however, need to be robust, reproducible and provide high chromatographic performance. As previous generations of commercial capillary columns were associated with high costs and short lifetimes, many labs, including ours, opted to produce their own in-house analytical columns.[57,58] With recent improvements in commercial column manufacturing, the trend however is moving towards a fully commercial plug-and-play setup from column producers such as PepSep, IonOpticks and Thermo Fisher Scientific. Apart from packed columns, micro-pillar array columns show great potential for applications in proteomics by reducing peak broadening.[59–61] The so called µPAC columns feature perfectly positioned and geometrically ordered micropillars, which are etched into silicon wafers and form separation channels.[62,63]

The drive for more reliable and consistent results has led to new choices in LC instruments. One example is the Evosep One LC system, which offers preset, short gradients for consistent results while maintaining high sensitivity. By operating at low pressure, it reduces equipment wear and extends operating time. The system performs multiple steps simultaneously between samples, allowing throughput of up to 500 samples per day.[64] It also uses disposable trap columns called Evotips that extend the main column's life and, in most cases, alleviate the need for additional sample clean up. These Evotips reduce sample handling and potential sample loss, which is especially valuable when working with small sample amounts.

Peptides at this stage are optimally separated; however they now need to be injected into the MS. A crucial step in MS-based proteomics that was revolutionized by the

introduction of electrospray ionization (ESI) source - lead by the team of John Fenn in 1989[65] - that earned John Fenn a joint Nobel Prize in chemistry in 2002. In ESI, analytes in solution are pumped through a capillary, which is maintained at a high voltage, and nebulized at the capillary tip (**Figure 3**). This leads to the dispersion of charged droplets, which are rapidly evaporated and undergo coulomb fission once the electrostatic repulsion outweighs the droplet surface tension, and the consequent transfer of residual charges to the analytes.[66,67] These ionized analytes are then moved into the high vacuum chamber of a mass spectrometer. Based on the initial principle, many improvements have been made to increase the efficiency, such as the introduction of a nanoESI source, which additionally enables the use of ESI for low flow gradients.[67–70] In a standard bottom-up MS run, the MS is operated in positive mode, meaning the emitter is maintained at a positive potential, and ionization of analytes happens through protonation.[67] The number of charges a peptide carries can depend on the experimental conditions as well as peptide length and amino acid sequence.[71–73] Tryptic peptides generally carry at least two charges, though non-tryptic digestion or specialized applications, such as immunopeptidomics, can also give rise to singly charged species.[33,74–76]

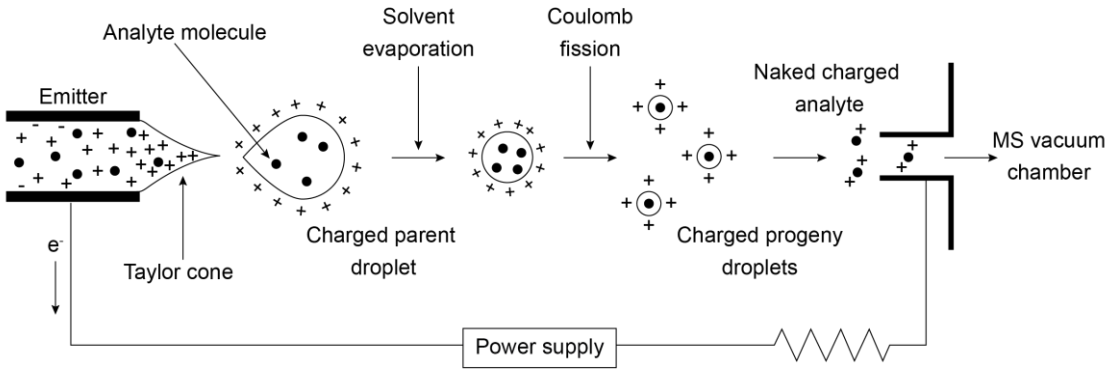


Figure 3 Schematic of electrospray ionization (ESI). Analytes in solution are exiting a capillary column or emitter in an electric field. At the orifice of the emitter, an electrospray or Taylor cone is formed and the analyte solution is nebulized. This leads to the formation of initially larger, charged parent droplets. Through solvent evaporation, the size of the droplet is reduced until the Rayleigh limit is reached, where electrostatic repulsion of like charges outweighs the droplet surface tension, and droplets undergo Coulomb fission to form smaller progeny droplets. This process continues until only the naked charged analyte remains. Adapted from ref.[67]

A mass spectrometer is essentially composed of three main components: an ion source to ionize the analytes, a mass analyzer to determine the “weight” or mass-to-charge

(*m/z*) ratio of the analytes, and a detector, which counts the number of ions at a given *m/z* value.[46,77] Having covered ESI as the most common ionization strategy in the previous section, the next is the mass analyzer, whose principle role is to separate analytes based on their *m/z* ratios. The most commonly used mass analyzer types include quadrupoles, linear ion traps, time-of-flight (TOF) analyzers, Fourier transform ion cyclotron resonance, and the Orbitrap analyzer.[78–85] Different analyzer types have their strengths and weaknesses with regards to analyzer performance factors, such as sensitivity, resolution, mass accuracy, and speed.[46,79,80] For optimal performance, different analyzer types are often combined in so called tandem mass analyzer approaches.[79,86] A common example being the combination of quadrupoles, for ion package selection, with more advanced analyzers such as TOF or Orbitrap analyzers. Regardless of the analyzer, precursor peptides are first profiled in a full MS or MS1 scan. Then precursors are selected for fragmentation and fragment ion scans (MS2) are recorded. Depending on the fragmentation technique used, different types of ions series are produced (**Figure 4**).[87,88]

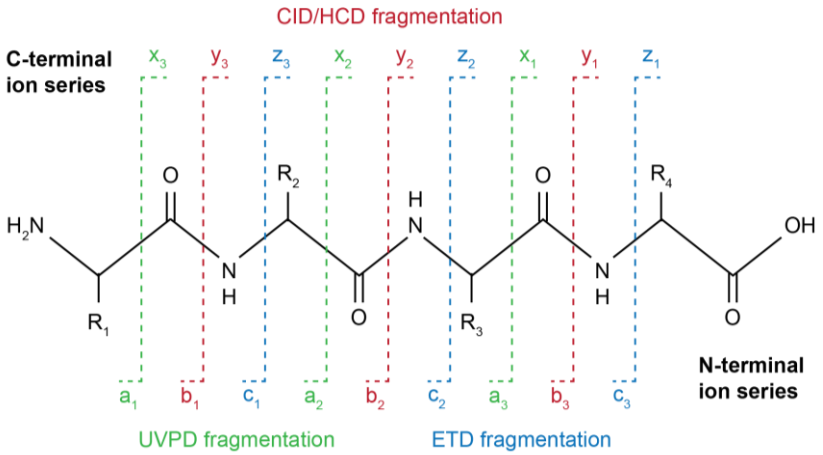


Figure 4 Peptide fragmentation pattern in mass spectrometry. Roepstorff-Fohlman nomenclature for the fragmentation of protonated peptides. The potential cleavage points along the peptide backbone are referred to as A, B, C or X, Y, Z depending on whether the charge retention is on the N- or C-terminal peptide respectively. Collision-induced dissociation (CID), including higher-energy collisional dissociation (HCD) produce b and y ions, while alternative fragmentation techniques such as electron-transfer dissociation (ETD) and ultraviolet photodissociation (UVPD) produce complementary c/z and a/x ion series, respectively. Adapted from ref.[87]

Bottom-up approaches generally rely on higher-energy collisional dissociation (HCD), a type of collision-induced dissociation (CID), for fragmentation, which yields b and y ions. Other fragmentation techniques such as electron-transfer dissociation (ETD) and ultraviolet photodissociation (UVPD) can produce complementary ions, which are particularly beneficial for top-down MS or the analysis of labile PTMs.[89–91]

The selection of precursors for fragment ion scans is a crucial step in mass spectrometry. In discovery proteomics, we can differentiate between data-dependent (DDA) and data-independent acquisition (DIA) (**Figure 5, left and middle**). As the name data-dependent acquisition suggests, precursor selection in DDA relies on information from MS1 scan. The n most abundant peptide precursors (topN) are sequentially isolated, subjected to fragmentation and MS2 scans of the corresponding fragment ions are recorded. This establishes a clear connection between a precursor and its fragments, but the stochastic nature of the precursor selection reduces reproducibility, which leads to a greater number of missing values across replicates. Moreover, coverage of the dynamic range of a mass spectrum is limited by the number of topN peaks that can be selected.[92,93]

In contrast to DDA, data-independent acquisition successively cycles through the entire mass range using a set of pre-defined isolation windows. Within these isolation windows, all detectable precursors are co-isolated and fragmented. This overcomes the dynamic range and reproducibility limitations of DDA, and can greatly increase proteomic depth.[94–97] However, these advantages come at the cost of the loss of the direct precursor-fragment relationship and increased spectral complexity, which requires more advanced search engines to process the obtained data.[94,95,98,99] In recent years multiple such software suites have been released to effectively and confidently analyze DIA data, each with their own advantages and disadvantages.[39,40,98,100–104]

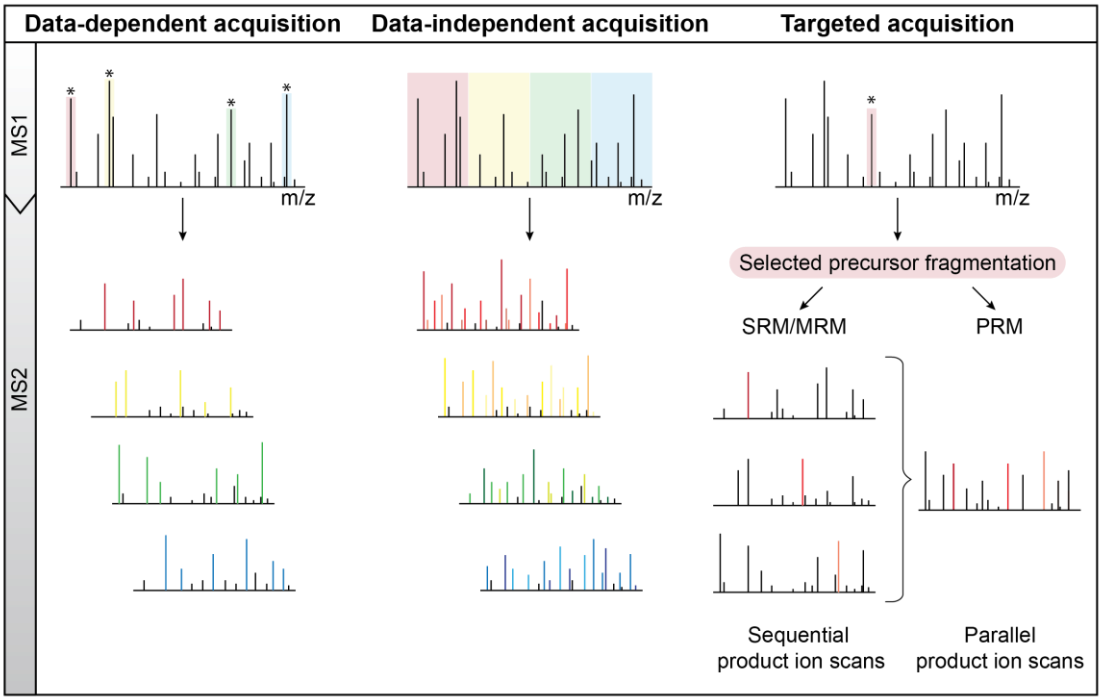


Figure 5 Overview of data acquisition modes. In discovery-based proteomics, the goal is to cover as wide a range of peptides and proteins as possible. This can be achieved with two data acquisition modes, data-dependent (DDA) and data-independent acquisition (DIA). In the former, the topN most abundant precursor of a given spectrum are sequentially isolated and subjected to fragmentation before MS2 scans of the corresponding fragment ions are recorded. In DIA the *m/z* range of the MS1 spectrum is divided into *m/z* isolation windows of predefined size. Within these windows all precursors are co-isolated and fragmented leading to a higher coverage of the precursors present at a given time at the expense of increasing spectral complexity. In contrast to discovery proteomics, targeted proteomics aims to specifically monitor a smaller number or peptides or proteins of interest. Based on a predefined target list, precursors are selected for fragmentation. Here we differentiate between selected or multiple reaction monitoring (SRM/MRM), in which a certain number of peptide fragments of a given precursor are analyzed separately, and parallel reaction monitoring (PRM), in which many or all fragment ions of a given precursor are analyzed in parallel.

Discovery proteomics can give comprehensive insights into the proteome and help identify proteins and peptides of interest for biological or clinical application, such as disease biomarkers. Once these proteins of interest have been identified, researchers can develop specialized MS-based assays to track these proteins – so-called targeted proteomics. In contrast to discovery proteomics, the goal in targeted proteomics is to specifically monitor a set of proteins or peptides of interest. Based on a pre-defined target list, precursors are selected for fragmentation and product ion scans are recorded (**Figure 5, right**). Fragment ions can either be analyzed sequentially (SRM) or in parallel (PRM), depending on the mass analyzer used.[105–110]

1.2.4 Computational mass spectrometry for data analysis

With the obtained mass spectrometry raw data in hand, we reach the data analysis step of the general bottom-up proteomics workflow. As protein-level information is lost during digestion, the identification of proteins from bottom-up proteomics samples is a complex task. It requires matching the experimental MS2 (fragment) spectra to theoretical, library or predicted fragment spectra to identify peptide precursors, inferring proteins (or protein groups) from the identified peptides and quantifying the assembled proteins.[46,111] Peptide identification can be achieved in three ways: *de novo* sequencing, database search approaches, or spectral library matching. In *de novo* sequencing peptide sequences are directly read out of the MS2 spectra without the use of a reference database. This is done using *de novo* sequencing algorithms, which by now often employ deep learning, and reconstruct the peptide sequence by interpreting mass differences between adjacent fragment ions in MS/MS spectra. Each mass increment corresponds to a specific amino acid residue mass, wherefore this systematic mass analysis along the peptide backbone enables sequential amino acid assignment.[112–115] While this can be of interest for studying proteoforms or proteomes of organisms without a complete reference genome, *de novo* sequencing has a lower accuracy and depth compared to database-assisted search strategies. Additionally, search parameters, such as peptide length, charge states and modifications, need to be limited as not to inflate the search space.[116]

As information content and spectral complexity between DDA and DIA differ, the data obtained by these two data acquisition strategies need to be handled differently. DDA data is conventionally analyzed using spectrum-centric approaches, in which MS2 spectra are matched against reference proteome databases or a spectral library.[37,117–120] Most common DIA analysis tools on the other hand employ targeted peptide-centric approaches, which query whether a predefined list of peptides from a spectral library are detectable in the extracted ion chromatograms of the experimental data.[101,121,122] This classical approach, however, requires the generation of an experimental spectral library by acquiring deep proteomes of the target organism using DDA, which can be tedious and time-intensive. So-called library-free approaches overcome the necessity for an experimental library. These strategies involve converting DIA data into pseudospectra that resemble DDA fragmentation patterns, enabling direct analysis with established database search algorithms. Among others, notable examples include DIA

Umpire or the directDIA search approach in Spectronaut.[98,101,121,123,124] In contrast, novel algorithms, which employ machine learning or deep learning for the prediction of peptides and peptide properties, allow the generation of tailored and fully *in-silico* predicted libraries. This enables the peptide-centric search of all possible peptides and precursors beyond the depth of experimental libraries. Moreover, these tailored libraries can reduce search space in specialized applications such as immunopeptidomics by decreasing the number of peptides in the spectral library to sequences likely present in the sample.[40,125–129]

Most of these software suites, however, are of a closed nature, meaning the source code and with it the details on how the search engine goes from MS raw data to a list of quantified proteins is unavailable to the user. With new version releases claiming ever higher identification rates from the same sample set, this can raise concerns about the accuracy and confidence in these identifications. While there have been many discussions recently about closed versus open-source proteomics tools, particularly in connection with academic software commercialization, open-source proteomics software has the potential to recover this trust in protein identifications. Moreover, an open-source concept invites contributions beyond the source lab, enabling a faster implementation of new features and functionalities. Since the code is freely available, developers can rapidly update these tools to process data from new and complex mass spectrometry scanning methods that traditional software cannot handle. One such example - AlphaDIA (Article 3) a modular framework for the analysis of DIA data[130] developed in our group - is highlighted in this thesis. Apart from being a fully open source DIA search implementation, its main advantages are a feature-free identification algorithm, which makes it particularly suitable for data produced by current state-of-the-art TOF analyzers, and its end-to-end workflow using AlphaPeptDeep for library prediction and directLFQ for quantification.[126,131]

During the protein inference, peptides are then assembled into proteins. As peptides are often not unique, but rather can be assigned to a few different proteins, it is necessary to introduce protein groups as not to inflate the number of identifications. If multiple proteins share the same peptides and no uniquely distinguishing sequences have been identified, these are assembled into a protein group.

Importantly, at both the peptide and protein identification level, the false discovery rates (FDR) should be controlled.[132] This is commonly done using target-decoy approaches that help estimate the FDR. For this, decoys, such as reversed or scrambled sequences

are added to a target list (e.g., the reference proteome of an organism of interest). Obtained mass spectrometry data is searched against this combined database and, based on the identification rate of target and decoy sequences, the FDR can be calculated. At both the peptide and protein level, an FDR cutoff of 1% is proposed for maximum confidence in identifications.[132,133]

The advantage of MS-based proteomics, however, is the ability to not just identify, but also quantify proteins. Protein quantification can either be achieved label-free (LFQ) or using isobaric or non-isobaric labels. For quantification, LFQ directly uses the integrated intensity of peptide peaks across the *m/z* and retention time (RT) dimensions. The core principle is that when measuring a given peptide multiple times, the relative proportions of its ions should remain consistent between multiple LC/MS runs. LFQ algorithms then compare the peptide signals across different experimental conditions and normalize the signals by using median-fold changes to calculate relative protein abundances or intensities.[131,134–136]

In comparison to LFQ, labeling or multiplexing strategies add distinct tags to proteins or peptides, depending on when the labeling step is incorporated in the sample preparation workflow. These tags create predictable mass differences between otherwise physio-chemically identical peptides and allow the differentiation and quantification of these peptides. Labeling strategies also allow sample multiplexing, where peptides from multiple samples are combined and analyzed in a single LC-MS run. This approach increases analytical throughput while reducing technical variability, as all experimental conditions are measured simultaneously. While these benefits often only hold true as long as all experimental conditions can be processed together in a so-called plex, one can also use one of the multiplexing channels to normalize between sets of multiplexed samples.

Labeling strategies for LC-MS usually fall into one of two categories, isobaric and non-isobaric labels. Isobaric labeling techniques, like TMT, iTRAQ, and EASI-tag, use chemical tags composed of reporter ions and balancing or equalizing groups.[137–140] During fragmentation, the reporter ions are released and used for quantification across the experimental conditions, while the balancing groups ensure identical precursor masses across different labels. While these are powerful tools with ever-increasing multiplexing capabilities, they suffer drawbacks in terms of ratio-compression and are

usually associated with high costs in comparison to LFQ sample preparation.[141–143] Non-isobaric labeling methods, such as SILAC, mTRAQ and dimethyl labeling do not rely on reporter ions, but rather use the inherent mass differences between the employed labels to distinguish between the differently labeled samples and quantify them from MS1 scans.[144–149] However, the addition of such labels can introduce shifts in RT, that need to be taken into the account by the analysis software. Moreover, they require near perfect labeling efficiency, as un-labeled peptides are not considered for downstream data processing. While some of these multiplexing strategies can only be used for DDA analysis, many of the mentioned approaches have been adapted for use with DIA in recent years.[150–155]

Once the MS raw data has been processed, the search algorithm results in a list of quantified protein groups and peptides, which can be used as the input for the last but certainly not least step of data exploration, interpretation and visualization. For this purpose, a plethora of bioinformatic tools have been developed that provide a framework for statistical data analysis and biological or clinical interpretation. Perseus, MSStats, and AlphaPeptStats, to name a few examples, are easing statistical analysis by having ready to use implementation of common statistical analysis in a user friendly graphical user interface or as assembled packages for coding languages such as R and Python.[156–159] For additional data visualization, a multitude of packages is available.[160,161] While these tools allow the visualization of proteomics data in a quantitative protein or peptide centric-view, it is also important to evaluate data quality at the level of MS raw data and peptide matching.[162–164] This not only enables to manually confirm peptide identifications on a spectrum level in case of low evidence, but is also important when evaluating novel MS acquisition methods or technology (software and hardware components) that directly impact the MS raw data. For one of the works presented in this study (Article 1), we used such a tool, called AlphaRaw, to analyze distances between neighboring peaks as a proxy for the resolving power provided by a novel raw data processing algorithm.[164,165]

1.3 Mass spectrometry technology

Mass spectrometry stands as one of the most important analytical technologies in proteomics and beyond. From the invention of what is now recognized as the first mass spectrometer by J.J. Thomson in 1912, to modern mass spectrometers that enable the analysis of full cellular proteomes in an hour, the field has undergone a remarkable evolution.[166–168] At the forefront of this evolution are the incredible technological advancements in mass spectrometry technology that are continuing today, striving to

make mass spectrometers ever more sensitive, fast, precise, and robust. Before we delve into the latest MS innovations, it is important to first introduce a few key MS terms (Table 1).

Table 1 Glossary of key mass spectrometry terms

Cycle time	Total time needed to complete a MS analysis cycle (MS1 + MS2 scans).
Duty cycle	Proportion of time the mass spectrometer spends collecting useful data.
Dynamic range	Range between the most and least abundant peak. Can be determined on an intra- or inter scan level.
Fill/Injection time	Time allowed for the accumulation of ions before analysis.
Mass accuracy	Difference between measured and theoretical *m/z* value of an ion. Typically expressed in parts-per-million [ppm].
Mass range	Range of *m/z* values that can be analyzed.
Mass resolution	Ability to distinguish between closely spaced peaks. Usually expressed as m/Δm (mass divided by peak distance).
Scan speed/rate	The number of spectra that can be acquired per unit time. Often expressed in hertz [Hz].
Sensitivity	Defines the minimum amount of sample needed for detection. Often expressed as a limit of detection or compared through signal-to-noise ratios.
Transmission efficiency	Percentage of ions that are successfully transferred through the instrument.

While in an ideal world, a mass spectrometer would combine the best available components, commercial instrument development is constrained by vendors' patents on specific technologies, including hardware designs, software solutions, and scan modes. For this reason, the instrument platforms from each vendor differs in core technologies, components and as a result in performance.

As mass spectrometry technologies from Thermo Fisher Scientific (TFS) played a pivotal role in my PhD and journey through MS-based proteomics, I will primarily focus on MS instrumentation and related innovations from this vendor in the next chapters.

1.3.1 A brief history of mass spectrometry technology in Bremen

Thermo Fisher Scientific sells a wide range of life science mass spectrometers from single, to triple quadrupole to linear ion trap instruments to their range of hybrid or Tribrid instruments. Hybrid instruments generally pair a quadrupole with the Orbitrap as a high-resolution accurate mass (HRAM) analyzer, while Tribrid instruments have a secondary mass analyzer in addition to the Orbitrap. In these pairings the quadrupole is generally only employed for mass selection, meaning it selects or filters ions based on their *m/z* values for downstream analysis. The available range of instruments is being manufactured across two factory sites in Germany and the USA. During my PhD, as well as during my master thesis, I had the opportunity to collaborate with the research teams at the Bremen factory.

With not just TFS, but also Bruker Daltonics having their factories in Bremen, mass spectrometry technology has a long history in there. Working on electromedical instrumentation, physicist Ludolf Jenckel decided to build a mass spectrometer in 1947, and was able to start a small division Atlas MAT with the aim to commercialize MS instruments in Germany and thereby starting a now 77-year long journey of MS innovation. Based on the quadrupole ion trap design of Wolfgang Paul, who would later go on to receive the Nobel prize in Physics in 1989 for his development of the ion trap technique, MAT introduced a commercial quadrupole analyzer in 1962.[169] Initially underestimating its potential, MAT was prompted by the success of the first quadrupole mass spectrometer from the US-based MS company Finnigan to introduce their own quadrupole MS, the MAT 44, in 1977. This instrument featured a, for the time unprecedented, resolution of 12,000, which could be attributed to the use of hyperbolic quadrupole rods instead of round ones as was custom at the time. After the fusion of MAT and Finnigan, these hyperbolic rods became a core technology in Finnigan's Quadrupole MS until the 1990s. This Fusion also marked the start of the scientific collaborations between San Jose (Finnigan) and Bremen (MAT), which continues to this day after Finnigan was acquired by ThermoElectron (later Thermo Fisher Scientific). Their main competitor at the time was Vacuum Generators, a British MS company in Manchester that after multiple changes in ownership was acquired by Thermo Instruments. This prompted the formation of HD Technologies, a new company which took over some of the former Vacuum Generators' operations in Manchester. The same

company employed Alexander Makarov, who developed the Orbitrap technology.[170] After HD Technologies was acquired by TFS in 2000, Alexander Makarov and the development of the Orbitrap technology relocated to Bremen. This lead to the release of the first commercial Orbitrap mass spectrometer, the LTQ Orbitrap, in 2005 and made the Orbitrap technology the foundation for TFS's high resolution mass spectrometers.[171–176]

Build upon the learnings of previous ion trap designs such as the paul trap (i.e. quadrupole), the kingdon and knight traps, as well as Yuri Golikov's theory of ion motion in quadro-logarithmic potential, the Orbitrap mass analyzer revolutionized the field of mass spectrometry.[81,169,170,177–179] The Orbitrap consists of two outer barrel-like and a central spindle-shaped electrodes that form an quadro-logarithmic electrostatic field between them (**Figure 6**). During injection, the direct current (DC) applied to the inner

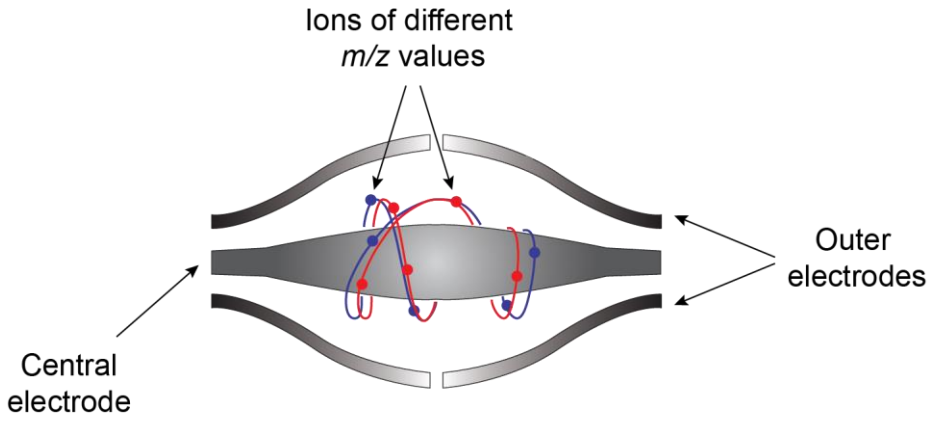


Figure 6 Schematic of the Orbitrap design and trapped ion movement. The Orbitrap mass analyzer features a spindle-shaped central electrode, which is surrounded by two split barrel-like outer electrodes that form an electrostatic field. Ions are injected tangentially and begin to orbit around the central electrode, while additionally oscillating back and forth along its length. The frequency of an ion's oscillation motion is proportional to its *m/z* value. The oscillation motion induces an image current at the split outer electrodes and the recorded raw image current is composed of the sine waves of all ions present in the Orbitrap. Fourier transformation is used to decompose the convoluted signal to create a frequency spectrum, which can be converted to a mass spectrum.

central electrode is ramped up quickly to contract the radius of the orbiting ions. A principle that is referred to as ion "squeezing" and prevents the ions from hitting the outer electrode at the opposite side of the Orbitrap during injection.[180] In the electrostatic field

between electrodes, the ions are then kept on an even distance from the central electrode due to an equilibrium of electrostatic attraction to the inner electrode and the centrifugal force. Ions are injected tangential to the central electrode and start orbiting it while harmonically oscillating back and forth along its length.[81,180–182] The frequency (ω) of these oscillations can be described as:

$$\omega = \sqrt{k\frac{z}{m}},$$

where *k* is a constant, *z* is the number of charges (or charge state), and *m* the mass of an ion. The frequency of ion motion therefore is a function of each ion's mass-to-charge or *m/z* value. If ions are introduced to the electrostatic field in a small temporal and spatial window, ions of the same *m/z* will oscillate together, while ions with different *m/z* values will oscillate at higher or lower frequencies. In all commercial Orbitrap instruments, this tight requirement on kinetic energy, as well as the temporal and spatial spread is achieved using the "C-trap", a curved radio frequency (RF)-only quadrupole with an opening in the electrode closest to the Orbitrap. In the C-trap ions are accumulated and subjected to collisional cooling. Effective injection of ions in small ion packages into the Orbitrap is achieved by rapidly ramping down the RF amplitude and applying direct current (DC) gradients across the C-trap.[176,180] The axial oscillations of ion rings are detected via image current, as the oscillating ions induce current on the outer split electrodes. All ions inside the Orbitrap at a given time induce current concurrently and the sum of these individual sine waves as a function of time produces the raw image current or "transient". Fourier transformation is used to deconvolute the raw image current into its various frequencies, providing a frequency spectrum that can be converted to a high-resolution mass spectrum.[183,184] Mass resolution (R) is defined as the minimum distance between two *m/z* values the analyzer can resolve and is therefore directly linked to the frequency resolution:

$$R = \frac{\omega}{2\Delta\omega} = \frac{1}{2\Delta\omega}\sqrt{k\frac{z}{m}} = \frac{\mathrm{m}}{\Delta m}$$

Two factors impact the resolution: the mass range and the timespan for which the transient is recorded. First, in Orbitrap mass spectrometry, the resolution is inversely proportional to the square root of *m/z*. The highest resolution can thus be achieved for low *m/z* ratios. As the resolution is not stable across the mass range, usually a nominal resolution at e.g., *m/z* of 200 is given. Second, for Fourier transform analysis, a longer transient, i.e., a longer time span in which the image current is recorded, enables a more fine-grained discrimination between frequencies and therefore *m/z* values. This can be

explained by the fact that sine waves of similar but not identical frequency will become increasingly out of sync the longer they are observed. From a practical point of view this means, longer transients equal higher mass resolution. Even in the initial publication from the year 2000, a resolution of 150,000 could be achieved.[81] However, one should note that in MS instruments with only a single mass analyzer, longer transients come at the cost of decreased duty cycle and scan rate.

Over the years, the Orbitrap technology has been continuously improved with regards to the achievable resolution, speed, and mass range. One such development was the introduction of the high-field Orbitrap. A reduced distance between inner and outer electrodes strengthened the electric field and increased the frequency of ion oscillations and with it the mass resolution at a given acquisition time (transient length).[185] The achievable speed was then gradually improved up to 40 Hz (MS2 acquisition) in the successively released Q-Exactive HF and Q-Exactive HF-X.[186,187] Improvements in the detection and processing steps could additionally improve resolution. So called "enhanced Fourier transformation (eFT)" could increase mass resolution by a factor of two for most experiments and a factor of 1.4 for rapidly decaying signals, such as the signals of intact proteins.[188] Overall, the Orbitrap marks a key invention in MS technology that launched MS into the modern era and considerably accelerated the pace of discovery. While today's Orbitrap analyzer appear as an elegant solution for mass analysis, Alexander Makarov's cabinet, or "museum" of failed prototype Orbitrap electrode assemblies at the TFS factory in Bremen, highlights the importance of perseverance in scientific innovation.

1.3.2 Modern mass spectrometry innovations

After delving into the history of mass spectrometry in Bremen, particularly the history of what today is Thermo Fisher Scientific, as well as introducing the Orbitrap, one of the key components of TFS's high resolution accurate mass MS instruments, I would like to focus on TFS's innovations in mass spectrometry and related technologies. While I had contact points with mass spectrometry and particularly mass spectrometry-based proteomics throughout my university studies, my personal hands-on journey in mass spectrometry began in 2019, which is roughly where I'd like to begin.

Field asymmetric waveform ion mobility mass spectrometry

Field asymmetric waveform ion mobility mass spectrometry (FAIMS) is a type of differential mobility spectrometry (DMS) that can be operated at atmospheric pressure. Ions in gas-phase are separated depending on their behavior in strong and weak electric fields.[189,190] FAIMS is often characterized by a curved or cylindrical electrode geometry, in contrast to generally planar DMS technologies. Interfaced with electrospray ionization, FAIMS can be used as an additional on-line orthogonal separation/fractionation between LC and MS. [191,192] Commercialized in the early 2000s, it was first used in the form of a front end accessory for SCIEX mass spectrometers, before a temperature controlled version was implemented for TFS's triple quadrupole MS in 2007.[192,193] While SCIEX moved forward with a planar geometry, TFS build upon the cylindrical design featuring an outer and inner electrode, where the asymmetric waveform is applied to the inner electrode (**Figure 7**), and released an updated commercial interface, the FAIMS Pro, for the use with their Tribrid MS instruments in 2018. Initially optimized for low flow applications, its functionality was extended to high flow application.[194,195]

Analytical Principle: Carried along by a carrier gas, ions enter the space between the two electrodes to which an asymmetric high-voltage alternating current, the so-called dispersion voltage, is applied. As the electric field continuous to alternate, ions transverse the space between electrodes in a "zigzag" motion (**Figure 7A**). If an ion exhibits differential mobility in the high vs. low field, it will eventually collide with one of the two electrodes. Therefore, only ions with the same mobility across the alternating field will be transmitted. For selective separation, an additional direct current termed compensation voltage (CV) is applied that offsets the dispersion voltage and stabilizes the flight path of specific ion packages (**Figure 7B**). Mobility in the FAIMS dimension is influenced by a multitude of factors, including peptide length, charge state, shape, center of mass.[196] As the optimal CV can differ between sample types as well as injection amount and instruments, predetermining the optimum is recommended.[197–199] While this is usually achieved through a so-called CV-sweep by injecting the sample multiple time and acquiring data at different CV values, prediction models to infer optimal CV values from peptide sequences have recently been proposed.[200]

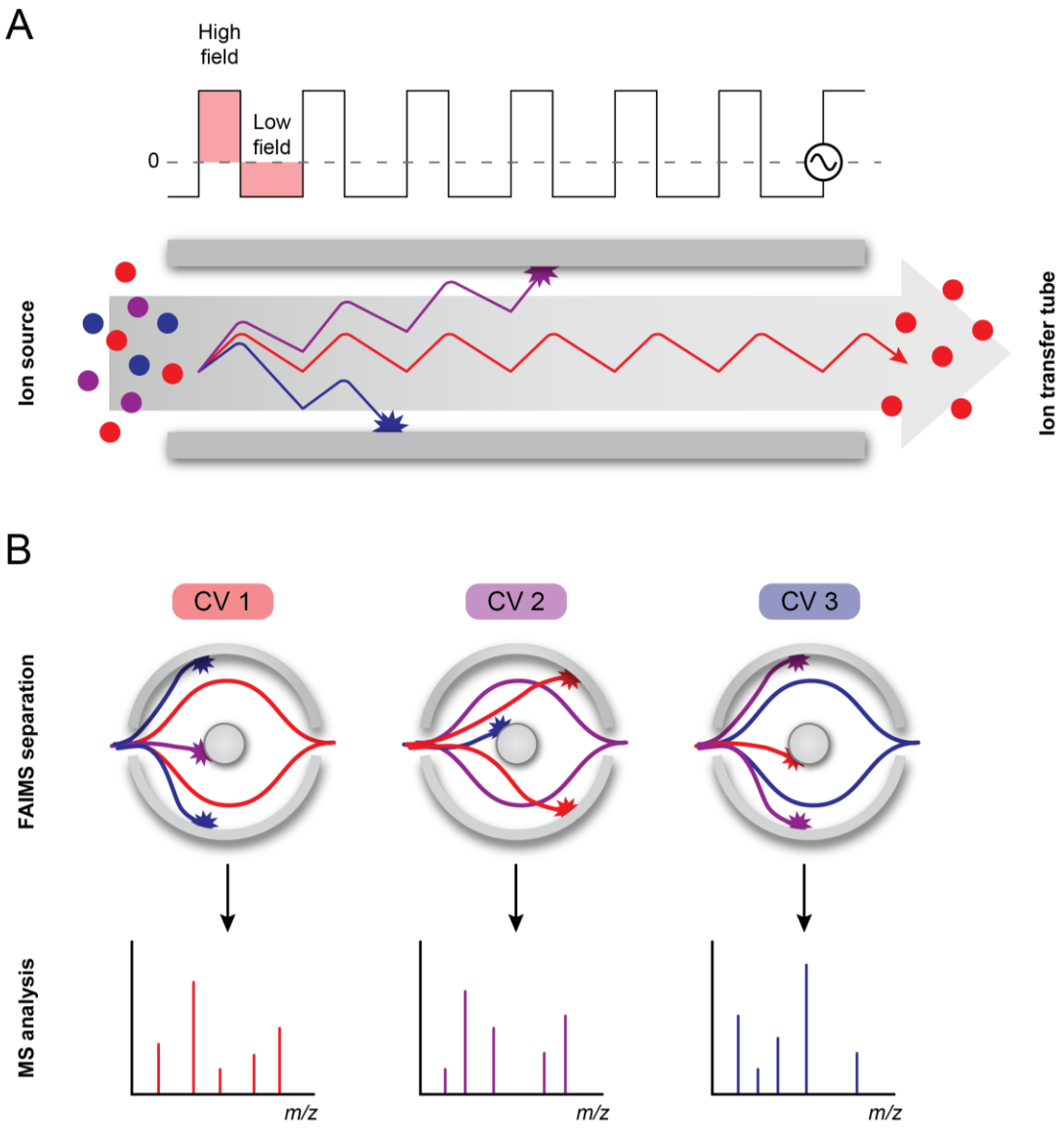


Figure 7 Field asymmetric waveform ion mobility mass spectrometry (FAIMS). (A) Ions enter the space between the two electrodes to which an asymmetric high-voltage alternating current, the so-called dispersion voltage, is applied. In this alternating field, the ions transverse the space between electrodes in a "zigzag" motion. If an ion exhibits differential mobility in the high vs. low field, it will eventually collide with one of the two electrodes. **(B)** To offset the dispersion voltage, an additional compensation voltage (CV) is applied. Through CV switching, either within a LC-MS run or between runs, the flight path of different ion packages can be stabilized. These can then be analyzed in the interfaced mass spectrometer. Adapted from references [198,199]

Benefits and application: The separation of ion packages using FAIMS can greatly reduce chemical noise in the form of singly charged ions, which leads to increased

sensitivity and protein identification. The cylindrical design blocks neutral molecules and removes singly charged background ions, reducing contamination. This results in better instrument performance and reliability. Since its release, FAIMS has been used for a plethora of applications, including the analysis of full proteomes, PTMs, PTM crosstalk, single cells, intact proteins, protein complexes, crosslinking mass spectrometry and even the characterization of monoclonal antibody oligomers. [194,198,201–213] Depending on the application, these make use of the ability to either increase proteomic depth by reducing noise or the specific selection of proteins and peptide species in the FAIMS dimension. In bottom-up proteomics, the application of FAIMS is particularly useful for the analysis of low input and single cell samples. Here the decreased chemical noise leads to visibly cleaner mass spectra and increased peptide and protein identifications.[214–217] While the increase of protein identification still holds true to a certain extent at higher sample load, the reduction of the total ion population can, however, lead to a lower number of peptides-per-protein, decreases protein sequence coverage and confidence in correct protein to peptide assignment.[201] The use of multiple FAIMS CVs alleviates this problem and additionally acts as a form of online fractionation tool for separating complex samples, protein and peptide isoforms and intact protein mixture analysis. This is achieved in either a single LC-MS run, through CV stepping, or in separate runs. While the former requires less sample material and MS time, acquiring MS1 and MS2 scans at two or three CVs more than doubles or triples the cycle time, respectively. As such, stepping through multiple CVs in a single LC/MS run, is more suitable for use with longer chromatographic gradients, where broader peaks result in higher tolerance for extended cycle times.[198]

Comparison to other commercial ion mobility implementations: Most prominent mass spectrometry manufacturers/vendors offer a commercial ion mobility (IM) implementation, though they build on different IM principles. SCIEX's SelexION device, similar to FAIMS, it is based on differential mobility spectrometry, albeit in a planar geometry. While the cylindrical electrode assembly blocks neutral and focuses the traversing ions, a planar geometry has the benefit that it allows ions to simply traverse the electrode assembly when no dispersion voltage is applied.[191,218]

Apart from DMS/FAIMS, three more IM implementations (**Figure 8**) have been coupled to MS: drift tube ion mobility spectrometry (DTIMS), traveling wave ion mobility spectrometry (TWIMS) and trapped ion mobility spectrometry (TIMS). In short, DTIMS utilizes a uniform, weak electric field and measures the amount of time an ion takes to traverse a pressurized, gas-filled drift region. The ion mobility (ions traveling slow or fast)

is influenced by collision events with the carrier gas and hence depends on the ion's shape-to-charge ratio. DTIMS uniquely enables the accurate measurement of collisional cross section (CCS) without the need for calibrant ions.[218,219] While the drift region of TWIMS is similar to that of DTIMS, it utilizes an oscillating electric field that pushes the analyte ions through the drift tube. Measurement of CCS values requires prior calibration with known ions, but ion focusing in the drift region increases ion transmission in comparison to DTIMS.[218,220–222] TIMS essentially reverses the separation principle of DTIMS by utilizing a moving gas phase and an electric field gradient. Analyte ions migrate through the electric field against the gas drag and are immobilized in the electric field gradient once the ion drift velocity and opposing gas velocity reach an equilibrium. Traversed distance is proportional to an ion’s mobility, with low CCS (high mobility) ions being trapped closer to the entrance, and ions with larger CCS values (lower mobility) residing closer to the exit of the TIMS device. Trapped ion packages can then be sequentially eluted from the TIMS device by reducing the electric field strength[223–226]

Both TWIMS and TIMS have been coupled to ESI-MS, with TWIMS being implemented on the Synapt and Select Series MS from Waters, and TIMS on Bruker’s timsTOF platform. In comparison to TWIMS and TIMS, FAIMS/DMS do not require pulsing ions into the ion mobility device, but rather operate in a continuous fashion, through which very high duty cycles can be achieved.[218] They do, however, lack the capability to measure CCS values and offer lower resolution in comparison to other IM approaches. The highest theoretical separation resolution can be achieved with TIMS, though at reduced scan speed. For proteomics applications separation resolution is usually balanced with speed. Through the capture and release mode utilized in TIMS, especially with ‘parallel accumulation – serial fragmentation’ (PASEF) acquisition mode available on the commercial timsTOF platform, ion utilization of up to 100% can be achieved.[85,227,228] The space-charge capacity of the TIMS device might, however, limit achievable dynamic range in comparison to DMS and TWIMS. Additionally, more complex tuning and calibration procedures might require higher levels of user training.[218,229] While this comparison only includes a selection of commercially available implementations, the

IMS-MS field is rapidly evolving, improving and reimagining the available technology, which might in future alleviate some of the limitations mentioned.[230–236]

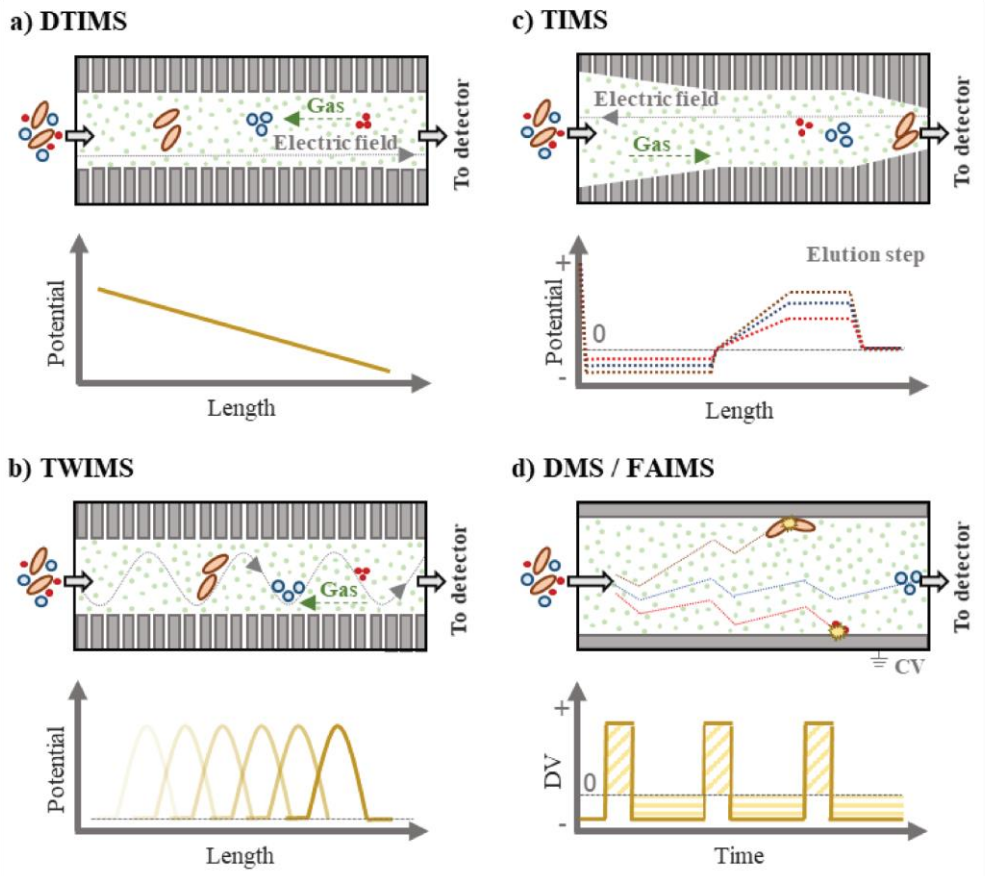


Figure 8 Overview of available ion mobility implementations. Schematic was adapted from reference[222] as permitted by the CC BY 4.0 international license.

Overall, FAIMS provides a sensitive orthogonal analyte separation, that offers many benefits for proteomic applications. Comparison to other IMS-MS implementation, however also reveal potential shortcomings and opportunities for improvement. Particularly, faster separation and CV switching will be crucial in enabling a higher ion utilization. Operated at a single CV value, FAIMS nonetheless provides superior signal-to-noise ratios for low input applications and, in our hand, extends instrument robustness in high load applications.

Modern Quadrupole-Orbitrap instruments

20 years after the initial introduction of the Orbitrap technology, TFS released a new instrument line, the Orbitrap Exploris (OE) series (**Figure 9**), which feature an atmospheric pressure ion source interfaced with electrodynamic ion funnel via a high-

capacity transfer tube, a quadrupole, a C-trap, an ion routing multipole, and an ultra-high field Orbitrap analyzer.[175,201,237,238] Three models were released, the Orbitrap Exploris 480, Orbitrap Exploris 240 and Orbitrap Exploris 120. Named after their maximum achievable resolution the three were supposed to serve different analytical purposes. With the lowest resolution, the OE 120, was optimized for environmental, food safety, and toxicology analysis, while the OE 240 and 480 were intended for high performance omics and pharmaceutical applications. During my master's thesis, I was part of a team of researchers evaluating the Orbitrap 480 mass spectrometer for proteomics applications.[201] This OE model has been widely adopted in the field and considered a workhorse instrument in many proteomics laboratories. In this thesis, I will focus on this OE model.

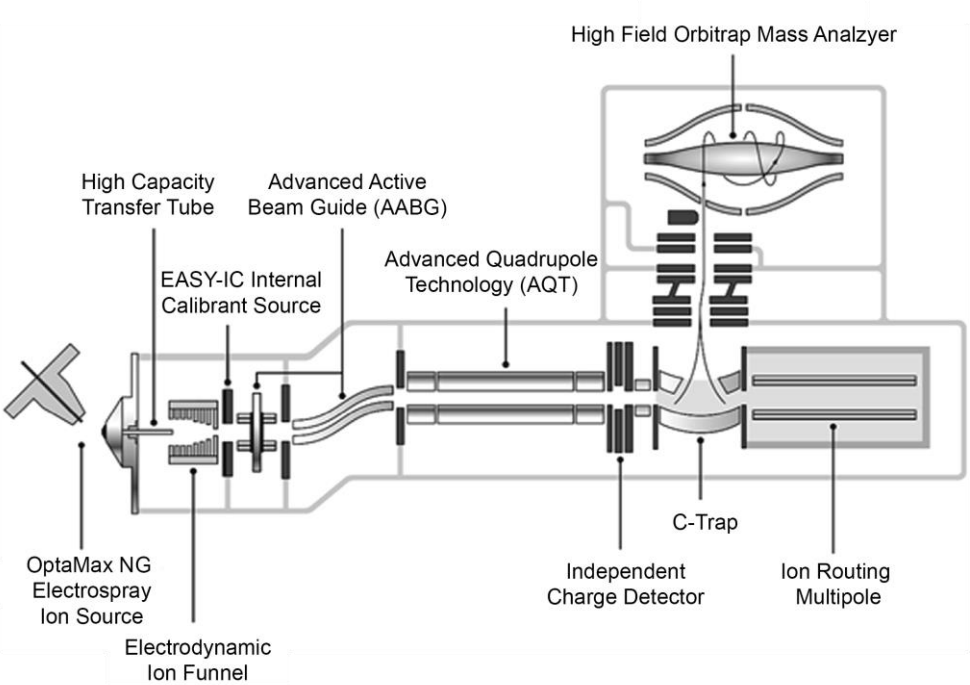


Figure 9 Schematic of the Orbitrap Exploris mass spectrometer series. The OE series of instruments features advanced quadrupole technology as well as a high field Orbitrap analyzer for mass resolution up to 480,000. In comparison to previous instrument platforms, it features a much smaller footprint, improved robustness and performance. Schematic was adapted from reference[238] as permitted by the CC BY 4.0 international license.

The most noticeable change in comparison to the Q-Exactive instrument platform is the greatly reduced footprint and volume of the instrument. This could be achieved by employing a single six-stage turbomolecular pump module instead of the previously necessary bulky multi-pump systems. Modularization and alignment of all ion optical components along a common axis in the new reduced footprint additionally increases

ease of access. Interestingly, already at this stage TFS highlights that the analyzer modularization and the instrument frame allow access for potential extension of the instrument beyond the ion routing multipole (IRM).[239] A comment that in hindsight seems to hint at instrumentation released 4 years later granting insight into the timeframe for instrument development.

Apart from the reduced footprint and new pump module, many changes in the hard- and software were implemented. Starting at the front of the instrument, the OE series marks the unification of TFS's instrument lines, by adapting the front-end or interface design used for the Tribrid and triple quadrupole instruments. This allows for full compatibility with ion sources designed for these instruments as well as Tribrid front-end options, including the FAIMS Pro interface, which enables FAIMS. Use of the (Tribrid) EASY-IC discharge ion source, allows for improved ppm-level mass stability by releasing a stable flow of fluoranthene ions that can be used as lock masses, which increases robustness of the system.[240] Moving on to the quadrupole, a so-called symmetrical ion loading was introduced that distributes filtered-out ions more evenly across the quadrupole rod pairs using automatic and regular polarity switches in the quadrupole. To ensure that both rod pairs provide equal transmission and isolation efficiency and quality, the quadrupole manufacturing has been improved. Overall, this can increase the time between instrument cleaning, and therefore decrease downtime, up to a factor of two. On the Orbitrap analyzer end, additional focusing lenses have been added to allow for a new C-trap design that only applies the pull-out pulse to the slotted C-trap electrode (closest to the Orbitrap), which increases ion focusing and reduces ion losses at the edge of the extraction slot. Improved pulse control on the Orbitrap central electrode further allows for electrodynamic ion squeezing of a much broader mass range.

In terms of software changes, the instrument control software was engineered to resemble the control software of the Tribrid and triple quadrupole platforms. This includes the harmonization of instrument setting such as collision energies between these instrument platforms. The OE series additionally marks the first instrument with a commercial implementation of the Phase-constraint spectrum deconvolution method (ΦSDM) to increase resolution, albeit it was only applied to the small *m/z* range of TMT reporter ions in an acquisition mode termed Turbo-TMT to ensure real-time computation directly on the instrument computer.[241–243]

Overall, this enables resolution of up to 480,000 at *m/z* 200, a scanning speed up to 40 Hz (as on the HF-X instruments), and a mass range up to 6,000 *m/z* (or 8,000 in the

biopharma version) all in an instrument of reduced footprint. This has led to a wide spread adaptation of the OE instruments for the analysis of proteomes, PTMs, and even single cells.[201,244–249] Our investigation of the effects of muscle loss on the human plasma proteome (Article 2) could additionally showcase its use for clinical proteomics.[250] This has been explored by many other labs as well, with notable examples being the identification of biomarkers for alcohol-related liver disease and the proteomic profiling of eczema, both of which could have clinical implications.[17,251] Mass spectrometers of the Exploris, and Q-Exactive series have even been used in combination with the MassSpec Pen, a liquid-extraction-based device, for intraoperative tissue analysis in clinics.[252–254] The instrument, however, is not exclusive to proteomics, but has been applied in peptidomics, metabolomics and lipidomics.[255–259]

While a fourth instrument in the Exploris series, the OE MX, has been released for pharmaceutical analysis of native proteins and oligonucleotides, there were no commercial hardware updates or upgrades of the original three instruments 2019.[260] However, many (so far non-commercial) options to extend the functionalities and performance metrics of the OE 480 have been explored. Focusing on the latter, TFS developers could show that a mass resolution of up to 2,000,000 at *m/z* of 200 is possible (4s transient time) on an OE 480 with a specifically selected Orbitrap assembly, manual mass calibration and fine tuning. If these tuning requirements can be reproduced in the serial instrument, this could enable the resolution of fine isotopic structure analysis in proteomics, metabolomics as well as trace and petrol analysis.[238] Another way of increasing mass resolution on the OE 480 is extending the use of ΦSDM to the full mass range, as described in this thesis (Article 1). Since transient time and mass resolution are inherently linked, ΦSDM can also be used to increase acquisition speed. Achieving the same mass resolution in half the transient time is particularly of interest for short chromatographic gradients, where peptide signals are compressed to increasingly more narrow peaks, that require MS acquisition methods with short cycle times for adequate quantification.

Extension of functionalities or information content seemingly focused on two topics: collisional cross section (CCS) analysis and targeted proteomics. The CCS of an ion reflects its size, shape, and charge and is generally used in structural characterization of intact proteins or as an additional metric for separation in ion mobility mass spectrometry (IMS).[261,262] While the analysis of CCS values usually requires a separate

ion mobility device, two ways of measuring CCS on the OE 480 have been proposed. Both utilize the ion decay rate in either time or frequency domain, with one primarily being used for analysis full protein CCS values[263], and the other to the analysis of peptide CCS in complex proteomic samples. The latter takes advantage of the decrease in full scan resolution observed when operating at elevated ultra-high vacuum pressure and high MS1 resolution. Switching between UHV pressure conditions, however, requires minor hardware modifications.[264,265]

Implementation of targeted mass spectrometry similarly has been achieved in different ways. On one hand, the use of an application programming interface (API) allows for the use of MaxQuant.Live for global targeting and control of data acquisition in real-time.[266] The TFS proprietary SureQuant workflow also offers real-time adjustment, but relies on synthetic peptide spike-ins to trigger quantification scans. Predefined template methods in the TFS method editor, additionally make this approach more user and beginner friendly.[267–269] Lastly, a hybrid-DIA approach, using an API for method customization, combines the benefits of targeted and discovery DIA. Triggered by the use of heavy-labeled peptides, DIA scans are interjected with multiplexed MS2 scan of the predefined peptides targets, which allows the targeted acquisition of peptide targets and DIA data acquisition in a single run.[270]

Lastly, I would like to highlight an implementation of ion pre-accumulation on a modified OE that allows for an ion trapping and accumulation step in the bent flatapole parallel to C-trap operations. In contrast to regular operations, the exit lens of the bent flatapole is set to trapping mode at the end of an ion injection to the IRM. While the first ion package is transferred from the IRM to the C-trap and subsequently to the Orbitrap, ions are accumulated in the bent flatapole. At an acquisition rate of 40 Hz, max. acquisition rate of an OE, the instrument sensitivity could effectively be doubled and a 100% duty cycle was achieved. Moreover, acquisition rate could be increased to over 80 Hz without a decreased duty cycle. This initial, albeit crude implementation showcases the potential for proteomics applications and could be especially of interest in combination with an inherently faster HRAM analyzer.[271]

Tribrid MS instrumentation

Since 2019 two members of the well-established Tribrid platform of TFS instruments have been released: the Orbitrap Eclipse and Orbitrap Ascend mass spectrometers. The Tribrid instrument platform utilizes the synergy of three different analyzers: the

quadrupole mass filter, the Orbitrap analyzer and a dual-pressure linear ion trap (LIT) analyzer (**Figure 10**).[240,272,273] This allows for the parallelization of MS1 and MS2 scans, where high resolution MS1 scans are recorded in the Orbitrap and fast, high sensitivity MS2 scans are acquired in the LT.

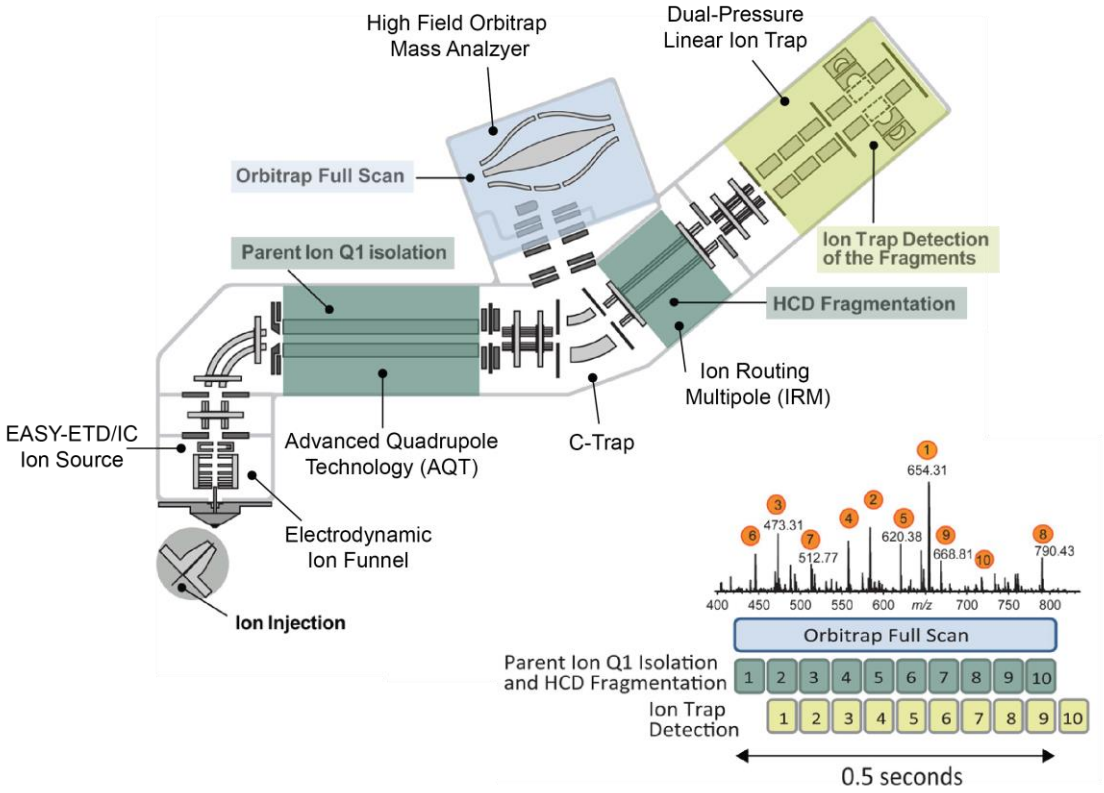


Figure 10 Schematic of a TFS Tribrid instrument. General instrument design as suggested at release of the first Tribrid instrument, which features a quadrupole for mass selectivity, an Orbitrap for high resolution MS1 scans and a linear ion trap (LIT) for fast and sensitive MS2 scans. Due to the mass analyzer duality, MS1 and MS2 scans be acquired in parallel as shown at the bottom right schematic for scan scheduling. Adapted with permission from reference[240]. Copyright (2013) American Chemical Society.

The dual-pressure LIT, was first introduced in the LTQ Velos instrument, back in 2009 and features, as the name suggests, two ion trapping cells, which are maintained at differential pressure levels and separated by a single aperture lens. First a high-pressure cell is used for ion trapping, isolation and fragmentation. Second, the low-pressure region is used for mass analysis. This dual-pressure design improves efficiency of ion trapping and fragmentation, scan rates and mass resolution. [272,274] It additionally allows the implementation of alternative fragmentation strategies to HCD. In modern Tribrid

instruments, such as the Orbitrap Eclipse and Ascend, these include CID, ETD, UVPD and even EThcD. As mentioned previously, these provide complementary ions to HCD and are highly beneficial for the analysis of full proteins or PTMs. First tested on a modified Orbitrap Fusion Lumos, the Eclipse and Ascend Tribrid MS also implement proton transfer charge reduction (PTCR), which is based on the proton transfer from multiply charged analyte cations to singly charged perfluoroperhydrophenanthrene anions. This leads to the a reduced charge of the analyte cation and shifts the cations to higher *m/z* value, effectively extending the analyte charge envelope.[275,276] Over the years, this has shown great potential for the targeted analysis of proteins as well as for middle down proteomics approaches.[275,277–280] In comparison to the Orbitrap Exploris series, a commercial upgrade allows for Orbitrap resolution of 1,000,000 at *m/z* of 200 on the Orbitrap Tribrid Eclipse MS and the Tribrid MS instruments are capable of MS3 or MS^n analysis, which has proven to be especially effective for TMT, crosslinking mass spectrometry and single cell analysis.[281–285] For more intelligent selection of ions for MS3 analysis, the instruments have an inbuilt implementation of the “Real-Time Search” algorithm. Active instrument control through an instrument API and the use of Comet, an open source search engine, allows to identify fragment spectra on the fly and to only trigger the acquisition of quantitative spectra after confident peptide identification.[286–289] As many of the mentioned features improve the analysis of isobaric, specifically TMT, labeled samples, it is not surprising that the instruments also feature a TurboTMT implementation of ΦSDM.[285]

In comparison to previous Tribrid instruments, the design of the Eclipse MS already featured improvements such as advanced quadrupole mass filter and higher-transmission ion optics, that lead to an increase in ion transmission of 25-50%. However, it still utilized the Orbitrap/C-trap assembly components and electronics from the Q-Exactive series.[285] In the state-of-the-art Tribrid instrument, the Orbitrap Ascend, this is updated to feature the improved Orbitrap/C-trap design of the Exploris series, mentioned before. Moreover, the updated instrument design includes a new ion funnel for gentler ion injection and, most notably, a second RF-only IRM in front of the C-trap.[290] Together this lead to increased ion transmission, sensitivity, and speed, which translates into increased identification rates for proteome and PTM analysis.[290,291] Additional fragmentation modes, MS^n functionality, and extended mass range make the Tribrid instruments particularly suitable for top or middle down proteomics, and the analysis of labile PTMs, where these functionalities are of higher value.[292]

Orbitrap Astral MS - a novel HRAM Orbitrap-TOF instrument

Overall, the Exploris, Tribrid, and other Orbitrap MS instruments, especially in combination with front end accessories, showcase the strengths of the Orbitrap technology, namely high resolution, mass accuracy and dynamic range. However, the technology also has its limitations. The Orbitrap has slower acquisition rates and sensitivity in comparison to other MS instrumentations, such as high-end time-of-flight (TOF) analyzers. While single-ion detection has been shown to be possible, this required transient times of multiple seconds.[293] With high resolution in FT-MS inherently being linked to the transient time, Orbitrap resolution additionally needs to be balanced with scanning speed for proteomics applications. Size-constraints additionally limit the charge capacity and too high ion load leads to space-charging effects, impacting resolution.[294,295] While the addition of a linear ion trap in TFS Tribrid instruments addresses some of these limitations, ion traps cannot provide the same level of mass resolution and accuracy as HRAM mass analyzers. While many labs still prefer Orbitrap-based instruments, recent improvements in TOF technology - like Bruker's timsTOF and SCIEX's ZenoTOF - have gained popularity due to their enhanced sensitivity, resolution, and speed.[85,292,296–299] Particularly, the timsTOF instrument series, with its implementation of TIMS and the PASEF acquisition mode, surpassed the Orbitrap technology in terms of speed, sensitivity and duty cycle.[85,94,300]

In 2023, TFS introduced some of the previously mentioned technical advances on their Exploris and Tribrid series and worked on new analyzer concepts, which ultimately lead to the introduction of a novel HRAM mass spectrometer, the Orbitrap Astral MS. The asymmetric track lossless (ASTRAL) analyzer is a multi-reflector (MR) type TOF comprised of two elongated, asymmetric ion mirrors, a pair of prism-shaped deflectors and specifically shaped electrodes, termed ion foils.[301,302]

A brief history of multi-reflector time of flight (MR-TOF) analyzer: In and of itself, MR-TOF is not a novel idea. As resolution in TOF MS is dependent on the total length of the ion flight path, it is no surprise that the idea to reflect ions using electrostatic mirrors first arose in the 1950s and was implemented in the 1970s.[303,304] In general, MR-TOF mass spectrometer utilize repeated ion reflections between electrostatic mirrors to achieve flight paths significantly longer that the instruments dimensions.[305–308] Over the years, many researchers have developed different versions of MR-TOF instruments,

each offering unique advantages and limitations.[309–313] Notably, Anatoly Verenchikov and his company MSC-CG Ltd made significant advances in MR-TOF technology. Their work contributed to Waters Corporation's development of a high-resolution MR-TOF analyzer, now used primarily in imaging mass spectrometry.[308,314] Building on these technological advances, multiple patent applications by TFS suggest that work on a MR-TOF type mass spectrometer has been ongoing for at least 10 years.[315–319] Before they arrived at the released Astral analyzer, other avenues, such as the concept of a so-called OrbiTOF analyzer were explored (**Figure 11**).[320]

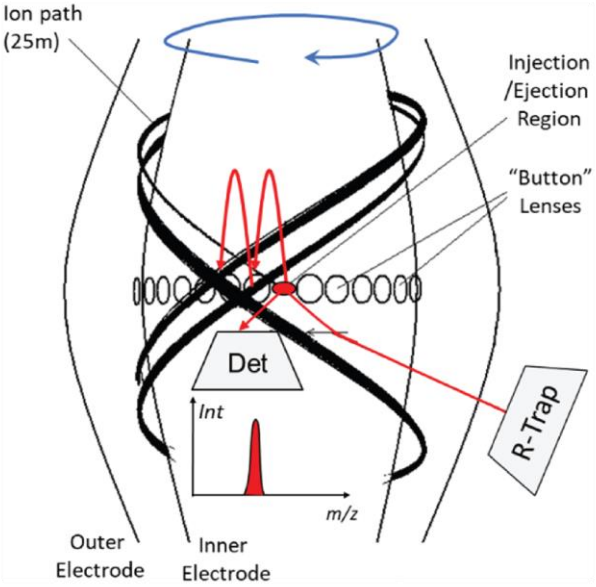


Figure 11 Schematic of the OrbiTOF design and ion motion. Reprinted with permission from ref[320], Copyright (2024) Elsevier under license 5901441070212.

As the name suggests the MR-type OrbiTOF analyzer is based on the Orbitrap technology with the addition of periodic lenses, termed “button” lenses, that are wrapped around the central electrode. Shortly, ions are accumulated in an ion trap, before being pulsed between the inner and outer electrode of the OrbiTOF analyzer. There, ions turn around the inner electrode as they additionally drift to the top of the analyzer, before being reflected back by a quadratic mirror potential. By the time the ions pass the injection slot at the equator, they have performed a single orbit around the inner electrode and are refocused by the first of a periodic series of button lenses. As ions continue to oscillate around the inner electrode they pass the subsequent button lenses, which prevents beam dispersion, and finally hit a multi-channel plate detector. While this approach could achieved mass resolution up to 70,000, because of limitations in ion

transmission as well as flaws concerning the button-lens based refocusing, the development was discontinued in favor of the Astral MR-TOF concept.[301,302,319,320]

The Orbitrap Astral MS components: The Orbitrap Astral (OA) MS marks the start of a new instrument line, which combines TFS advanced quadrupole and Orbitrap technologies with the novel Astral analyzer, and is a step up in sensitivity, resolution and speed in comparison to previous instrument generations. [301,302] While OA components up until the IRM are kept consistent with the OE 480 MS, the IRM is then interfaced with the secondary instrument part through an octupole ion guide. For optimal instrument performance, the novel Astral analyzer is complemented by advanced ion optics, a novel so-called ion processor and a custom-design high dynamic range (HDR) detector (**Figure 12**).[321,322]

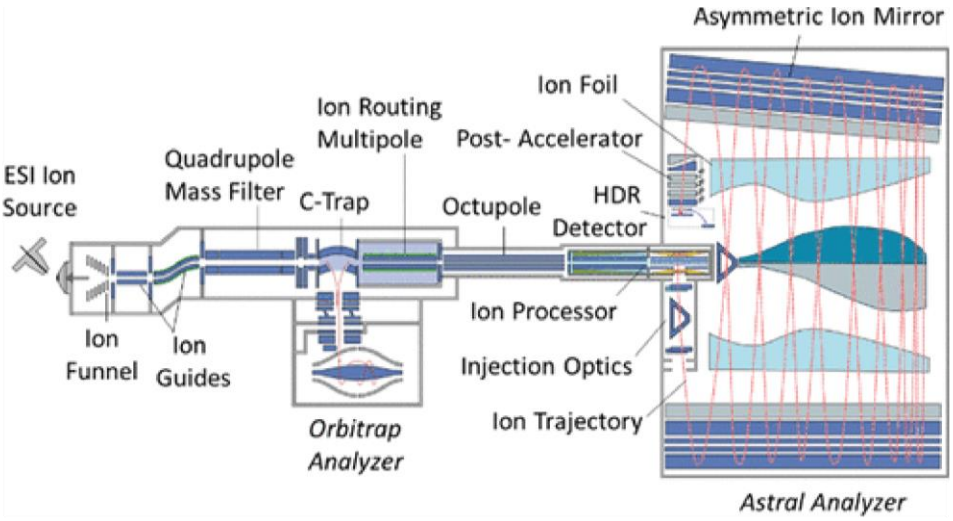


Figure 12 Schematic representation of the Orbitrap Astral mass spectrometer. The OA instrument design is based on the Orbitrap Exploris series and all instrument components up until the IRM are kept consistent with the OE 480 MS. The IRM is then interfaced to the ion processor, a dual-pressure trap, through an octupole ion guide. The ion processor accumulates, fragments and thermalizes ions prior to orthogonal pulsed extraction into the Astral analyzer. Ions are traverse the space between two asymmetric ion mirrors on a multireflection path until they reach the high dynamic range (HDR) detector. Reprinted from reference[302] as permitted by the CC-BY-NC-ND 4.0 international license.

Ion processor: The novel ion processor, a dual-pressure linear quadrupole ion trap, serves the purpose of ion accumulation, fragmentation, and extraction for subsequent

analysis in the Astral analyzer (**Figure 13**). In the high-pressure region of the ion processor, ions are first accelerated and undergo HCD fragmentation.

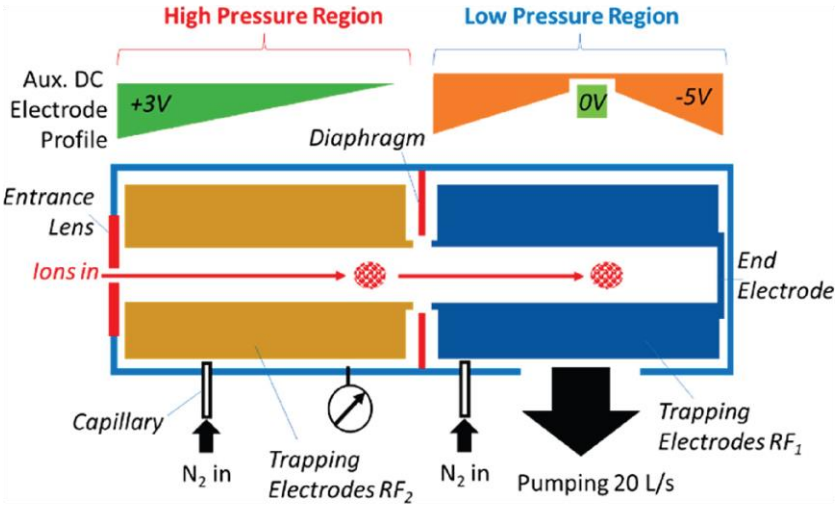


Figure 13 An ion processor for parallelized ion processing. Schematic representation of the ion processor featuring a high- and low-pressure region for accumulation, fragmentation, thermalization and orthogonal ion extraction. Reprinted from ref[322] as permitted by CC-BY-NC-ND 4.0 international license.

They are subsequently moved to the far end of the high-pressure region by a DC gradient, where they accumulate and are subjected to thermal cooling, before being transferred to the low-pressure region by an increase in the DC offset. Here auxiliary DC electrodes move the ions along RF ion guides to an axial potential well in the center of the low-pressure region, where there are stored and thermalized for subsequent orthogonal pulsed extraction into the mass analyzer. Ejection of ions is achieved by raising the low-pressure region to a higher potential, which accelerated the accumulated ions towards the Astral analyzer (**Figure 14**). Parallel to the pulsed ion extraction, the high-pressure region of the ion processor is reopened for accumulation and fragmentation of a second ion package. Overall, the dual pressure design of the ion processor enables high ion transmission, as well as the parallelization of ion processing steps for maximum instrument utilization.[302,322]

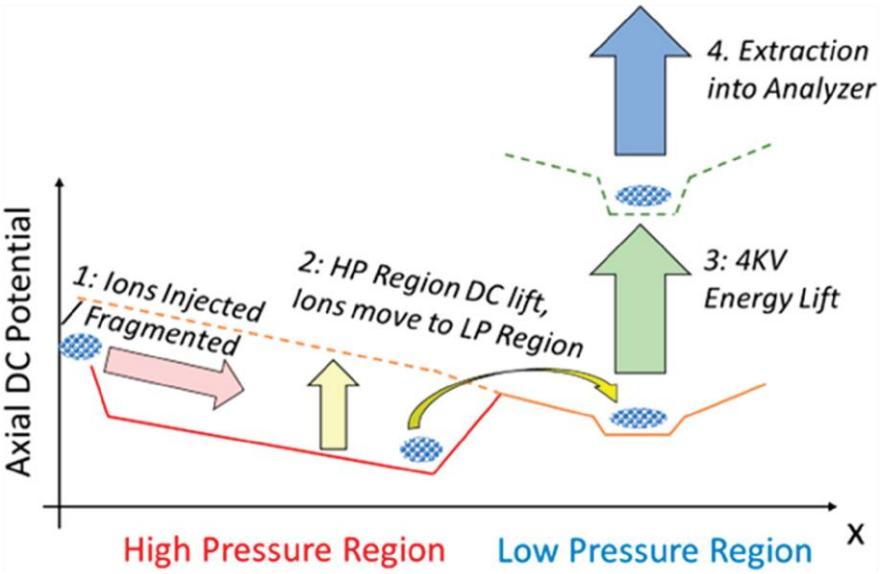


Figure 14 Potential and ion processing sequence in the ion processor. Ions are injected into the high-pressure region, fragmented and subsequently accumulated and "cooled" at the rear end of the high-pressure region. An increase in DC potential allows the ions to move to the low-pressure region, where the ions are thermalized in an axial potential well. Ions are then lifted to higher potential for orthogonal extraction. Reprinted from ref[322] as permitted by CC-BY-NC-ND 4.0 international license.

Astral analyzer: The Astral mass analyzer is a multi-reflector-type TOF analyzer that utilizes two elongated asymmetric gridless ion mirrors and ion foils to create a multi-reflection ion flightpath of ~30 m. After extraction from the beforementioned ion processor, ions packages pass through the injection optics, comprised of a pair of lenses and two electrostatic prisms, and are shaped, focused, and then deflected at an optimal injection angle. Ions now oscillate between the ion mirrors and drift towards the rear end of the mirror length. Over the course of 12-13 reflections, the ion drift rate is decelerated by the slight, converging mirror tilt and ultimately reversed. This reversion is primarily achieved by a returning electrostatic potential, which is formed by a combination of mirror tilt and refraction on a set of specially shaped electrodes, termed ion foils. The ion foils additionally compensate for temporal aberrations and potential misalignment of the asymmetric ion mirrors. After another 12-13 reflections, the ions pass the second electrostatic prism and are deflected to the HDR detector, which is located at the proximal end of the ion mirrors. Over the full course of 24-26 oscillations, and a total flight path of ~30 m, the ion packages are separated based on the *m/z* values. Drift expansion, during the first set of oscillations towards the distant mirror end, lead to a

spatial dispersion of up to 5 cm. While this is essential for decreasing Coulomb repulsion forces, it also leads to overlapping of oscillations of different ion populations. The drift spread, however, is reduced on their returning oscillations and the ions are refocused spatially as a single ion package for before reaching the detector.[301,302,323] While multiple options for dispersion control were tested, the described implementation outperformed them.[324] Overall, the combination of optimal injection optics, gridless design, and spatial refocusing allows for very high ion transmission through the Astral analyzer, which inspired the inclusion of "lossless" in the Astral abbreviation.[301,325] Though it should be noted that this is to be considered "relative lossless" in comparison to other TOF analyzers. While the Astral analyzer has a reduced charge capacity in comparison to the Orbitrap, the sensitivity and low noise levels in combination with advanced detector technology enable single ion detection. Moreover, the long flight path routinely enables mass resolution of over 80,000.

High dynamic range (HDR) detector: To fulfil the speed, dynamic range and resolution requirements of the Astral analyzer, a novel HDR detector was designed and manufactured in a cooperation between TFS and El-Mul Technologies Ltd.[321] The detectors features a unique combination of 10 kV post-acceleration with an integrated correction for ion package tilt, BxE (crossed magnetic and electrostatic field) focusing, an optically coupled detector, pre-amplification and dual channel acquisition (**Figure 15**). After ions completed their oscillations between the asymmetric ion mirrors, they are

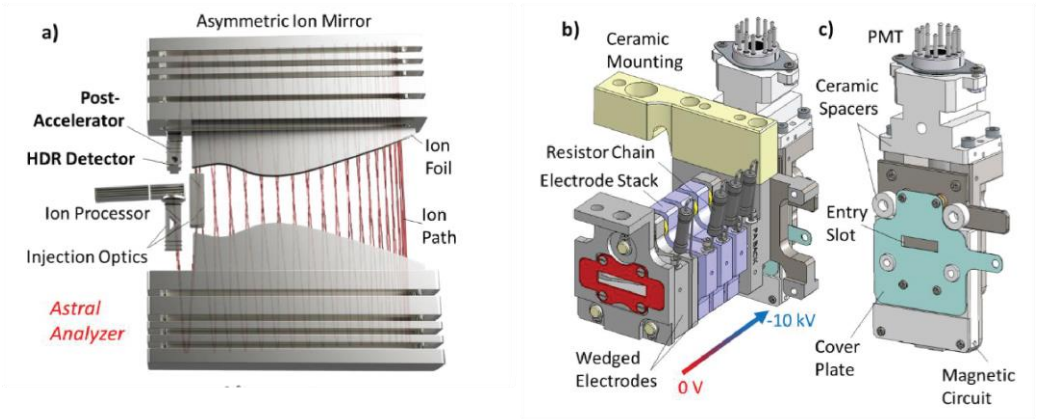


Figure 15 A novel high dynamic range detector. **(a)** Arrangement of ion optics in the Astral analyzer. **(b)** Schematic of the high dynamic range (HDR) detector assembly, including the post accelerator stack and insulating ceramic division. **(c)** Schematic of the HDR detector. Adapted from ref[321] as permitted by CC-BY-NC 4.0 international license.

deflected towards the detector (**Figure 15a**). A post-accelerator accelerates the ions from 4 kV to 14 kV and focuses them on the detector plate (**Figure 15b, c**). A "deflector"

in close proximity to the detector allows for controlled tilting of the focal plane to align the ion trajectory with the detector surface and compensates potential small mechanical errors. Ions enter the detector through an entry slot, strike a conversion dynode, and produce secondary electrons. These focus in the BxE field and produce photons when they converge with a scintillator. The photons in turn travel to the photomultiplier tube (PMT), where the photon signal is amplified. To improve dynamic range of the detector, the PMT output is split into two channels, each with their own amplifier, where one channel gets amplified fivefold, while the other gets reduced to half its original level. The high and low gain signals are then directed to separate analogue-to-digital-convertor channels within a dual-channel digitizer. After noise thresholding in the digitizer, the data from both channels is transferred to the instrument embedded PC and combined in a single mass spectrum. Overall, the HDR detector achieves an intra-scan dynamic range of 4 orders of magnitude, resolution of over 100,000, effective single ion detection and a relative immunity to detector aging for an increased lifetime.[321]

Mode of operation: Similar to the Tribrid series, the OA is operated in Orbitrap/Astral mode, meaning MS1 scans are recorded in the Orbitrap, while simultaneous MS2 scans are recorded in the Astral analyzer. While the instrument can additionally be operated in Orbitrap/Orbitrap mode, the Astral analyzer can, at the moment, be exclusively used for MS2 acquisition. For MS1 scans, the quadrupole transmits ions of a wide *m/z* range to the C-trap, before they continue to the IRM for trapping and accumulation. During the accumulation in the IRM, ions are “cooled” before being transferred back to the C-trap and from there into the Orbitrap analyzer, once the desired number of charges is reached. For Orbitrap MS2 scans, selected precursor ions are subjected to HCD fragmentation in the IRM prior to mass analysis. For Astral MS2 scans, the selected precursor ions are routed through the C-trap to the far end of the IRM, where they are set to accumulate for a defined amount of time. The ion package is then transmitted through the octupole ion guide to the ion processor, where the ions are subjected to HCD fragmentation, thermalized and injected into the Astral analyzer for mass analysis. When both analyzers are being utilized, MS1 and MS2 scans can be acquired in parallel. In this case, the advanced ion control enables the simultaneous handling of five ion packages. While the Orbitrap is performing an MS1 scan using the first ion package, a second ion package is accumulated in the ion routing multipole. The segmented, dual pressure nature of the ion processor allows for the handling of two additional ion packages, one in the high-pressure region, where peptides are being fragments prior to

MS2 analysis, and the other in the low-pressure region, where the ion package is being focused prior to injection into the Astral analyzer. The fifth and final ion package is, therefore being analyzed in the Astral analyzer. Fast Astral scanning speeds of up to 200 Hz make the Orbitrap Astral analyzer ideal for DIA applications using narrow, DDA-like DIA isolation windows.[302,326,327] The fast scanning speed in combination, with high resolution (>80,000 at *m/z* 524), mass accuracy (<5ppm) and sensitivity (single ion detection), make it one of the highest performing MS for proteomics applications at the moment. In the less than 1.5 years since its release, ~100 publications - peer-reviewed or preprinted - have been published covering a wide range of applications for proteomics and beyond.[214,326–329]

Comparison to state-of-the-art TOF analyzers: While TFS primarily relied on the Orbitrap technology, in combination with a quadrupole (Q) and in the case of the Tribrid series the LIT, for their HRAM mass spectrometers, many other vendors have advanced their TOF analyzer technology. One break-through on this front was the introduction of Bruker's high-resolution Q-TOF, the Impact II.[296] Building on this technology, they later introduced a TIMS device for an added ion mobility dimension and increased ion usage.[85,223] However, many other vendors also have high resolution Q-TOF instruments in their portfolio. Notable examples include Agilent's 6546 Q-TOF and 6560 IM Q-TOF, Waters' Synapt XS and Select and Select cyclic IMS series, and SCIEX's ZenoTOF instruments.[330–335]

TOF design and resolution: Most TOF instruments share a basic design: an orthogonal accelerator pushes ions into a long flight tube, where they travel up to a reflectron, bounce back, and hit a detector at the bottom of the tube. Similar to the Astral analyzer, the Waters' instrument lines, however, also feature a MR-TOF using gridless ion mirrors. While the instruments achieve impressive resolution of > 300,000 at *m/z* 785, they are comparatively slow with a scan speed of 30 Hz, and are as of yet primarily used for imaging mass spectrometry. It does, however, highlight the key benefit of MR-TOF designs, which is high resolution without the need of an extensively long flight tube. In line with this, the Astral analyzer offers the highest achievable resolution in comparison to the other ESI-Q-TOF instruments discussed, followed by the Bruker timsTOF instruments.

Scan speed: In terms of speed, both the Agilent and Waters instruments have comparatively low scan rates with up to 50 Hz and 30 Hz respectively, however both of their primary applications lie outside of the analysis of complex bottom-up proteomics

samples. As mentioned before, the Waters instruments are interfaced with ion sources for matrix-assisted laser desorption ionization (MALDI) and desorption electrospray ionization (DESI) and used for imaging MS, while the mentioned Agilent instrument find application in metabolomics and food safety. In these cases, the reduced scan speed might have less of an impact. Bruker TimsTOF Pro 2 instruments reach a scan rate up to 120 Hz in dda-PASEF, while their HT, SCP and Ultra models can reach up to 300 Hz. The SCIEX ZenoTOF 7600 can reach scan speeds of up to 133 Hz, though their newest release the ZenoTOF 7600+ promises up to 640 Hz. The Astral analyzer can reach up to 200 Hz.

Sensitivity: The sensitivity of LC/MS instruments generally refers to their ability to identify and quantify low concentrations of analytes. This is expressed as a signal-to-noise ratio, which infers that sensitivity can be improved either through increased signal intensity or by reducing noise. On the other hand, sensitivity is decreased by ion losses and poor ion utilization. For conciseness of the sensitivity comparison, I will limit it to the ESI-Q-TOF instruments with application in proteomics, namely from Bruker, TFS, and SCIEX. Overall, their state-of-the-art Q-TOF instruments are all highly sensitive and enable the proteomics analysis of low input samples down to single cells.[217,297,336] In the Orbitrap Astral MS, ion transmission is exceedingly high. However, when using narrow window DIA only a fraction of the ion beam is actually used for each subsequent MS2 scan. In comparison to this, both the ZenoTOF and timsTOF MS implement a pre-accumulation of ions, followed by transmission of these ion packages. In the case of ZenoTOF instruments the accumulation is performed in a so-called ZenoTrap, which ejects the accumulated ions into the TOF analyzer. This greatly enhances ion utilization and consequently sensitivity. While ion transmission for MS1 scans across the instrument platforms is generally >90%, Q-TOF instruments generally suffer significant ion losses in the orthogonal accelerator and in the flight tube. In the Bruker Impact II (2015), this culminated in the ion loss between the quadrupole and flight tube of ~40%. Overall transmission in the flight tube was reduced to 74% due to the restricted grid transmission of the reflectron.[296] While improved ion optics and ion focusing techniques can improve ion transmission between the quadrupole and TOF flight tube, this remains a bottleneck for sensitivity.[297] Likewise optimization of the reflectron grids or gridless TOF designs can further increase sensitivity.[325,337] This is highlighted by the single ion resolution that can be achieved with the gridless ion mirror design of the Astral analyzer.[301] Ion losses are further reduced by optimal injection optics, initial drift

expansion (to reduce Coulomb forces) followed by spatial refocusing on of the ions, as well as by operating the analyzer at a pressure below 10^{-8} mbar (reduced collisional ion losses), leading to a (near) lossless ion path to the Astral analyzer.[301,302]

Overall, each of these instruments achieve high performance LC/MS analysis, has their advantages and limitations. While vendor patents often limit the dissemination of novel technological advancements, they can none the less serve as inspiration for further improvements.

1.3.3 Potential future directions of the Orbitrap Astral platform

As the Orbitrap Astral instrument was only introduced in June 2023, we can expect the instrument platform to mature over the next years. An updated software release, expected for release in June 2025, might reintroduce functionalities available for the previous instrument platforms, such as a SureQuant-like targeting, and stepped collision energies. Without direct intel from the mass spectrometry manufacturers, it can be difficult to guess what improvements or innovations will be released next, but practical considerations as well as literature and patent review might give some insights.

As a first, it could be of interest to enable Astral MS1 scans, which would allow acquisition rates that surpass the Orbitrap's capabilities. A patent from TFS, covering tandem MS1 acquisition in two different analyzers, one being the Orbitrap, and the other being referred to as a TOF, suggests that that Astral MS1 is actively investigated.[338] For this, one should however consider the lower resolution and dynamic range of the Astral analyzer in comparison to the Orbitrap. Dynamic range limitations might be addressed by further improvements in detector technology, whereas increased mass resolution relies on a longer TOF ion path. While increasing the size of the Astral analyzer might neither be feasible nor desirable, multi-pass methods, where ions traverse the Astral analyzer for more than one pass, have recently shown the ability to potentially double the achievable mass resolution.[339]

In comparison to other state-of-the-art TOF instruments, the ion utilization and duty cycle of the Astral analyzer is comparatively low. While advantageous for selectivity, dynamic range and deeper proteome coverage, the current operations using narrow window DIA discard a large percentage of the ion cloud. The implementations of pre-accumulation step could overcome these limitations. Ion pre-accumulation in the IRM has already been proposed for the Exploris platform and could, if initial results are confirmed, be adapted to the Orbitrap Astral instrument. A so-called "ion guide" patent, showing an

different ion guide concepts that are supposed to facilitate ion path length differences, as well as their implementation in an Orbitrap MS between bent flatapole and quadrupole, could be interpreted as an inbuild mode of separation in addition to the quadrupole.[340]

While TFS seems to be working on incorporating the measurement of CCS values in the Astral analyzer, the current design does not include an IM device. Based on the description, the idea seems to be utilizing the same principle as suggested for the Orbitrap Exploris, where a spectrum is acquired at two different pressure levels to infer CCS.[265,341] It will be interesting to see how the CCS information will be utilized in such an approach. Nonetheless, a future implementation of an integrated IM dimension could be a valuable addition, especially in combination with a trapping function similar to TIMS.

For direct transfer from targets identified with Astral discovery DIA to a targeted assay, implementation such as SureQuant could improve easy-of-use. With the recent developments in targeted proteomic instrumentation in mind, the transfer of the adaptive real-time retention time alignment technology, from the TFS Stellar MS to the Astral, additionally would be highly beneficial.[342,343]

In line with the Tribrid instrument series, another interesting direction could be the integration of fragmentation techniques beyond HCD. This would allow the generation of complementary fragment ions and enable more detailed analysis of PTMs, intact proteins, as well as increased sequence coverage. While a biopharma option of the Orbitrap Astral was available at the release in 2023, this only includes extended mass range up to *m/z* of 8,000 for the Orbitrap analyzer. In line with the potential for Astral-based MS1 analysis, it might be needed to increase the covered mass range to allow for the analysis of intact proteins and biomolecules.

As many of the discussed applications and functionalities might require hardware and electronics changes, one will have to wait until the reveal of a 2.0 version of the Orbitrap Astral to find out which, if any, of the mentioned patents have been implemented.

1.4 Applications of MS technology for clinical and spatial proteomics

Although MS technology, innovation and MS method development were the central to my PhD research, the continuous improvements in resolution, sensitivity, speed and robustness were implemented with the intension to further our knowledge and find answers to biological or clinical questions. In other words, they were made and meant to be applied. In the following sections, I briefly highlight a few applications relevant to the projects presented in my thesis.

1.4.1 Clinical proteomics

The proteome, with its dynamic changes in protein abundance, localization, and diverse proteoforms, is our closest proxy to understanding cellular phenotype. Its high adaptability to both intrinsic and extrinsic changes make it an invaluable window into cellular function. As such dysregulation and the manifestation of disease often occurs on the protein level, making proteins ideal candidates for disease biomarkers or potential therapeutic targets.[14–17,344–347] Here, MS-based proteomics allows for the systematic evaluation of protein-level changes caused by disease manifestation, progression, and treatment administration. Moreover, once potential biomarkers have been identified, targeted mass spectrometry could offer a high throughput solution for the identification and quantification of protein markers in clinical testing.[109,342,348–351]

The study of proteomic changes in health and disease can utilize various sample types, each with unique advantages and limitations. While cell cultures and model organisms offer accessible approaches to studying human diseases, they often struggle to fully replicate *in-vivo* disease phenotypes. Patient-derived cell culture or organoid models, especially in combination with xenotransplantation can alleviate some of these limitations.[352–359] On the other hand, clinics routinely collect patient material in the form of body fluids, punch biopsies and surgical tissue specimens that offer direct insights into disease pathology in the human body.

When working with patient material, a key consideration for clinical proteomics is the assembly of well stratified clinical cohorts. Balancing potential confounders, such as age, biological sex, life style, etc., between case and control cohorts can reduce the chance study biases and misinterpretation of biomarker relevancy.[360–363] Larger cohort sizes, additionally enable higher statistical power particularly in the study of diseases with low effect size, and decrease the effect of cofounders (**Figure 16**).[361,362] While this was previously limited by long sample acquisition times, advances in LC and MS

technology now enable the routine analysis of thousands of samples.[64,364–366] Another important aspect is the standardization of sample collection to prevent batch effects and sample contamination.[367] This is particularly crucial for blood plasma proteomics, where contaminations, for instance with cellular blood components, greatly affects sample integrity and the validity of potential biomarker discoveries.[344]

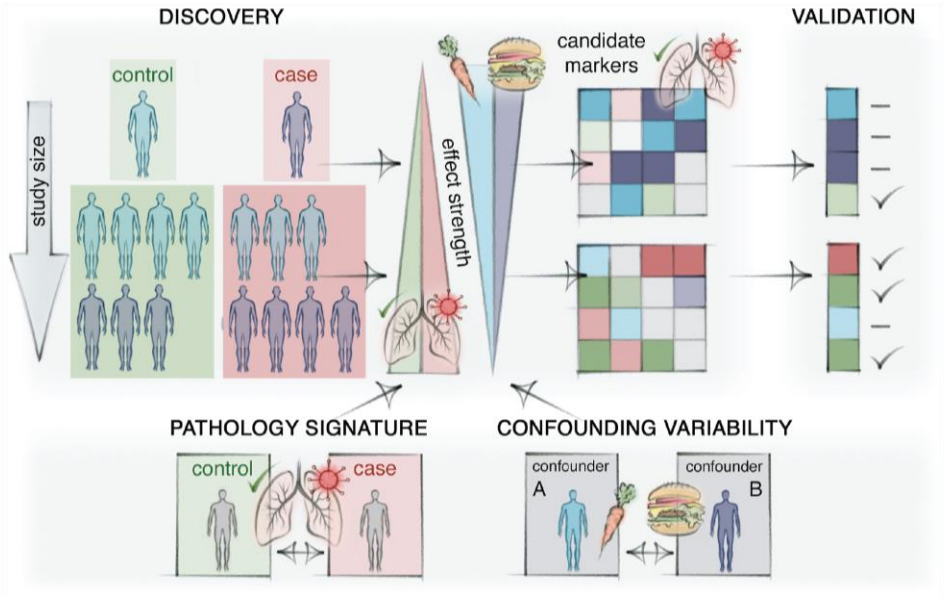


Figure 16 Clinical cohort design. Larger cohort size for discovery proteomics offers the double benefit of increasing disease effect size and reducing the influence of cofounders, such as patient age, biological sex, and lifestyle. This enables more confident identifications of potential biomarker. Adapted from ref[361] as permitted by the CC-BY 4.0 international license.

The analysis of blood plasma offers a minimally invasive strategy to evaluate proteomic changes in health and disease. Obtained through the removal of cellular blood components, blood plasma generally contains three classes of proteins: Classical plasma proteins that are generally produced in the liver, secreted and have an active function in the blood, immunoglobulins, and so-tissue leakage proteins that are released into the blood stream after tissue damage, and could potentially serve as biomarkers.[362,368–371] A notable example for this are cardiac troponins, which are routinely used in clinical analysis for the diagnosis of acute myocardial infarction.[372] In depth analysis of the plasma proteome by MS, however, is limited due to the large abundance range in plasma. From the most abundant to the least abundant protein, plasma is expected to span at least 12 orders of magnitude. Additionally, about half of the protein content in plasma is comprised of the classical plasma protein serum albumin, and the 15-20 top most abundant proteins account for about 99% of the

biomass.[362,369,370,373,374] As modern mass spectrometers can often only cover about 5 orders of magnitude within a single scan, these high abundant proteins limit the identification of lower abundant proteins. More advanced data acquisition methods, such as narrow window DIA, or additional orthogonal peptide separation, e.g., through ion mobility implementations, can alleviate some of these difficulties. Moreover, multiple upstream sample processing steps for selective enrichment/depletion of plasma proteins, e.g., using antibodies, acid precipitation or nanoparticles, have been implemented. In combination with highly sensitive MS instrumentation, such as the Orbitrap Astral of timsTOF HT mass spectrometers and optimized acquisition strategies, these methods now enable identification of 1000-2000 proteins in a single LC/MS run.[327,366,373,375–380]

Fresh frozen or FFPE patient tissue samples, albeit more invasive than plasma, allow for greater reflection of disease manifestation and progression by directly analyzing proteins in the affected tissues. Higher concentration of disease-relevant proteins and less dynamic range issues than plasma, might aid in the discovery of disease biomarkers or therapeutic targets. Over the years, many atlases for in depth characterization of organ-specific proteomes in health and disease have been published and serve as great resources for the continued investigation of disease phenotypes.[381–388] Additionally, analysis of patient tissue retains the cellular context and allows for studying cell-type dependent effects, the analysis of cell-cell interactions and signaling. In combination with microscopy, histological staining approaches, and laser-microdissection this moreover provides a spatial aspect by enabling the analysis of different regions within a tissue, including the tissue microenvironment, or even more fine-grained proteomic analysis at cell type resolution. This offers unique insights into disease progression and heterogeneity.[389–395]

1.4.2 Deep Visual Proteomics

Many diseases manifest in the tissues of our bodies and change their normal morphology. These observed changes can be traced back to innumerable molecular changes on the level of single cells, each contributing to a heterogenous mosaic of cells in the unique tissue architecture. Conventional proteomics approaches, however, often lose this spatial information through the analysis of bulk tissue or even sorted cells. To overcome these limitations, several spatial proteomics techniques have been developed over the years. MS imaging, for instance, directly maps the spatial distribution of proteins using a focused ionization beam to acquire mass spectra of defined tissue regions within a sample.[396–398] Multiplexed ion beam imaging and imaging mass cytometry on the other

hand utilize metal-labeled antibodies for protein mapping.[399,400] Another approach is the systematic antibody-based analysis of proteins to determine their cell type and tissue specificity, as well as their subcellular localization. One such effort is the Human Protein Atlas, a spatial proteomics resource that aims to map the entire human proteome.[384,401,402]

While these approaches have their benefits, our group aimed to combine several layers of technology to generate molecular proteomic maps at single cell (type) resolution. Termed Deep Visual Proteomics (DVP), this recent innovation combines high content imaging, machine-learning assisted image-based cell classification and segmentation, with laser-microdissection and high-sensitivity LC/MS (**Figure 17**).[395,403] DVP can be used to investigate cell-type resolved proteomes from fresh frozen as well as FFPE tissue, while preserving the spatial context. Briefly, tissue sections, mounted on membrane slides[404], are stained by immunohistochemistry or immunofluorescence to define cellular features, such as the cell shape, diameter or granularity, and differentiate between different cell types of interest by targeted staining. Extensions of this protocol now allow for highly multiplexed staining.[405] After high-resolution microscopy images of the stained tissue sections are acquired, pre-trained deep learning-based models are applied for image-based cell segmentation using the BIAS software, Cellpose or other tools for biological segmentation.[395,406,407] Cells of interest are then excised using laser-microdissection at single cell resolution. The excised cells are directly collected in 96- or 384-well plates and subjected to automated sample processing in low volumes for optimal protein retention. Digested proteins are separated using liquid chromatography and measured on a high-sensitivity mass spectrometer. In a standard DVP experiment, multiple hundred cell shapes per cell type or morphological feature are pooled to obtain high proteomic depth and a robust cellular phenotype. While the original manuscript analyzed 500-1000 cell shapes per cell type and achieved a proteomic depth of ~5000 protein groups, recent innovation in MS technology maintain high proteomic depth at greatly reduced input material (~100 shapes).[408–410]

Pushing this to the next level, colleagues in the department developed a workflow to enable the analysis of single cells.[297,411] Based on other single cell omics, such as single cell transcriptomics, single cell proteomics aims to analyze individual cells to capture cell heterogeneity and reveal cellular dynamics, among others. Due to the limited sample amount, achieving biologically relevant proteomics depth has been challenging.

However, the rapid advances in cell isolation, sample preparation and MS technology now enable to measure up to multiple thousand protein groups from single cells.[61,214,215,217,248,297]

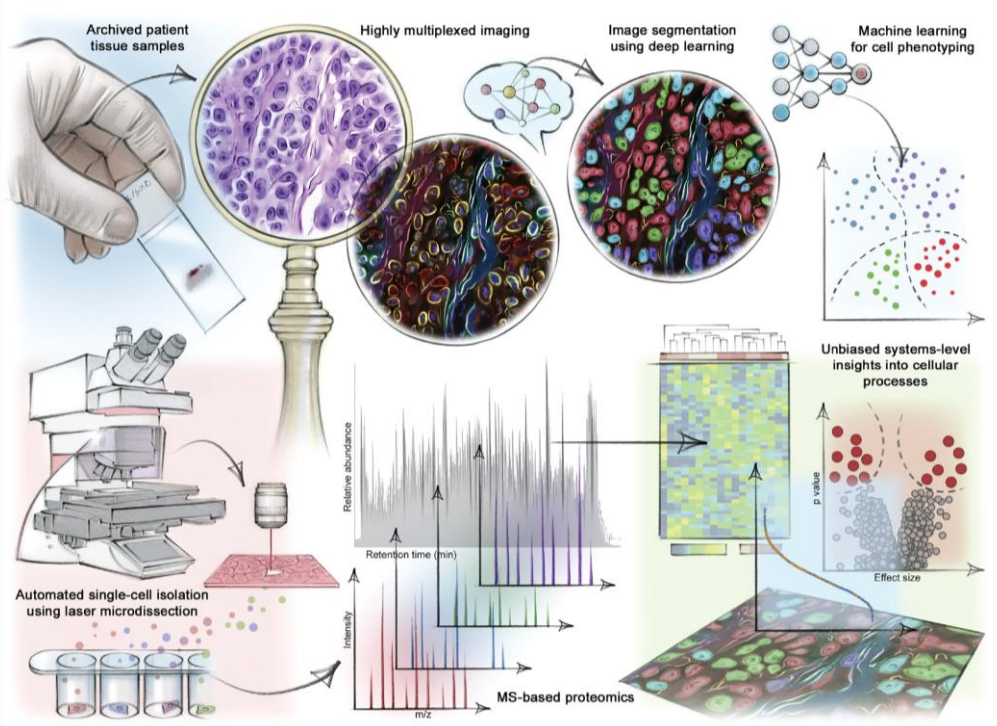


Figure 17 The Deep Visual Proteomics (DVP) workflow. DVP combines high content imaging, machine-learning assisted image-based cell classification and segmentation, with laser-microdissection and high-sensitivity LC/MS. This allows the analysis of cell type-resolved proteomes while preserving their spatial context in the analyzed tissue. Adapted from ref[403]. Copyright (2022) Elsevier under license 5902730019602.

For classic single cell proteomics approaches cultured cells are usually dissociated and sorted, for instance using the cellenONE cell sorter.[412,413] In comparison to this, single cell deep visual proteomics (scDVP) extends single cell proteomics to the intact tissue context. Initially applied to the analysis of single hepatocytes to evaluate proteomics changes along the portal to central vein axis in murine liver, scDVP allows to map the protein abundances of single cells back to their spatial location in tissue samples.[411] Apart from the study of tissue zonation effects, this is also advantageous for the analysis of disease pathologies with a defined spatial component. One such example is the distribution of cells with alpha-1 antitrypsin aggregates in fibrotic liver sections of patients with alpha-1 antitrypsin deficiency presented in this thesis (Article 6).

Since its introduction in 2022, DVP has been applied to study a multitude of pathological conditions, such as colorectal adenoma[414], borderline ovarian cancer[394], and Hodgkin's lymphoma.[415] In this thesis, we additionally showcase its potential for personalized

medicine in signet-ring cell carcinoma, the evaluation of model systems for human disease and elucidating proteotoxic stress signals in Alpha-1 antitrypsin deficiency (Articles 4-6). Most notable, DVP demonstrated a breakthrough recently by revealing the involvement of the JAK/STAT signaling pathway in a lethal skin disease called toxic epidermal necrolysis, which lead to the successful treatment of ten patients with already FDA-approved JAK inhibitors. This highlights the translatability of DVP for clinical applications.[410]

2. Aims of the thesis

As discussed in the introduction, advances in MS-based proteomics are greatly driven by innovations in MS technology. In recent years, novel instruments have further pushed the boundaries on sensitivity, acquisition speed, and accuracy, which now enables the routine analysis of thousands of clinical samples, near full proteomes and even single cells. While state-of-the-art MS technologies provide a basis for these applications, it is necessary to optimize instrument parameters and acquisition strategies to utilize them to their fullest potential. Having been involved in the technical evaluation of state-of-the-art instrumentation[201] during my master thesis, I developed a great interest in MS technology itself and further explored this topic during my PhD. Overall, the overarching focus of my thesis was to evaluate MS technologies, contribute optimal MS acquisition strategies, and apply them to clinical and spatial proteomics. This goal also included, facilitating a number of projects within the lab by giving introductions to MS technology, designing DIA methods or directly advising on and designing acquisition strategies.

Translating the previously used setup and method to the state-of -the-art LC/MS setup, I optimized a faster acquisition strategy for plasma proteomics, which we applied to studying the effects of muscle loss in humans undergoing bedrest (**Article 2**). Focusing on further extending the functionality of existing hardware, I collaborated with Thermo Fisher Scientific to evaluate the full mass range application of a computation approach to either increase the mass resolution or decrease the acquisition time of Orbitrap mass spectrometers (**Article 1**). As MS applications are increasingly moving to shorter gradients, acquisition speed is particularly limiting. Here, ΦSDM could significantly increase the performance of Orbitrap instrumentation without having to upgrade the existing hardware.

Through our long-standing collaboration with Thermo Fisher Scientific, I then had the opportunity to gain pre-access to the Orbitrap Astral MS and used my obtained knowledge to optimize DIA acquisition methods for the application in our lab, including full proteome analysis, multiplexed DIA, and low input applications. To fully make use of this data and to establish a framework for the analysis of upcoming acquisition strategies, I contributed to a modular, open-source framework, for the analysis of DIA data, which is particularly suitable for data produced on state-of-the art time-of-flight (TOF) analyzers (**Article 3**). The sensitivity, acquisition speed, and resolution of the Orbitrap Astral MS has shown to be particularly advantageous for low input applications and is broadening the applicability of deep visual proteomics (DVP). Three such DVP

applications are included in this thesis. Focusing first on tissues from a single patient with signet ring cell carcinoma (SRCC), we showcased the potential of DVP for personalized medicine and were able to propose a treatment option that effectively halted tumor progression (**Article 4**). We next used DVP in combination with the Orbitrap Astral MS to evaluate the phenotypic shifts after xenotransplantation of organoid models. In a human mucosa model, we could show that xenotransplanted tissue was closer to human physiology and regained its functional profile in comparison to *in-vitro* organoid cultures highlighting the potential of this approach for studying human disease (**Article 5**). Lastly, we extended the previously described single cell DVP (scDVP) workflow to formalin-fixed paraffin-embedded (FFPE) tissue, increasing the proteomic depth by 50% using optimized variable window methods, and applied it to study proteotoxic stress in alpha-1-antitrypsin deficiency (**Article 6**).

3. Publications

3.1 Expanding the usability of MS technology

Paired with fast LC systems, modern MS has especially shown potential for applications with clinical application, such as the identification of biomarkers in human health and disease.[361] While novel mass spectrometers offer great potential, multiple factors can limit the usability of novel instrumentation. One of these is the cost associated with novel high-resolution MS instruments, which make upgrading to the latest releases a privilege of well-funded institutions. To bridge this performance gap, we evaluated a computational solution to increase the acquisitional speed or resolution of existing Orbitrap MS instruments (**Article 1**). As part of the MARS-PRE project, funded by the Italian Space Agency, we evaluated the effects of muscle loss, caused by bed rest or cancer cachexia, on the plasma proteome (**Article 2**) using the at the time state-of-the-art Exploris480 MS. Another limitation can be the available analysis software suits, as these might not be able to handle novel acquisition methods or the amount of data produced my modern mass spectrometers. To overcome this, we introduced AlphaDIA, a modular, open-source framework for DIA analysis (**Article 3**).

Article 1: Full Mass Range ΦSDM Orbitrap Mass Spectrometry for DIA Proteome Analysis

Molecular and Cellular Proteomics 23(2), 100713 (2024)

Sophia Steigerwald[1], Ankit Sinha[1], Kyle L. Fort[2], Wen-Feng Zeng[1], Lili Niu[3], Christoph Wichmann[4], Arne Kreutzmann[2], Daniel Mourad[2], Konstantin Aizikov[2], Dmitry Grinfeld[2], Alexander Makarov[2], Matthias Mann[1,3], and Florian Meier[1,5]*

[1]Department Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany; [2]Thermo Fisher Scientific (GmbH), Bremen, Germany; [3]Department Clinical Proteomics, NNF Center for Protein Research, University of Copenhagen, Copenhagen, Denmark; [4]Department Computational Systems Biochemistry, Max Planck Institute of Biochemistry, Martinsried, Germany; [5]Functional Proteomics, Jena University Hospital, Jena, Germany

**Corresponding author*

Since their commercialization in 2005, Orbitrap mass analyzers have become one of the most widely used mass analyzers in the field of proteomics. This can mainly be attributed to their high mass accuracy and resolving power. Over the years multiple improvements in term of mass resolution, such as the high-field Orbitrap geometry or so-called

enhanced Fourier transformation (eFT), have been made. However, while mass resolution scales with transient time, or the duration for which the image current of the trapped ions is recorded, practicality often limits the use of very long transients in favor of higher quantitative accuracy. Multiple computational approaches have been proposed to overcome these limitations, but only recently were able to provide additional spectral content and enable high resolution at lower transient times.[242,416,417]

In this study we apply one of these approaches, termed phase-constrained spectrum deconvolution method (ΦSDM) to the full mass range, evaluate its performance and highlight it's benefits for proteomic applications, particularly short gradient DIA. In theory ΦSDM is able to either at least double the mass resolving power at a given transient half the transient at a given resolving power in comparison to eFT.[242] Here, I was able to show that this theoretical principle indeed translates to an at least doubled mass resolution in complex proteomes with minimal scan overhead time. This required me to extend my analysis to the raw data level and systematically evaluate all observed peak distances in the ΦSDM spectra in comparison to the resolution limits imposed in eFT as a proxy for ΦSDM resolving power. Overall, the improved resolution, significantly increased signal to noise ratios and was especially beneficial in areas of high peptide density. As proteomics applications are gradually moving to higher through-puts, increased resolving power and faster acquisition speeds become more and more vital. In line with this, we found that ΦSDM signal processing is particularly advantageous for increasingly shorter gradient times. While we focused on constant transient times (equals increased resolution), short gradient DIA applications could additionally benefit from ΦSDM's possibility to shorten the transient time at a given resolution to either increase quantitative accuracy or decrease spectral complexity. We hypothesize that ΦSDM could be a useful addition to extend the potential of existing Orbitrap mass spectrometers and should be applicable for a wide range of proteomics applications beyond the label-free DIA acquisitions shown in this manuscript.

Contribution:
First-authorship. Under the guidance of Florian Meier-Rosar in Matthias Mann's group and in close collaboration with Thermo Fisher Scientific. Florian and I conceptualized this study. I conducted the experiments and analyses presented in this paper, made all figures and wrote the first draft of the manuscript. Florian Meier-Rosar and I edited the manuscript with input from Matthias Mann and our collaboration partners at Thermo Fisher Scientific.

MCP TECHNOLOGICAL INNOVATION AND RESOURCES

Full Mass Range ΦSDM Orbitrap Mass Spectrometry for DIA Proteome Analysis

Authors

Sophia Steigerwald, Ankit Sinha, Kyle L. Fort, Wen-Feng Zeng, Lili Niu, Christoph Wichmann, Arne Kreutzmann, Daniel Mourad, Konstantin Aizikov, Dmitry Grinfeld, Alexander Makarov, Matthias Mann, and Florian Meier

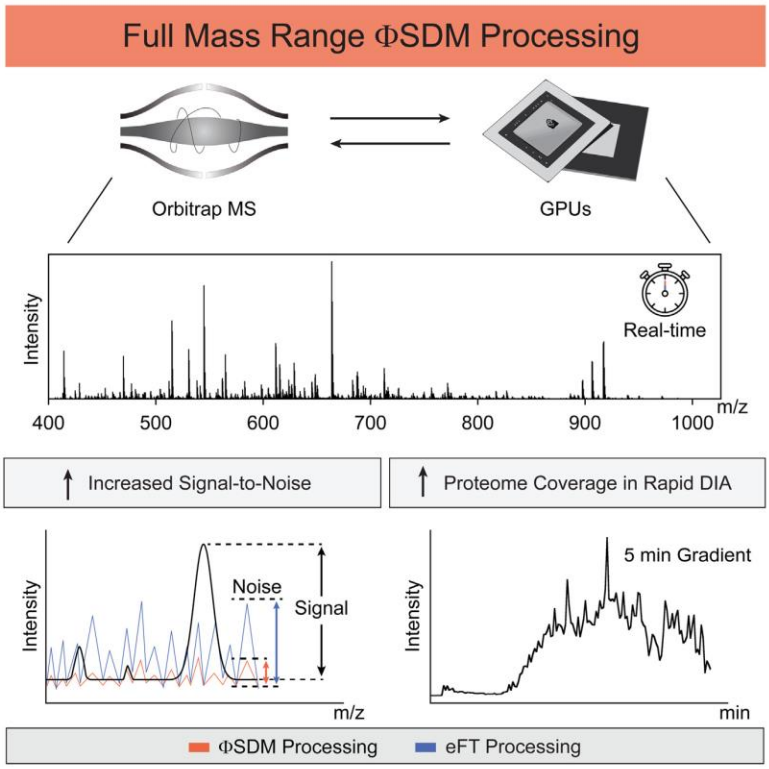
Correspondence

fmeier@biochem.mpg.de; florian.meier@med.uni-jena.de

In Brief

We describe the full mass range application of the ΦSDM signal processing algorithm for Orbitrap mass spectrometry in real time and evaluate its performance for label-free data-independent acquisition. ΦSDM increases the mass resolving power beyond the limits imposed by Fourier transformation, with advantages in areas of high spectral complexity and for fast chromatographic gradients. Our results suggest that it will be interesting to explore full mass range, real-time ΦSDM signal processing also for other applications of Orbitrap MS in proteomics research.

Graphical Abstract



Highlights

- ΦSDM signal processing increases Orbitrap mass resolution (or speed) >2-fold.
- GPUs enable real-time ΦSDM processing of full mass range spectra.
- ΦSDM resolves interfering signals in complex DIA spectra.
- Increased identification rates in short gradients.

2024, Mol Cell Proteomics *23*(2), 100713

https://doi.org/10.1016/j.mcpro.2024.100713

MCP TECHNOLOGICAL INNOVATION AND RESOURCES

Full Mass Range ΦSDM Orbitrap Mass Spectrometry for DIA Proteome Analysis

Sophia Steigerwald[1], Ankit Sinha[1], Kyle L. Fort[2], Wen-Feng Zeng[1], Lili Niu[3], Christoph Wichmann[4], Arne Kreutzmann[2], Daniel Mourad[2], Konstantin Aizikov[2], Dmitry Grinfeld[2], Alexander Makarov[2], Matthias Mann[1,3], and Florian Meier[1,5,*]

Optimizing data-independent acquisition methods for proteomics applications often requires balancing spectral resolution and acquisition speed. Here, we describe a real-time full mass range implementation of the phase-constrained spectrum deconvolution method (ΦSDM) for Orbitrap mass spectrometry that increases mass resolving power without increasing scan time. Comparing its performance to the standard enhanced Fourier transformation signal processing revealed that the increased resolving power of ΦSDM is beneficial in areas of high peptide density and comes with a greater ability to resolve low-abundance signals. In a standard 2 h analysis of a 200 ng HeLa digest, this resulted in an increase of 16% in the number of quantified peptides. As the acquisition speed becomes even more important when using fast chromatographic gradients, we further applied ΦSDM methods to a range of shorter gradient lengths (21, 12, and 5 min). While ΦSDM improved identification rates and spectral quality in all tested gradients, it proved particularly advantageous for the 5 min gradient. Here, the number of identified protein groups and peptides increased by >15% in comparison to enhanced Fourier transformation processing. In conclusion, ΦSDM is an alternative signal processing algorithm for processing Orbitrap data that can improve spectral quality and benefit quantitative accuracy in typical proteomics experiments, especially when using short gradients.

LC–MS has become the method of choice for the investigation of protein sequences and complex proteomes (1, 2). One of the most widely used mass analyzers for MS-based proteomics is the Orbitrap analyzer, first described in 2000 (3–5). In Orbitrap MS, the image current of trapped ions is recorded ("transient") and converted into a high-resolution accurate mass spectrum using Fourier transformation (FT). As with other FT mass spectrometry (MS) analyzers, mass resolution scales with the transient duration, and even though enhanced FT (eFT) calculations enabled a twofold increase in mass resolving power using the same transient (6, 7), the mass resolution is inherently limited by the Fourier uncertainty. Interpolation techniques have been proposed to address this limitation; however, they lack the power to increase the spectral information content (8, 9). Only more recently, several approaches in ion cyclotron resonance MS have succeeded and are able to provide the required mass resolution at shorter transients (10–14). In particular, a novel computational strategy for processing Orbitrap transients, termed phase-constrained spectrum deconvolution method (ΦSDM), has the potential to double the mass resolving power at a given Orbitrap transient and could thereby significantly improve spectral quality and acquisition speed (15, 16). ΦSDM has already been implemented in the acquisition software of the most recent Orbitrap mass spectrometers (17, 18); however, because of the computational cost associated with the processing algorithm, its application has so far been limited to a narrow *m/z* region, such as the *m/z* range of tandem mass tag reporter ions (19, 20).

Here, we reasoned that a full mass range implementation of ΦSDM should be highly beneficial for data-independent acquisition (DIA), which has become a key driver of advancements in MS-based proteomics in recent years (21, 22). First popularized on a quadrupole time-of-flight instrument (21), DIA strategies have now been established on a multitude of mass analyzers (23–29). Unlike data-dependent acquisition (DDA), DIA does not sequentially fragment the top N most abundant peaks but cycles through the entire *m/z* range using isolation windows of defined width to simultaneously fragment all detectable precursors in each window. However, optimizing DIA methods often requires a compromise between spectral complexity and cycle time associated with a tradeoff between proteome coverage and quantitative accuracy (22, 24). In Orbitrap MS, narrow isolation windows and high mass resolution reduce complexity, improving spectral deconvolution, but this comes at the cost of longer cycle times and

From the [1]Department Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany; [2]Thermo Fisher Scientific (GmbH), Bremen, Germany; [3]Department Clinical Proteomics, NNF Center for Protein Research, University of Copenhagen, Copenhagen, Denmark; [4]Department Computational Systems Biochemistry, Max Planck Institute of Biochemistry, Martinsried, Germany; [5]Functional Proteomics, Jena University Hospital, Jena, Germany
*For correspondence: Florian Meier, fmeier@biochem.mpg.de or florian.meier@med.uni-jena.de.

https://doi.org/10.1016/j.mcpro.2024.100713

therefore a decrease in the ability to accurately quantify chromatographic peaks. To address this, here we investigated the potential of full mass range ΦSDM for DIA proteomics. In particular, we tested the compatibility with high-throughput DIA MS strategies using short LC gradients.

EXPERIMENTAL PROCEDURES

Sample Preparation

Human cervix carcinoma (HeLa) cells were cultured in Dulbecco's modified Eagle's medium (Life Technologies Ltd) containing 20 mM glutamine, 10% fetal bovine serum, and 1% penicillin–streptomycin. After harvest, the cells were resuspended in PreOmics lysis buffer and incubated at 95 °C for 10 min to reduce disulfide bridges, alkylate cysteine residues, and denature proteins. Samples were sonicated using a rod sonicator (Branson SFX 250 Digital Sonifier) and subsequently incubated at 95 °C for an additional 5 min. HeLa cell lysates were diluted with an equal volume of water and digested overnight using equal amounts of LysC and trypsin (1:100 ratio at protein level). Following digestion, peptides were acidified to a final concentration of 1% TFA and purified on StrataTM-X-C (Polymeric Strong Cation) cartridges. Peptides were eluted in 80% acetonitrile (ACN)/1.25% NH_4OH and subsequently dried using a SpeedVac (Eppendorf). Samples were resuspended in buffer A* (0.1% TFA, 2% ACN, or buffer A [0.1% formic acid (FA)]), for measurement with the Thermo Scientific EASY-nLC 1200 system or the Evosep LC system, respectively. Peptide concentrations were estimated by measuring absorbance at 280 nm on a Thermo Scientific NanoDrop 2000 spectrophotometer. For online MS injection using the Evosep One (LC) system, peptides were loaded onto Evotips according to the manufacturer's instructions.

High-pH Reverse-Phase Fractionation for Spectral Library Generation

For the short-gradient DIA experiments, gradient-specific spectral libraries were generated from 48 high-pH reverse-phase fractions for each gradient (5, 12, and 21 min) using a "spider" low-flow fractionator (30). The fractions were dried using a SpeedVac and resuspended in buffer A for Evotip loading and subsequent LC–MS analysis using the Evosep One system. We chose the peptide input amount for fractionation based on the injection amounts used for each gradient length. Peptide concentrations were estimated using a NanoDrop 2000 spectrophotometer, and 200, 100, and 50 ng per fraction were loaded on Evotips for the 60 samples per day (SPD), 100 SPD, and 200 SPD LC methods.

LC–MS

All data were acquired on a Thermo Scientific Orbitrap Exploris 480 mass spectrometer (17). Standard LC measurements were performed using a Thermo Scientific EASY-nLC 1200 system, and an Evosep LC system (31) was used for preprogrammed short gradients with gradient lengths of 21, 12, and 5 min (60, 100, and 200 SPD). For the EASY-nLC chromatography system, we used an in-house packed 50 cm, 75 μm i.d. capillary column with 1.9 μm Reprosil-Pur C18 beads (Dr Maisch) and a laser-pulled electrospray emitter. The column temperature was maintained at 60 °C (sonation column oven). For the 120 min nLC gradient, mobile phase A was water with 0.1% FA, and mobile phase B was 80% ACN and 0.1% FA in water. Peptides were separated at a constant flow rate of 300 nl/min with a linear gradient of 5 to 30% mobile phase B within 95 min, followed first by a linear increase from 30 to 65% mobile phase B within 5 min and then a linear increase from 65 to 95% within another 5 min, where it was kept for 5 min before re-equilibration. Evosep measurements for 60 and 100 SPD (preprogrammed gradients) were performed using an in-house packed 8 cm, 150 μm i.d. capillary column with 1.9 μm Reprosil-Pur C18 beads (Dr Maisch). Column temperature was maintained at 20 °C. For the 200 SPD method, a commercial Evosep capillary column (EV1107) of 4 cm, 150 μm i.d. with 1.9 μm Reprosil-Pur C18 beads (Dr Maisch) was connected to an Evosep 30 μm i.d. stainless steel emitter (EV1086). Column temperature was maintained at 40 °C using a butterfly oven (Phoenix S&T). For both LC setups (EASY-nLC and Evosep One LC), in-house packed columns were interfaced with the Thermo Scientific NanoSpray Flex Ion Source, whereas the commercial column and emitter setup (for Evosep 200 SPD) was interfaced with the Thermo Scientific EasySpray Ion Source. For all measurements, spray voltage was set to 2400 V, RF level was set to 40, and the heated capillary temperature was set to 275 °C.

For EASY-nLC DIA, Orbitrap full MS scans were acquired from 400 to 1000 *m/z* at a resolution of 60,000 at *m/z* 200 with a normalized automated gain control (AGC) target of 200% and a maximum ion injection time of 45 ms. For MS/MS scans, the collision energy was set to 30%, the resolution to 15,000 at *m/z* 200, the normalized AGC target to 3000%, whereas the maximum injection time was set to "auto," and the mass range was *m/z* 400 to 1000. For a theoretical cycle time of 3 s, 82 DIA windows of 7.3 *m/z* and an overlap of 1 *m/z* were used. For Evosep One LC DIA measurements, we designed gradient-specific methods. The general method settings for full MS and MS/MS were as aforementioned, except for the full MS AGC target, which was set to 300%. Cycle times and window placement were optimized according to the expected peak width (as reported by Spectronaut (Biognosys) based on 1.7 * full width at half maximum) of the different gradient lengths at 21, 12, and 5 min for 60, 100, and 200 SPD, respectively. For the 60 SPD method, 53 DIA windows of 11.3 *m/z* with an overlap of 1 *m/z* were used (~2 s cycle time). For the 100 and 200 SPD methods, 38 DIA windows of 15.4 *m/z* with an overlap of 1 *m/z* were chosen (~1.5 s cycle time). Experiments to generate Evosep gradient-specific spectral libraries were performed using a DDA top12 method. Full MS scans were acquired from 400 to 1000 *m/z* at a resolution of 60,000 at *m/z* 200 with a normalized AGC target of 300% and a maximum injection time of 25 ms. Precursor ions were isolated in a 1.3 Thomson window, normalized AGC target was set to 200% with a maximum injection time of 22 ms, and the normalized collision energy was set to 30%. Precursors with charge states of 1+ or above 5+ were excluded from sequencing, and the exclusion time for previously targeted precursors was set to 30 s. All Orbitrap mass spectra were recorded in centroid mode.

Real-Time and Full Mass Range ΦSDM Signal Processing

The ΦSDM has previously been described and applied successfully to small *m/z* areas for improved mass resolution of tandem mass tag reporter ions (15, 19). In brief, the algorithm is capable of resolving spectral features beyond the limitation imposed by the Fourier uncertainty by deconvolving an observed standard eFT spectrum on a multiply refined frequency grid with the sinc function as its basis functions. The sinc function reflects the finite length of a transient signal and is completely characterized by its length (*i.e.*, known *a priori*). The ΦSDM spectrum is a solution that minimizes discrepancy between the model and the observed signals in sense of L2 norm, being subject to a phase constraint in a narrow interval around the precalibrated phase. To avoid overdetermination, the phase constraint is relaxed to form a cone. For the full mass range implementation of ΦSDM, we interfaced the instrument internal PC with additional graphics processing units (GPUs). ΦSDM settings were accessed through a research prototype Tune, version (3.1.279.9, Thermo Fisher Scientific). Before measurements, ΦSDM phase and noise levels were calibrated. ΦSDM processing was performed on the external GPUs

("on box"), the number of iterations was limited to 150, the noise threshold was set to 1.41, and version 2 of the backfilling approaches was applied.

Raw Data processing

DDA raw files for the spectral library were analyzed, and the libraries were generated using the Pulsar algorithm in Spectronaut, version 15.6 with default settings. The 5 min library consisted of 26,822 precursor and 4196 protein groups, the 12 min library of 61,111 precursor and 6824 protein groups, and the 21 min library of 92,865 precursor and 8173 protein groups. Targeted data extraction from DIA raw files was performed with Spectronaut, version 15.6 (32). The "Protein LFQ Method" was set to MaxLFQ, "Data Filtering" to Q-value, the "Normalization Strategy" to local normalization, and "Row Selection" was based on Q-value percentile with a "Fraction" setting of 0.2. For library generation and direct-DIA analysis, raw files were searched against a target/decoy database of the human proteome (UniProt, September 2021) with and without isoforms (80,426 and 20,588 entries). Trypsin/P was selected to generate peptides, and a maximum number of two missed cleavages were allowed. For all searches, carbamidomethyl (C) was set as a fixed modification, and acetyl (protein N-term) and oxidations (M) were set as variable modifications. For the MS1 and MS2 mass tolerance, we used the default value for Orbitrap MS in Spectronaut (40 ppm). A 1% false discovery rate cutoff at precursor and protein levels was applied.

Data analysis

Statistical analysis and data visualization of the Spectronaut output tables was performed in Python (version 3.8.8) using matplotlib, pandas, and seaborn. For the manual inspection of close proximity peptide signals, we used a custom Python script based on alpharaw to read RAW data, alphabase to process peptides and fragments, and alphaviz (33, 34) to visualize peptide to spectrum matches (https://github.com/MannLabs).

For the analysis of neighboring peaks, because the resolving power in Orbitrap MS is inversely proportional to $\sqrt{m/z}$, we first calculated a theoretical tolerance window as a function of m/z assuming a nominal resolution of 30,000 at m/z 200. The resolving power is calculated as $R = (m/z)/(\Delta m/z)$, with m/z being the m/z value of a given peak and $\Delta m/z$ being the smallest peak-to-peak distance still resolvable at a given resolving power. We used this tolerance window to select peaks in close proximity to all peaks in all MS2 spectra of a given LC–MS experiment. The neighboring peak pairs were then filtered for noise using 4% relative to the base peak as an abundance threshold and retaining only pairs for which one of the peaks was not greater than four times more abundant than the other one. The resulting peak neighbor pairs represent peak pairs that require a nominal resolving power of ≥30,000 to be resolved, and their m/z and interpeak distance can therefore be considered as a measure of resolving power (3, 35–37).

Signal-to-noise ratio (SNR) scatter plots were filtered for outliers with log2 SNRs of 13 and 14 or higher for the x- and y-axis, respectively. This was necessary because these outliers (supplemental Fig. S4*A*) represent instances, for which Spectronaut could not determine an empirical noise value for a given extracted ion chromatogram (XIC), resulting in an overestimation of the SNR.

Experimental Design and Statistical Rationale

All experiments were performed using aliquots of the same HeLa digest to minimize confounders from preanalytical steps. The 2 h HeLa experiment for the analytical evaluation was performed in quadruplicates, whereas all short-gradient experiments were performed in triplicates. Evaluation of the effects of ΦSDM on spectral quality, however, was performed on a per-spectrum level over the averaged information of thousands of spectra in a single run. To benchmark the two alternative signal processing algorithms, we kept the MS method settings identical for each comparison, except for activating ΦSDM or not (eFT).

RESULTS

Full Mass Range ΦSDM Computation

The ΦSDM can resolve signals in the mass spectrum that are closer than the limitation imposed by the Fourier uncertainty. This is achieved by iteratively fitting the observed signal to a refined frequency grid (15). To enable this computationally expensive method for the full mass range, we interfaced an Orbitrap mass spectrometer with GPUs for highly parallelized processing (Fig. 1). In our setup, the image current induced on the outer electrode of the Orbitrap analyzer (transient) is marshaled from the instrument's internal computer to the GPUs. We reasoned that four Titan Xp Nvidia graphic cards installed on an auxiliary computer should provide sufficient resources to process multiple signals in parallel with an optimized CUDA C++ implementation of the ΦSDM algorithm. The calculated frequency spectrum is centroided and marshaled back to the instrument computer, where it is converted into a mass spectrum (4) and stored in the proprietary Thermo Fisher RAW file data format.

The key feature of ΦSDM is that it uses the phase as a constraint for signal deconvolution. To speed up the computation, making use of the very high stability of the MS electronics, we precalibrated the phase function externally as part of our weekly instrument maintenance routine. Furthermore, based on preliminary experiments, we parametrized the ΦSDM algorithm as detailed in the Experimental Procedures section and set the number of iterations to 150, which yielded a good compromise between processing speed and resolving power.

Resolving Power of Full Mass-Range ΦSDM

Having established an experimental setup that should be capable of processing full mass range spectra with ΦSDM in real time, we first inspected the resulting mass spectra with complex proteomics samples. For this, we analyzed the HeLa cell line proteome with 2 h gradients with DIA using either ΦSDM or eFT signal processing (Fig. 2*A*). Our DIA method comprised 82 equidistant isolation windows from m/z 400 to 1000 resulting in a cycle time of ~3 s with transient times of 128 and 32 ms for full MS and MS/MS scans. These correspond to a nominal eFT resolution of 60,000 and 15,000 at m/z 200. Figure 2*B* shows two representative mass spectra for eFT (*upper panel*) and ΦSDM (*lower panel*) with matching retention time and isolation window between the two raw files. As expected, both spectra appeared very similar (Fig. 2*B* and supplemental Fig. S1). Upon closer inspection, we observed additional peaks in the ΦSDM spectrum in close proximity to peaks that ΦSDM and eFT had in common. To investigate the

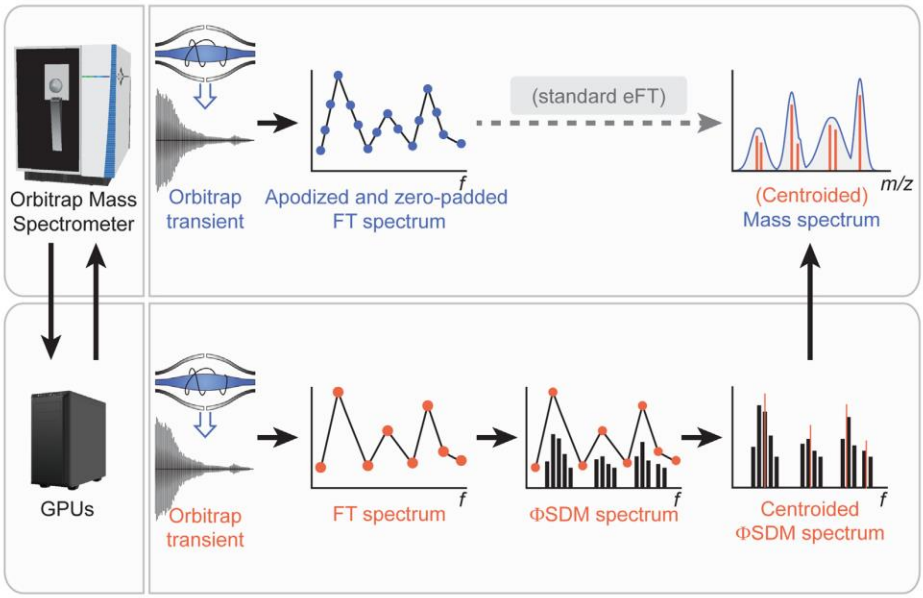


FIG. 1. **ΦSDM for Orbitrap signal processing.** The image current induced on the detection plates of the Orbitrap by the oscillating ions is amplified and recorded as a transient signal followed by Fourier transformation (FT). With the assistance of an array of GPU cards to compensate for added computation costs, the resolution of the FT frequency spectrum is further enhanced by processing it with the ΦSDM. The ΦSDM spectrum is centroided, converted to the mass spectrum, and then stored in RAW format on the MS internal computer. ΦSDM, phase-constrained spectrum deconvolution method; GPU, graphics processing unit.

nature of these signals systematically, we parsed all MS2 spectra from a full LC–MS experiment with ΦSDM to find all neighboring peak pairs. Here, we defined close neighbors as *m/z* peak pairs with a distance that requires a resolving power ≥30,000 at *m/z* 200 to be resolved (see the Experimental Procedures section). In total, we observed >100,000 such peak pairs across the active part of the LC gradient (between scan #12,500 and #148,000) covering an *m/z* range between 100 and 1700. For these, we then calculated the theoretical resolving power required to distinguish them in a mass spectrum at full width half maximum (Fig. 2*C* and supplemental Table S1). The pairwise peak resolution across the *m/z* range in bins of 100 *m/z* followed the expected inverse proportionality between resolving power and $\sqrt{m/z}$, while exceeding the nominal eFT resolution by more than twofold. To further illustrate this point, we selected multiple peak pairs in a small *m/z* window of *m/z* 984 to 992 in the ΦSDM MS/MS spectrum #35,938 at a retention time of 25.5 min (supplemental Fig. S2). With eFT processing at a 32 ms transient, the resolving power in this *m/z* range is ~7000, which means that two signals of equal abundance need to be at least 0.15 *m/z* apart to be resolved by eFT. Strikingly, all but one peak pair in this part of the ΦSDM spectrum were closer than 0.07 *m/z*, which equates a resolving power >13,000 in this *m/z* range or >30,000 at *m/z* 200.

Next, we investigated whether ΦSDM signal processing introduces extra scan overhead times. Comparing the empirical average cycle times with either eFT or ΦSDM processing to the sum of all Orbitrap transient times revealed overhead times of 0.39 and 0.54 s per scan cycle (Fig. 2*D*). This means that, even at an MS/MS scan rate of about 30 Hz, the additional data transfer to and back from the auxiliary computer as well as the iterative signal deconvolution caused only a minimal increase in cycle time of 0.15 s per 83 spectra. The comparison to eFT processing suggests that most of the overhead time can be attributed to AGC prescan events and ion routing. The ΦSDM processing time is mainly determined by the number of iterations to minimize the difference between modeled and observed signal. In our default setting, we limited the number of iterations to 150. To refine this, we varied the number of iterations from 100 to 200 in steps of 25, using the same 2 h LC gradient (supplemental Fig. S3). We observed a nearly linear increase in cycle time of ~0.03 s for every additional 25 iterations, from 0.08 s for 100 iterations, to 0.20 s for 200 iterations. As the difference in cycle time between the 100 iterations and 150 iterations is negligible on the chromatographic time scale, all remaining datasets used 150 iterations.

SNR and Mass Accuracy of Full Mass Range ΦSDM

Having confirmed that ΦSDM achieves an at least twofold higher resolving power across the full mass range with minimal impact to the acquisition rate, we asked whether this benefits mass accuracy and SNR in a practical proteomics setting. We

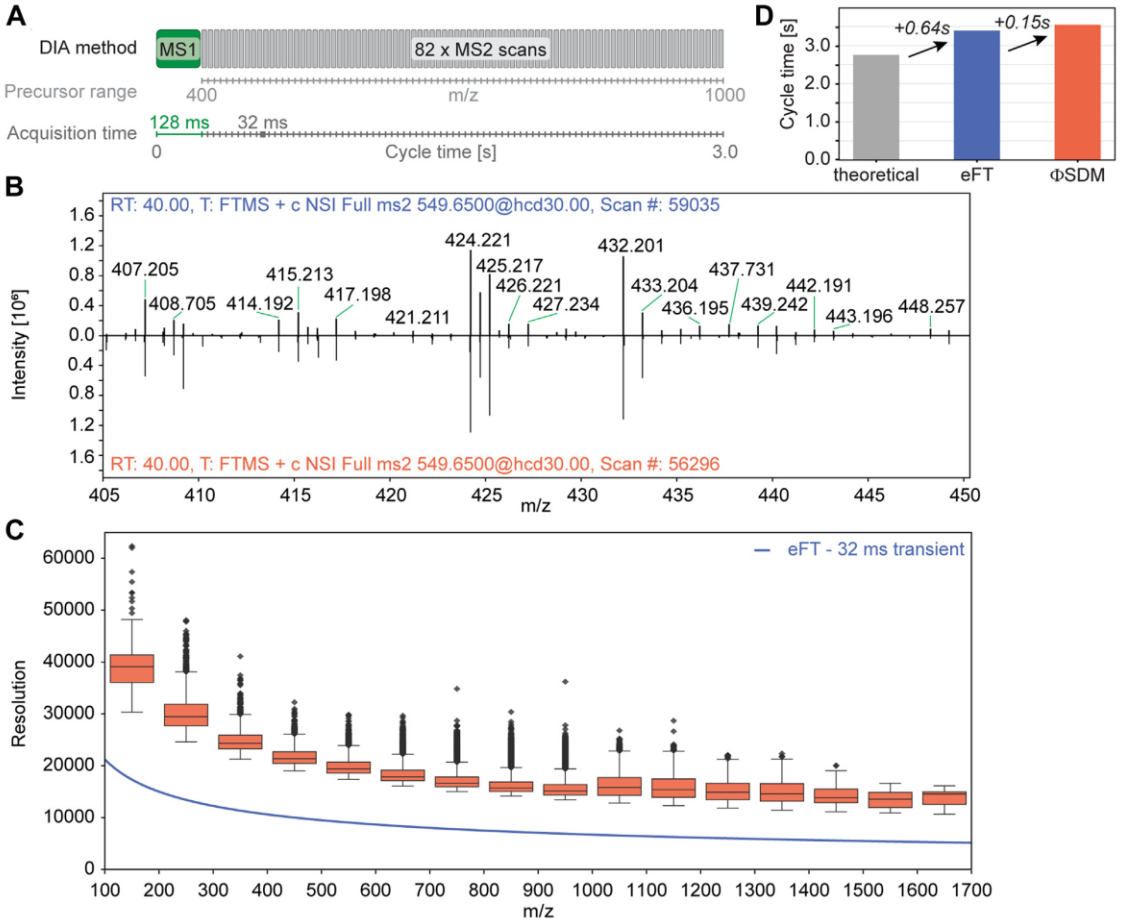


FIG. 2. **Spectrum quality with ΦSDM in complex proteomics samples.** Analytical evaluation of ΦSDM and eFT signal processing using quadruplicate injections of a HeLa full proteome digest with a 2 h EASY-nLC gradient. *A*, data-independent acquisition (DIA) schema used to acquire both standard eFT and ΦSDM data. *B*, spectrum comparison for a representative eFT (*top*) and ΦSDM (*bottom*) DIA MS2 scan at a matching retention time and DIA isolation window. For inspection of areas of lower abundance ions, the *m/z* region 405 to 450 is shown. Full range spectra are provided in supplemental Fig. S1. *C*, Box–Whisker plot showing pairwise resolution of neighboring peaks with ΦSDM as compared with the nominal eFT resolution for an Orbitrap transient of 32 ms (*solid line*). See text for more details. *D*, comparison of summed transient time (*gray*) to experiment DIA cycle times for eFT (*blue*) and ΦSDM (*orange*). ΦSDM, phase-constrained spectrum deconvolution method; DIA, data-independent acquisition; eFT, enhanced Fourier transformation; MS, mass sprectrometry.

first analyzed the data with a 'directDIA' spectrum library and extracted SNRs. The Spectronaut software computes SNRs for identified peptides based on XICs, where signal is the maximum intensity of the summed fragment XICs within the chromatographic peak boundaries and noise is the average summed fragment XICs outside the peak boundaries. Figure 3*A* shows the logarithmized SNR for peptides shared between quadruplicate eFT and ΦSDM injections (see also the Experimental Procedures section). Our analysis revealed a substantial shift toward higher SNRs with ΦSDM (median ΦSDM to eFT ratio of 1.5, supplemental Fig. S4), suggesting that ΦSDM successfully resolves interfering signals from fragment ion traces (chemical noise). Figure 3*B* visualizes this effect for one example chosen from Figure 3*A* (*red dot*, ΦSDM:eFT ratio 2.0). The fragment XICs for the triply charged precursor ion of VDINTPDVDVHGPDWHLK showed low CVs in-between replicates and similar intensities in eFT (Fig. 3*B*, *upper panel*) and ΦSDM (Fig. 3*B*, *lower panel*), whereas the interfering signals were markedly reduced with ΦSDM in all four replicates (supplemental Fig. S5).

Next, we investigated the mass accuracy (after nonlinear recalibration) for ΦSDM in comparison to eFT both on the fragment (Fig. 3*C*) and precursor (Fig. 3*D*) ion level (supplemental Table S2). The mass error distribution was

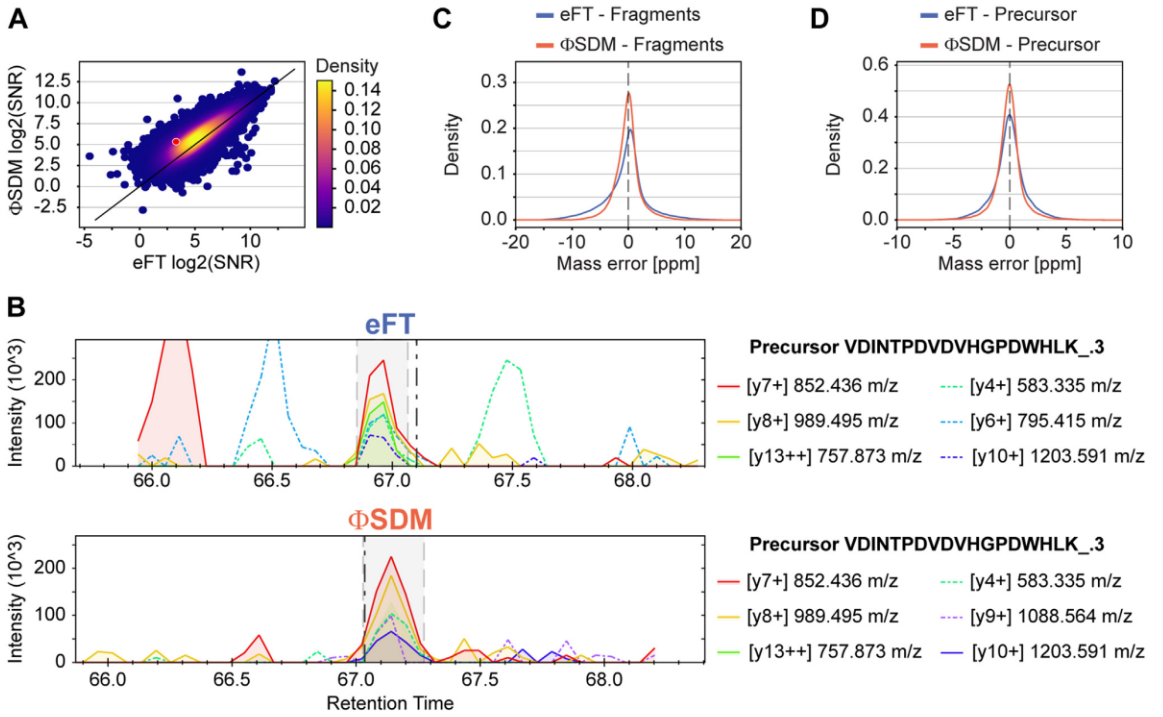


FIG. 3. **Signal-to-noise (SNR) and mass accuracy in complex samples.** *A*, scatter plot representing the log2 SNR comparison between eFT and ΦSDM. *Diagonal* indicated in *black* represents line of origin, and S/N distribution is colored based on density. Position of VDINTPDVDVHGPDWHLK_.3 peptide highlighted in *red*. *B*, comparison between extracted ion chromatograms (XICs) for precursor VDINTPDVDVHGPDWHLK_.3 from an eFT (*upper panel*) or ΦSDM (*lower panel*) run. *C*, comparison of calibrated mass error for all fragments identified in eFT (*blue*) and ΦSDM (*orange*). *D*, comparison of calibrated mass error for all precursors identified in eFT (*blue*) and ΦSDM (*orange*). ΦSDM, phase-constrained spectrum deconvolution method; eFT, enhanced Fourier transformation.

centered on 0 for both, and we observed only minor differences in shape and standard deviation between ΦSDM and eFT processing (supplemental Fig. S6). This confirms that ΦSDM signal processing does not affect mass accuracy, whereas the precision of mass spectral peak centroiding in proteomics practice appears primarily limited by the transient length rather than resolving power (38).

Effect of ΦSDM on Identification Rates in Complex DIA Spectra

Having established the analytical figures of merit, we investigated the influence of ΦSDM on peptide identification rates and label-free quantification accuracy in a typical DIA experiment (Fig. 4). In the quadruplicate 2 h HeLa experiments, on average, 47,883 and 55,607 peptides for eFT and ΦSDM were identified with 'directDIA' (Fig. 4*A*, *left panel*). This translated into an 8% improvement on the protein group level and over 6000 identified protein groups per replicate with ΦSDM (Fig. 4*A*, *right panel*). Irrespective of the signal processing method, we achieved an excellent quantitative reproducibility with median CV <8% on the peptide and <4% on the protein group level (Fig. 4*B*). Comparing only the subset of shared peptide identifications, we found similar median CVs of 7.1% and 6.6% for ΦSDM and eFT, respectively.

To delineate the higher identification rates with ΦSDM, we plotted the distribution of peptide ions in *m/z* and retention time. Figure 4*C* shows a consistent increase in the number of identified peptides throughout the binned precursor *m/z* range (bin size of 50 *m/z*). Interestingly, the largest relative increase of up to 12% was in the range of *m/z* 400 to 600, where most peptides were identified in absolute numbers. In contrast, in the higher *m/z* range with fewer peptides, the increase by ΦSDM was moderate. This result indicates that ΦSDM outperforms eFT particularly in areas of high peptide density. This is further supported by a comparison of identification rates along the retention time dimension (Fig. 4*D*). Again, the highest gains were in the center of the chromatographic gradient in RT bins with the overall highest number of identifications.

Peptide abundances with both eFT and ΦSDM spanned more than five orders of magnitude (Fig. 4*E*). Peptides

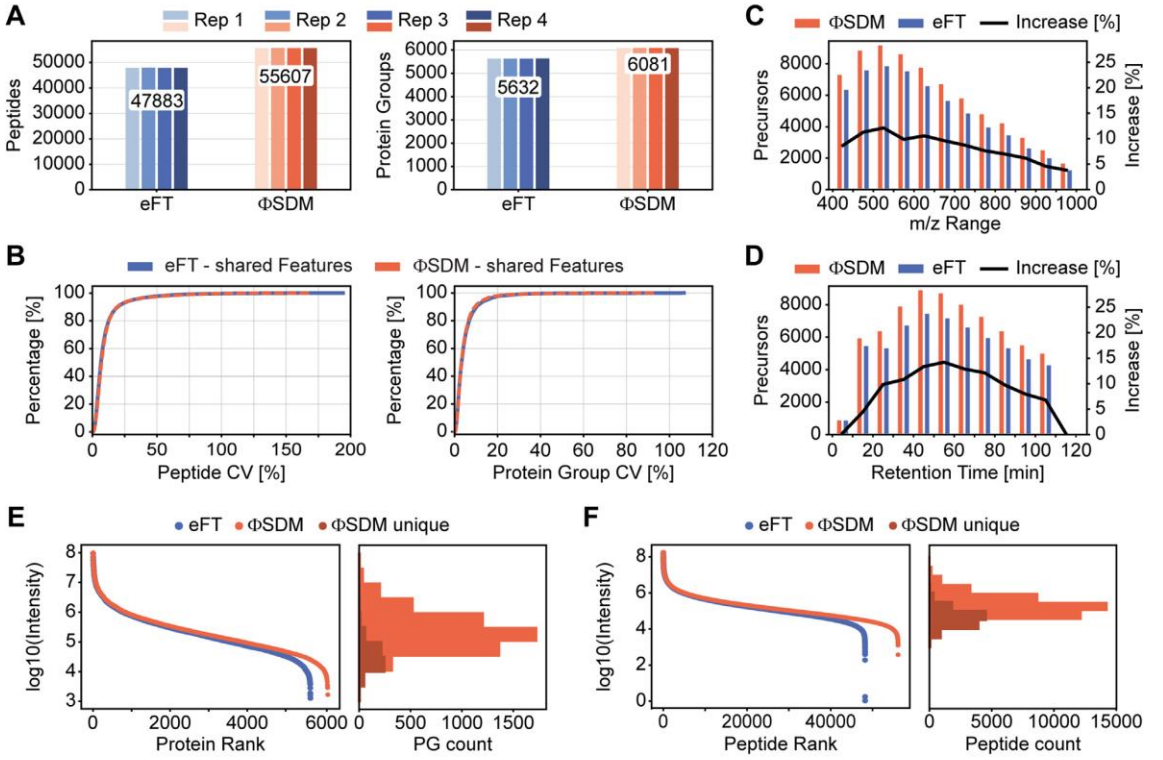


FIG. 4. **Influence of ΦSDM on identification and quantification**. Analytical analysis of the influence of the ΦSDM and eFT processing method on spectral quality was performed using quadruplicate HeLa measurements on a 2 h nLC gradient. *A*, bar plots comparing the number of peptides (*left*) and protein groups (PGs; *right*) identified in quadruplicate measurements of 200 ng HeLa digest in the eFT (*blue*) and ΦSDM (*orange*) dataset. Mean number of identifications indicated. *B*, comparison of cumulative CV values for shared peptides (*left*) and PGs (right). CVs for peptides or PGs identified in eFT and ΦSDM are represented in *blue* and *orange*, respectively. *C*, bar chart of precursor identification for eFT (*blue*) and ΦSDM (*orange*) along the retention time dimension with a bin size of 10 min. Increase in identification (in percent [%]) for ΦSDM in comparison to standard eFT is indicated in *black*. *D*, bar chart of precursor identification for eFT (*blue*) and ΦSDM (*orange*) along the retention mass-to-charge (*m/z*) range with a bin size of 50 (*m/z*). Increase in identification (in percent [%]) for ΦSDM in comparison to standard eFT is indicated in *black*. *E*, abundance distribution (*left side*) of proteins identified in the eFT (*blue*) and ΦSDM (*orange*) datasets. Abundance is represented as the log10 scale median protein intensities. The slight shift toward lower abundance for proteins uniquely identified in the ΦSDM dataset (*red*) in comparison to those that are common between the ΦSDM and eFT datasets (*orange*) is highlighted in the histogram. *F*, abundance distribution (*left side*) of peptides identified in the eFT (*blue*) and ΦSDM (*orange*) datasets. The slight shift toward lower abundance for peptides uniquely identified in the ΦSDM dataset (*red*) in comparison to those that are common between the ΦSDM and eFT datasets (*orange*) is highlighted in the histogram. Abundance is represented as the log10 scale median peptide intensities. ΦSDM, phase-constrained spectrum deconvolution method; eFT, enhanced Fourier transformation.

uniquely identified in the ΦSDM experiments were distributed across the entire abundance range, even though a comparison with peptides that were in common between ΦSDM and eFT revealed a bias toward the mid-to-lower abundance range (histogram in Fig. 4*E*). Consequently, the protein groups uniquely identified in ΦSDM runs were distributed over the entire abundance range of about five orders of magnitude, but with a higher density in the lower abundance range (Fig. 4*F*). From this, we concluded that ΦSDM—while keeping all other experimental parameters constant—facilitates the detection of lower-abundance signals in complex samples such as full proteome digests.

Rapid DIA Experiments With ΦSDM

The field of MS-based proteomics is currently pushing for increasing throughput to facilitate large experimental designs and clinical studies (39–43). However, shortening LC gradients entails increased spectrum complexity as more peptides coelute and, in addition, accurate quantification of narrower chromatographic peaks requires fast detection systems. The most common strategies to accommodate this in (Orbitrap) DIA methods are to either decrease the number of DIA windows and thus increase the number of cofragmented peptides for a fixed total precursor mass range or to lower the mass resolution to achieve faster cycle

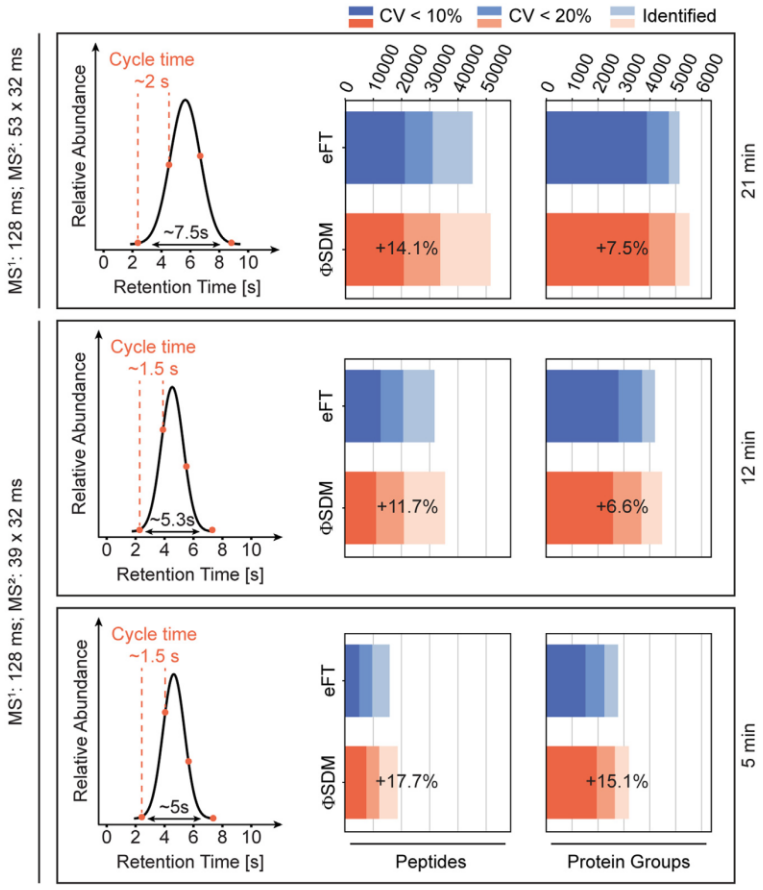


FIG. 5. **ΦSDM for rapid DIA proteomics.** DIA acquisition schemas were optimized for each gradient to guarantee at least three datapoints per peak. Average peak width and chosen cycle time are shown in the *left panel*. Full MS, MS/MS transients, and number of DIA windows used to achieve the different cycle times (1.5 s for 200 and 100 SPD methods, 2 s for the 60 SPD method) are indicated on the *left*. Number of protein and peptide identification (*right panel*) in triplicates. HeLa measurements using the Evosep 5 min (200 SPD, *top*), 12 min (100 SPD, *middle*), and 21 min (60 SPD, *bottom*) gradient for eFT (*blue*) and ΦSDM (*orange*). Total identifications across the triplicates shown in *light blue* and *light orange* for eFT and ΦSDM, respectively. Proteins and peptides quantified with CV values <20% are shown in medium *blue/orange*, whereas those with CV values <10% are shown in *dark blue/orange* for eFT and ΦSDM, respectively. For the triplicate measurements, 100 ng HeLa were injected for the 5 min and 12 min gradients each, whereas 200 ng were used per injection for the 21 min gradient. ΦSDM, phase-constrained spectrum deconvolution method; DIA, data-independent acquisition; eFT, enhanced Fourier transformation; SPD, sample per day.

times (22, 24). Our aforementioned results indicate that ΦSDM is most beneficial in dense regions of LC gradients. To test this hypothesis further, we turned to gradually shorter LC gradients compressing the peptide elution window. We used the Evosep One LC system to run preconfigured gradients for a throughput of 60, 100, and 200 SPD and designed DIA methods aiming for at least three data points per peak on average. The three gradients resulted on average in chromatographic peak widths of 5, 5.3, and 7.5 s (Fig. 5, *left column*). Accordingly, we adapted the number of DIA isolation windows in the *m/z* range 400 to 1000 to achieve cycle times around 1.5 s for the 200 and 100 SPD methods, and 2 s for the 60 SPD method, recording 128 and 32 ms transients for MS and MS/MS scans (supplemental Fig. S7*A*).

As our objective was to maximize the proteome coverage, we generated gradient-specific libraries with DDA from 48 high-pH reverse-phase fractionated HeLa samples per gradient. A database search using the Pulsar search engine integrated in the Spectronaut software resulted in 4196, 6824, and 8173 protein groups for the 200, 100, and 60 SDP gradients, respectively. Matching triplicate single-run measurements of 200 ng HeLa digest with both eFT and ΦSDM to the respective library, we observed an overall increase in peptide and protein group identifications by ΦSDM (Fig. 5). In line with our results for the 2 h gradient, we observed increasing SNRs

even though this effect was attenuated for shorter gradients (supplemental Fig. S7*B*).

Consistently for all short gradients, ΦSDM increased the number of identified peptides over conventional eFT signal processing particularly in retention time and isolation bins with high peptide density (supplemental Fig. S8). From the 60 SPD gradient, we identified 45,201 peptides with eFT and 52,558 peptides with ΦSDM, from which 5151 and 5536 protein groups were inferred. Likely because of the still relatively long cycle time, the fraction of peptides and proteins quantified with a CV <10% remained constant, whereas we quantified slightly more proteins with a CV <20% in the ΦSDM experiment. Using the 100 SPD gradient and a DIA method with wider isolation windows resulted in 11.7% and 6.6% more peptide and protein group identifications with ΦSDM. In line with our starting hypothesis, we observed the highest benefits of ΦSDM for the 5 min gradient (200 SPD) with a 17.7% increase in peptide and 15.1% increase in protein group identifications. Here, we identified over 3000 protein groups (out of 4200 in the library) from triplicate injections of 100 ng, while maintaining a very good quantitative reproducibility with median CVs of 10% and 7% for peptides and protein groups.

DISCUSSION

The ΦSDM signal processing method for Orbitrap MS can achieve a more than twofold higher mass resolution than conventional eFT for the same transient length but was previously limited to narrow *m/z* ranges because of its high computational cost (19). Here, we have implemented ΦSDM on an auxiliary computer to parallelize data acquisition and signal processing in real time. This setup allowed us to extend ΦSDM to the full mass range with only minimal impact on the acquisition rate in DIA proteomics experiments and maintaining the high mass accuracy of the Orbitrap mass analyzer. Analyzing fragment ion peak pairs in complex spectra, we confirmed that ΦSDM increases the mass resolving power by more than twofold over conventional eFT in the full mass range. In DIA experiments of a human cancer cell lysate, this resulted in 50% increased SNRs, facilitating peptide identification and label-free quantification. Furthermore, we found increased identification rates in dense areas of chromatographic gradients, making the combination of DIA with ΦSDM particularly attractive for short LC gradients. While we here focused on increasing resolving power (keeping transient length constant), in such applications, it can be desirable to shorten the transient length (keeping resolving power approximately constant). The faster scan rate would then allow for more data points per peak (shorter cycle time) or lower spectral complexity by increasing the number of DIA windows per cycle.

Similarly, while we focused on label-free quantification in this study, we note that workflows using nonisobaric labeling or isobaric labeling with high-mass reporter ions should directly benefit from higher mass resolution (44–48). Moreover, faster scan rates open up opportunities for advanced DIA acquisition schemes that, for example, include BoxCar (49) scans for high dynamic range MS1 scans or cycle through multiple compensation voltages with field asymmetric ion mobility spectrometry (17, 50–53). We also envision that ΦSDM could be even more beneficial for top–down proteomics as ion decay in the Orbitrap analyzer limits the practical maximum transient length. We thus conclude that full mass range and real-time ΦSDM signal processing is attractive for a wide range of MS-based proteomics applications.

DATA AVAILABILITY

The MS proteomics data have been deposited at the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) *via* the PRIDE partner repository (54) and are available with the dataset identifier PXD044292.

Supplemental data—This article contains supplemental data (Supplemental Figs. S1–S8; Supplemental Tables S1 and S2).

Acknowledgments—We thank the current and former colleagues in the department of Proteomics and Signal Transduction for help and fruitful discussions, especially Igor Paron, Johannes Müller-Reif, Philipp Geyer, Sebastian Virreira-Winter, Patricia Skowronek, and Maximilian Zwiebel. We thank Oliver Lange (Thermo Fisher Scientific) for the valuable input. In addition, we thank Lukas Reiter, Oliver Bernhardt, and Tejas Gandhi (Biognosys) for the input on Spectronaut data processing.

Funding and additional information—This study was supported by the Max-Planck Society for the Advancement of Science, the European Union's Horizon 2020 research and innovation program under grant agreement number 686547 (MSmed project), and by the Bavarian State Ministry of Health and Care through the research project DigiMed Bayern (www.digimed-bayern.de).

Author contributions—S. S., K. L. F., F. M., and M. M. conceptualization; S. S., A. K., D. M., K. A., D. G., A. M., and F. M. methodology; K. L. F., A. K., D. M., K. A., D. G., and A.M. software; S. S., A. S., K. L. F., W.-F. Z., L. N., C. W., A. M., and F. M. investigation; K. L. F., A. K., D. M., K. A., D. G., and A. M. resources; S. S., A. S., W.-F. Z., K. L. F., A. M., and F. M. formal analysis; S. S. and F. M. writing–original draft.

Conflict of interest—K. L. F., A. K., D. M., K. A., D. G., and A. M. are employees of Thermo Fisher Scientific, the manufacturer of Orbitrap instrumentation used in this research. M. M. is an indirect investor in Evosep Biosystems. All other authors declare no competing interests.

Abbreviations—The abbreviations used are: ΦSDM, phase-constrained spectrum deconvolution method; ACN,

acetonitrile; AGC, automated gain control; DDA, data-dependent acquisition; DIA, data-independent acquisition; eFT, enhanced Fourier transformation; FA, formic acid; GPU, graphics processing unit; MS, mass spectrometry; SNR, signal-to-noise ratio; SPD, sample per day; XIC, extracted ion chromatogram.

Received August 30, 2023, and in revised form, December 21, 2023
Published, MCPRO Papers in Press, January 4, 2024, https://doi.org/10.1016/j.mcpro.2024.100713

REFERENCES

1. Aebersold, R., and Mann, M. (2016) Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355
2. Makarov, A., and Scigelova, M. (2010) Coupling liquid chromatography to Orbitrap mass spectrometry. *J. Chromatogr. A* **1217**, 3938–3945
3. Makarov, A. (2000) Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal. Chem.* **72**, 1156–1162
4. Hu, Q., Noll, R. J., Li, H., Makarov, A., Hardman, M., and Graham Cooks, R. (2005) The Orbitrap: a new mass spectrometer. *J. Mass Spectrom.* **40**, 430–443
5. Zubarev, R. A., and Makarov, A. (2013) Orbitrap mass spectrometry. *Anal. Chem.* **85**, 5288–5296
6. Lange, O., Damoc, E., Wieghaus, A., and Makarov, A. (2014) Enhanced Fourier transform for Orbitrap mass spectrometry. *Int. J. Mass Spectrom.* **369**, 16–22
7. Michalski, A., Damoc, E., Hauschild, J. P., Lange, O., Wieghaus, A., Makarov, A., *et al.* (2011) Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole orbitrap mass spectrometer. *Mol. Cell Proteomics* **10**, M111.011015
8. Comisarow, M. B., and Marshall, A. G. (1974) Fourier transform ion cyclotron resonance spectroscopy. *Chem. Phys. Lett.* **25**, 282–283
9. Savitski, M. M., Ivonin, I. A., Nielsen, M. L., Zubarev, R. A., Tsybin, Y. O., and Håkansson, H. P. (2004) Shifted-basis technique improves accuracy of peak position determination in Fourier transform mass spectrometry. *J. Am. Soc. Mass Spectrom.* **15**, 457–461
10. Kozhinov, A. N., and Tsybin, Y. O. (2012) Filter diagonalization method-based mass spectrometry for molecular and macromolecular structure analysis. *Anal. Chem.* **84**, 2850–2856
11. Aushev, T., Kozhinov, A. N., and Tsybin, Y. O. (2014) Least-squares fitting of time-domain signals for fourier transform mass spectrometry. *J. Am. Soc. Mass Spectrom.* **25**, 1263–1273
12. Aizikov, K., and O'Connor, P. B. (2006) Use of the filter diagonalization method in the study of space charge related frequency modulation in Fourier transform ion cyclotron resonance mass spectrometry. *J. Am. Soc. Mass Spectrom.* **17**, 836–843
13. Leach, F. E., Kharchenko, A., Vladimirov, G., Aizikov, K., O'Connor, P. B., Nikolaev, E., *et al.* (2012) Analysis of phase dependent frequency shifts in simulated FTMS transients using the filter diagonalization method. *Int. J. Mass Spectrom.* **325-327**, 19–24
14. Martini, B. R., Aizikov, K., and Mandelshtam, V. A. (2014) The filter diagonalization method and its assessment for Fourier transform mass spectrometry. *Int. J. Mass Spectrom.* **373**, 1–14
15. Grinfeld, D., Aizikov, K., Kreutzmann, A., Damoc, E., and Makarov, A. (2017) Phase-constrained spectrum deconvolution for Fourier transform mass spectrometry. *Anal. Chem.* **89**, 1202–1211
16. Makarov, A., Grinfeld, D., and Ayzikov, K. (2019) Chapter 2 - fundamentals of Orbitrap analyzer. In: Kanawati, B., Schmitt-Kopplin, P., eds. Fundamentals and Applications of Fourier Transform Mass Spectrometry, Elsevier, Amsterdam, Netherlands: 37–61
17. Bekker-Jensen, D. B., Martínez-Val, A., Steigerwald, S., Rüther, P., Fort, K. L., Arrey, T. N., *et al.* (2020) A compact quadrupole-orbitrap mass spectrometer with FAIMS interface improves proteome coverage in short LC gradients. *Mol. Cell Proteomics* **19**, 716–729
18. Kelstrup, C. D., Bekker-Jensen, D. B., Arrey, T. N., Hogrebe, A., Harder, A., and Olsen, J. V. (2018) Performance evaluation of the Q exactive HF-X for shotgun proteomics. *J. Proteome Res.* **17**, 727–738
19. Kelstrup, C. D., Aizikov, K., Batth, T. S., Kreutzman, A., Grinfeld, D., Lange, O., *et al.* (2018) Limits for resolving isobaric tandem mass tag reporter ions using phase-constrained spectrum deconvolution. *J. Proteome Res.* **17**, 4008–4016
20. Phaneuf, C. G., Aizikov, K., Grinfeld, D., Kreutzmann, A., Mourad, D., Lange, O., *et al.* (2023) Experimental strategies to improve drug-target identification in mass spectrometry-based thermal stability assays. *Commun. Chem.* **6**, 64
21. Gillet, L. C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., *et al.* (2012) Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell Proteomics* **11**. https://doi.org/10.1074/mcp.O111.016717
22. Ludwig, C., Gillet, L., Rosenberger, G., Amon, S., Collins, B. C., and Aebersold, R. (2018) Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* **14**, e8126
23. Skowronek, P., and Meier, F. (2022) High-throughput mass spectrometry-based proteomics with Dia-PASEF *Methods in Molecular Biology*. Humana Press Inc, New York, NY: 15–27
24. Pino, L. K., Just, S. C., MacCoss, M. J., and Searle, B. C. (2020) Acquiring and analyzing data independent acquisition proteomics experiments without spectrum libraries. *Mol. Cell Proteomics* **19**, 1088–1103
25. Sidoli, S., Simithy, J., Karch, K. R., Kulej, K., and Garcia, B. A. (2015) Low resolution data-independent acquisition in an LTQ-orbitrap allows for simplified and fully untargeted analysis of histone modifications. *Anal. Chem.* **87**, 11448–11454
26. Moseley, M. A., Hughes, C. J., Juvvadi, P. R., Soderblom, E. J., Lennon, S., Perkins, S. R., *et al.* (2018) Scanning quadrupole data-independent acquisition, Part A: qualitative and quantitative characterization. *J. Proteome Res.* **17**, 770–779
27. Borràs, E., Pastor, O., and Sabidó, E. (2021) Use of linear ion traps in data-independent acquisition methods benefits low-input proteomics. *Anal. Chem.* **93**, 11649–11653
28. Messner, C. B., Demichev, V., Bloomfield, N., Yu, J. S. L., White, M., Kreidl, M., *et al.* (2021) Ultra-fast proteomics with scanning SWATH. *Nat. Biotechnol.* **39**, 846–854
29. Meier, F., Brunner, A.-D., Frank, M., Ha, A., Bludau, I., Voytik, E., *et al.* (2020) diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nat. Methods* **17**, 1229–1236
30. Kulak, N. A., Geyer, P. E., and Mann, M. (2017) Loss-less nano-fractionator for high sensitivity, high coverage proteomics. *Mol. Cell Proteomics* **16**, 694–705
31. Bache, N., Geyer, P. E., Bekker-Jensen, D. B., Hoerning, O., Falkenby, L., Treit, P. V., *et al.* (2018) A novel LC system embeds analytes in pre-formed gradients for Rapid, ultra-robust proteomics. *Mol. Cell Proteomics* **17**, 2284–2296
32. Bruderer, R., Bernhardt, O. M., Gandhi, T., Miladinović, S. M., Cheng, L. Y., Messner, S., *et al.* (2015) Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell Proteomics* **14**, 1400–1410
33. [preprint] Eugenia, V., Patricia, S., Wen-Feng, Z., Maria, C. T., Andreas-David, B., Marvin, T., *et al.* (2022) AlphaViz: visualization and validation of critical proteomics data directly at the raw data level. *bioRxiv*. https://doi.org/10.1101/2022.07.12.499676
34. [preprint] Maximilian, T. S., Isabell, B., Wen-Feng, Z., Eugenia, V., Constantin, A., Julia, S., *et al.* (2021) AlphaPept, a modern and open framework for MS-based proteomics. *bioRxiv*. https://doi.org/10.1101/2021.07.23.453379
35. Makarov, A., Denisov, E., Kholomeev, A., Balschun, W., Lange, O., Strupat, K., *et al.* (2006) Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal. Chem.* **78**, 2113–2120
36. Murray, K. K. (2022) Resolution and resolving power in mass spectrometry. *J. Am. Soc. Mass Spectrom.* **33**, 2342–2347
37. Murray, K. K., Boyd, R. K., Eberlin, M. N., Langley, G. J., Li, L., and Naito, Y. (2013) Definitions of terms relating to mass spectrometry (IUPAC Recommendations 2013). *Chem. Internat.* **85**, 1515–1609
38. Makarov, A., Denisov, E., Lange, O., and Horning, S. (2006) Dynamic range of mass accuracy in LTQ orbitrap hybrid mass spectrometer. *J. Am. Soc. Mass Spectrom.* **17**, 977–982
39. Niu, L., Thiele, M., Geyer, P. E., Rasmussen, D. N., Webel, H. E., Santos, A., *et al.* (2022) Noninvasive proteomic biomarkers for alcohol-related liver disease. *Nat. Med.* **28**, 1277–1287

40. Wewer Albrechtsen, N. J., Geyer, P. E., Doll, S., Treit, P. V., Bojsen-Møller, K. N., Martinussen, C., *et al.* (2018) Plasma proteome profiling reveals dynamics of inflammatory and lipid homeostasis markers after Roux-en-Y Gastric bypass surgery. *Cell Syst.* **7**, 601–612.e603
41. Bruderer, R., Muntel, J., Müller, S., Bernhardt, O. M., Gandhi, T., Cominetti, O., *et al.* (2019) Analysis of 1508 plasma samples by capillary-flow data-independent acquisition profiles proteomics of weight loss and maintenance. *Mol. Cell Proteomics* **18**, 1242–1254
42. Geyer, P. E., Wewer Albrechtsen, N. J., Tyanova, S., Grassl, N., Iepsen, E. W., Lundgren, J., *et al.* (2016) Proteomics reveals the effects of sustained weight loss on the human plasma proteome. *Mol. Syst. Biol.* **12**, 901
43. Johnson, E. C. B., Dammer, E. B., Duong, D. M., Ping, L., Zhou, M., Yin, L., *et al.* (2020) Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. *Nat. Med.* **26**, 769–780
44. Minogue, C. E., Hebert, A. S., Rensvold, J. W., Westphall, M. S., Pagliarini, D. J., and Coon, J. J. (2015) Multiplexed quantification for data-independent acquisition. *Anal. Chem.* **87**, 2570–2575
45. Zhong, X., Frost, D. C., Yu, Q., Li, M., Gu, T.-J., and Li, L. (2020) Mass defect-based DiLeu tagging for multiplexed data-independent acquisition. *Anal. Chem.* **92**, 11119–11126
46. Thielert, M., Itang, E. C. M., Ammar, C., Rosenberger, F. A., Bludau, I., Schweizer, L., *et al.* (2023) Robust dimethyl-based multiplex-DIA doubles single-cell proteome depth via a reference channel. *Mol. Syst. Biol.* **19**, e11503
47. Derks, J., Leduc, A., Wallmann, G., Huffman, R. G., Willetts, M., Khan, S., *et al.* (2023) Increasing the throughput of sensitive proteomics by plexDIA. *Nat. Biotechnol.* **41**, 50–59
48. Pino, L. K., Baeza, J., Lauman, R., Schilling, B., and Garcia, B. A. (2021) Improved SILAC quantification with data-independent acquisition to investigate bortezomib-induced protein degradation. *J. Proteome Res.* **20**, 1918–1927
49. Mehta, D., Scandola, S., and Uhrig, R. G. (2022) BoxCar and library-free data-independent acquisition substantially improve the depth, range, and completeness of label-free quantitative proteomics. *Anal. Chem.* **94**, 793–802
50. Meier, F., Geyer, P. E., Virreira Winter, S., Cox, J., and Mann, M. (2018) BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat. Methods* **15**, 440–448
51. Barnett, D. A., Ells, B., Guevremont, R., and Purves, R. W. (2002) Application of ESI-FAIMS-MS to the analysis of tryptic peptides. *J. Am. Soc. Mass Spectrom.* **13**, 1282–1291
52. Hebert, A. S., Prasad, S., Belford, M. W., Bailey, D. J., McAlister, G. C., Abbatiello, S. E., *et al.* (2018) Comprehensive single-shot proteomics with FAIMS on a hybrid orbitrap mass spectrometer. *Anal. Chem.* **90**, 9529–9537
53. Saba, J., Bonneil, E., Pomiès, C., Eng, K., and Thibault, P. (2009) Enhanced sensitivity in proteomics experiments using FAIMS coupled with a hybrid linear ion trap/orbitrap mass spectrometer. *J. Proteome Res.* **8**, 3355–3366
54. Perez-Riverol, Y., Bai, J., Bandla, C., García-Seisdedos, D., Hewapathirana, S., Kamatchinathan, S., *et al.* (2022) The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* **50**, D543–D552

Article 2: Plasma proteome profiling of healthy subjects undergoing bed rest reveals unloading-dependent changes linked to muscle atrophy

Journal of Cachexia, Sarcopenia and Muscle 14,439–451 (2023)

Marta Murgia[1,2]*, Lorenza Brocca[3], Elena Monti[1], Martino V. Franchi[1,4], Maximilian Zwiebel[2], **Sophia Steigerwald[2]**, Emiliana Giacomello[5], Roberta Sartori[1,6], Sandra Zampieri[1,4,7], Giovanni Capovilla[7], Mladen Gasparini[8], Gianni Biolo[5], Marco Sandri[1,6], Matthias Mann[2,9] & Marco V. Narici[1,4]

[1]Department of Biomedical Sciences, University of Padova, Padua, Italy
[2]Max-Planck-Institute of Biochemistry, Martinsried, Germany
[3]Department of Molecular Medicine, University of Pavia, Pavia, Italy
[4]CIR-MYO Myology Center, Padua, Italy
[5]Department of Medicine, Surgery and Health Sciences, University of Trieste, Trieste, Italy
[6]Veneto Institute of Molecular Medicine, Padova, Italy
[7]Department of Surgical, Oncological and Gastroenterological Sciences, Padova University Hospital, Padua, Italy
[8]Izola General Hospital, Izola, Slovenia
[9]NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark
**Corresponding author*

Muscle atrophy, the weakening and decreasing of muscle mass, can be caused, among others, by inactivity, by old age (sarcopenia) or cancer (cancer cachexia).[418–420] Inactivity-induced muscle atrophy is particularly relevant for patients undergoing long hospitalization, chronic disease or also for astronauts in microgravity conditions, as the onset of muscle atrophy has been observed after two days of unloading.[421] The extend of muscle loss during inactivity, however, can be subject to patient-based heterogeneity.[422] Establishing a minimally invasive measure to predict or monitor the patient-based muscle loss during inactivity or other muscle atrophy inducing conditions could therefore be of great use.

Here we leverage MS-based proteomics to evaluate the effects of two conditions causing skeletal muscle atrophy, namely bedrest and cachexia, on the serum/plasma proteome and identify potential biomarkers correlated with muscle loss. Due to the high dynamic range, plasma or serum poses a unique challenge in MS-based proteomics and requires optimized MS acquisition. With this we quantified 500 and 400 plasma proteins in the bedrest and cachexia cohort respectively. In the initial cohort of healthy individuals undergoing voluntary bedrest for ten days, we identified 30 proteins that show significant abundance changes during bedrest (timepoint 0 vs 10 days). Notably, the tissue-leakage protein teneurin-4 showed a 1.6-fold increase at the bedrest endpoint on day 10, while the abundance of extracellular matrix protein lumican decreased during unloading and remained low in the recovery. Evaluating differences in individual

offloading response, we additionally identified six proteins differentiating between individuals that maintain muscle mass and those developing unloading-mediated muscle loss. Four of which, haptoglobin-related protein (HPR), transthyretin and two apolipoproteins were more abundant in atrophy-resistant subjects. Looking at cancer cachexia, comparison of cancer patients with cachexia to the controls lead to the identification of two significant proteins. Importantly, haptoglobin-related protein, was twofold more abundant in non-cachexia controls. Together this indicates that levels of circulating HPR correlate with the maintenance of muscle mass in both bed rest and cancer cachexia and its potential use as a biomarker.

Contribution:

Co-authorship. This study was primarily conceptualized and conducted by the first author Marta Murgia. The plasma cohort shown in the study was part of the MARS-PRE project, funded by the Italian Space Agency in 2019. The aim of this consortium of nineteen groups with multidisciplinary background was the identification of biochemical functional biomarkers to characterize the adaptation of the human body to spaceflight and variations in gravitational conditions. Bed rest was used as a ground-based model for space missions. To verify whether the plasma proteins whose abundance changed during bed rest were also affected in other types of muscle atrophy, we added to the study a cohort of cancer patients with and without muscle wasting (cachexia). This second cohort was measured on the Thermo Scientific Exploris platform. I familiarized Marta Murgia with the Exploris 480 MS instrument and optimized the MS acquisition strategy. I took part in data acquisition and analysis. Alongside the other co-authors, I also contributed to revising and editing the manuscript.

ORIGINAL ARTICLE

Journal of Cachexia, Sarcopenia and Muscle 2023; **14**: 439–451
Published online 14 December 2022 in Wiley Online Library (wileyonlinelibrary.com) **DOI:** 10.1002/jcsm.13146

Plasma proteome profiling of healthy subjects undergoing bed rest reveals unloading-dependent changes linked to muscle atrophy

Marta Murgia[1,2*], Lorenza Brocca[3], Elena Monti[1], Martino V. Franchi[1,4], Maximilian Zwiebel[2], Sophia Steigerwald[2], Emiliana Giacomello[5], Roberta Sartori[1,6], Sandra Zampieri[1,4,7], Giovanni Capovilla[7], Mladen Gasparini[8], Gianni Biolo[5], Marco Sandri[1,6], Matthias Mann[2,9] & Marco V. Narici[1,4]

[1]*Department of Biomedical Sciences, University of Padova, Padua, Italy;* [2]*Max-Planck-Institute of Biochemistry, Martinsried, Germany;* [3]*Department of Molecular Medicine, University of Pavia, Pavia, Italy;* [4]*CIR-MYO Myology Center, Padua, Italy;* [5]*Department of Medicine, Surgery and Health Sciences, University of Trieste, Trieste, Italy;* [6]*Veneto Institute of Molecular Medicine, Padova, Italy;* [7]*Department of Surgical, Oncological and Gastroenterological Sciences, Padova University Hospital, Padua, Italy;* [8]*Izola General Hospital, Izola, Slovenia;* [9]*NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark*

Abstract

Background Inactivity and unloading induce skeletal muscle atrophy, loss of strength and detrimental metabolic effects. Bed rest is a model to study the impact of inactivity on the musculoskeletal system. It not only provides information for bed-ridden patients care, but it is also a ground-based spaceflight analogue used to mimic the challenges of long space missions for the human body. In both cases, it would be desirable to develop a panel of biomarkers to monitor muscle atrophy in a minimally invasive way at point of care to limit the onset of muscle loss in a personalized fashion.

Methods We applied mass spectrometry-based proteomics to measure plasma protein abundance changes in response to 10 days of bed rest in 10 young males. To validate the correlation between muscle atrophy and the significant hits emerging from our study, we analysed in parallel, with the same pipeline, a cohort of cancer patients with or without cachexia and age-matched controls. Our analysis resulted in the quantification of over 500 proteins.

Results Unloading affected plasma concentration of proteins of the complement cascade, lipid carriers and proteins derived from tissue leakage. Among the latter, teneurin-4 increased 1.6-fold in plasma at bed rest day 10 (BR10) compared with BR0 (6.E9 vs. 4.3E9, $P = 0.02$) and decreased to 0.6-fold the initial abundance after 2 days of recovery at normal daily activity (R + 2, 2.7E9, $P = 3.3E\text{-}4$); the extracellular matrix protein lumican was decreased to 0.7-fold (1.2E9 vs. 8.5E8, $P = 1.5E\text{-}4$) at BR10 and remained as low at R + 2. We identified six proteins distinguishing subjects developing unloading-mediated muscle atrophy (decrease of >4% of quadriceps cross-sectional area) from those largely maintaining their initial muscle mass. Among them, transthyretin, a thyroid hormone-binding protein, was significantly less abundant at BR10 in the plasma of subjects with muscle atrophy compared with those with no atrophy (1.6E10 vs. 2.6E10, $P = 0.001$). Haptoglobin-related protein was also significantly reduced in the serum of cancer patients with cachexia compared with that of controls.

Conclusions Our findings highlight a combination or proteomic changes that can be explored as potential biomarkers of muscle atrophy occurring under different conditions. The panel of significant proteomic differences distinguishing atrophy-prone and atrophy-resistant subjects after 10 days of bed rest need to be tested in a larger cohort to validate their potential to predict inactivity-triggered muscle loss in humans.

Keywords Proteomics; Plasma; Skeletal muscle; Bed rest; Atrophy; Cachexia

Received: 12 June 2022; Revised: 4 November 2022; Accepted: 10 November 2022
**Correspondence to: Marta Murgia, Department of Biomedical Sciences, University of Padova, Via Ugo Bassi, 58/B, 35131 Padua, Italy. Email: mmurgia@biochem.mpg.de*

Introduction

Muscle atrophy can be triggered by immobility and nutrients deprivation and is a severe co-morbidity for patients suffering from debilitating chronic diseases or undergoing long hospitalizations.[1] Sarcopenia is a crucial factor in the loss of autonomy of the elderly population and, together with weight loss, is part of the diagnostic criterion for cancer cachexia,[2] a multifactorial syndrome associated with poor outcomes in cancer patients.[3]

Atrophy causes detrimental changes to the morphology and function of skeletal muscles. The onset of muscle atrophy caused by unloading is observed in just 2 days[4] accompanied by alterations of contractile properties within the same timeframe.[5,6] This atrophic state develops when hyperactivation of proteolysis and organelle degradation exceed rates of protein synthesis and organelle biogenesis. Proteolysis occurs via calcium-dependent proteolytic pathways and ubiquitin-mediated proteasomal and autophagic lysosomal processes. These are potentiated when cellular signalling events promote transcription of genes controlling protein degradation, which are controlled by Forkhead box protein O (FoxO)-dependent pathways.[7,8] FoxO dephosphorylation induces the ubiquitin proteasome system through the activation of E3 ubiquitin ligases[9,10] and can directly enhance the autophagy system.[11] Mitochondrial alterations (for instance, mitochondrial fusion/fission machinery imbalance)[12] and reactive oxygen species (ROS) can activate FoxO pathways as well as systemic signals, such as the pro-inflammatory cytokines IL1, IL6 and TNFα and myostatin.[13]

As muscles are the major site of glucose uptake through glucose transporter type 4 (GLUT4), and the largest amino acid reservoir,[14] loss of muscle mass has systemic consequences on metabolism. Blood plasma, by circulating through all organs, is expected to relay this information dynamically, through changes in the abundance of its ions, small molecules and protein composition. With this in mind, we set out to use plasma proteomics as a tool to convey first-hand information on skeletal muscle trophism and monitor muscle atrophy. Our goals were to provide a system view of the changes caused by muscle atrophy to the plasma proteome and to highlight single proteins and protein signatures whose plasma abundance correlates with the loss of muscle mass. If a pool of plasma biomarkers of muscle atrophy existed, one could use a minimally invasive 'liquid biopsy' to monitor muscle mass at point of care, in combination with indirect proxies such as body weight and grip strength. This would be instrumental for frail sarcopenia patients as well as for astronauts during long space missions on the International Space Station, where they experience severe muscle atrophy and loss of force despite intensive physical training on board.[15]

To this aim, we used state-of-the-art mass spectrometry (MS)-based proteomics to analyse the blood plasma of a cohort of 10 young healthy subjects undergoing 10 days of continuous bed rest. In the same cohort, we had analysed in parallel muscle atrophy in great detail, showing a median 5.2% loss of the quadriceps volume and 13% of maximum isometric voluntary contraction of the knee extensors.[16] We here measure the plasma proteome of these subjects before bed rest (BR0) and at the endpoint of the unloading phase at day 10 (BR10). Our data reveal changes in the abundance of 34 proteins after 10 days of bed rest, comprising both canonical plasma components and proteins possibly originating from tissue leakage. Our parallel analyses had unexpectedly shown that three subjects in our cohort were largely resistant to bed rest-induced muscle atrophy, whereas the other seven had lost both mass and force at BR10. Exploring this serendipitous observation, we could find proteins distinguishing the plasma proteome of subjects undergoing no or minor muscle atrophy from that of subjects undergoing extensive atrophy after 10 days of bed rest.

To carry out an initial validation of our findings, we analysed by MS-based proteomics the serum of a second cohort comprising gastrointestinal cancer patients, with and without cachexia, and age-matched patients hospitalized for non-neoplastic diseases. Although muscle atrophy is a common feature, there are profound differences between the two cohorts. However, it is well established that various types of atrophy share a common set of transcriptional adaptations acting through the regulation of proteasome activity. FoxO-controlled atrophy-related genes were discovered as commonly regulated in conditions as diverse as cachexia, starvation, diabetes and kidney disease.[13]

Aiming at downstream common changes, in line with previous studies,[17] we thus explored by proteomics two conditions causing skeletal muscle atrophy. We highlight a group of potential biomarkers that can be explored for their correlation to muscle atrophy in different pathological states.

Methods

Patient cohorts

The bed rest study was approved by the National Ethical Committee of the Slovenian Ministry of Health on 17 July 2019, with reference number 0120-304/2019/9. The study involving cancer patients was approved by the Ethical Committee for Clinical Experimentation of Provincia di Padova (protocol number 3674/AO/15). The bed rest cohort has been previously described.[16] Ten young healthy volunteers (*Table* S1) were housed in a horizontal lying position for 10 full days in standard hospital rooms without interruption and were not allowed to carry out any form of exercise on their beds. They were given an individually controlled eucaloric diet during the whole hospital stay. Blood was sampled right before the begin of bed rest (BR0), at day 10

right before the subject was allowed to stand up (BR10). We also analysed plasma drawn after 2 days of monitored recovery in the hospital (R + 2). This was the endpoint of the study carried out under strictly controlled diet and activity conditions, after which the subjects were discharged from the hospital. We used part of a cachexia cohort, which has been previously described.[18] Patients were stratified into 'cachectic' and 'pre-cachectic' subgroups[2] and compared with patients undergoing surgery for non-neoplastic noninflammatory diseases (control). The subgroup of patients from that cohort used in this study is described in *Table* S2.

Plasma and serum sample processing

The 12 most abundant plasma protein components comprise about 95% of the total protein mass, making the quantification of proteins in the low abundance range extremely challenging.[19] We therefore used a highly sensitive analytical workflow, with one-buffer sample preparation combined with novel MS acquisition modes and computational methods. For peptide preparation, 5 µl of plasma or serum was diluted in 50 µl of LYSE buffer (PreOmics), heated at 95°C for 5 min and sonicated in a water-bath sonicator (Diagenode) for 5 min with a 50% duty cycle. Proteolytic digestion was carried out by addition of 2 µg of endoproteinase LysC and 2 µg of trypsin. After overnight digestion at 37°C under continuous shaking, samples were acidified to a final concentration of 0.1% trifluoroacetic acid (TFA) and loaded onto StageTip plugs of SDB-RPS. Purified peptides were eluted with 80% acetonitrile-1% ammonia and dried. For the library used for match between runs (see below) peptides from all samples were pooled and eluted into 24 fractions using a Spider Fractionator. Concentration of HDL, LDL cholesterol and triglycerides was measured in plasma of bed rest subjects using an automated hospital clinical chemistry pipeline. Contamination from erythrocytes, platelets and coagulation factors was calculated using a custom-made R script based on an online resource (www.plasmaproteomeprofiling.org) derived from recent findings.[20]

Liquid chromatography and tandem mass spectrometry

Peptides were separated on 50-cm columns (75 µm inner diameter) of ReproSil-Pur C18-AQ 1.9 µm resin (Dr Maisch GmbH) packed in-house. The columns were kept at 60°C using a column oven. Liquid chromatography–mass spectrometry (LC-MS) analysis was carried out on an EASY-nLC-1200 system (Thermo Fisher Scientific) coupled through a nanoelectrospray source to a mass spectrometer. Samples were analysed in technical triplicates. Samples S8–S10 at time R + 2 were analysed in technical duplicates. For the bed rest cohort, the analysis was carried out on a Q Exactive HF mass spectrometer (Thermo Fisher Scientific). Peptides were loaded in buffer A (0.1% (v/v) formic acid) applying a non-linear 45-min gradient of 3–75% buffer B (0.1% (v/v) formic acid, 80% (v/v) acetonitrile) at a flow rate of 450 nL/min. For the cancer patient cohort, samples were analysed on an Orbitrap Exploris 480 mass spectrometer (Thermo Fisher Scientific). Peptides were loaded in buffer A applying a non-linear 120-min gradient of 0–65% buffer B at a flow rate of 300 nL/min. Data acquisition switched between a full scan and 15 data-dependent tandem mass spectrometry (MS/MS) scans. Multiple sequencing of peptides was minimized by excluding the selected peptide candidates for 30 s.

Computational proteomics and data deposition

The MaxQuant software (versions 1.6.10.43 and 2.0.3.0) was used for the analysis of raw files searching against the human UniProt databases (UP000005640_9606, UP000005640_9606_additional) and a common contaminants database.[21] The false discovery rate (FDR) was set to 1% for peptides and proteins and was determined by searching a reverse database. Peptide identifications by MS/MS were matched between the samples and the library files with a 0.7-min retention-time match window. Peptides with a minimum length of seven amino acids were considered for the search including N-terminal acetylation and methionine oxidation as variable modifications and cysteine carbamidomethylation as fixed modification. Enzyme specificity was set to trypsin cleaving C-terminal to arginine and lysine. A maximum of two missed cleavages was allowed. In our dataset, the number of quantified peptides per proteins varied from 258 (Apolipoprotein B) to 1. In the bed rest dataset, of the 44 proteins out of 535 that were quantified with only one peptide, only teneurin-4 (TENM4, 112 MS/MS quantifications) was considered for further analysis. All other proteins quantified with only one peptide were not further analysed. The MS-based proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD032969 and are publicly available as of the date of publication.

Bioinformatic and statistical analysis

Analyses were performed with the Perseus software (version 1.6.15.0), part of the MaxQuant environment[22] and with the R software environment (https://www.R-project.org). Label-free quantification values with a minimum ratio of 1 were used throughout the analysis for protein abundance using the feature implemented in MaxQuant.[23] Categorical annotations were supplied in the form of UniProt Keywords, Corum, KEGG and Gene Ontology terms. Annotation enrichments

were calculated by Fisher's exact test using the Benjamini–Hochberg method for FDR truncation at a cutoff of 2% and the UniProt human proteome as background. For longitudinal comparisons, we used paired Student's *t*-tests, with significance threshold set at 5% using permutation-based FDR with 250 randomizations. Technical replicates were averaged (N = 10). For comparison between different subjects within both cohorts, we used Welch tests with significance cut-off set at 5% employing permutation-based FDR with 250 randomizations and one-way ANOVA ($P < 0.05$) with Tukey's honestly significant difference post hoc tests. Principal component analysis (PCA) was carried out after filtering the dataset for 60% valid values and imputing missing values, assuming a Gaussian distribution and with a downshift of 1.8 the standard deviation of valid values.

Results

Proteomic workflow and features of the bed rest plasma dataset

We carried out a longitudinal proteomic analysis of the blood plasma of 10 young healthy volunteers undergoing ten days of continuous bed rest (*Tables* S1 and S3). Samples were taken immediately before bed rest (BR0), at day 10 (BR10) and after 2 days of free re-ambulation post bed rest (R + 2). Our second cohort consisted of 14 cancer patients, seven with cachexia and seven without, and 14 controls (*Tables* S2 and S4). All samples were analysed by liquid chromatography coupled to MS/MS followed by computational analysis. The two datasets were measured separately and the protein abundance results were cross-analysed (*Figure* 1A; see also Materials and methods).

Our proteomic analysis of the bed rest cohort resulted in an average Pearson correlation of 0.95 among all subjects and time points without clear outliers and of >0.96 for technical replicates (*Figure* S1A). We quantified 535 proteins in total and 360 per subject on average, of which 286 were quantified in all subjects (*Figure* S1B). We carried out a quality control assessment of our plasma samples by measuring the intensities of known marker proteins from three contamination panels, derived from other blood components and occurring in plasma due to improper sample handling.[20] Contamination from red blood cells was consistently below a recommended intensity threshold of 2.5% (*Figure* S1C) and from platelets below 0.5% (*Figure* S1D). Coagulation markers were mostly below 10% in all samples, except for triplicates of one subject at one time point (28%), likely resulting from incorrect sample handling (*Figure* S1E).

The quantitative dynamic range of intensity in our plasma dataset spans five orders of magnitude from highly abundant albumins to the lowest intense protein quantified, the cytoskeleton-associated protein Profilin-1. We crossed our plasma dataset to the 'secreted to blood' protein list of the Human Protein Atlas, containing 784 proteins.[24] Proteins in the highest expression quartile predominantly originated from plasma. Proteins of other origin, possibly deriving from tissue leakage, were progressively more prevalent in the lower intensity quartiles. We could detect nuclear and mitochondrial proteins, likely deriving from cell damage (*Figure* 1B). The highest intensity quartile 1 was significantly enriched in GO terms of apolipoproteins and acute-phase proteins, whereas the lowest quartile 4 was enriched in intracellular proteins, indicating a tissue leakage origin (*Figure* 1C).

To verify that unloading was the major source of variability within the samples, we carried out PCA. This procedure separated the BR0 (black squares) from the BR10 (orange dots) samples diagonally along components 1 and 2 (*Figure* 1D), based on differences in both canonical plasma proteins, like APP and SERPINA1, and tissue leakage proteins, like the mitochondrial ATP5B and the chaperone HSP90AB1 (*Figure* 1E). This result shows that the differences in the plasma proteome correlating with unloading are larger than the individual variability among subjects. Samples taken at R + 2 were clearly separated by PCA from both BR0 and BR10 (*Figure* S1F,G).

Loading-dependent changes in the plasma proteome

Blood was drawn from all subjects right before bed rest (BR0), after 10 days of continuous bed rest (BR10) and 2 days after re-ambulation (R + 2) before hospital discharge. We constructed a global correlation map containing pairwise relationships between all proteins quantified in the dataset. In our case, there were up to 87 abundance values for each plasma protein (10 individuals; three time points, two to three technical replicates). Unsupervised hierarchical correlation clustering of the expression profiles across all samples yielded clusters of highly co-regulated proteins (cluster mean >0.8) involved in the immune response, complement and coagulation cascades, lipid metabolism and integrin signalling (*Figure* S2).

To highlight proteins whose abundance in plasma changes at the different loading states of this sequence, we compared samples taken in the unloading phase (BR10) with those drawn in the loading phase pre-bed (BR0) and post-bed rest (R + 2). We carried out paired *t*-tests for all 10 subjects and three technical replicates, retrieving 22 proteins with significantly different abundance between BR10 and BR0 and 32 between BR10 and R + 2. Unsupervised hierarchical clustering of the median expression of these proteins in 10 subjects grouped the plasma at BR0 and R + 2, separating it from that at BR10. Eight proteins were significant hits in both comparisons. (*Figure* 2A and *Table* S5). The plasma proteins whose

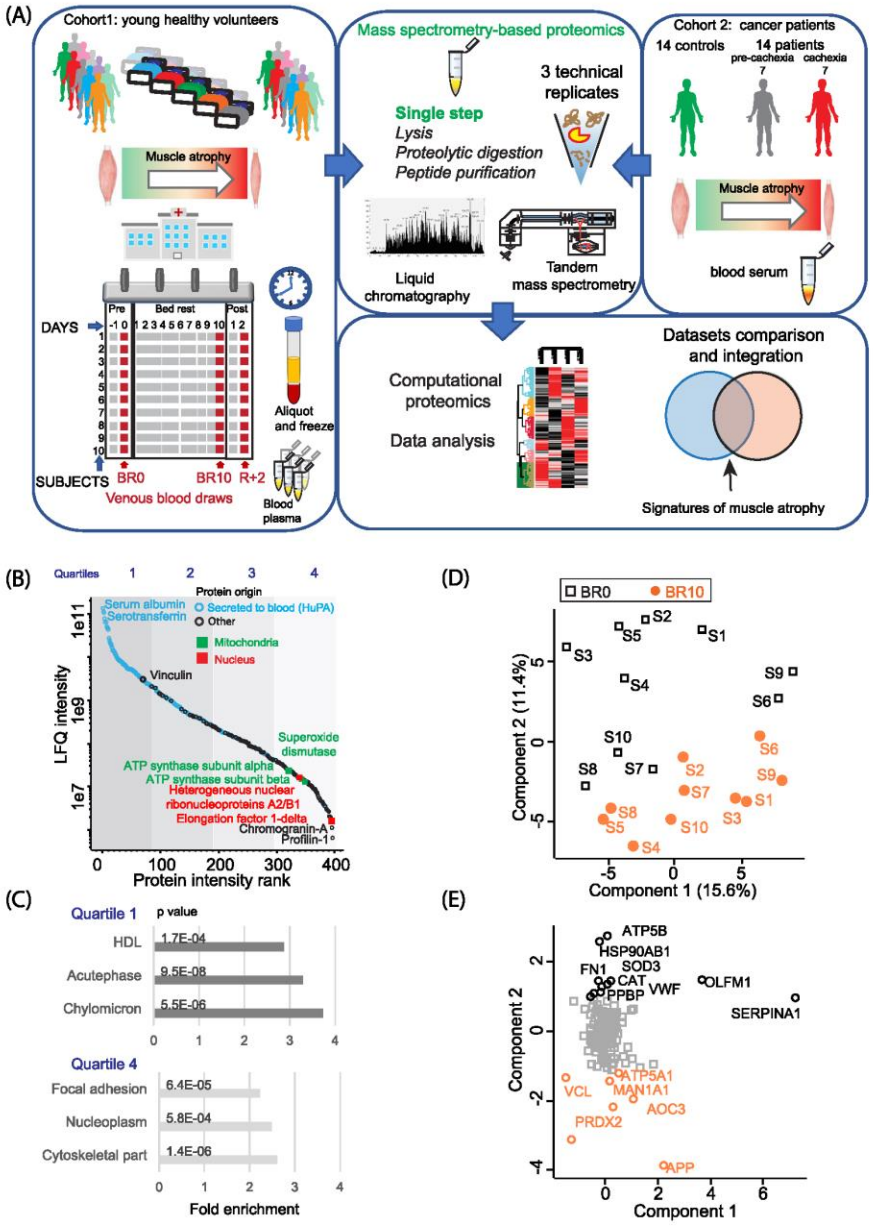


Figure 1 Study design and main features of the dataset. (A) Sample collection, preparation and proteomic analysis workflow. The study cohort involved 10 subjects who underwent 10 days of continuous bed rest that caused muscle atrophy. BR0, before bed rest. BR10, after 10 days of continuous bed rest, R + 2, after 2 days of recovery at weight-bearing conditions. A second cohort used for validation comprised serum from 14 cancer patient with or without cachexia and 14 controls. Frozen samples were proteolytically digested and analysed by liquid chromatography coupled to mass spectrometry, followed by computational proteomics and data analysis. (B) Intensity rank of the proteins quantified in the plasma of the subjects at three time points. Proteins in light blue are annotated as 'secreted to blood', a list of 784 proteins in the Human Protein Atlas (HuPA) repository. The two most and least abundant proteins are labelled. Representative proteins derived from intracellular compartments are marked over the abundance range (red and green squares). Abundance quartiles are visualized with different shades of grey. (C) Top annotation enrichments among highly abundant proteins in the first quartile (top) among low abundance proteins in the fourth quartile (bottom). Fisher exact test, Benjamini–Hochberg FDR for truncation with threshold set at 0.02. (D) Principal component analysis (PCA) separating the plasma proteome of 10 subjects at BR0 (black squares) from that at BR10 (orange dots). (E) PCA loadings, with the proteins driving the separation between the groups labelled in the corresponding colours.

Journal of Cachexia, Sarcopenia and Muscle 2023; **14**: 439–451
DOI: 10.1002/jcsm.13146

abundance changes significantly between different loading states can be assigned several distinct functions. Most proteins were annotated to GO terms blood coagulation, immunity and lipid transport (*Figure* 2B). Four hits, namely, Lumican, Teneurin 4, Proteoglycan 4 and IgGFc-binding proteins, were not among the plasma-secreted to blood gene set. Based on their characterized interactors (STRING, see Supporting Information), they may be interacting with the extracellular matrix, and one of them, Teneurin4, has a synaptic localization.[25] Proteoglycan4/Lubricin is a lubricating glycoprotein localized at the cartilage surface with a key function in the biomechanical properties of the tissue[26] (*Figure* 2C).

We confirmed the decreased abundance of Lumican in plasma during bed rest by western blot, analysing also an intermediate timepoint of the sequence, BR5. Our results showed a consistently higher Lumican signal in the plasma from BR0 in all subjects analysed with this method, matching the results obtained by MS-based proteomics (*Figure* 2D). Depletion from plasma of the 12 most abundant proteins allowed a clearer visualization of this effect (*Figure* 2D, right panel).

Functional interaction networks of plasma proteins changing in abundance with body loading state

We divided the plasma proteins significantly changing in abundance between body loading states into two groups, namely proteins with (i) lower abundance at BR10 and (ii) higher abundance at BR10 unloading compared with both BR0 and R + 2. We then visualized both groups as functional interaction networks, based on physical interaction, co-expression and data mining (see Supporting Information). Proteins whose abundance in plasma decreased significantly during bed rest and increased again upon reloading included proteins involved in coagulation and were significantly enriched in the annotation term extracellular matrix organization (P = 1.4E-5) (*Figure* 3A). We calculated the BR10/BR0 abundance ratio in each of the 10 subjects separately. The decrease in abundance could be measured in a majority of subjects for most proteins, and it was especially large (>4-fold) in the case of fibronectin, platelet basic protein and von Willebrand factor (*Figure* 3B). Proteins whose abundance increased at BR10 compared with both BR0 and R + 2 formed a tight functional network specifically enriched in the annotation term protease inhibitor (P = 2.5E-10) and lipoproteins (P = 2.7E-8) (*Figure* 3C). The latter might be correlated with the changes in lipid metabolism measured in these subjects during bed rest, which included a decrease in plasma cholesterol and an increase in triglycerides, concomitant with an increase in insulin resistance (*Figure* S3). Interestingly, three proteins did not show any functional association with the network under our conditions, namely, Teneurin 4 (see *Figure* 2), Proteoglycan 4, a component of the extracellular matrix of cartilage, and Attractin, a protein expressed in many tissues including skeletal muscle.[27] Detailing the changes in each subject (BR10/BR0) IgGFc-binding protein, secreted phosphoprotein 24 and Teneurin 4 had the largest increases in abundance (>4-fold) (*Figure* 3D).

Subject-centric correlation between muscle atrophy and plasma proteome

We previously measured muscle atrophy at BR10 in this bed rest cohort.[16] We observed that seven subjects developed muscle atrophy amounting from 4% up to 12.2% loss of quadriceps volume during 10 days of bed rest. Three subjects were relatively atrophy resistant, displaying a corresponding quadriceps volume change of 0.5–1.9% at BR0 compared with BR0 (*Figure* 4A). This was confirmed measuring the difference in fibre cross-sectional area (CSA) in the muscle biopsies of these two groups of subjects. A 7.7% decrease of the median fibre CSA was measured in the fibres of the seven subjects developing muscle atrophy. Conversely, the fibres of the three atrophy-resistant subjects showed essentially no median CSA decrease at BR10 (*Figure* 4B). The heterogeneity of individual responses to bed rest in terms of muscle atrophy and bone loss has been documented in previous studies.[28] The plasma proteome of atrophy-prone and atrophy-resistant subjects at BR0 showed no significant difference, although complement factor H-related protein 3 (CFHR3) had a clear tendence to over two-fold higher expression in atrophy-prone subjects (*Figure* 4C). However, the same comparison at BR10 highlighted four proteins expressed at higher level in atrophy-resistant subjects and two expressed at higher level in atrophy-prone subjects. Haptoglobin-related protein (HPR), apolipoproteins AI and AIV (APOA1, APOA4) and transthyretin (TTR) were more abundant in the plasma of atrophy-resistant subjects, suggesting that higher plasma abundance of these proteins may have a positive correlation with the preservation of muscle mass. Conversely, inter-alpha-trypsin inhibitor H3 (ITIH3) and complement factor H (CFH) displayed higher abundance in the plasma of subjects undergoing larger loss of muscle mass during bed rest, indicating a negative correlation with muscle trophism (*Figure* 4D).

We reasoned that, if these proteins relay the loss vs maintenance of muscle mass occurring in these two groups of subjects, they might also be common to other contexts in which muscle atrophy occurs. For this purpose, we analysed the serum proteome of a cohort of 14 cancer patients with or without cachexia and 14 age-matched controls. We quantified 390 proteins in total, ranging from 223 to 278 in different subjects (*Table* S3). Contamination from red blood cells and platelets were minor (*Figure* S4A,B). We carried out ANOVA and post hoc tests comparing control subjects with cancer patients with and without cachexia. This analysis retrieved 24

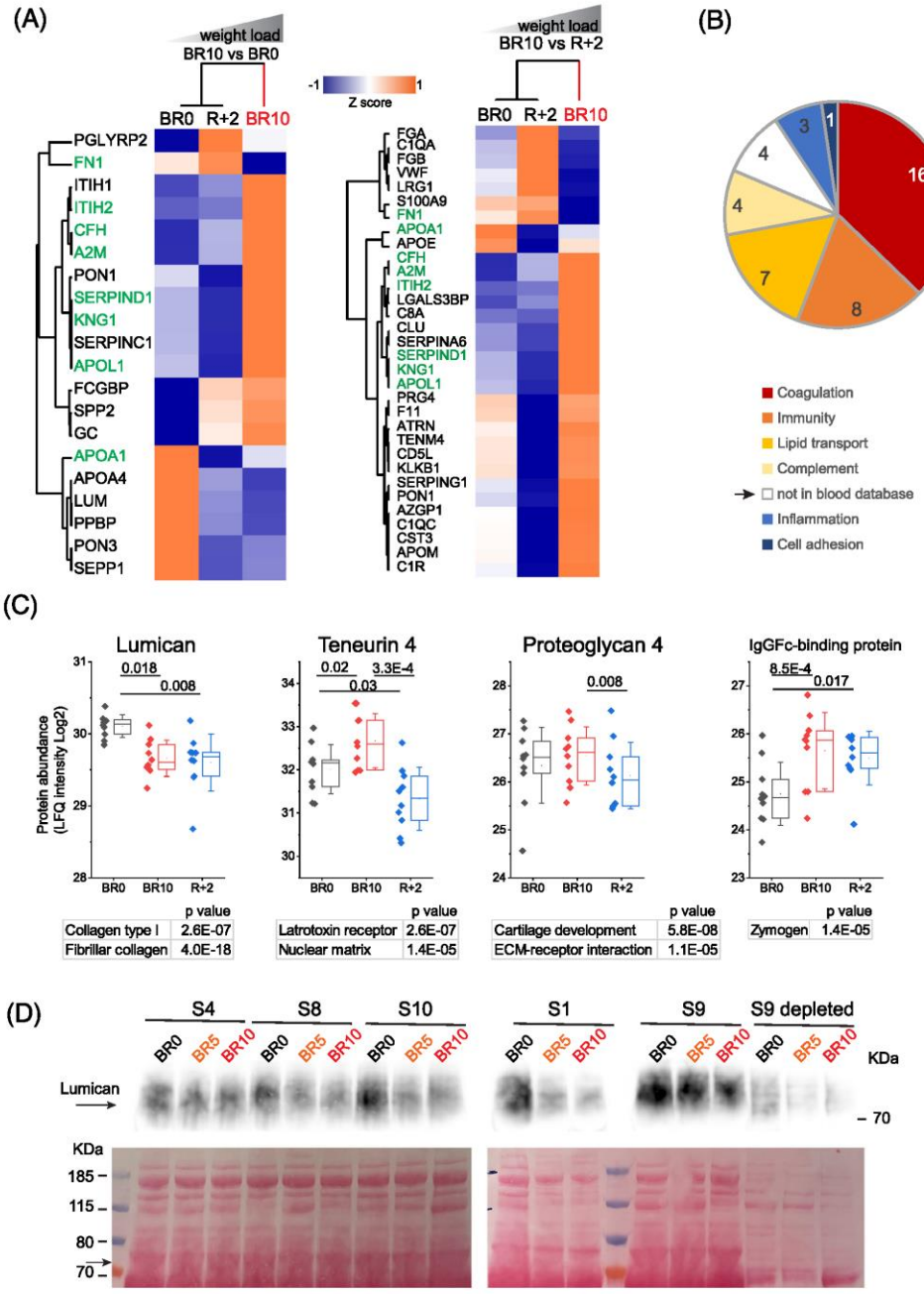


Figure 2 Proteins whose abundance in plasma varied at different loading conditions. (A) Unsupervised hierarchical clustering of proteins that changed significantly in different loading conditions between at least two time points (paired *t*-test, *N* = 10 subjects, permutation-based FDR = 0.05). (B) Main cell compartment or functional class distribution of among ANOVA significant proteins (in percent). From GO terms, manually curated. (C) Plasma abundance changes at different phases of the bed rest protocol of four proteins not of blood origin (see arrows in C) likely originating from tissue leakage. Student's *t*-test, *N* = 10. (D) Top, western blots showing a decreased abundance of Lumican in whole plasma of five subjects (S, see label on top) at BR5 (not analysed by MS-based proteomics) and BR10, matching the results in (C). For S9, Lumican expression after depletion of the 12 most abundant proteins is also shown (top right, last three lanes. Bottom, Ponceau S staining of the upper part of the corresponding membrane). The position of the Lumican band is indicated by an arrow.

Journal of Cachexia, Sarcopenia and Muscle 2023; **14**: 439–451
DOI: 10.1002/jcsm.13146

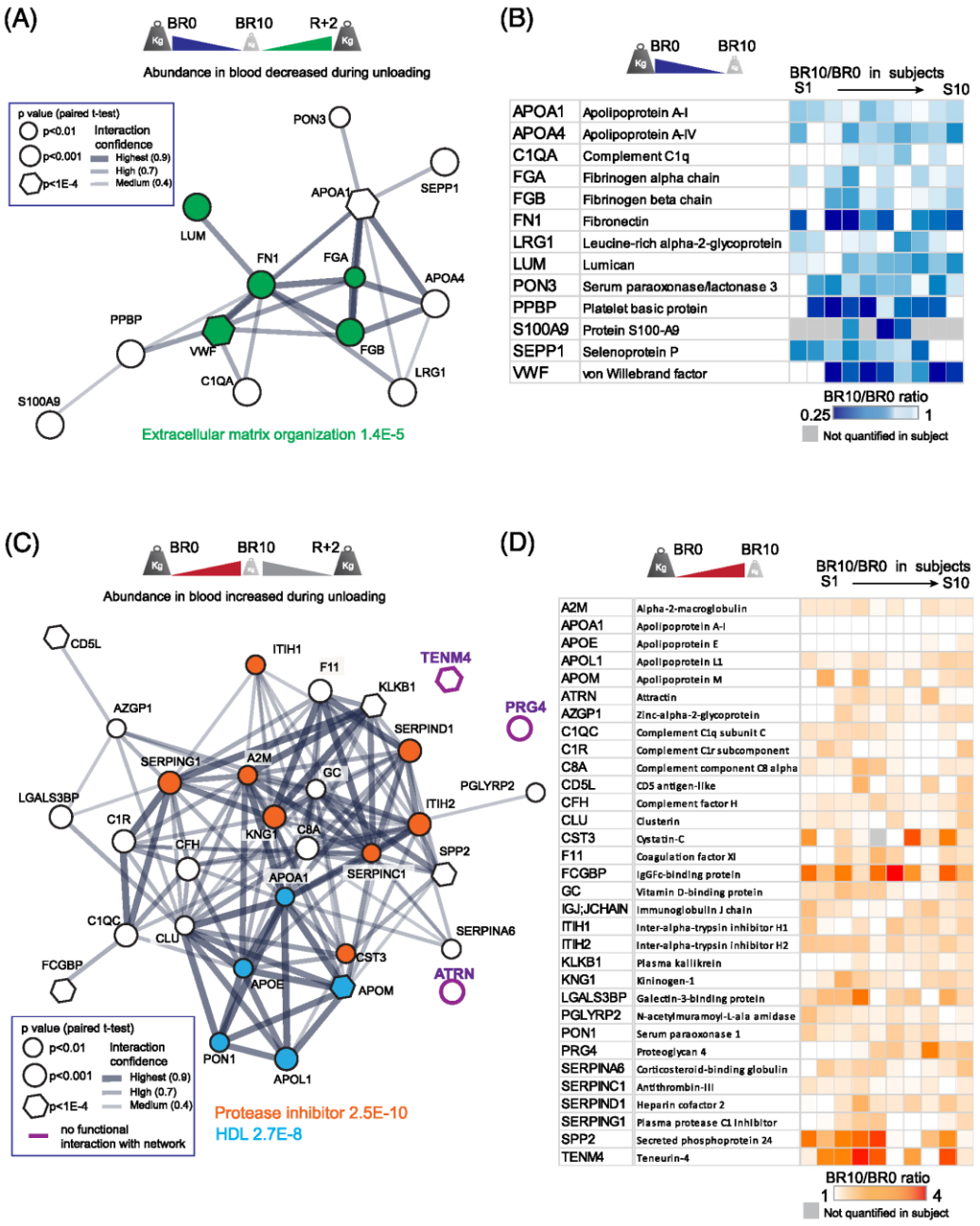
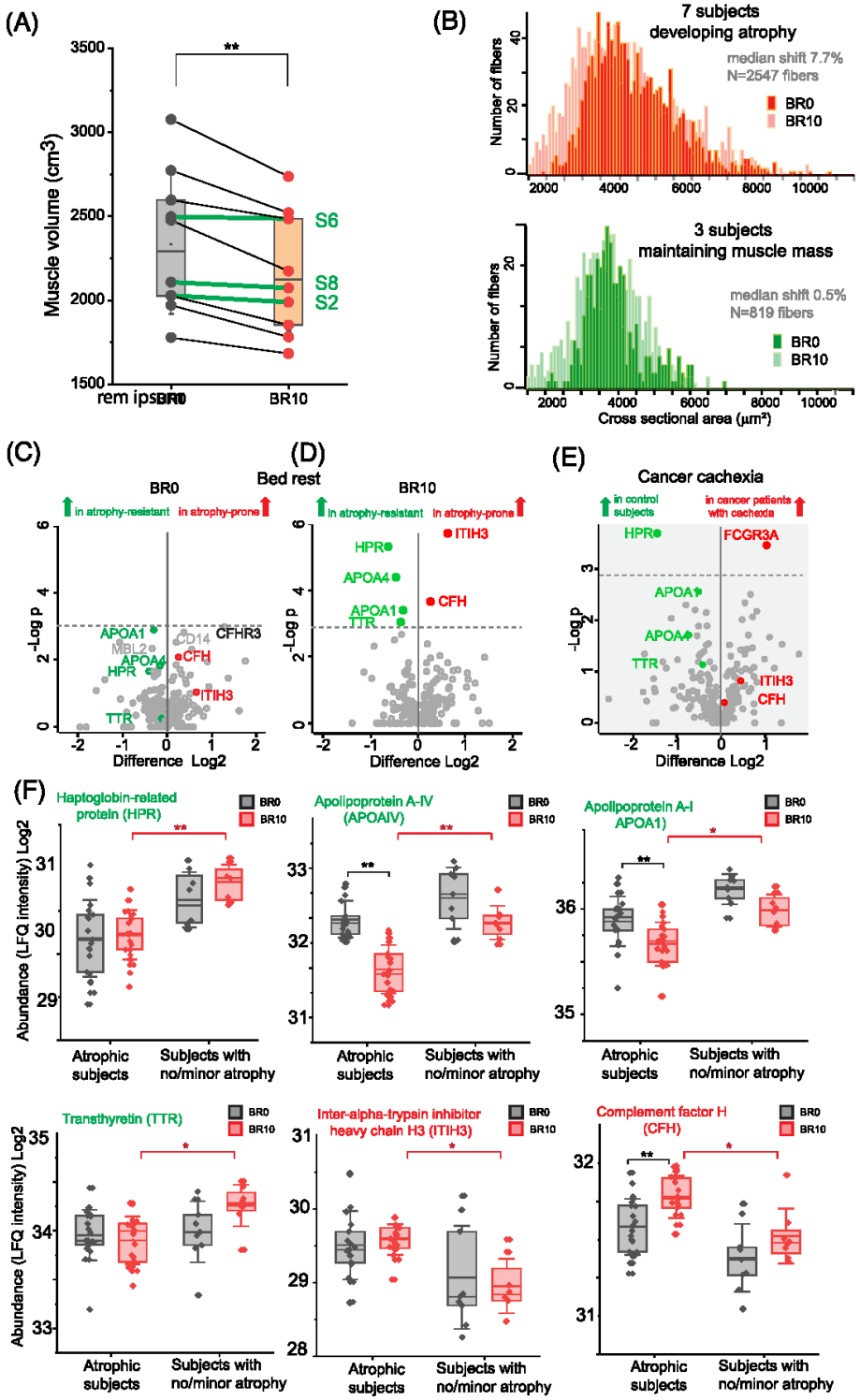


Figure 3 Functional interactions of plasma proteins changing in abundance according to loading conditions. (A) Functional interaction network of proteins whose abundance in plasma decreases during unloading (schematically represented on top). Thickness of connecting lines represents interaction confidence. Node shape is determined by the *P* value of ANOVA. Colours refer to the two most enriched annotations as indicated. See methods for details. (B) Subject-centric heatmap analysis of the decrease in abundance between BR10 and BR0 (BR10/BR0 ratio). Each square shows the median ratio (BR10/BR0) in each of the 10 subjects ordered from S1 to S10, left to right as indicated on top. (C) Functional interaction network of proteins whose abundance in plasma increases during unloading (see A for description). (D) Subject-centric heatmap analysis of the increase in abundance between BR10 and BR0.

Journal of Cachexia, Sarcopenia and Muscle 2023; **14**: 439–451
DOI: 10.1002/jcsm.13146



Journal of Cachexia, Sarcopenia and Muscle 2023; **14**: 439–451
DOI: 10.1002/jcsm.13146

Figure 4 Common features of muscle atrophy in bed rest and cancer cachexia. (A) Bed rest-dependent volume changes of quadriceps femoris at BR0 (black dots) and BR10 (red dots) measured by magnetic resonance imaging (MRI). Three subjects developing no or minor atrophy are indicated by green lines. N = 10 subjects. Box shows median, 75th and 25th percentile; whiskers show standard deviation. (B) Distribution of muscle fibre cross-sectional area overlaying BR0 (dark) and BR10 (light). Top, in red, all fibres measured in the muscle biopsies of seven subjects that developed muscle atrophy during bed rest (N = 2547 fibres). Bottom, in green, the same analysis for subjects S2, S6 and S8 (see A) that were largely atrophy-resistant (N = 819 fibres). (C) Volcano plot comparing the plasma proteome of atrophy-prone and atrophy-resistant bed rest subjects at BR0. In red and green, proteins with significant abundance difference between the subjects at day 10. N = 3 atrophy-resistant and 7 atrophy-prone subjects, technical triplicates. Dashed line, P value 0.05. Threshold, permutation-based FDR = 0.05. (D) Same analysis as in (C), BR10. Proteins with significantly different abundance between atrophy-resistant and atrophy-prone subjects are labelled with a filled circle. (E) Volcano plot comparing the serum proteome of cancer patients with cachexia (N = 7) with that of controls (N = 14). Proteins with significantly different abundance between these groups are labelled in colour with a filled circle above the dashed line marking P value = 0.05. Proteins with significantly different abundance in the bed rest dataset at day 10 are labelled in colour. (F) Expression of the six proteins with differential expression at BR10 between the seven atrophy-prone (left side of each graph, technical triplicates) and the three atrophy-resistant subjects (right side of each graph, technical triplicates). The red line with asterisks shows the significant differences at BR10 between the two groups of subjects. The black line shows significant differences between BR0 (in black) and BR10 (in red) within the two subject groups. N = 10 subjects with two to three technical replicates. $*P < 0.05$, $**P < 0.01$. Box shows median, mean, 75th and 25th percentile; whiskers show standard deviation.

proteins whose abundance in plasma differed between at least one of the three groups. Unsupervised hierarchical clustering separated control from cancer patients and the significant hits formed three distinct clusters (*Figure* S4C).

We then compared the serum proteome of control subjects with that of cancer patients that had developed cachexia, leading to the loss of over 5% of their body weight (*Table* S2). Similar to the results obtained in bed rest, HPR was significantly more abundant in the serum of controls compared with that of cachectic cancer patients (*Figure* 4E). This shows that HPR abundance in plasma/serum decreased in subjects losing muscle mass both because of mechanical unloading in young healthy subjects and of cachexia in cancer patients, two scenarios with very few common aspects. Cancer patients significantly up-regulated the receptor for the invariable Fc fragment of immunoglobulin gamma FCGR3A, which was part of a regulated protein cluster in bed rest (see *Figure* 3D and S2). Interestingly, the remaining proteins with higher abundance in atrophy-resistant bed rest subjects also tended to be more abundant in controls compared with cancer patients (compare *Figure* 4D,E, labelled in green). A similar analysis of the serum of controls and cancer patients classified clinically as having pre-cachexia yielded a different set of significant proteins, and the expression difference between the two groups was small (*Figure* S4D). Interestingly, ITIH3, a member of inter-alpha-trypsin inhibitor protein family, was more abundant in the plasma of bed rest subjects at BR10 as well as in the serum of cancer patients with cachexia, correlating in both cohorts with an atrophy state (*Figure* 4D,E).

A plasma protein with significantly different abundance at BR0 between atrophy-prone and atrophy-resistant subjects could be explored as a predictive biomarker. No plasma protein had this behaviour (at a P value cut-off of 0.05) in our bed rest cohort (*Figure* 4C). We then focused on the analysis of the six plasma proteins correlated with maintenance or loss of muscle mass at BR10. Interestingly, all but TTR had a tendence to different median expression levels at BR0 in subject developing muscle atrophy compared with those essentially resistant to it (*Figure* 4F; compare grey boxes). All of them had significantly different expression at BR10 as expected (see also *Figure* 4D).

Discussion

We applied MS-based proteomics to the analysis of plasma samples from a cohort of 10 participants in a bed rest study, undergoing muscle atrophy varying from 12 to 0.4% (quadriceps femoris volume) in 10 days.[5,16] With this approach, we aimed at correlating the loss of muscle mass with changes in abundance of plasma proteins, which could be used to monitor the state of skeletal muscle in a minimally invasive way. We could quantify over 500 proteins in total, amounting to 360 on average in each subject.

Our results revealed over 30 proteins undergoing abundance changes in plasma comparing BR10, the endpoint of mechanical unloading, with BR0, the time point immediately before bed rest. Interestingly, four of the significant proteins were not typical blood components but possibly deriving from tissue leakage. One of them, Teneurin 4, is part of an evolutionarily conserved protein family located predominantly at the synapse.[25] We have previously shown that the subjects of this cohort showed neuromuscular instability, as indicated by the up-regulation of neural cell adhesion molecule 1 (NCAM) in skeletal muscle, a marker of denervation/re-innervation events.[16] It will be of interest to test Teneurin 4 as a readout for NMJ instability. Proteoglycan 4 (PRG4)/lubricin is a cartilage protein whose serum abundance increases in patients with active inflammatory cartilage disease.[29] We detected a minor increase in the plasma abundance of PRG4 during bed rest, but a significant twofold decrease 2 days after reloading. Lumican, a protein enriched in the extracellular matrix of articular cartilage, was more abundant at BR0 than at both BR10 and BR5, as we could

show in validation experiments using western blot. It could be speculated that variations in loading cause extensive remodelling of cartilage, leading to changes in plasma abundance of extracellular matrix proteins.[30]

In addition, we found significant decrease in abundance at BR10 for proteins involved in interactions with the extracellular matrix and in blood coagulation. Long hospitalizations are linked to a hypercoagulable state and to increased risk of thromboembolytic complications. However, in line with our findings, previous bed rest studies in healthy young subjects have observed no increase in major coagulation parameters during 21 days[31] or 60 days of head down tilt bed rest.[32] Indeed, both studies reported a tendence to a hypocoagulable state during bed rest, which would work as a compensatory mechanism. Our results show that the abundance of some plasma apolipoproteins was higher at BR10 compared with BR0, including APOA1, whose plasma concentration was not modified by inactivity in other studies.[33] These changes may be due to the inactivity-linked insulin resistance that we and others have consistently observed starting in the early phases of bed rest.[34] Insulin resistance causes lipoprotein lipase inhibition and activation of hepatic triglyceride synthase, which are known to cause significant changes in blood lipid profile.[35] Inhibitors of different protease families, including anti-trypsin, anti-thrombin and anti-C3, were more abundant at BR10 compared with BR0. Members of the inter-alpha-trypsin family have been recently suggested to associate with mortality in COVID-19. ITIH3 and ITIH1/2 showed opposite differences in abundance between survivors and non-survivors.[36] Our data confirm opposite changes of different members of this protein family, both in subjects undergoing bed rest and in cancer patients (see below).

Interindividual differences in the response to intervention (e.g. lifestyle of pharmacological) are the theoretical basis for personalized medicine, which is rapidly developing with the support of large throughput data generated with omics technologies. The ability to predict different impacts of inactivity with minimally invasive methods would be of great interest to monitor community health and design early intervention, particularly in the elderly population. In the future, biomarkers predicting a muscle atrophy-resistant phenotype might be of paramount importance for the selection of astronauts for long space missions, where body unloading due to microgravity represents a severe challenge for human health.[37] A serendipitous finding of our previous analysis of this bed rest subject cohort was significant inter-individual heterogeneity in the susceptibility to unloading-induced muscle atrophy, consistent with previous reports.[28] Whereas seven subjects lost between 4% and 12.2% of their quadriceps volume in 10 days, three of them had minor decreases, from 1.9 to 0.4%. Comparing the abundance of plasma proteins in atrophy-prone and atrophy-resistant subjects at BR10, we highlighted six proteins showing significant differences between the two groups. Two proteins were more abundant in subjects developing atrophy during bed rest, namely, the protease inhibitor ITIH3 and complement factor H (CFH). Four proteins, haptoglobin-related protein (HPR), transthyretin (TTR) and the apolipoproteins APOA1 and APOA2 were more abundant in atrophy-resistant subjects. Interestingly from a biomarker perspective, the abundance difference was the same at BR10 as at BR0, though only the samples at BR10 reached statistical significance under our conditions (compare *Figure* 4C,D). It will be of interest to further evaluate the ability of these proteins, alone or in combination, to predict the proneness to muscle atrophy in different subjects.

To further evaluate the relationship between loss of muscle mass and changes in circulating proteins, we analysed the serum of seven cancer patients with cachexia, leading to over 8% loss of total body weight. The comparison between the serum of cachectic patients and that of controls yielded two significant proteins. The receptor for the invariable Fc fragment of immunoglobulin gamma FCGR3A/CD16A, a cytotoxicity receptor of human natural killer (NK) cells,[38] was more abundant in cancer patients with cachexia. This might be linked to the disease phenotype, though the functional annotation FCGR activation was also regulated in bed rest (*Figure* S2). Interestingly, haptoglobin-related protein/HPR was over twofold more abundant in the serum of controls compared with cancer patients with cachexia.

Crossing the results of the bed rest and cancer cachexia cohort, we thus show that the level of circulating haptoglobin-related protein/HPR correlates with the maintenance of muscle mass in both conditions inducing skeletal atrophy, despite the large differences characterizing the two subject groups. Although HPR has been proposed as a serum marker of lymphoma,[39] the abundance of HPR does not result different when we compare the serum cancer patient without cachexia with that of controls (*Figure* S4D). Our analysis points to a positive correlation between circulating HPR and muscle mass.

Despite suggesting a number of circulating potential biomarkers of muscle atrophy, our study presents several limitations that need to be taken into account. The bed rest dataset lacks an intermediate time point. As a consequence, our proteomic data do not show how these potential biomarkers change over time and whether they occur in the early phase of bed rest, where most of the signal transduction controlling atrophy unfolds, or whether they manifest towards the end of the bed rest sequence, where atrophy is most pronounced. For Lumican, we could show by western blot that the plasma abundance is already decreased at BR5 and maintained at a low level at BR10. Our pilot study is also limited in sample size, so our findings will need further validation in larger cohorts. At this stage, our result cannot yet contribute practical predictive power to the indirect methods used to assess muscle atrophy, like grip strength or body weight measurements. However, this detailed quantification of the plasma

\proteome, together with the characterization of the skeletal muscle of the same bed rest cohort from our parallel studies,[16,40] will allow to draw correlations and perform data mining once larger validation cohorts have been analysed.

In conclusion, we found changes in the plasma proteome of healthy subjects undergoing voluntary bed rest that accompany and may be linked to the mechanical loading/activity state of the body and to muscle trophism. In the future, this type of studies validated in large cohorts will lead to the definition of biomarkers panels relaying information on skeletal muscle trophism, contributing to the development of point-of-care diagnostics for human health.

Acknowledgements

We thank Igor Paron for his assistance with MS; Katharina Zettl and Bianca Splettstoesser for their help with the validation experiments; and Isabell Bludau and Wen-Feng Zeng for the discussion of statistical approaches (all at MPI of Biochemistry). We are grateful to the bed rest team and Izola General Hospital personnel and to the volunteers enrolled this study. This work was funded by the Italian Space Agency (ASI), MARS-PRE Project, No. DC-VUM-2017-006 (to MMu and MN) and by the Louis-Jeantet Foundation and EU 7th Framework Programme (grant agreement HEALTH-F4-2008-201648/PROSPECTS) (to MMa). The authors of this manuscript certify that they comply with the ethical guidelines for authorship and publishing in the *Journal of Cachexia, Sarcopenia and Muscle*.[41]

Open Access funding enabled and organized by Projekt DEAL.

Conflict of interest

The authors declare no conflict of interest.

Online supplementary material

Additional supporting information may be found online in the Supporting Information section at the end of the article.

References

1. Cruz-Jentoft AJ, Sayer AA. Sarcopenia. *Lancet* 2019;**393**:2636–2646.
2. Fearon K, Strasser F, Anker SD, Bosaeus I, Bruera E, Fainsinger RL, Jatoi A, Loprinzi C, MacDonald N, Mantovani G, Davis M, Muscaritoli M, Ottery F, Radbruch L, Ravasco P, Walsh D, Wilcock A, Kaasa S, Baracos VE. Definition and classification of cancer cachexia: an international consensus. *Lancet Oncol* 2011;**12**:489–495.
3. Argiles JM, Busquets S, Stemmler B, Lopez-Soriano FJ. Cancer cachexia: understanding the molecular basis. *Nat Rev Cancer* 2014;**14**:754–762.
4. Kilroe SP, Fulford J, Jackman S, Holwerda A, Gijsen A, van Loon L, Wall BT. Dietary protein intake does not modulate daily myofibrillar protein synthesis rates or loss of muscle mass and function during short-term immobilization in young men: a randomized controlled trial. *Am J Clin Nutr* 2021;**113**:548–561.
5. Franchi MV, Sarto F, Simunic B, Pisot R, Narici MV. Early changes of hamstrings morphology and contractile properties during 10 days of complete inactivity. *Med Sci Sports Exerc* 2022;**54**:1346–1354.
6. Šimunič B, Koren K, Rittweger J, Lazzer S, Reggiani C, Rejc E, Pišot R, Narici M, Degens H. Tensiomyography detects early hallmarks of bed-rest-induced atrophy before changes in muscle architecture. *J Appl Physiol* 2019;**126**:815–822.
7. Sartori R, Romanello V, Sandri M. Mechanisms of muscle atrophy and hypertrophy: implications in health and disease. *Nat Commun* 2021;**12**:330.
8. Bodine SC, Latres E, Baumhueter S, Lai VK, Nunez L, Clarke BA, Poueymirou WT, Panaro FJ, Na E, Dharmarajan K, Pan ZQ, Valenzuela DM, DeChiara TM, Stitt TN, Yancopoulos GD, Glass DJ. Identification of ubiquitin ligases required for skeletal muscle atrophy. *Science* 2001;**294**:1704–1708.
9. Sandri M, Sandri C, Gilbert A, Skurk C, Calabria E, Picard A, Walsh K, Schiaffino S, Lecker SH, Goldberg AL. Foxo transcription factors induce the atrophy-related ubiquitin ligase atrogin-1 and cause skeletal muscle atrophy. *Cell* 2004;**117**:399–412.
10. Hughes DC, Baehr LM, Driscoll JR, Lynch SA, Waddell DS, Bodine SC. Identification and characterization of Fbxl22, a novel skeletal muscle atrophy-promoting E3 ubiquitin ligase. *Am J Physiol Cell Physiol* 2020;**319**:C700–C719.
11. Masiero E, Agatea L, Mammucari C, Blaauw B, Loro E, Komatsu M, Metzger D, Reggiani C, Schiaffino S, Sandri M. Autophagy is required to maintain muscle mass. *Cell Metab* 2009;**10**:507–515.
12. Tezze C, Romanello V, Desbats MA, Fadini GP, Albiero M, Favaro G, Ciciliot S, Soriano ME, Morbidoni V, Cerqua C, Loefler S, Kern H, Franceschi C, Salvioli S, Conte M, Blaauw B, Zampieri S, Salviati L, Scorrano L, Sandri M. Age-associated loss of OPA1 in muscle impacts muscle mass, metabolic homeostasis, systemic inflammation, and xepithelial senescence. *Cell Metab* 2017;**25**:1374–1389.e6.
13. Cohen S, Nathan JA, Goldberg AL. Muscle wasting in disease: molecular mechanisms and promising therapies. *Nat Rev Drug Discov* 2015;**14**:58–74.
14. Richter EA, Hargreaves M. Exercise, GLUT4, and skeletal muscle glucose uptake. *Physiol Rev* 2013;**93**:993–1017.
15. Rittweger J, Albracht K, Flück M, Ruoss S, Brocca L, Longa E, Moriggi M, Seynnes O, di Giulio I, Tenori L, Vignoli A, Capri M, Gelfi C, Luchinat C, Franceschi C, Bottinelli R, Cerretelli P, Narici M. Sarcolab pilot study into skeletal muscle's adaptation to long-term spaceflight. *NPJ Microgravity* 2018;**4**:18.
16. Monti E, Reggiani C, Franchi MV, Toniolo L, Sandri M, Armani A, Zampieri S, Giacomello E, Sarto F, Sirago G, Murgia M, Nogara L, Marcucci L, Ciciliot S, Šimunic B, Pišot R, Narici MV. Neuromuscular junction instability and altered intracellular calcium handling as early determinants of force loss during unloading in humans. *J Physiol* 2021;**599**:3037–3061.
17. Lim S, Dunlap KR, Rosa-Caldwell ME, Haynie WS, Jansen LT, Washington TA, et al. Comparative plasma proteomics in muscle atrophy during cancer-cachexia and disuse: the search for atrokines. *Physiol Rep* 2020;**8**:e14608.
18. Sartori R, Hagg A, Zampieri S, Armani A, Winbanks CE, Viana LR, Haidar M, Watt KI, Qian H, Pezzini C, Zanganeh P, Turner BJ,

Larsson A, Zanchettin G, Pierobon ES, Moletta L, Valmasoni M, Ponzoni A, Attar S, da Dalt G, Sperti C, Kustermann M, Thomson RE, Larsson L, Loveland KL, Costelli P, Megighian A, Merigliano S, Penna F, Gregorevic P, Sandri M. Perturbed BMP signaling and denervation promote muscle wasting in cancer cachexia. *Sci Transl Med* 2021;**13**.

19. Hortin GL, Sviridov D, Anderson NL. High-abundance polypeptides of the human plasma proteome comprising the top 4 logs of polypeptide abundance. *Clin Chem* 2008;**54**:1608–1616.
20. Geyer PE, Voytik E, Treit PV, Doll S, Kleinhempel A, Niu L, et al. Plasma proteome profiling to detect and avoid sample-related biases in biomarker studies. *EMBO Mol Med* 2019;**11**:e10427.
21. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 2008;**26**:1367–1372.
22. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* 2016;**13**:731–740.
23. Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* 2014;**13**:2513–2526.
24. Uhlén M, Karlsson MJ, Hober A, Svensson AS, Scheffel J, Kotol D, Zhong W, Tebani A, Strandberg L, Edfors F, Sjöstedt E, Mulder J, Mardinoglu A, Berling A, Ekblad S, Dannemeyer M, Kanje S, Rockberg J, Lundqvist M, Malm M, Volk AL, Nilsson P, Månberg A, Dodig-Crnkovic T, Pin E, Zwahlen M, Oksvold P, von Feilitzen K, Häussler RS, Hong MG, Lindskog C, Ponten F, Katona B, Vuu J, Lindström E, Nielsen J, Robinson J, Ayoglu B, Mahdessian D, Sullivan D, Thul P, Danielsson F, Stadler C, Lundberg E, Bergström G, Gummesson A, Voldborg BG, Tegel H, Hober S, Forsström B, Schwenk JM, Fagerberg L, Sivertsson Å. The human secretome. *Sci Signal* 2019;**12**.
25. Tucker RP. Teneurins: domain architecture, evolutionary origins, and patterns of expression. *Front Neurosci* 2018;**12**:938.
26. Chawla K, Ham HO, Nguyen T, Messersmith PB. Molecular resurfacing of cartilage with proteoglycan 4. *Acta Biomater* 2010;**6**:3388–3394.
27. Ehara A, Taguchi D, Nakadate K, Ueda S. Attractin deficiency causes metabolic and morphological abnormalities in slow-twitch muscle. *Cell Tissue Res* 2021;**384**:745–756.
28. Bocker J, Schmitz MT, Mittag U, Jordan J, Rittweger J. Between-subject and within-subject variation of muscle atrophy and bone loss in response to experimental bed rest. *Front Physiol* 2022;**12**:743876.
29. Kisla Ekinci RM, Balcı S, Coban F, Bisgin A. Serum lubricin levels in patients with juvenile idiopathic arthritis. *Reumatologia* 2021;**59**:373–377.
30. Yokota H, Leong DJ, Sun HB. Mechanical loading: bone remodeling and cartilage maintenance. *Curr Osteoporos Rep* 2011;**9**:237–242.
31. Cvirn G, Waha JE, Ledinski G, Schlagenhauf A, Leschnik B, Koestenberger M, Tafeit E, Hinghofer-Szalkay H, Goswami N. Bed rest does not induce hypercoagulability. *Eur J Clin Invest* 2015;**45**:63–69.
32. Venemans-Jellema A, Schreijer AJ, le Cessie S, Emmerich J, Rosendaal FR, Cannegieter SC. No effect of isolated long-term supine immobilization or profound prolonged hypoxia on blood coagulation. *J Thromb Haemost* 2014;**12**:902–909.
33. Yanagibori R, Suzuki Y, Kawakubo K, Kondo K, Iwamoto T, Itakura H, Makita Y, Sekiguchi C, Gunji A, Kondou K. The effects of 20 days bed rest on serum lipids and lipoprotein concentrations in healthy young subjects. *J Gravit Physiol* 1997;**4**:S82–S90.
34. Mikines KJ, Richter EA, Dela F, Galbo H. Seven days of bed rest decrease insulin action on glucose uptake in leg and whole body. *J Appl Physiol* 1991;**70**:1245–1254.
35. Bey L, Hamilton MT. Suppression of skeletal muscle lipoprotein lipase activity during physical inactivity: a molecular reason to maintain daily low-intensity activity. *J Physiol* 2003;**551**:673–682.
36. Vollmy F, van den Toorn H, Zenezini Chiozzi R, Zucchetti O, Papi A, Volta CA, et al. A serum proteome signature to predict mortality in severe COVID-19 patients. *Life Sci Alliance* 2021;**4**.
37. Garrett-Bakelman FE, Darshi M, Green SJ, Gur RC, Lin L, Macias BR, McKenna MJ, Meydan C, Mishra T, Nasrini J, Piening BD, Rizzardi LF, Sharma K, Siamwala JH, Taylor L, Vitaterna MH, Afkarian M, Afshinnekoo E, Ahadi S, Ambati A, Arya M, Bezdan D, Callahan CM, Chen S, Choi AMK, Chlipala GE, Contrepois K, Covington M, Crucian BE, de Vivo I, Dinges DF, Ebert DJ, Feinberg JI, Gandara JA, George KA, Goutsias J, Grills GS, Hargens AR, Heer M, Hillary RP, Hoofnagle AN, Hook VYH, Jenkinson G, Jiang P, Keshavarzian A, Laurie SS, Lee-McMullen B, Lumpkins SB, MacKay M, Maienschein-Cline MG, Melnick AM, Moore TM, Nakahira K, Patel HH, Pietrzyk R, Rao V, Saito R, Salins DN, Schilling JM, Sears DD, Sheridan CK, Stenger MB, Tryggvadottir R, Urban AE, Vaisar T, van Espen B, Zhang J, Ziegler MG, Zwart SR, Charles JB, Kundrot CE, Scott GBI, Bailey SM, Basner M, Feinberg AP, Lee SMC, Mason CE, Mignot E, Rana BK, Smith SM, Snyder MP, Turek FW. The NASA twins study: a multidimensional analysis of a year-long human spaceflight. *Science* 2019;**364**.
38. Zhu H, Blum RH, Bjordahl R, Gaidarova S, Rogers P, Lee TT, Abujarour R, Bonello GB, Wu J, Tsai PF, Miller JS, Walcheck B, Valamehr B, Kaufman DS. Pluripotent stem cell-derived NK cells with high-affinity noncleavable CD16a mediate improved antitumor activity. *Blood* 2020;**135**:399–410.
39. Epelbaum R, Shalitin C, Segal R, Valansi C, Arselan I, Faraggi D, Leviov M, Ben-Shahar M, Haim N. Haptoglobin-related protein as a serum marker in malignant lymphoma. *Pathol Oncol Res* 1998;**4**:271–276.
40. Murgia M, Ciciliot S, Nagaraj N, Reggiani C, Schiaffino S, Franchi MV, et al. Signatures of muscle disuse in spaceflight and bed rest revealed by single muscle fiber proteomics. *PNAS Nexus* 2022;**1**.
41. von Haehling S, Morley JE, Coats AJS, Anker SD. Ethical guidelines for publishing in the Journal of Cachexia, Sarcopenia and Muscle: update 2021. *J Cachexia Sarcopenia Muscle* 2021;**12**:2259–2261.

Article 3: AlphaDIA enables End-to-End Transfer Learning for Feature-Free Proteomics

Pre-print published online: bioRxiv (2024), doi: 10.1101/2024.05.28.596182.

Georg Wallmann[1], Patricia Skowronek[1], Vincenth Brennsteiner[1], Mikhail Lebedev[1], Marvin Thielert[1], **Sophia Steigerwald[1]**, Mohamed Kotb[1], Tim Heymann[1], Xie-Xuan Zhou[1], Magnus Schwörer[1], Maximilian T. Strauss[2], Constantin Ammar[1], Sander Willems[1], Wen-Feng Zeng[1]*, Matthias Mann[1,2]*

[1]Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany
[2]Proteomics Program, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
**Corresponding authors*

Data-independent acquisition (DIA) strategies have become increasingly more powerful and popular over the recent years, surpassing the performance of data dependent acquisition (DDA).[94,346,423–425] In contrast to DDA, DIA is not limited to the selection of only the most abundant precursor and therefore allows for a higher dynamic range and depth. However, DIA strategies come with their own set of challenges, mainly the increased spectral complexity caused by co-isolation and co-fragmentation of precursor and peptide ions. This requires more advanced and computationally heavy search algorithms able to deconvolute this data, especially as data acquisition strategies and MS instrumentation become more advanced.

In this study Georg Wallmann, in a collaboration across our bioinformatics and method development team, developed a modular open-source framework for DIA analysis which features a feature-free identification algorithm particularly suitable for data produced on state-of-the art time-of-flight (TOF) analyzers. Building on the scientific python stack and alphaX ecosystem[117], accessible through a number of interfaces, such as python API, command line or GUI, and running on the most common operating systems, AlphaDIA is setting a new standard for accessibility and transparency. Unlike other DIA search engines, which rely on predefined feature boundaries, AlphaDIA's feature-free identification algorithm does not reduce the data and processes the raw MS signal by aggregating all relevant information, such as RT, IM and fragment intensities before the identifications step. This enhances sensitivity, identification accuracy and makes AlphaDIA particularly adept at handling "noisy" TOF data as AlphaDIA's convolution kernels can aggregate evidence across multiple dimensions to confidently identify peptides and precursors even at low fragment intensities. While AlphaDIA can be used with empirical (experimental) libraries, it also features an end-to-end workflow using AlphaPeptDeep as a basis for fully predicted libraries. These predicted libraries can then

be fine-tuned for the specific experimental conditions via transfer learning, boosting identification by 48% and 25% on precursor and protein group level respectively in comparison to the standard models. Whether using empirical or AlphaPeptDeep predicted libraries, AlphaDIA shows competitive or superior performance for identification, quantitative accuracy and FDR in comparison to popular search engines, such as DIA-NN and Spectronaut. This is especially true for high-sensitivity platforms such as the Orbitrap Astral, where AlphaDIA was able to identify more than 120,000 precursors and 9,500 protein groups in a 21 min LCMS acquisition. Moreover, it supports novel and complex acquisition strategies, such as synchro-PASEF, and provides the flexibility to process PTMs, labelled proteomics samples as well as increasingly more complex acquisition strategies as MS instrumentation continues to evolves.

Contribution:
Co-authorship. The study was conceptualized by Georg Wallmann, Wen-Feng Zeng and Matthias Mann. I initially optimized the Orbitrap Astral acquisition methods and gave input on data acquisition for this study. Alongside the other co-authors, I contributed to revising and editing the manuscript.

AlphaDIA enables End-to-End Transfer Learning for Feature-Free Proteomics

Georg Wallmann[1] Patricia Skowronek[1], Vincenth Brennsteiner[1], Mikhail Lebedev[1], Marvin Thielert[1], Sophia Steigerwald[1], Mohamed Kotb[1], Tim Heymann[1], Xie-Xuan Zhou[1], Magnus Schwörer[1], Maximilian T. Strauss[2], Constantin Ammar[1], Sander Willems[1], Wen-Feng Zeng[1]*, Matthias Mann[1,2]*

1 Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany

2 Proteomics Program, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

* Correspondence: mmann@biochem.mpg.de, wzeng@biochem.mpg.de

Abstract

Mass spectrometry (MS)-based proteomics continues to evolve rapidly, opening more and more application areas. The scale of data generated on novel instrumentation and acquisition strategies pose a challenge to bioinformatic analysis. Search engines need to make optimal use of the data for biological discoveries while remaining statistically rigorous, transparent and performant. Here we present alphaDIA, a modular open-source search framework for data independent acquisition (DIA) proteomics. We developed a feature-free identification algorithm particularly suited for detecting patterns in data produced by sensitive time-of-flight instruments. It naturally adapts to novel, more efficient scan modes that are not yet accessible to previous algorithms. Rigorous benchmarking demonstrates competitive identification and quantification performance. While supporting empirical spectral libraries, we propose a new search strategy named end-to-end transfer learning using fully predicted libraries. This entails continuously optimizing a deep neural network for predicting machine and experiment specific properties, enabling the generic DIA analysis of any post-translational modification (PTM). AlphaDIA provides a high performance and accessible framework running locally or in the cloud, opening DIA analysis to the community.

Introduction

Proteomics entails the study of key players of life – proteins – and their translation, composition of isoforms, post-translational modification and degradation[1]. As proteomes are composed of thousands of different proteoforms, which produce hundreds of thousands of peptides in bottom-up proteomics, handling complexity is central to MS based proteomics acquisition and bioinformatic analysis.

Until recently, data dependent acquisition (DDA) was the acquisition method of choice. The direct relationship between selected precursors and relatively pure fragmentation spectra, combined with its mature ecosystem of search engines, results in confident peptide identifications[2–5]. Due to the straightforward relationship between precursor and fragment spectrum, this also holds for challenging cases such as complex patterns of post-translational modifications or the interpretation inter-protein cross-links[6,7]. Yet, selecting only a single peptide at a time comes at the cost of increased data acquisition time and stochastic sampling of precursors across liquid chromatography (LC)-MS runs[8].

In contrast to DDA, Data Independent Acquisition (DIA), allows the selection of multiple peptides in parallel, originally in the form of cycles of fixed-width, relatively wide selection windows[9,10]. This results in systematic sequencing of all available peptides only limited by sensitivity. Importantly, repeated scanning of the same

mass range yields complete elution profiles of both the precursors and the fragments. This increases dynamic range, allows for faster acquisition and deeper proteome characterization down to the single cell level[11,12]. The principal challenge of DIA is the increased spectral complexity as multiple peptides fragment together leading to convoluted spectra. Thus, DIA data by its nature requires algorithms to deconvolute overlapping fragmentation patterns and assign peptide identifications.

Initially, DIA involved generating an empirical, sample specific spectral library, usually acquired by offline fractionation of samples and DDA acquisition, or spectrum centric processing[13,14]. Different algorithms have been designed to process DIA data. Deconvolution of co-isolated peptides into individual spectra effectively reduces them to DDA like data, amenable to the plethora of proven DDA methods. However, peptide-centric approaches, in which each spectrum of the library is matched to the complex DIA data, achieve higher performance especially if paired with deep-learning based scoring of identifications as pioneered by Demichev et al. [15–17]. Deep learning also allows the prediction of libraries in silico, obviating the need for sample specific empirical libraries [18–20]. However, for optimal performance this has so far required DDA data on the same MS platform and experimental method. This is in particular the case for spectra of post-translationally modified peptides[21,22].

Despite the enormous potential of DIA, the fact that spectra are not easily manually interpretable has hindered full acceptance, especially as researchers must generally rely on few closed source algorithms. Flexible and open algorithms would clearly be beneficial to extend the reach, transparency, and acceptance of DIA. This becomes especially necessary as the most recent generation of instrument employs time-of-flight (TOF) detectors which are sensitive down to the single molecule level[23,24]. Raw files easily contain billions of detector events, often with no clearly visible peaks and up to four dimensions (4D) of separation[25]. Handling this data has usually required data reduction such as centroiding of the ion mobility, introducing feature boundaries or centroiding[26,27], which may all lead to loss of information. We have found that this presents formidable challenges when implementing novel scan modes that make data processing even more demanding[28], especially when the underlying algorithms and source code are not available.

To enable open, performant, and extensible processing of high complexity DIA data, we therefore propose a new processing framework which builds on technology driving the current breakthroughs in artificial intelligence, especially deep learning. Our algorithms view a DIA experiment as high-dimensional snapshot of the peptide spectrum space. This representation is amenable to DIA methods on all major instrument platforms and naturally covers simple DIA methods as well as ion mobility, variable windows, sliding quadrupole windows and yet to be developed acquisition modes. Integral to this generalized representation, the data is processed without reduction of retention time or mobility resolution. Instead, our feature-free approach performs machine learning directly on the raw signal, combing all available information before making discrete identifications. Furthermore, we propose an end-to-end deep transfer learning strategy based on our recently published alphaPeptDeep library. Transfer learning adapts the peptide library directly to the instrument and sample workflow. We showcase performance and versatility by extending DIA arbitrary PTMs, closing the gap between the versatility of DDA and the performance of DIA

Results

We present alphaDIA, a modular, open-source, next generation framework for DIA search. It builds on the scientific python stack and the alphaX[29] ecosystem allowing flexible search strategies as well as default workflows accessible through a Python API, Jupyter notebooks, a command line interface or an easily installable graphical user interface (**Fig. 1, a, Methods**). AlphaDIA covers the entire workflow from raw files to reporting protein quantities and can process files and proprietary formats from all major vendors. It was designed for 'one stop processing' of large cohorts and arbitrary data sizes, running natively on Windows, Linux and Mac or in a distributed fashion in the cloud with Slurm or Docker.

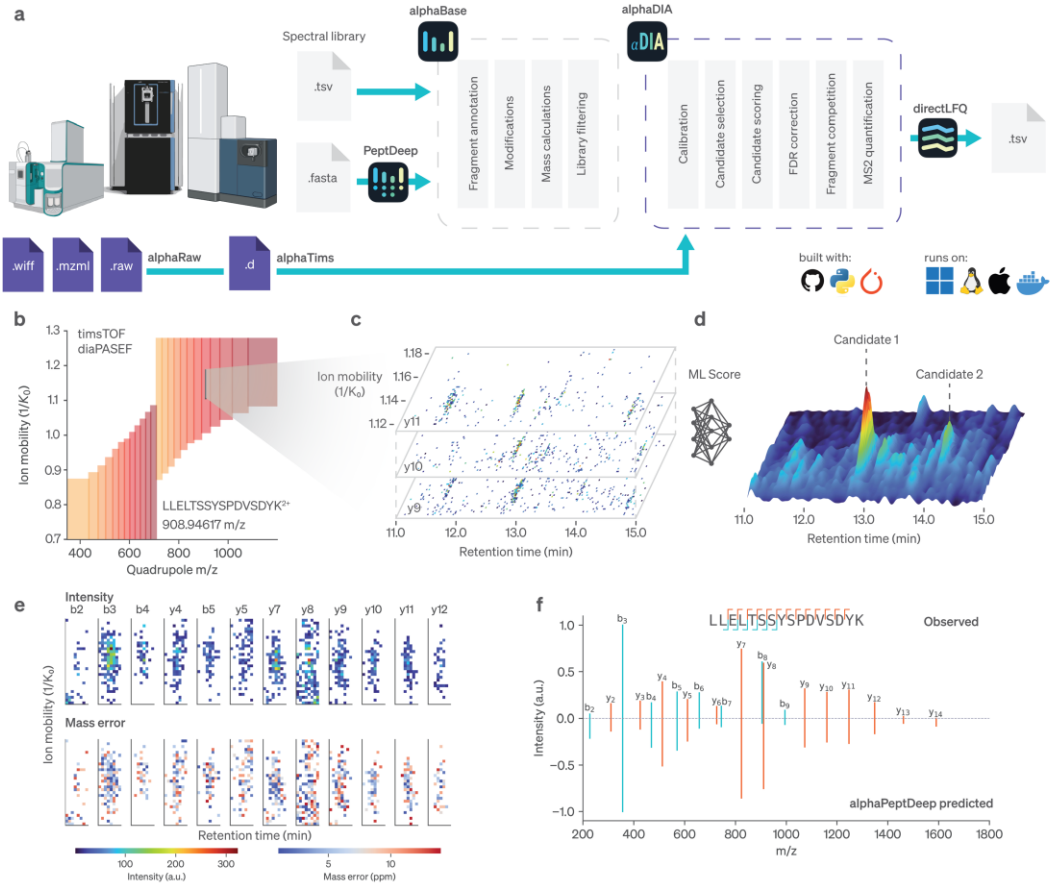


Fig. 1 | Overview of the alphaDIA framework. a, Components of alphaDIA and the integration into the alphaX ecosystem. AlphaDIA uses alphaRaw and alphaTims[30] for accessing raw data from all major vendors. Importing as well as prediction of spectral libraries is facilitated by alphaBase and alphaPeptDeep[20]. After successful search, label free quantification is performed using directLFQ[31]. AlphaDIA uses best software engineering practices and builds on modern open architectures (GitHub, Python, PyTorch) **b-f**, TIMS DIA data acquired using optimal dia-PASEF[32] is searched using a peptide centric algorithm. **b**, The library entry for a single peptide sequence is selected for search **c**, Fragment spectra containing the precursor of interest are extracted and converted into a dense matrix in spectrum space. **d**, Information from fragments mapping to the precursor of interest are combined in a continuous score. **e**, AlphaDIA defines candidate peak groups with discrete integration boundaries (top row: intensities, bottom row: mass deviation from theoretical mass. **f**, Aggregating signal across the integration boundaries in ion mobility and retention time reveals the peptide spectrum. For further scoring, AlphaPeptDeep spectrum predictions are used.

Feature-free processing for high dimensional TOF data

Apart from state-of-the-art DIA processing, the impetus for alphaDIA was the shift towards fast, sensitive but also stochastic TOF detectors, presenting novel algorithmic challenges and opportunities. AlphaDIA's feature-free and peptide-centric search is illustrated by the identification of the peptide LLELTSSYSPDVSDYK^{2+} from timsTOF Ultra dia-PASEF data (**Extended Data Fig. 1**). First, we select all MS1 and MS2 spectra that contribute evidence for this precursor (**Fig. 1 ,b**). A dense representation of the spectrum space is used to score potential peak group candidates, which does not involve feature building or centroiding (**Fig. 1 ,c-d**). Instead, signals are aggregated across retention time, ion mobility and fragments using learned convolution kernels. Only after all this evidence has been collected, discrete peak groups are determined (**Fig. 1, e**). In this way noisy TOF data in which individual fragment signals are not distinguishable from background can still be processed (**Extended Data Fig. 2**). After the signals in the peak groups are integrated it becomes evident that they correspond to a confidently identified peptide, given the agreement with the predicted spectrum (**Fig. 1 ,f**).

Deep learning based search allows for whole proteome characterization

AlphaDIA uses deep learning based target-decoy competition and iterative calibration to search complex proteomes with spectral libraries. For each target precursor entry with a given sequence and charge state, a paired decoy peptide is created using a mutation pattern (**Methods**). Each peak group is scored by a collection of up to 47 features using a fully connected neural network (NN) (**Fig. 2, a**). False precursor identifications are controlled using a count-based FDR, calculated from the probabilities predicted by the NN (**Fig. 2, b-c**). Measured properties like retention time, ion mobility and m/z ratios are iteratively calibrated to the observed data on a high confidence subset of precursors, using non-linear LOESS regression with polynomial basis functions (**Fig. 2, d-f, Extended Data Fig. 3**). AlphaDIA uses spectrum centric fragment competition to ensure that fragment information is only used for a single precursor identification, even when multiple library entries match the same observed signal (**Methods**). On a 21 minute, 60 samples per day (SPD) gradient of HeLa cell lysate measured on a timsTOF Ultra with dia-PASEF, our algorithm identified more than 73,000 precursors with unique sequence and charge, corresponding to almost 6,800 protein groups (**Fig. 2, g-i**). For label free quantification (LFQ) we integrated the recently developed directLFQ algorithm[31], which resulted in a median coefficient of variation of 7.7% for protein groups and a Person R > 0.99 across replicates (**Fig. 2, j-k**). This suggests that alphaDIA can search and quantify complex protein mixtures with excellent depth and quantitative precision.

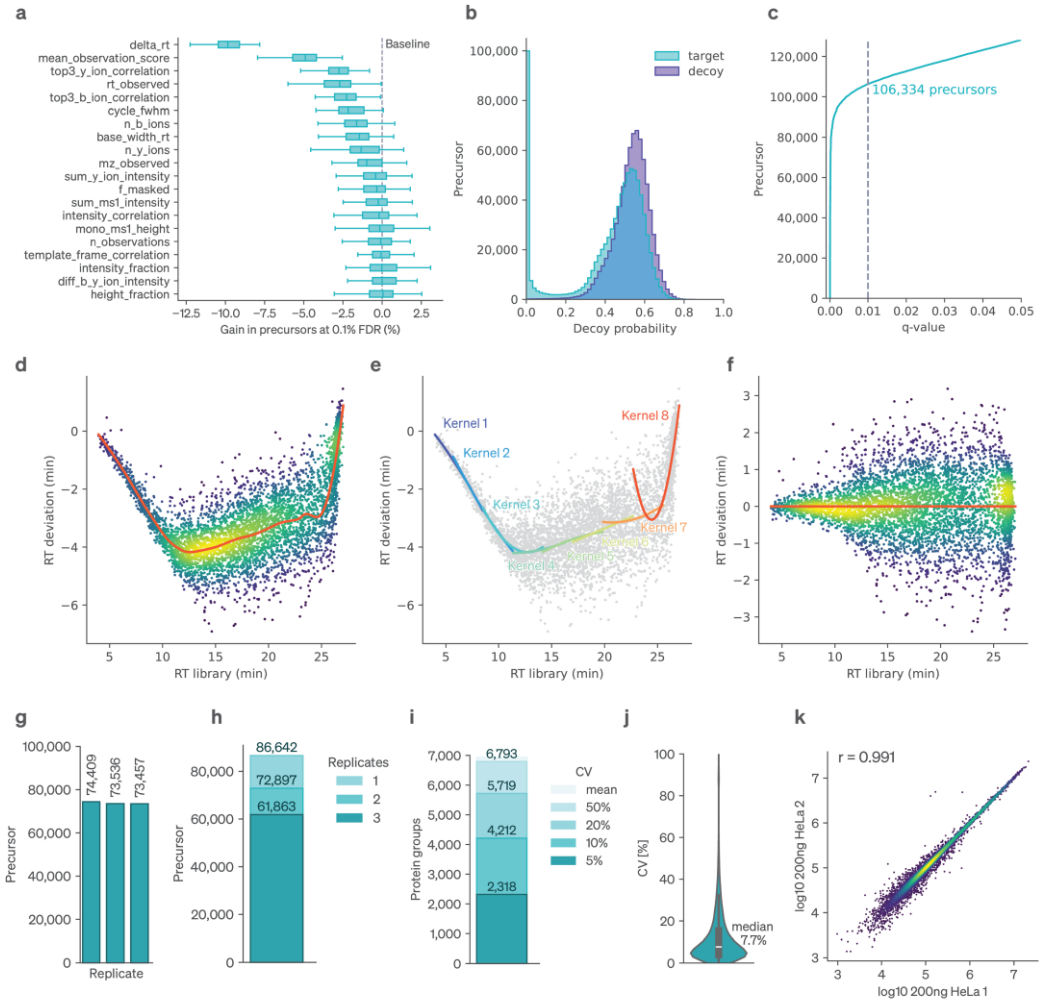


Fig. 2 | Central search engine components. a Classifier features and their importance for the supervised target decoy competition. a, Feature importance is defined as percentage drop of precursor identifications at 0.1% FDR. **b**, Deep neural network output probability for decoy peptides. **c**, Number of precursors identified as a function of the q-value cutoff. **d**, Non-linear calibration of retention times using LOESS regression (**Extended Data Fig. 3** and **Methods**). **e**, Collection of polynomial basis-functions combined using local kernels. **f**, Retention time deviation after calibration. **g-k**, Results for the library-based search of HeLa lysates measured with dia-PASEF. **g**, Number of precursors identified at a 1% FDR in three replicates. **h**, Precursors shared across replicates. **i**, Protein groups identified at given coefficient of variants (CVs). **j**, Distribution of protein group CVs. **k**, Pearson correlation of precursor intensities across samples.

AlphaDIA adapts to different instruments and enables new acquisition methods

Recently, DIA has been coupled to sophisticated data acquisition schemes where the quadrupole isolation window scans nearly continuously through the m/z or m/z and ion mobility space[11,24,27]. The methods, termed synchro-PASEF or midia-PASEF hold the promise of much improved precursor specificity and quantitative accuracy, which, however, has been difficult to realize due to lack of flexible algorithms handling the thousands of individual isolation windows per DIA cycle. AlphaDIA's processing algorithm and alphaRaw's efficient data handling allows to use all synchro scans which contribute signal for a given precursor, considering its isotope distribution as a prior (**Fig. 3, a**). Using the masses and abundance of the precursor isotopes we model the behavior of the quadrupole, resulting in a template with the expected intensity distribution across synchro scan observations (**Fig. 3, b**). This template includes the slicing of the isotope distribution by the quadrupole which must be recapitulated in the intensity profiles of the fragments (**Fig. 3, c**). This comparison of the fragment profile with the template contributes to our deep-learning based identification score and enables analysis of complex proteomes (**Fig. 3, d, Extended Data Fig. 4**). This first processing algorithm for sliding quadrupole data could be extended from synchro-PASEF to similar acquisition schemes such as midia-PASEF or scanning SWATH.

Next, we wanted to extend the reach of alphaDIA to other proteomic platforms and methods. For instance, our algorithms adapted naturally to fixed as well as variable window DIA data from quadrupole Orbitrap analyzers. The absence of ion mobility reduces the search space to a one-dimensional search across retention time while still utilizing all valid MS2 observation for a given precursor (**Fig. 3, e**). As before, after discrete peak group candidates have been identified (**Fig. 3, f**) the spectrum centric view allows detailed scoring utilizing alphaPeptDeep predicted spectra (**Fig. 3, g**). Additionally, alphaDIA can process Orbitrap and Orbitrap Astral data with wide, narrow, variable or overlapping DIA windows. It can likewise process Sciex SWATH data (**Extended Data Fig. 5**).

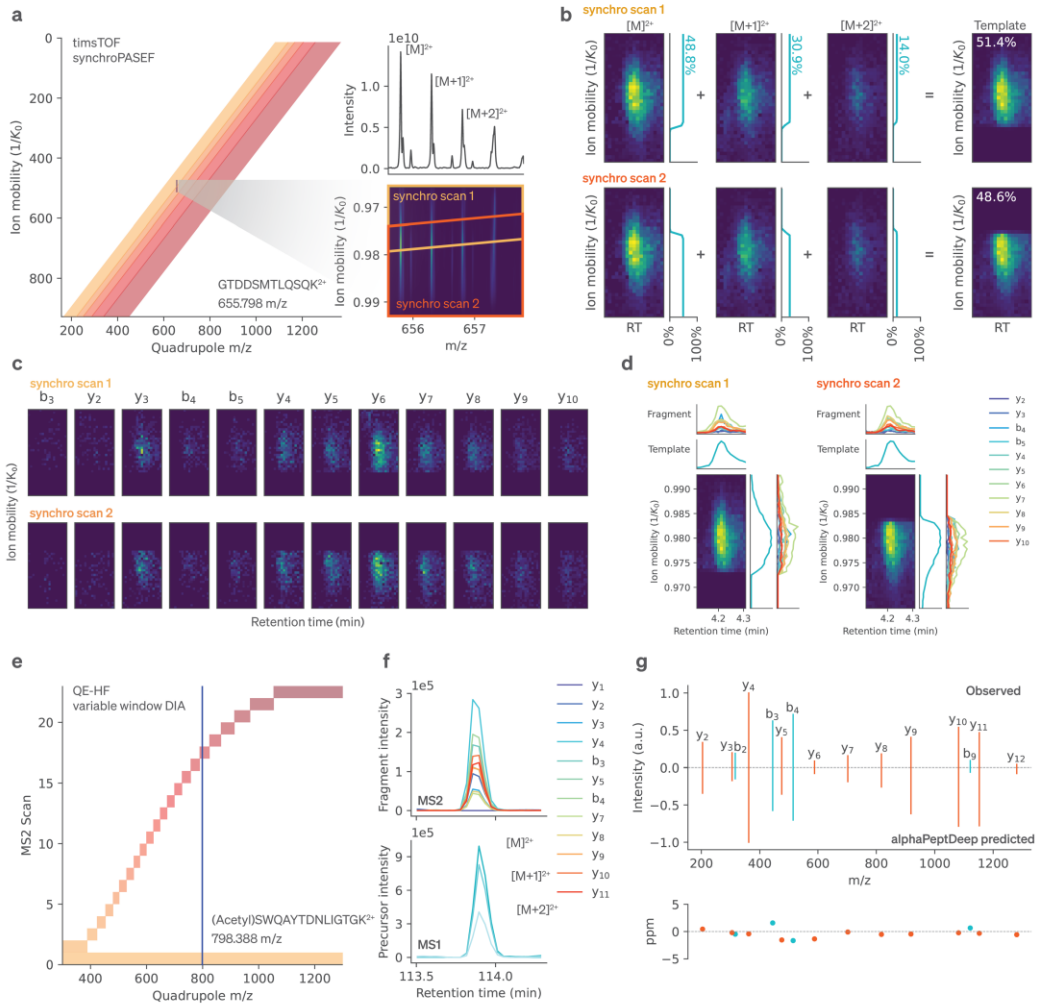


Fig. 3 | AlphaDIA enables flexible processing for different acquisition methods a, Variable window synchro-PASEF acquisition on the timsTOF. The quadrupole mass filter moves as precursors are released from the TIMS trap. The precursor with sequence GTDDSMTLQSQK is sliced by the quadrupole, resulting in fragment signal across two synchro scans. **b**, Slicing patterns are resolved by calculating the expected distribution of fragment signal in form of a template matrix. The template matrix is calculated by transforming the individual precursor isotope signal with the quadrupole transmission function of the synchro scans. **c**, Observed fragment signal across the two synchro scans. **d**, For each of the two synchro scans the elution and ion mobility XICs are compared. Comparison of the fragment signal (rainbow colors) to the template (blue) provides evidence of the identification of peptides. **e**, Application of the processing algorithm to variable window DIA data without ion mobility separation on a quadrupole Orbitrap analyzer (QE-HF). For the given precursor (Acetyl)SWQAYTDNLIGTGK all valid MS2 scans contributing evidence are selected. **f**, Elution profile of MS2 (top) and MS1 (bottom) ions for the precursor of interest. **g**, Observed and predicted fragment intensities after integration of the peak area (top) and mass accuracy for the same precursor (bottom).

AlphaDIA matches or exceeds popular packages in empirical library-based search

Having established the ability of alphaDIA for in-depth analysis of complex proteomes and its adaptability to diverse platforms, we next wanted to directly benchmark its performance against other common DIA search engines. To avoid potential bias, we build upon a recently published benchmarking study from the Shui group, in which mouse brain membrane isolates were spiked into a complex background of yeast proteins in varying ratios and measured on a quadrupole orbitrap (QE-HF) and a timsTOF[33]. The authors generated empirical libraries with MS Fragger[4] and optimized search parameters for DIA-NN, Spectronaut and MaxDIA (**Fig. 4, a**).

Based on the provided libraries alphaDIA identified up to 50,600 mouse peptides in the QE data across all samples and up to 81,500 on the timsTOF (**Extended Data Fig. 6**). Inferring proteins from uniquely identified peptide involves considerations that can influence the number of reported protein groups[34]. AlphaDIA allows strict (maximum parsimony) or commonly used 'heuristic' grouping (**Methods**). With the latter, we identified 5,366 proteins (QE-HF) and 7,649 (timsTOF) protein groups across all samples, matching and even exceeding the other algorithms (**Fig. 4, b-c**). This is also reflected across replicates for single conditions. AlphaDIA quantified the most protein groups in at least 3 out 5 replicates for most ratios while maintaining comparable coefficients of variation (CV) and accuracy as judged by the proteome mixing ratios (**Fig. 4, d, Extended Data Fig. 6-9**).

To prevent over-reporting by sophisticated DIA database searching strategies based on internal target decoy FDR estimates, results can be externally validated by including additional proteome databases from species not present in the sample[35]. As in the benchmarking study, we performed an entrapment search with an Arabidopsis library added in increasing proportions to the target library. On both MS platforms, even for 100% entrapment Arabidopsis identifications matched the chosen target FDR of 1% at the protein level **Fig. 4**, e-f). At this protein FDR, false positive precursors are even less likely appearing only at 0.1% globally. This contrasted with some of the other tested tools, which reported up to three-fold more false positive Arabidopsis identifications than intended at the chosen FDR target (**Extended Data Fig. 8, a-d**). Importantly, the increased library size only minimally decreased overall identifications for alphaDIA (**Extended Data Fig. 8, e-h**). We conclude that for library-based search alphaDIA provides at least competitive performance with common search engines while maintaining a reliable and conservative FDR.

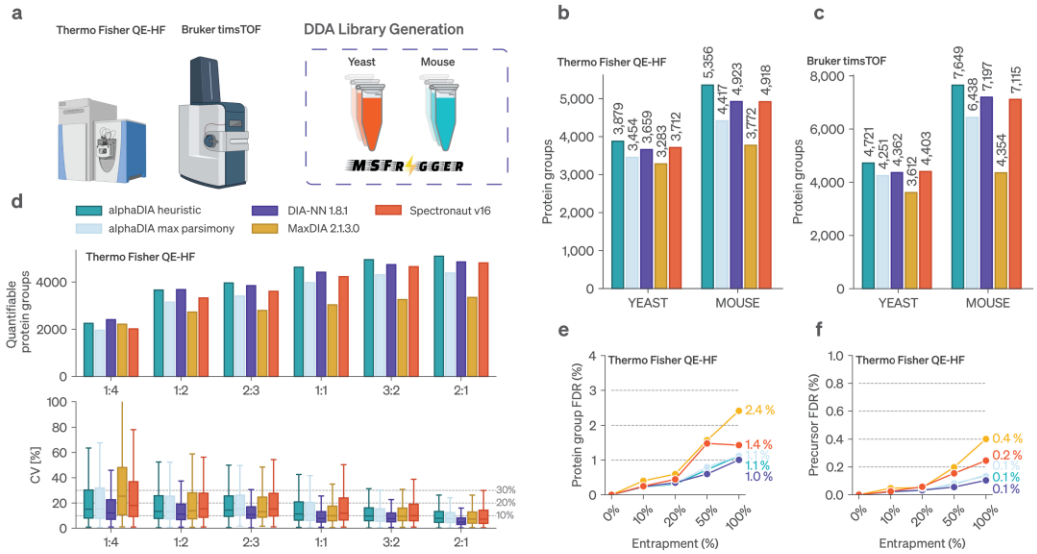


Fig. 4 | Benchmarking alphaDIA against established software for library bases DIA search. a, Overview of the benchmarking dataset[33] for empirical library based search acquired on the quadrupole orbitrap QE-HF platform and the timsTOF. Fractionated bulk samples are analyzed using DDA to generate sample specific libraries using MSFragger. Mouse brain membrane isolates are spiked into a complex yeast background at different ratios and analyzed in 5 replicates using DIA on both platforms. **b**, Number of Mouse protein groups identified at 1% FDR across all replicates on the QE-HF. **c**, Same as b but on the timsTOF platform. **d**, Quantified Mouse protein groups between different spike ins and a reference sample. Proteins we're deemed quantifiable if they were observed in at least 3 out of 5 replicates. The coefficient of variation (CV) is shown for each set of identifications. **e**, Benchmarking of false discoveries using increasing amounts of Arabidopsis entrapments compared to the Yeast / Mouse spectral library. The false discovery rate on the protein level is shown for the QE-HF platform. **f**, Same as e, but on the precursor level.

Combining alphaDIA and alphaPeptDeep allows fast search of fully predicted libraries

While empirical libraries benefit from implicitly capturing instrument and workflow specific properties, the key advantage of deep-learning predicted libraries of the entire proteome database is that it eliminates cumbersome library measurement altogether. We recently introduced alphaPeptDeep, an open source, transformer-based deep learning framework for predicting all MS-relevant peptide properties from their sequences[20].

With these state-of-the art predicted libraries, we devised a two-step search workflow in alphaDIA consisting of library refinement and quantification (**Fig. 5 a**). Furthermore, we reasoned that our feature-free search should adapt well to the high sensitivity TOF data generated by the Orbitrap Astral mass spectrometer. For benchmarking, we acquired and searched bulk Hela samples with an alphaPeptDeep predicted library containing 3.6 million tryptic precursors. AlphaDIA identified on average more than 120.000 precursors, matching or exceeding the performance of all other tested search engines (**Fig. 5 b**). Remarkably, in this 60 SPD method (21 min) this corresponded to the identification of 9,500 protein groups of which 8,200 had a CV less than 20% (**Fig. 5 d**). The great depth of proteome characterization was also reflected in the data completeness across replicates (**Extended Data Fig. 10**). Search times stayed below the rapid acquisition time (**Fig. 5 e**). We validated the FDR control of this more complex two step workflow using the entire Arabidopsis library, which externally confirmed rigorous control of false positive identifications (1.08% at protein level and 0.2% at precursor level, **Fig. 5, f**).

To compare identified proteins across search engines, we mapped peptide sequences to the UNIPROT reference proteome, discarding ambiguous peptides mapping to multiple proteins. Reassuringly, more than 70,000 peptides and close to 8,000 proteins were jointly identified by all tested tools (**Fig. 5 g**). AlphaDIA had the highest number of uniquely identified peptides among search engines, manifesting in higher sequence coverage (median of 8 peptides per protein, **Fig. 5 h**).

To assess the accuracy of label-free quantification (LFQ), we used the established strategy[36] of three species proteomes mixed in defined ratios, acquired on the Orbitrap Astral. Fully predicted library search combined with directLFQ recapitulated the expected ratios with excellent precision and accuracy (**Fig. 5 i, Extended Data Fig. 11**).

Multiplexed DIA has recently shown great potential to increase throughput and depth[37,38]. To analyze such data, identifications must be transferred between the channels which involves an additional channel FDR. Due to the modular nature of alphaDIA this functionality was readily incorporated. We benchmarked it on a DIA dataset in which HeLa cells were heavy and light SILAC labeled and analyzed on a QE-HFX[39] (**Extended Data Fig. 12**). In proportions of identifications in 'light only', 'heavy only' and 'light and heavy' were very similar to the previous DDA and DIA results, validating our channel FDR. Interestingly, on the same data the absolute number of identified peptides was threefold higher than in the original paper, reflecting advances in DIA search over the last years in general, and specifically in alphaDIA.

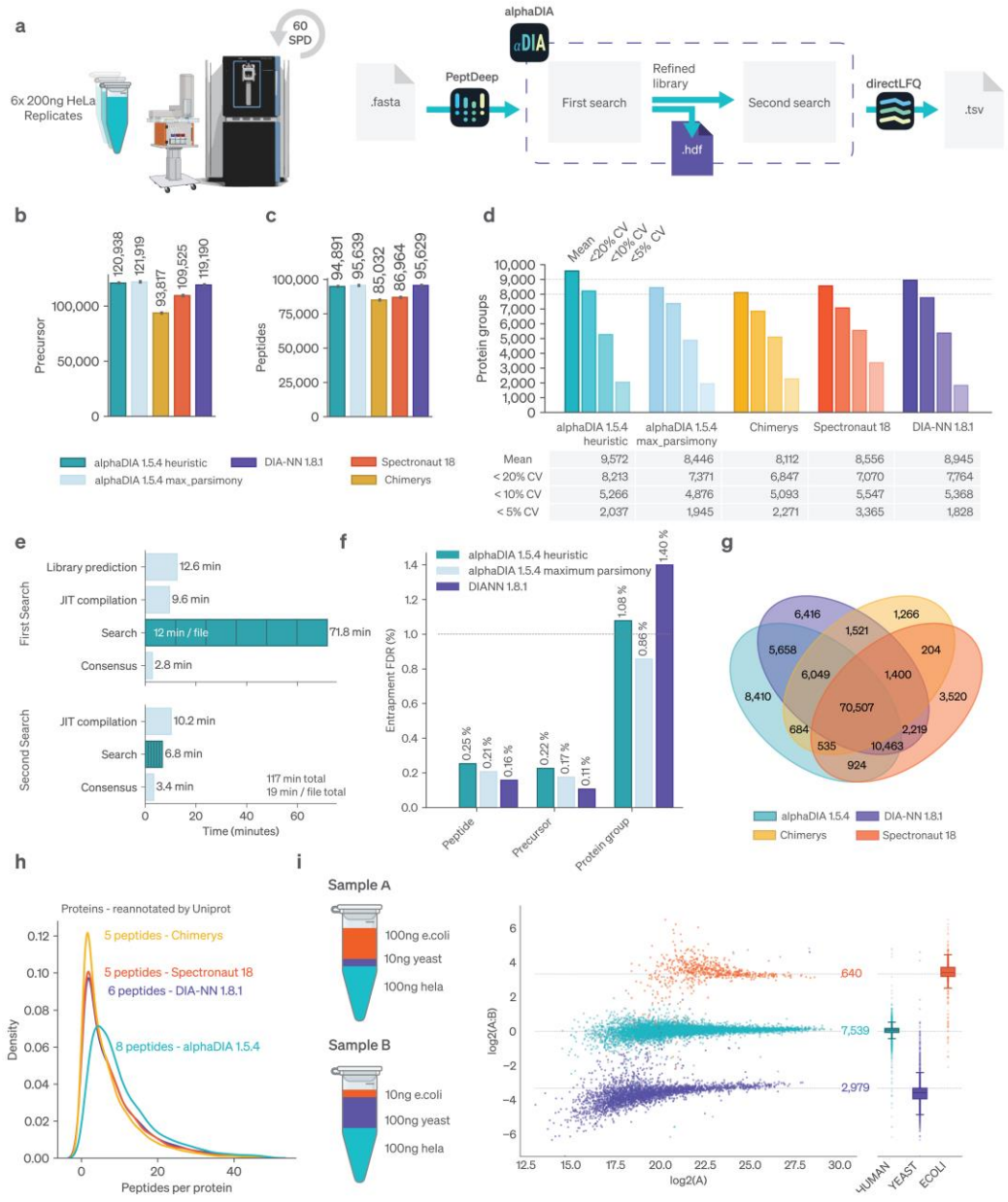


Fig. 5 | Searching complex proteomes acquired on the Orbitrap Astral with fully predicted spectral libraries. a, Six replicates of 200ng HeLa bulk data were analyzed on the Orbitrap Astral with a 60SPD (21 min) gradient. A fully predicted alphaPeptDeep library was used for a two-step search in alphaDIA. Different search engines were used for comparison. **b**, Mean precursors identified across search engines **c**, Mean protein groups identified across processing methods **d**, Protein groups identified at given coefficients of variation (CV) cutoffs. **e**, Analysis time for different processing steps when analyzed with on a 32 core machine. **f**, Arabidopsis entrapment search using the fully predicted library workflow. The share of identified Arabidopsis proteins at 1% target decoy FDR is shown. **g**, Venn diagram showing the overlap of proteotypic peptides across processing methods. **h**, Analysis of protein overlap between different processing methods. Peptides were mapped back to the same reference proteome, discarding ambiguous matches. Number of peptides identified per protein. The median number of peptides per protein is shown. **i**, Mixed

species experiment for establishing quantitative accuracy. Human, Yeast and E.coli proteomes were combined in defined ratios. Plotting the ratio between species-unique protein groups recapitulates the expected ratio (dashed lines).

DIA transfer learning generalizes DIA search to unseen modifications

To date, fully predicted libraries address many of the needs of DIA workflows but their pretrained prediction models are still best suited to the sample and instrument types that were used in training. This makes it necessary to train custom models for different situations - for example PTMs, as they generally change retention and fragmentation behavior compared to the unmodified peptide. We reasoned that close integration of prediction by deep learning and the search engine might have the potential learn to adapt to such differences, an approach that we call *end-to-end transfer learning*. Following search with alphaDIA confidently identified precursors and their spectra are first collected into a training data set. The general pretrained models for retention time, fragmentation spectra and charge state provided with alphaPeptDeep are then finetuned using transfer learning on the experiment specific training data set (**Fig. 6**, a, b). This results in a custom model, reflecting the behavior of peptides on the individual LCMS setup. A held-out test data set ensures generalization and prevents overfitting.

To assess the potential of this end-to-end transfer learning concept, we first applied it to a dataset of dimethylated HeLa peptides, an example of a modification that is known to alter retention times and fragmentation behavior (**Methods, Fig. 6, c**). We found that transfer learning accurately modeled the effects of the lysine and N-terminal dimethylation on retention time behavior, improving R^2 from 0.69 to 0.99 (**Fig. 6**, d-i).

Using the transfer learned model resulted in a total of 96,000 unique precursor and 8,613 protein identifications, a 48% increase over the 65,000 precursors identified without transfer learning and a 25% increase in protein groups (**Fig. 6, d,e**; **Extended Data Fig. 14**). This gain in identifications is driven additively by both improved predictions of retention times from a median prediction error of 317 s down to only 11 s and an increase in the median correlation to predicted spectra from 0.5 to 0.85 (**Fig. 6, g,h**).

Given these drastic improvements, we wished to ascertain that they were not the result of overfitting, despite the use of a holdout test dataset. Similarly to before, we used entrapment with the Arabidopsis proteome library followed by transfer learning with all precursors, including false positive Arabidopsis hits (**Extended Data Fig. 13,a**). Remarkably, even successive rounds of transfer learning led to more confident precursors identifications and less than 0.5% false Arabidopsis identifications at 1% FDR (**Extended Data Fig. 13, b-d**). Upon inspection, we found that predictions of target hits showed substantial improved agreement with observed data, whereas the opposite was true of false positive Arabidopsis hits (**Extended Data Fig. 13, e-g**). This implies that end to end transfer learning generalizes to the peptide behavior in the actual experiment improving identifications and control of false discoveries at the same time.

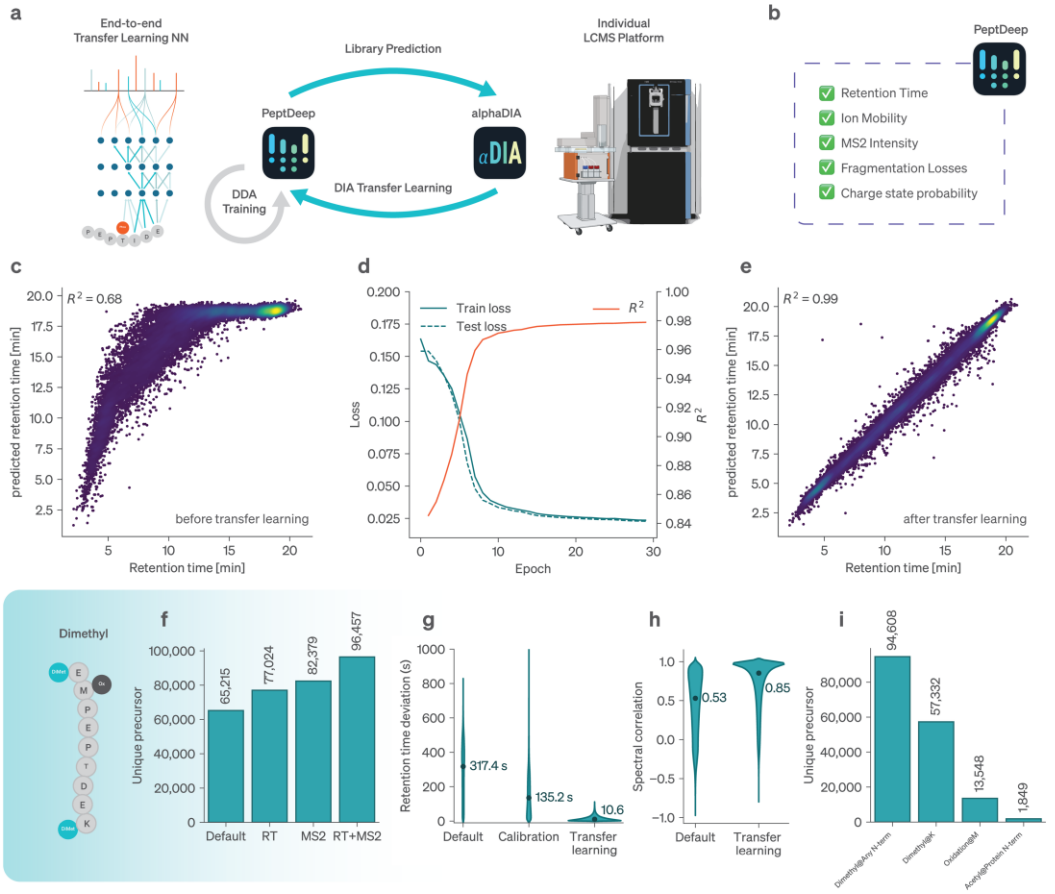


Fig. 6 | DIA transfer learning for discovery of modified peptides a, A custom deep learning model is trained for every experiment using the identifications from the DIA search engine. **b**, Multiple properties are being optimized resulting in smaller and better matching spectral libraries. **c**, Observed and predicted retention times for dimethylated precursors before transfer learning. **d**, DIA transfer learning for the retention times of dimethylated peptides. During training by stochastic gradient descent, a 20% test set of precursors is held out to mitigate overfitting and ensure generalization to the peptide space of interest. **e**, Retention times after transfer learning. **f**, Comparison of the number of unique peptides identified with the pretrained base model (Default) to the transfer learned model after RT and MS2 transfer learning. **g**, Distribution of absolute retention time errors for the pre trained base model (Default), the non-linear calibration within alphaDIA and after transfer learning. **h**, Comparison of spectral correlation before and after MS2 transfer learning. **i**, Number of unique observed modifications by type.

Discussion

The development of alphaDIA addresses several critical challenges inherent to DIA, such as the complexity of spectral data and the need for robust, adaptable algorithms capable of handling high-dimensional data from advanced instrumentation. Our results demonstrate that already the first public version of alphaDIA matches and in many cases surpasses existing software tools in terms of performance and versatility, making it a valuable addition to the proteomics toolkit.

AlphaDIA's feature-free processing method is central to its performance and flexibility. Traditional DIA processing methods often rely on predefined feature boundaries, which can lead to information loss, especially with the high sensitivity and stochastic nature of TOF detectors. By contrast, alphaDIA's approach aggregates signals across multiple dimensions, ensuring that all relevant data is utilized before making discrete identifications. This results in higher accuracy and sensitivity, as evidenced by our ability to confidently identify peptides even in noisy datasets. Additionally, alphaDIA extends the reach of DIA to novel acquisition modes. Together with its open-source architecture this enables the community to quickly loop between experimental innovations and their algorithmic implementation.

Our benchmarking against established tools using both empirical and predicted libraries showcases alphaDIA's equal or superior performance. This holds true across platforms and experimental designs including the Orbitrap Astral, where alphaDIA identified over 120,000 precursors and 9,500 protein groups in a 60 SPD format.

One of the most innovative and promising aspects of alphaDIA is its end-to-end transfer learning capability. Based on integration with the transformer models of alphaPeptDeep, alphaDIA closes the loop between spectral library prediction and DIA search. Our approach allows the model to adapt to experiment-specific conditions, enhancing the accuracy of peptide identifications. We showcased this on a dataset of dimethylated HeLa peptides demonstrating dramatic improvements in retention time prediction and spectral correlation, resulting in a 48% increase in unique precursor identifications and a 25% increase in protein groups compared to using pretrained models alone. This allows the application of DIA search to hitherto inaccessible areas such as post-translationally modified proteins without PTM specific pretraining or to the better identification of HLA peptides. Importantly we demonstrated that transfer learning not only improves overall identifications but even improves FDR control, ensuring reliable results.

The advancements presented by alphaDIA pave the way for more comprehensive and accurate proteomic analyses which will be important as MS technology continues to evolve. This will be especially important in clinical and translational research, where ever increasing cohorts and data require large scale, distributed processing.

The framework's open-source nature ensures that it can be continuously improved and extended by the scientific community, fostering innovation and collaboration. We therefore aim to establish alphaDIA as a cornerstone for the next generation of DIA analysis, closely coupled to the developments in artificial intelligence.

Acknowledgments

We thank Mann Labs members and Isabell Bludau for insightful discussions. This work is funded by the Max Planck Society for the Advancement of Science, and by the Bavarian State Ministry of Health and Care through the research project DigiMed Bayern (www.digimed-bayern.de). It was supported by European Union's Horizon 2020 research and innovation program under grant agreement No. 874839 (ISLET).

Potential conflicts of interest

MM is an indirect investor in Evosep.

Contributions

Conceptualization: G.W., WF.Z and M.M. Bioinformatic method development G.W., M.L., V.B., M.K., C.A., WF.Z. Architecture of ecosystem algorithms & software WF.Z., C.A., G.W., M.K., M. Sch., M.St. S.W. Proteomics method development and data acquisition T.H., P.S., M.T., S.S., Writing - original draft: G.W. and M.M. Writing - review and editing: all authors; Resources: all authors. Supervision: M.M.; Funding acquisition: M.M.

Code Availability

All code presented herein as part of alphaDIA is free software accessible under the permissive Apache license. **AlphaDIA** can be found at www.github.com/MannLabs/alphadia, **alphaRaw** can be found at www.github.com/MannLabs/alpharaw , **alphaBase** is found at www.github.com/MannLabs/alphabase.

Data Availability

All data will be made available upon publication of the manuscript.

Methods

Calibration of retention time, ion mobility and m/z

During search retention time, ion mobility, precursor m/z and fragment m/z are calibrated to the measured values. Starting with initial default settings of 15ppm MS1 and MS2 tolerance, 300 seconds rt tolerance and 0.04 mobility tolerance the library is iteratively calibrated within a minimum of three epochs. Every epoch, batches of precursors are searched and scored with an exponential batch plan (2000, 4000, 8000, etc.) until a minimum number of precursors has been identified at 1% FDR. The number of target precursors is increase with every epoch (default: 200 precursors/epoch). If one epoch has accumulated enough confident target precursors, they are calibrated to the measured values using locally estimated scatterplot smoothing (LOESS) regression. For calibration of fragment m/z values, up to 5000 (but at least 500) of the best fragments according to their XIC correlation are used. Following a single calibration pass, all tolerances are updated to the 95 percentile error after calibration but not below the chosen target level.

LOESS regression using uniformly distributed kernels is used for each property which should be calibrated (**Extended Data Fig. 3**). Regression is performed on first and second degree polynomials basis functions of the calibratable property. For m/z and ion mobility, two local estimators with tricubic kernels are used. For retention time prediction, six estimators with tricubic kernels are used. The architecture is built on the scikit-learn package and can be configured to use different hyperparameters and arbitrary predictors for calibration.

Scoring of precursors and decoys using convolution kernels and supervised classification

AlphaDIA employs a two-step scoring machine learning algorithm to identify the best potential peak group for every library entry. The first step builds on a collection of weighted convolution kernels, learned during optimization and calibration of the spectral library. For every precursor of interest, MS1 scans and MS2 scans contributing information towards the identification are identified from the DIA cycle pattern of the acquisition method. Based on a certain number of highest intensity fragments in the library (default: 12), dense representations of the search space in ion mobility and retention time dimension are assembled. To identify putative peak groups for each precursor, a set of convolution kernels, reflecting the expected distribution in retention time, ion mobility and fragment intensity are learned during calibration and optimization. The convolution of the search space is performed in Fourier space for fast processing, and a single score is calculated as log sum across kernels and fragments. Local maxima are identified using a simple peak picking algorithm and retention time and ion mobility boundaries of the peak group of interest are defined from the joint scoring function. These candidates are subsequently rescored for FDR estimation.

As second step, AlphaDIA uses target decoy competition for scoring the quality of precursor spectrum matches. Upon library import, paired known false positive decoy peptides are created for every target. By default, a mutation pattern GAVLIFMPWSCTYHKRQENDBJOUXZ => LLLVVLLLLTSSSSLLNDQEVVVVVV is used. For every library entry, target and decoy, the best high scoring matches from the convolution kernel score are used for supervised classification. Up to 47 features are calculated for each peak-group match, reflecting the merit of the identification. A multi-layer perceptron (MLP) deep neural network with layer sizes 100, 50, 20, 5 and 47 input dimensions (10,810 parameters) is trained to predict the probability of being a false decoy identification. Training is performed with stochastic gradient descent for 10 epochs with a batch size of 5000 and learning rate of 0.001. While training on an 80% training set a 20% test set is held-out to mitigate overfitting. Based on the final score, the best (lowest) decoy probability peak group is retained for every library entry and a count based FDR is calculated.

False discovery rate calculation

AlphaDIA uses a count based FDR on the level for assigning confidence to precursor, peptide, protein and channels. Identifications are given as a set of target and decoy identifications $P = \{p_0, p_1, \dots p_i\}$ all associated with a ground truth decoy status $decoy$: $P \rightarrow \{true, false\}$ and a deep-learning derived decoy score $\hat{y}$: $P \rightarrow \mathbb{R}$. For every precursor with index i the number of targets with lower or equal decoy probability

$$n_{target} = |\{ p \mid \hat{y}(p) \leq \hat{y}(p_i), decoy(p) = false\}|$$

and the number of decoys with lower or equal decoy probability

$$n_{decoy} = |\{ p \mid \hat{y}(p) \leq \hat{y}(p_i), decoy(p) = true\}|$$

are calculated. Furthermore, the total number of targets and decoys in the set are calculated as:

$$N_{target} = |\{ p \mid decoy(p) = false\}|$$

$$N_{decoy} = |\{ p \mid decoy(p) = true\}|$$

The local count-based q value is given as:

$$q_i = \frac{n_{decoy}}{n_{target}} \times \frac{N_{target}}{N_{decoy}}$$

This is converted to a false discovery rate (FDR) by using the minimum q-value where a precursor was accepted:

$$FDR_i = min(q_i, \{q \mid \hat{y}(p) > \hat{y}(p_i)\})$$

By default, all identifications are filtered on a run-level 1% FDR precursor threshold and global 1% protein group-level threshold.

Spectrum centric fragment competition

Competition of precursors for fragment ion is used as spectrum centric element to mitigate double use of fragments for multiple identifications from the same spectra. Following initial FDR calculation, precursor candidates are filtered at 5% FDR and split into groups of potentially fragment sharing. This is determined by the quadrupole cycle pattern. Then, precursor candidates and their elution width at half maximum are compared so that precursors with overlapping elution width at half maximum have no more than $k_{max} = 1$ shared fragment masses within the chosen MS2 mass accuracy δ_{MS2}. If two or more precursor candidates share more fragments than permitted the precursor candidate with the lowest decoy score is used.

Protein inference

Reporting all proteins whose sequence can be matched to any identified peptide can lead to drastic inflation of false discoveries on the protein level[40]. Following the approach outlined by Nesvizhskii et al. [41], we consider a precursor as a single piece of evidence, and the task of protein inference is then to assemble these precursors into proteins while controlling the accumulation of spurious protein identifications. AlphaDIA aims to implement a simple and transparent inference approach, allowing for three inference modes: library, maximum_parsimony and heuristic. Apart from the library mode which uses the inference performed during empirical library creation, protein inference is based on an implementation of the “greedy set cover” algorithm with grouping by default (heuristic) and without grouping for strict inference (maximum_parsimony).

In brief, alphaDIA’s protein inference starts with a table of identified precursors. Each precursor is associated with a set of genes and proteins and based on user choice, the inference is performed on the gene or protein

level (default: gene). While a common peptide precursor may match many proteins, a proteotypic peptide will match one single protein. During grouping, the precursor and protein arrays are reshaped into a protein-centric view, where each protein is associated with one set of precursors. Then, proteins are sorted by the length of their precursor set in descending order, and the protein with the largest number of precursors removed from the lists as the first query. The query is compared to all remaining subject proteins. From each subject precursor set, all precursors matching the query set are removed. If a protein's precursor set becomes empty, it is considered redundant and dropped. After all precursor sets have been compared, the process repeats by reordering the list and extracting the next query. After completion, retained queries are denoted master proteins, necessary to explain all discovered precursors. In strict maximum_parsimony mode all master proteins are simply reshaped to precursor-centric format, linking each precursor to one single protein ID. In the heuristic mode, the list of master proteins is used to remove all non-master proteins from the initial precursor table, effectively leaving each precursor with a set of associated proteins comprised solely of master proteins. Thereby, the same precursor can be claimed by different proteins, creating protein groups (see also the tutorial notebook in the GitHub repository).

Protein FDR

Protein FDR is performed on the protein groups (PGs) calculated during protein inference. For all target and decoy protein groups, 7 features are calculated: the total number of precursors across runs for the PG; the mean decoy score for precursors across runs for the PG; the number of unique peptides for the PG; the number of unique precursors for the PG; the number of runs the PG was found in; the lowest decoy score across precursors for the PG; the highest decoy score across precursors for the PG. We use a multi-layer-perceptron (MLP) to classify decoy PGs from target PGs. Correct training is ensured by a 20% held-out test set. PG FDRs are calculated on a global level using the FDR mechanism described just above.

Library refinement for fully predicted libraries

AlphaDIA uses an established two step-search strategy for library refinement[15]. Following an initial search of all or a subset of raw files, protein inference and FDR is performed as configured by the user. All precursors are automatically filtered at 1% local precursor FDR and global 1% protein group FDR and accumulated into a spectral library and finally saved to the project folder. For each precursor, the identification with the best (lowest) decoy probability is used. By default, MS2 quantities are used as annotated in the original library. If transfer learning accumulation is used, custom user specified fragment types can be selected and observed MS2 intensities are extracted. This spectral library is then used for the second search with full MS2-based target decoy scoring without any relaxed FDR parameters. For protein inference and FDR, library annotated protein groups are used.

Transfer learning

To create transfer learning libraries, precursors identified at 1% precursor and protein FDR are selected for requantification. Precursors are requantified for user defined fragment ion types (a, b, c, x, y, z, modification loss, etc.) and a user-defined maximum charge (default: 2). Extracted fragment quantities are accumulated across samples and ordered by their decoy probability. For each unique modified precursor, the observations with the three lowest decoy scores are selected. AlphaDIA also creates a high quality subset where only precursors with a median fragment correlation greater than 0.5 are included. For these precursors we only retain fragments whose correlation values exceed 75% of the median fragment correlation of the respective precursor. The implementation of transfer learning library is globally sequential. At any given time, we can limit the implementation to only parallelize across a limited number of processes. This approach allows the process to scale without storing all runs in memory.

For transfer learning, we prioritized robustness to ensure performance instead of requiring users to define hyperparameters. The transfer learning dataset is split into a training (80%) and test set (20%) and trained for a maximum of 50 epochs. After each training epoch, we run a test epoch for assessing the test loss and data specific test metrics. AlphaDIA uses a custom learning rate scheduler with two phases. The first phase is a warm-up period (default 5 epochs) during which the learning rate gradually increases to a maximum value (default: 0.005). After this warm-up phase the learning rate scheduler halves the learning rate if the training loss does not significantly improve (default: >5% test loss) within a patience period (default: 3 epochs). Additionally, we use a simple early stopping mechanism that interrupts training if the validation loss starts to diverge or does not significantly improve (default: 12 epochs).

After training, the deep learning model is stored on disk, and can be loaded as necessary. Retention time and ion mobility finetuning are supervised by calculating the L1 loss, R2, 95th percentile of the absolute error on the training data. MS2 finetuning is supervised by calculating the L1 loss, Pearson correlation coefficient, spectral angle, Spearman correlation on the test data. Charge finetuning is supervised by calculating the cross entropy loss, accuracy, precision, recall on the test data. All training and test metrics are reported to the user. The specific implementation and details of the test metrics can be found in the open-source code on GitHub (see **Code Availability**).

Sample preparation of HeLa bulk digests

HeLa S3 cells (ATCC) were cultured in Dulbecco's modified Eagle's medium (Life Technologies Ltd) supplemented with 20 mM glutamine, 10% fetal bovine serum, and 1% penicillin-streptomycin. After washing the cells in PBS and cell lysis, the proteins were reduced, alkylated, and digested by trypsin (Sigma-Aldrich) and LysC (WAKO) (1:100, enzyme/protein, w/w) in one step. The peptides were dried, resuspended in 0.1% TFA/2% acetonitrile (ACN), and 200 ng digest was loaded onto Evotips (Evosep). The Evotips were prepared by activation with 1-propanol, washed with 0.1% formic acid (FA)/99.9% ACN, and equilibrated with 0.1% FA. After loading the samples, tips were washed once with 0.1% FA.

Sample preparation of dimethylated peptides for transfer learning

HeLa cells were cultured as describe above. A HeLa cell pellet was lysed by boiling for 10 min in 1 % SDC in 60 mM TEAB pH 8.5, followed by sonication in a Branson type instrument, Heinemann Sonifier 250 (Schwäbisch Gmünd), operating at 20% duty cycle and 3-4 output for 1 min, and boiling for 5 min again. After cooling down to room temperature, the protein concentration was determined using the tryptophan fluorescence based, WF-assay in the microtiter plate format using white Nunc 96-well plates with a flat bottom (Thermo Fisher Scientific, 136101). After diluting the lysate to 1 ug/uL in lysis buffer, disulfide bonds were reduced by adding Tris(2-carboxyethyl)phosphine (TCEP) to a final concentration of 10 mM TCEP and briefly incubating for 10 min. Denatured protein lysate was digested by Arg-C Ultra (Promega) and Lys-C (WAKO) at a 1:250 and 1:100 (enzyme/protein) ratio to the lysate at 37°C for 3 h, respectively. The peptides were labeled with a dimethyl group by using a 100 uL of 1 ug/uL digested peptides and adding 4 uL of 4 % formaldehyde and 4 uL of a 0.6 M NaBH3CN solution. The mixture was incubated at room temperature and every 10 minutes 2.8 uL (2 ug peptides) were sampled until 60 minutes and added to 17.2 uL of a 1 % solution of trifluoro acetic acid to quench the reaction.

Sample preparation for the mixed species experiments

For the mixed species experiment, three different mixtures with varying mixing ratios of HeLa tryptic digest (Pierce #1862824), S. cerevisiae tryptic digest (Promega V746A), and E. coli tryptic digest (Waters #186003196) were prepared: Sample A (10:1:10 Human(H):Yeast(Y):E. coli(E)), Sample B (10:10:1 H:Y:E), and Sample C (10:4:7 H:Y:E). Five replicates containing 210 ng were loaded per condition.

Peptide loading onto C-18 tips

C-18 tips (Evotip Pure, Evosep) were loaded with the Bravo robot (Agilent), by activation with 1-propanol, washing two times with 50 µl buffer B (99.9% ACN, 0.1% FA), activation with 1-propanol and two wash steps with 50 µl buffer A (99.9% H2O, 0.1% FA). In between, Evotips were spun at 700 g for 1 min. For sample loading, Evotips were prepared with 70 µl buffer A and a short spin at 700 g. Samples were loaded in 20 µl with the indicated concentration into the remaining buffer A and spun at 700 g for 1 min, if not described differently. After sample loading, Evotips were washed with 50 µl buffer A and stored with 150 µl buffer A after a short spin at 700 g at 4 °C until MS acquisition.

MS data acquisition of dia-PASEF and synchro-PASEF data

We used the Evosep One liquid chromatography system to separate peptide mixtures at varying throughputs using standardized gradients. These gradients consisted of 0.1% formic acid (FA) and 99.9% water (v/v), and 0.1% FA with 99.9% acetonitrile (v/v) as mobile phases. For the 60 SPD runs, peptides were separated on a Pepsep column (8 cm x 150 µm ID, 1.5 µm C18, Bruker Daltoniks) connected to a 10 µm ID fused silica emitter (Bruker Daltoniks). For the whisper40 SPD runs, we utilized an Aurora Elite nanoflow column (15 cm x 75 µm ID, 1.7 µm C18, IonOpticks).

The system was coupled with a timsTOF mass spectrometer (Bruker Daltoniks) to acquire data in dia-PASEF and synchro-PASEF modes. Sample loads above 25 ng were analyzed using a timsTOF Pro2, and those below 25 ng with a timsTOF Ultra. The dia-PASEF and synchro-PASEF methods were optimized using our Python tool, py_diAID[32]. This tool maximizes precursor coverage by optimally positioning the acquisition scheme over the precursor cloud and enhances sampling efficiency by adjusting the isolation window widths according to precursor density.

The dia-PASEF method covers an m/z range from 300 to 1200 with eight dia-PASEF scans and two isolation window positions per scan (cycle time 0.98 s). The synchro-PASEF method covers an m/z range from 140 to 1350 with four diagonal synchro scans (cycle time 0.53 s). The method files are deposited in the data repository. In both modes, the fragment scans were acquired with an m/z range from 100 to 1700. Furthermore, ions were accumulated and ejected at 100 ms intervals from the TIMS tunnel. The methods cover an ion mobility range from 1.3 to 0.7 V cm^−2, calibrated with Agilent ESI Tuning Mix ions (m/z, $1/K_0$: 622.02, 0.98 V cm^−2; 922.01, 1.19 V cm^−2; 1221.99, 1.38 V cm^−2). The collision energy was linearly decreased in relation to the ion mobility elution: from 59 eV at an ion mobility of 1.6 Vs cm^−2 to 20 eV at 0.6 V cm^−2.

MS data acquisition of SWATH data on the ScieX 7600

Triplicates of 200ng HeLa bulk digest were loaded onto C-18 tips as described above and analysed using an Evosep One system (Evosep) coupled to a 7600 ZenoTOF mass spectrometer (Sciex) using Sciex OS (version 3.3 or higher). Peptides were separated by the 60 SPD method gradient (Evosep) on a PepSep 8cm x 150 µm reverse-phase column packed with 1.5 µm C18-beads (Bruker Daltonics) at 50 °C connected to the low micro electrode for 1-10 µL/min. The mobile phases were 0.1% formic acid in LC–MS-grade water (buffer A) and 99.9% ACN/0.1% FA (buffer B). The ZenoTOF mass spectrometer was equipped with the Optiflow ion source using a spray voltage of 4.5 kV, ion source gas 1 of 15 psi, ion source gas 2 of 60 psi, curtain gas of 35 psi, CAD gas of 7 and a temperature of 200 °C. SWATH data was acquired using the following parameters: TOF MS start mass of 400 Da, a stop mass of 1500 Da, TOF MS accumulation time of 50 ms, TOF MSMS start mass 140 Da, stop mass 1750 Da, accumulation time 13 ms with dynamic collision energy turned on, a charge state of 2, Zeno pulsing enabled, and 60 variable SWATH windows covering the mass range of 400-900 m/z.

MS data acquisition of mixed species samples fostering innovation and collaboration on the Orbitrap Astral

For mixed species experiments, five replicates of samples A, B and C were loaded onto C-18 tips as described above. Samples were analyzed using an Evosep One system (Evosep) coupled to a Orbitrap Astral mass spectrometer (Thermo Scientific) using Thermo Tune software (version 1.0 or higher). Peptides were separated by the 60SPD method gradient (Evosep) on a PepSep 8 cm × 150 µm reverse-phase column packed with 1.5 µm C18-beads (Bruker Daltonics) at 50 °C. The analytical column was connected to a stainless-steel emitter with inner diameter of 30 µm (EV1086). The mobile phases were 0.1% formic acid in LC–MS-grade water (buffer A) and 99.9% ACN/0.1% FA (buffer B). The Orbitrap Astral mass spectrometer was equipped with a FAIMS Pro interface and an EASY-Spray source (both Thermo Scientific). A compensation voltage of -40V and a total carrier gas flow of 3.5 L/min was used as well as an electrospray voltage of 2.0 kV was applied for ionization. The MS1 spectra was recorded using the Orbitrap analyzer at 120k resolution from m/z 380-980 using an automatic gain control (AGC) target of 500% and a maximum injection time of 3 ms. The Astral analyzer was used for MS/MS scans in data-independent mode with 3 Th non-overlapping isolation windows with a scan range of 150-2000 m/z. The precursor accumulation time was 3ms and an AGC target of 500%. The isolated ions were fragmented using HCD with 25% normalized collision energy.

MS data acquisition of HeLa bulk data on the Orbitrap Astral

For analysis of HeLa bulk digest, 200ng of lysate was loaded onto C-18 tips in six replicates as described above. Samples were analyzed using an Evosep One system (Evosep) coupled to a Orbitrap Astral mass spectrometer (Thermo Scientific) using Thermo Tune software (version 1.0 or higher). Peptides were separated by the 60SPD method gradient (Evosep) on an Aurora Rapid 80 mm × 0.15 mm reverse-phase column packed with 1.7 µm C18-beads (IonOpticks) at 50 °C. The mobile phases were 0.1% formic acid in LC–MS-grade water (buffer A) and 99.9% ACN/0.1% FA (buffer B). The Orbitrap Astral mass spectrometer was equipped with a FAIMS Pro interface and an EASY-Spray source (both Thermo Scientific). A compensation voltage of -40V and a total carrier gas flow of 3.5 L/min was used as well as an electrospray voltage of 1.9 kV was applied for ionization. The MS1 spectra was recorded using the Orbitrap analyzer at 120k resolution from m/z 380-980 using an automatic gain control (AGC) target of 500% and a maximum injection time of 3 ms. The Astral analyzer was used for MS/MS scans in data-independent mode with 2 Th non-overlapping isolation windows with a scan range of 150-2000 m/z. The precursor accumulation time was 3ms and an AGC target of 500%. The isolated ions were fragmented using HCD with 25% normalized collision energy.

MS data acquisition of dimethylated peptides on the Orbitrap Astral

MS data acquisition was performed as described for mixed species samples on the Orbitrap Astral, if not described otherwise. For each of the six timepoints, triplicates of 50 ng of labeled peptide were injected. Samples were separated by the Whisper 40SPD method gradient (Evosep) on an Aurora Elite TS 15 cm and 75 µm ID (AUR3-15075C18-TS, IonOpticks) at 50 °C. The An electrospray voltage of 1.9 kV was applied. The MS1 resolution was 240 k with a maximum injection time of 100 ms and 6 ms for MS/MS.

Data Analysis

All data analysis was performed with python 3.11 using Numpy, Pandas, Seaborn and Matplotlib.

Search and analysis of dia-PASEF and synchro-PASEF data with alphaDIA

Data was searched with version 1.5.5 of alphaDIA using a previously published[32] empirical HeLa library. A default single step search was used with the following parameters: *target_ms1_tolerance = 15ppm*, *target_ms2_tolerance = 15 ppm*, *target_candidates = 5*. For synchro-PASEF *quant_all = true* was set and a

quant_window of 6 scans was used. All precursors with run-level FDR of 1% and protein groups with global FDR of 1% we're accepted. Coefficients of variation we're calculated on non-log transformed directLFQ normalized quantities.

Search and analysis of ZenoTOF data with alphaDIA

Data was searched with version 1.5.5 of alphaDIA using the HeLa library mentioned above. A default single step search was used with the following parameters: *target_ms1_tolerance = 15ppm, target_ms2_tolerance = 15 ppm, target_candidates = 3, target_rt_tolerance = 300*. All precursors with run-level FDR of 1% and protein groups with global FDR of 1% we're accepted. Coefficients of variation we're calculated on non-log transformed directLFQ normalized quantities.

Search and analysis of empirical library data from Lou et al.

Raw files, libraries and fasta files were used as provided in the original publication[33]. All data was searched with alphaDIA 1.5.5 using default parameters. For timsTOF data the following parameters were changed: *target_ms1_tolerance = 15ppm, target_ms2_tolerance = 15 ppm, target_candidates = 5, quant_window = 6, group_level = genes, scans, target_rt_tolerance = 500* seconds. For QE-HF data search was performed with *target_ms1_tolerance = 5ppm, target_ms2_tolerance = 10 ppm, target_candidates = 5, quant_window = 6, group_level = genes, scans, target_rt_tolerance = 600* seconds. Data for benchmarked tools was used as provided in the original publication. Analysis was performed as described in the original publication except for reassignment of proteins. Instead, search engine specific protein grouping was used. For alphaDIA, precursor passing local 1% FDR and protein groups passing a global 1% FDR were accepted.

Search and analysis of HeLa bulk data with fully predicted spectral libraries

For fully predicted library benchmarking, Spectronaut v18.6.231227.55695, DIA-NN 1.8.1, Chimerys on Ardia in Proteome discoverer and alphaDIA 1.5.4 was used. All analysis was performed using the same fasta file of reviewed human proteins without isoforms (01.12.2023). On all platforms, search was performed for tryptic precursors with carbamidomethyl modification at cysteine as fixed modification and variable methionine oxidation and protein N-terminal acetylation with maximum of two occurrences. Charge states 2 to 4 were included with sequence lengths between 7 and 35 amino acids with a single missed cleavage. For Chimerys, only peptides with up to 30 amino acids were used as the tool didn't support 35 amino acids. For alphaDIA automatic library prediction by alphaPeptDeep was used using the Lumos model for a NCE of 25. AlphaDIA used default parameters for a two-step search with the following changes: *target_ms1_tolerance = 4 ppm, target_ms2_tolerance = 7 ppm, target_rt_tolerance = 300s* in the first pass and *target_rt_tolerance = 100s* for the second pass. All data was analyzed at a 1% FDR threshold as enforced by the search engine. Coefficients of variation we're calculated on non-log intensities as provided by the search engine for all proteins. For Chimerys, quantification was only available on the protein level and not protein group level.

For Entrapment analysis, an Arabidopsis fasta with reviewed sequences and no isoforms was downloaded from Uniprot (02.02.2024). Search was performed as described above with heuristic inference. Following search all shared precursors, including isoleucine – leucine pairs were identified. Protein groups with shared precursors were discarded.

Search and analysis of mixed species data with fully predicted spectral libraries

For all three species, reviewed non-isoform proteomes were downloaded from Uniprot (21.02.2024). Proteins were in-silico digested using tryptic cleavage with carbamidomethyl modification at cysteine as fixed modification and variable methionine oxidation and protein N-terminal acetylation with maximum of two occurrences. Charge states 2 to 4 were included with sequence lengths between 7 and 35 amino acids with a

single missed cleavage. The Library was predicted using the alphaPeptDeep Lumos model at 25 NCE. AlphaDIA 1.5.4 was used with default parameters for a two-step search with the following changes: *target_candidates = 5, target_ms1_tolerance = 5 ppm, target_ms2_tolerance = 10 ppm, target_rt_tolerance = 200s* in the first pass and *target_rt_tolerance = 100s* for the second pass. Heuristic protein inference was used on the gene level. Proteins with shared sequences were removed as described above. For benchmarking accuracy, the median LFQ ratio was calculated for protein groups identified in at least three replicates.

Search and analysis of SILAC data with fully predicted spectral libraries

A fully predicted human library was generated with alphaPeptDeep as described above but for a NCE of 27. The library was multiplexed across the light channel without additional modifications and a heavy channel with isotopic labeling of Arginine (+10.008269) and Lysine (+8.014199). A single step search was performed with alphaDIA default parameters apart from: *target_ms1_tolerance = 5ppm, target_ms2_tolerance = 20ppm, target_rt_tolerance = 600 seconds, channel_wise_fdr = True.*

Search and analysis of dimethylated samples using transfer learning

A fully predicted human library was generated based on a reviewed human uniprot library (01.12.2023) with the general pretrained alphaPeptDeep model not trained on dimethylated peptides. The peptides were modified with Methionine oxidation and protein N-terminal acetylation as variable modifications with a maximum of two. N-Terminal and Lysine dimethylation were set as fixed modifications. Transfer search was performed using alphaDIA 1.5.5 with default parameters and *target_candidates = 1, target_ms1_tolerance = 4 ppm, target_ms2_tolerance = 7 ppm and target_rt_tolerance = 1200*. Transfer learning quantification was enabled and set to b and y ions with a maximum charge of 2 and the top 3 occurrences for every modified sequence. The generated transfer learning library was used for training with the default training scheme described above. For evaluation, the original pretrained model, the transfer learned retention time model, the transfer learned MS2 model and the fully transfer learned model were evaluated for search. All searches were performed with the same parameters as the transfer search apart from a *target_rt_tolerance = 100* for searches with the updated model.

Search and analysis of transfer learning entrapments

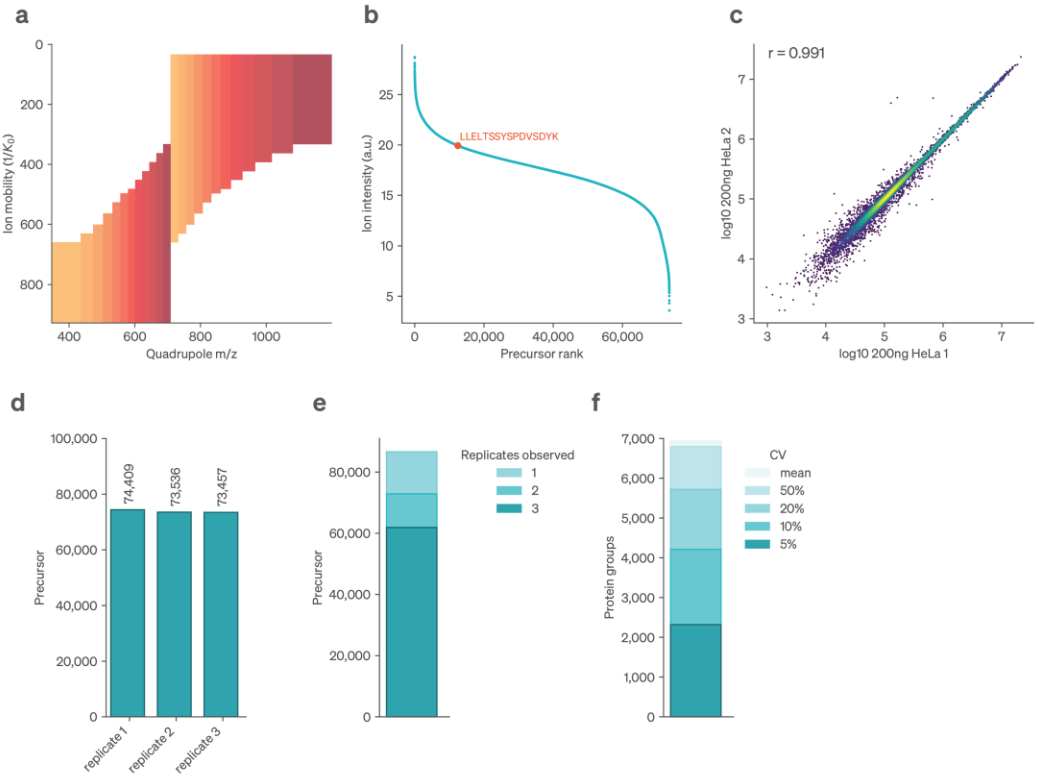
For evaluation of transfer learning on FDRs, entrapment experiments with known false positive Arabidopsis peptides were performed on the unmodified HeLa bulk samples acquired on the Orbitrap Astral. The entrapment library was generated as described above for the two step search with added N-terminal glutamate and glutamine to pyroglutamate conversion as variable modification. Raw files were searched with alphaDIA 1.5.5 using default parameters and *target_candidates = 1, target_ms1_tolerance = 4 ppm, target_ms2_tolerance = 7 ppm and target_rt_tolerance = 1200*. Transfer learning quantification was enabled and set to b and y ions with a maximum charge of 2 and the top 3 occurrences for every modified sequence. Transfer learning was performed utilizing all human and Arabidopsis precursors identified at 1% FDR cutoff. The transfer learning model was then reused for a second search with updated *target_rt_tolerance = 150* seconds. The process was repeated twice and the identifications after every search were analyzed for the number of false positive Arabidopsis identifications as described above.

References

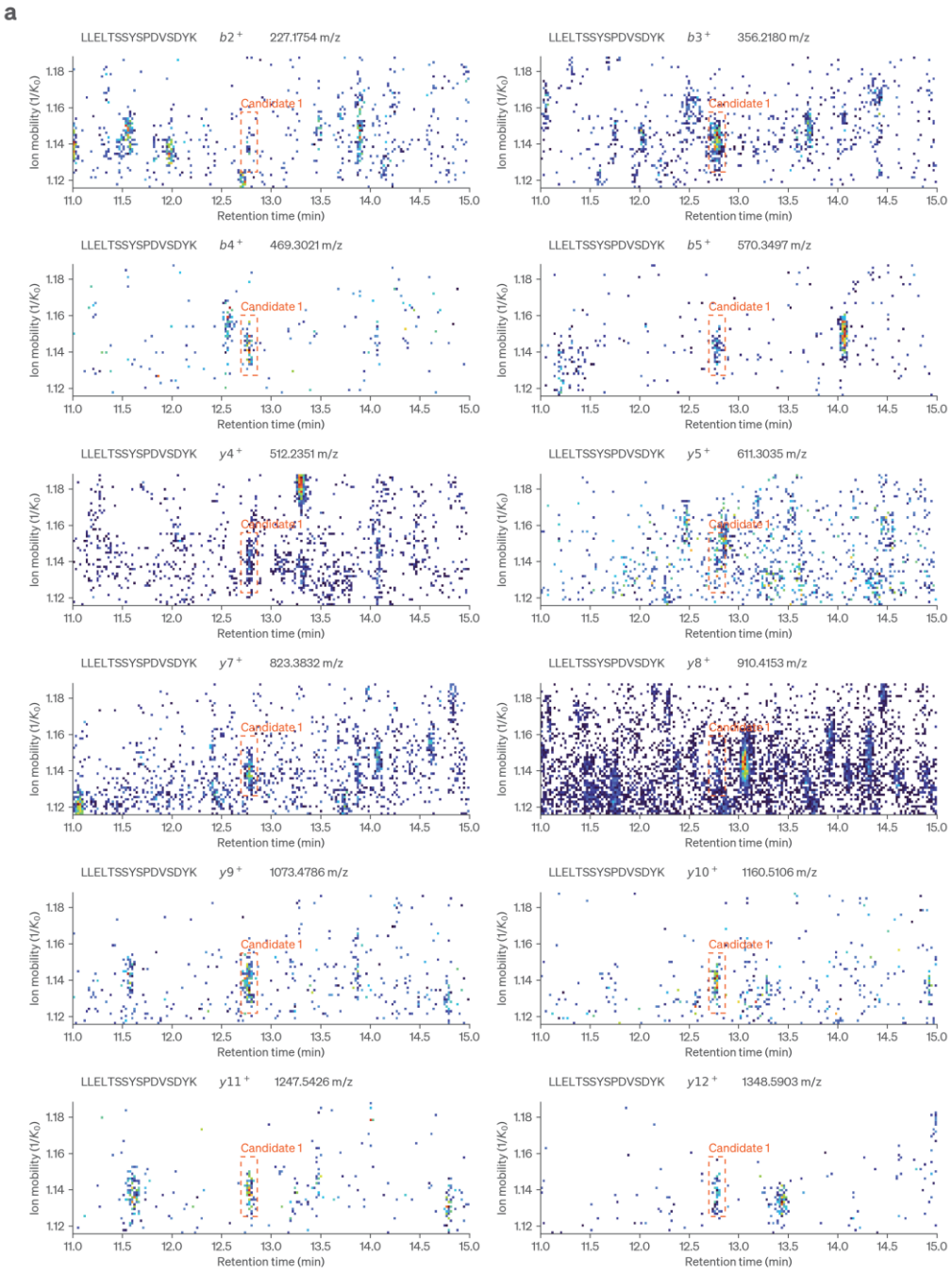
1. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
2. Navarro, P. *et al.* A multicenter study benchmarks software tools for label-free proteome quantification. *Nat. Biotechnol.* **34**, 1130–1136 (2016).
3. Cox, J. & Mann, 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).
4. 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. *Nat. Methods* **14**, 513–520 (2017).
5. Lazear, M. R. Sage: An Open-Source Tool for Fast Proteomics Searching and Quantification at Scale. *J. Proteome Res.* **22**, 3652–3659 (2023).
6. O'Reilly, F. J. & Rappsilber, J. Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology. *Nat. Struct. Mol. Biol.* **25**, 1000–1008 (2018).
7. Virág, D. *et al.* Current Trends in the Analysis of Post-translational Modifications. *Chromatographia* **83**, 1–10 (2020).
8. Liu, H., Sadygov, R. G. & Yates, J. R. A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics. *Anal. Chem.* **76**, 4193–4201 (2004).
9. 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. *Mol. Cell. Proteomics* **11**, O111.016717 (2012).
10. Collins, B. C. *et al.* Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nat. Commun.* **8**, 291 (2017).
11. Messner, C. B. *et al.* Ultra-fast proteomics with Scanning SWATH. *Nat. Biotechnol.* **39**, 846–854 (2021).
12. Brunner, A. *et al.* Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. *Mol. Syst. Biol.* **18**, e10798 (2022).
13. Bernhardt, O. *et al.* Spectronaut: a fast and efficient algorithm for MRM-like processing of data independent acquisition (SWATH-MS) data. in (2014).
14. Tsou, C.-C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **12**, 258–264 (2015).
15. 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. *Nat. Methods* **17**, 41–44 (2020).
16. Searle, B. C. *et al.* Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nat. Commun.* **9**, 5128 (2018).
17. Sinitcyn, P. *et al.* MaxDIA enables library-based and library-free data-independent acquisition proteomics. *Nat. Biotechnol.* **39**, 1563–1573 (2021).
18. Cox, J. Prediction of peptide mass spectral libraries with machine learning. *Nat. Biotechnol.* **41**, 33–43 (2023).
19. Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518 (2019).
20. Zeng, W.-F. *et al.* AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics. *Nat. Commun.* **13**, 7238 (2022).
21. Bekker-Jensen, D. B. *et al.* Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries. *Nat. Commun.* **11**, 787 (2020).
22. Steger, M. *et al.* Time-resolved in vivo ubiquitinome profiling by DIA-MS reveals USP7 targets on a proteome-wide scale. *Nat. Commun.* **12**, 5399 (2021).
23. Guzman, U. H. *et al.* Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition. *Nat. Biotechnol.* (2024) doi:10.1038/s41587-023-02099-7.
24. Wang, Z. *et al.* High-throughput proteomics of nanogram-scale samples with Zeno SWATH MS. *eLife* **11**, e83947 (2022).
25. Meier, F. *et al.* diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nat. Methods* **17**, 1229–1236 (2020).

26. Demichev, V. *et al.* dia-PASEF data analysis using FragPipe and DIA-NN for deep proteomics of low sample amounts. *Nat. Commun.* **13**, 3944 (2022).
27. Distler, U. *et al.* midiaPASEF maximizes information content in data-independent acquisition proteomics. Preprint at https://doi.org/10.1101/2023.01.30.526204 (2023).
28. Skowronek, P. *et al.* Synchro-PASEF Allows Precursor-Specific Fragment Ion Extraction and Interference Removal in Data-Independent Acquisition. *Mol. Cell. Proteomics* **22**, 100489 (2023).
29. Strauss, M. T. *et al.* AlphaPept: a modern and open framework for MS-based proteomics. *Nat. Commun.* **15**, 2168 (2024).
30. Willems, S., Voytik, E., Skowronek, P., Strauss, M. T. & Mann, M. AlphaTims: Indexing Trapped Ion Mobility Spectrometry–TOF Data for Fast and Easy Accession and Visualization. *Mol. Cell. Proteomics* **20**, 100149 (2021).
31. 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. Proteomics* **22**, 100581 (2023).
32. Skowronek, P. *et al.* Rapid and In-Depth Coverage of the (Phospho-)Proteome With Deep Libraries and Optimal Window Design for dia-PASEF. *Mol. Cell. Proteomics* **21**, 100279 (2022).
33. Lou, R. *et al.* Benchmarking commonly used software suites and analysis workflows for DIA proteomics and phosphoproteomics. *Nat. Commun.* **14**, 94 (2023).
34. Huang, T., Wang, J., Yu, W. & He, Z. Protein inference: a review. *Brief. Bioinform.* **13**, 586–614 (2012).
35. Granholm, V., Noble, W. S. & Käll, L. On Using Samples of Known Protein Content to Assess the Statistical Calibration of Scores Assigned to Peptide-Spectrum Matches in Shotgun Proteomics. *J. Proteome Res.* **10**, 2671–2678 (2011).
36. Cox, J. *et al.* Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).
37. Derks, J. *et al.* Increasing the throughput of sensitive proteomics by plexDIA. *Nat. Biotechnol.* **41**, 50–59 (2023).
38. Thielert, M. *et al.* Robust dimethyl-based multiplex-DIA doubles single-cell proteome depth via a reference channel. *Mol. Syst. Biol.* **19**, e11503 (2023).
39. Pino, L. K., Baeza, J., Lauman, R., Schilling, B. & Garcia, B. A. Improved SILAC Quantification with Data-Independent Acquisition to Investigate Bortezomib-Induced Protein Degradation. *J. Proteome Res.* **20**, 1918–1927 (2021).
40. Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* **73**, 2092–2123 (2010).
41. Nesvizhskii, A. I. & Aebersold, R. Interpretation of Shotgun Proteomic Data. *Mol. Cell. Proteomics* **4**, 1419–1440 (2005).

Extended Data Figures

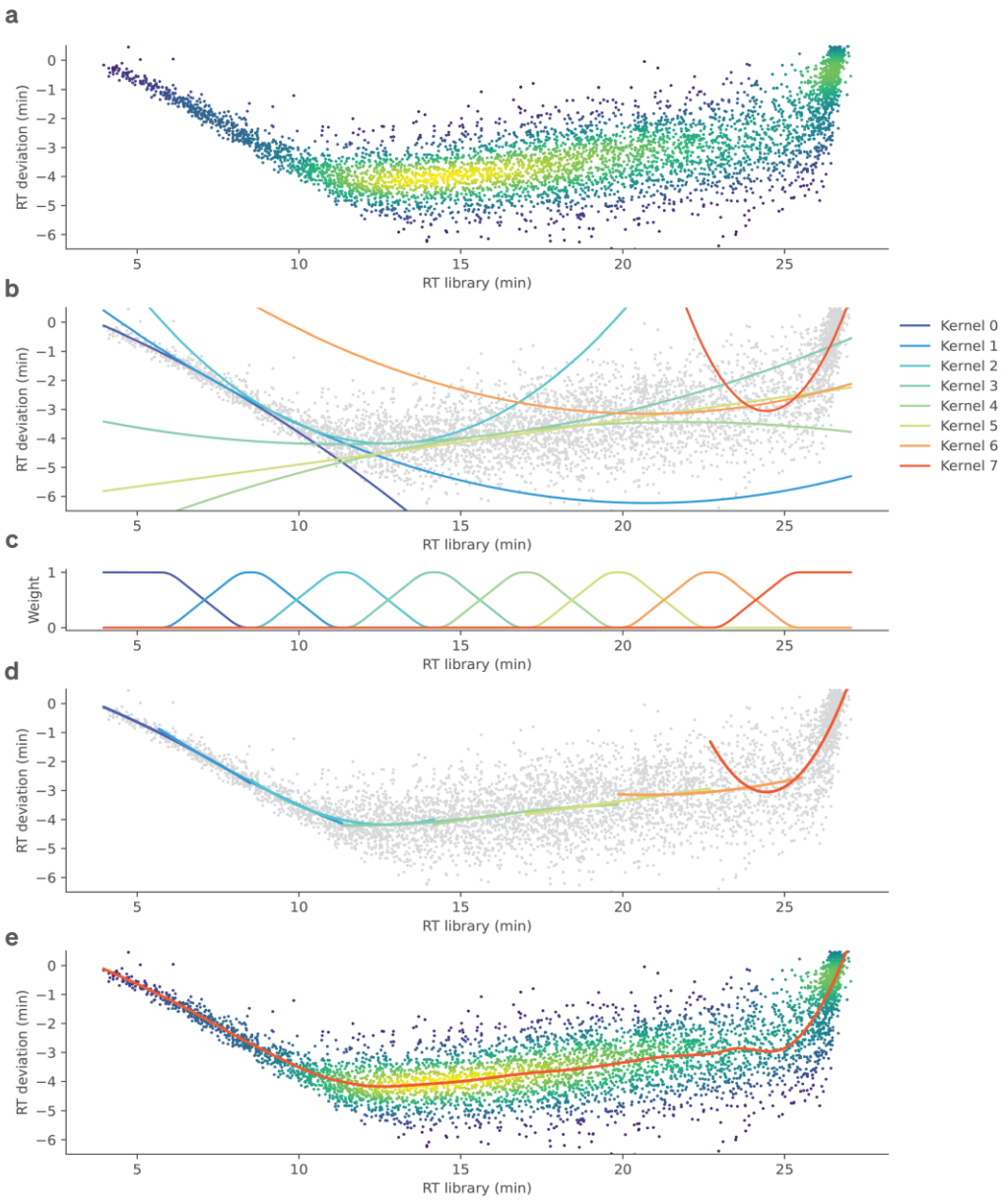


Extended Data Fig. 1 | alphaDIA search results for library-based search of triplicate bulk HeLa dia-PASEF data. Data was acquired at 60SPD (21min) on the timsTOF Ultra. **a**, Overview of the MS2 window distribution scheme of optimal dia-PASEF. **b**, Precursor selected as example in **Fig. 1** b-f. **c**, Correlation of LFQ protein quantities across replicates. **d**, number of precursors identified in each replicate at 1% FDR. **e**, Reproducibility of precursor identification across replicates. Number of precursors identified in at least 1, 2 or 3 replicates **f**, Precision of protein quantification. Number of protein groups for given CV cutoffs.

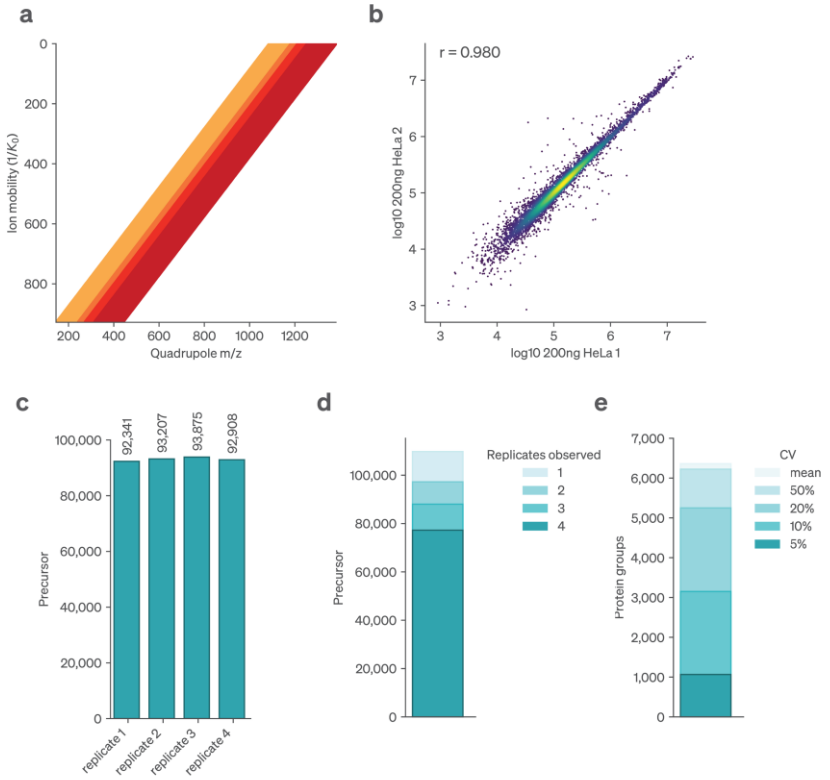


Extended Data Fig. 2 | Fragment signal across ion mobility and retention time for the precursor LLELTSSYSPDVSDYK2+. a, For each fragment all signal within the 15ppm of calibrated mass tolerance is shown and the final integration boundaries of the identified precursor are highlighted in red. Due to the high sensitivity of time-of-flight detectors fragment signal might only correspond to few ion copies. This leads to stochastic sampling of ions and discontinuous signal across retention time and ion

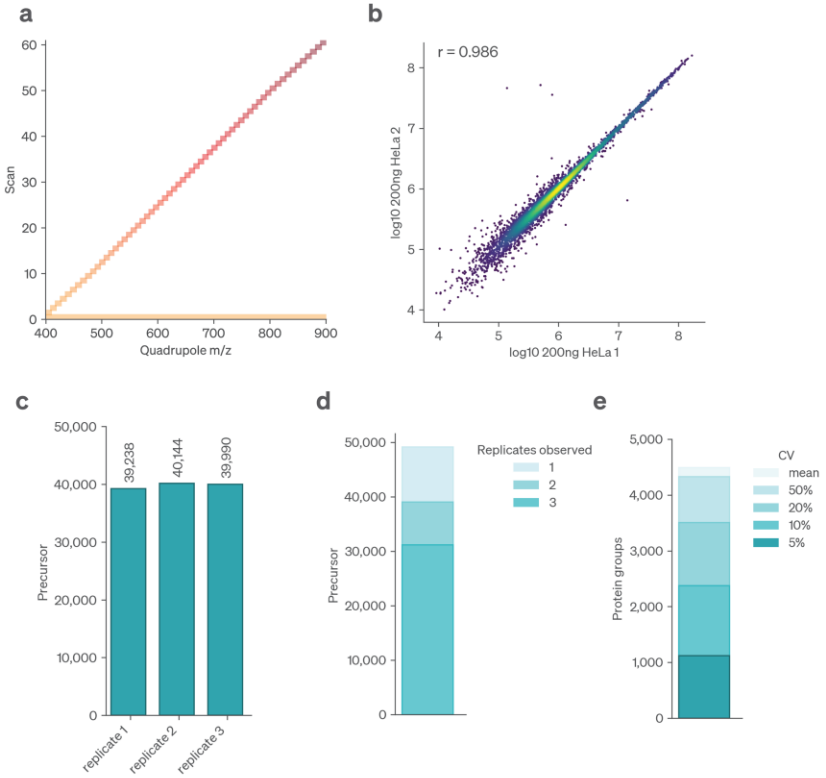
mobility. Distinguishing fragment signal from other ion species is challenging and prevents to determine clear peak boundaries. This requires an algorithm which does not need a minimum number of datapoints or certain peak shape. It's likewise important to combine evidence across fragments for determination of peak group boundaries.



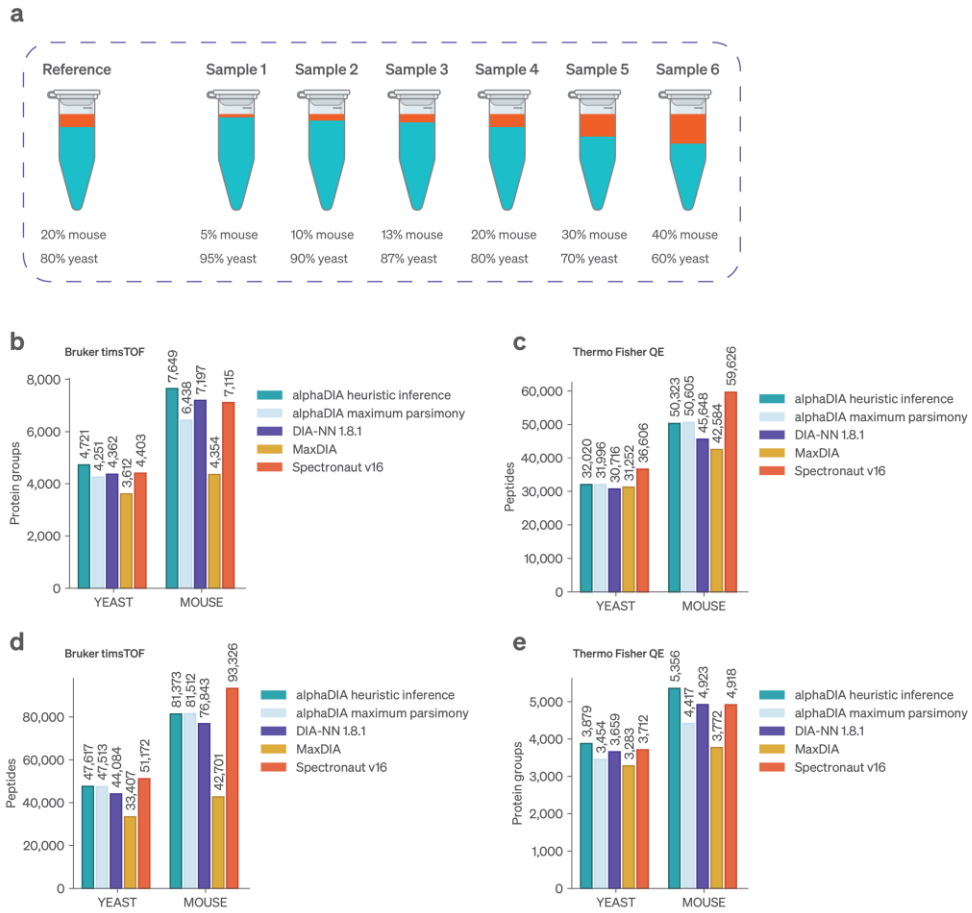
Extended Data Fig. 3 | Calibration of library properties to observed data using locally estimated scatterplot smoothing (LOESS) regression. a, Observed retention times of confidently identified precursors compared with the library annotated values. The absolute deviation in minutes is shown. **b**, A collection of polynomial kernels is fitted to uniformly distributed subregions of the data. **c**, The functions are combined and smoothed using tricubic weights. **d**, Combining the kernels with their weighting functions allows to approximate the systematic deviation of the data locally. **e**, The sum of the weighted kernels can then be used for continuous approximation and calibration of retention times.



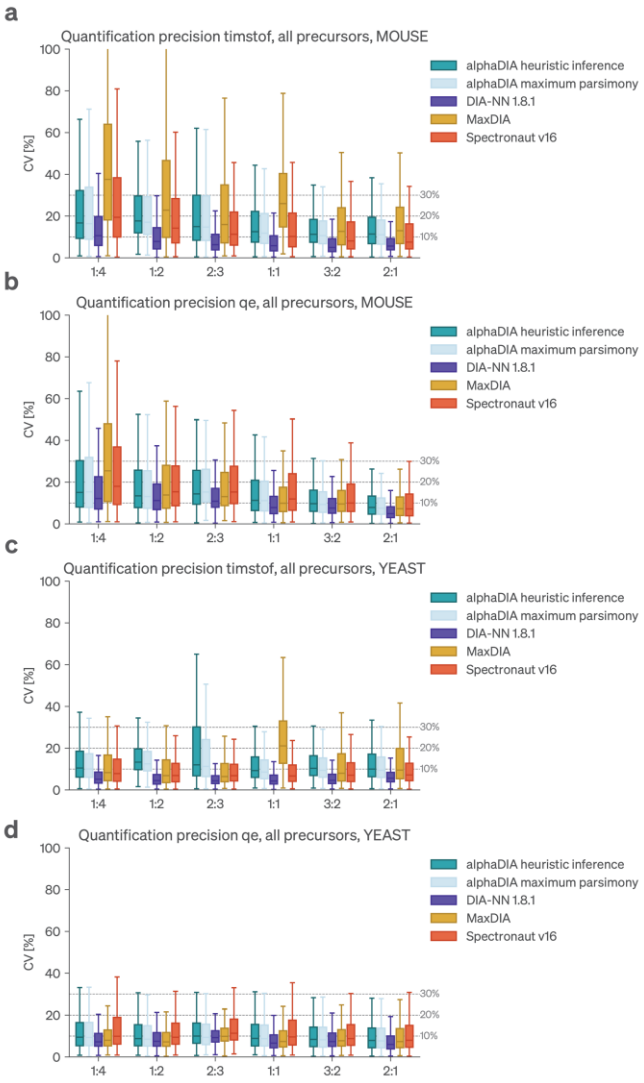
Extended Data Fig. 4 | Processing of synchro-PASEF data with alphaDIA. Analysis of bulk HeLa lysate with synchro-PASEF on the timsTOF Ultra. **a,** In synchro-PASEF the quadrupole is continuously scanning across the mass range while ions elute from the TIMS trap. In this method, four synchro scans of variable width are being used. **b**, Correlation of protein groups quantified between two replicates of HeLa lysate **c**, Number of precursors identified at 1% FDR per replicate. **d**, Data completeness given by precursors identified in a minimum number of replicates. **e**, Coefficient of variation (CV) for protein groups.



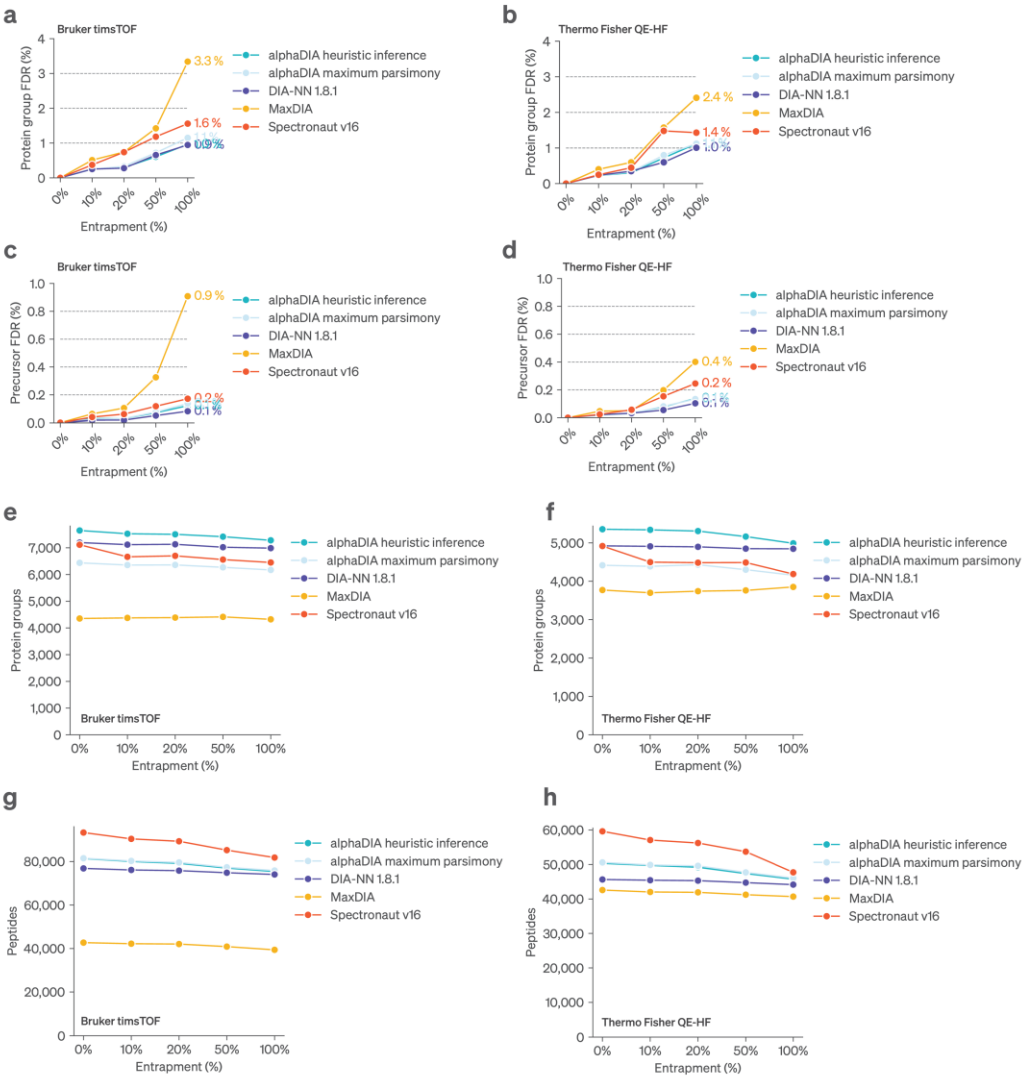
Extended Data Fig. 5 | Analysis of Sciex swath data acquired on the ZenoTOF 7600. Bulk HeLa lysate was analyzed with 21minutes of active gradient. **a**, Overview of the acquisition method used for data acquisition. The position of MS2 quadrupole windows is shown for a single DIA cycle. **b**, Correlation of protein groups quantified between two replicates of HeLa lysate **c**, Number of precursors identified at 1% FDR per replicate. **d**, Data completeness given by precursors identified in a minimum number of replicates. **e**, Coefficient of variation (CV) for protein groups.



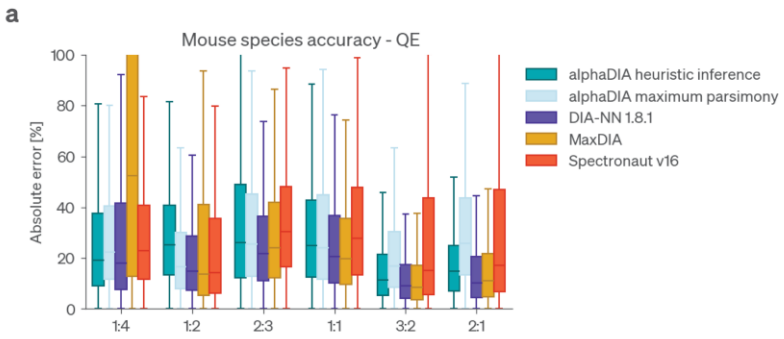
Extended Data Fig. 6 | Benchmarking library based search in a complex background a, Experimental setup as described by Lou et al.[33] Mouse brain isolate digests were spiked into a complex yeast proteome background in different ratios. **b**, Protein groups identified at 1% FDR on the Bruker timsTOF. **c**, Protein groups identified at 1% FDR on the Thermo Fisher QE-HF. **d**, Unique modified peptides identified 1% FDR across replicates on the Bruker timsTOF. **e**, Unique modified peptides identified 1% FDR across replicates on the Thermo Fisher QE-HF.



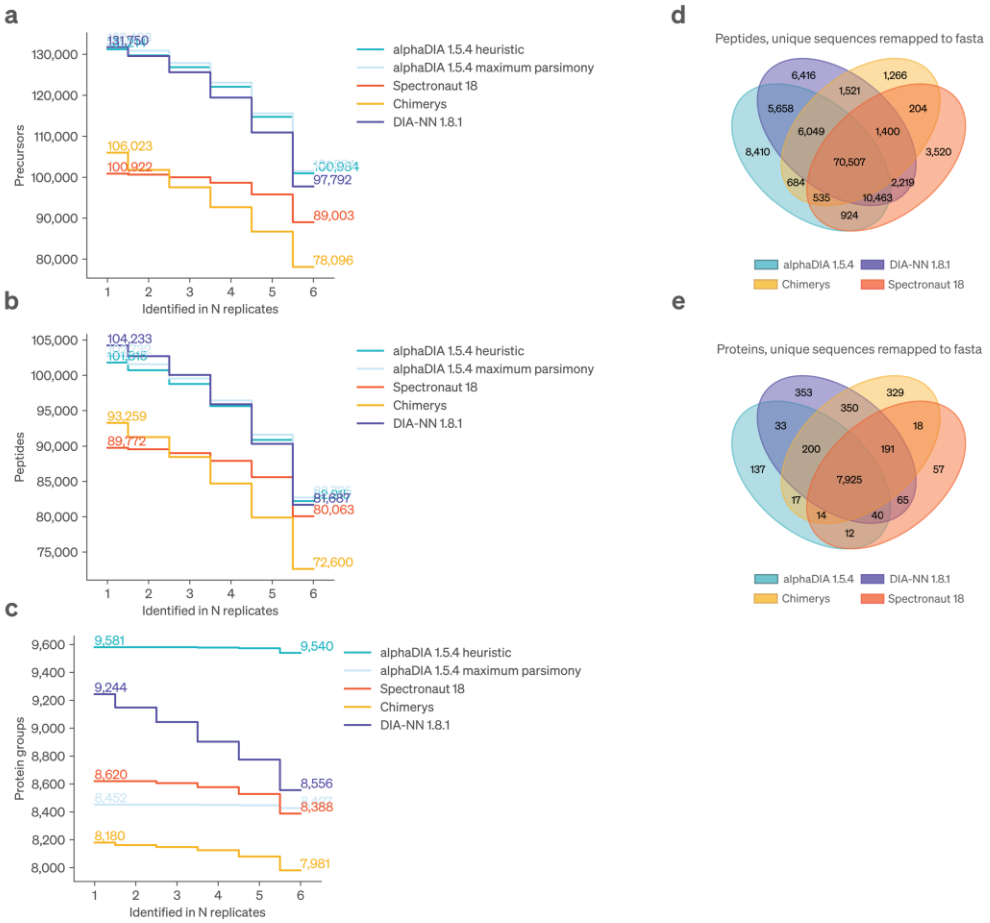
Extended Data Fig. 7 | Coefficient of variation for proteins in the empirical library benchmark. The quantitative precision was assessed by calculating the coefficient of variation for quantifiable protein abundances, identified in at least three out of five replicates. **a**, Mouse proteins identified on the timsTOF. **b** Mouse proteins identified on the QE-HF **c**, Yeast proteins identified on the timsTOF. **d**, Yeast proteins identified on the QE-HF.



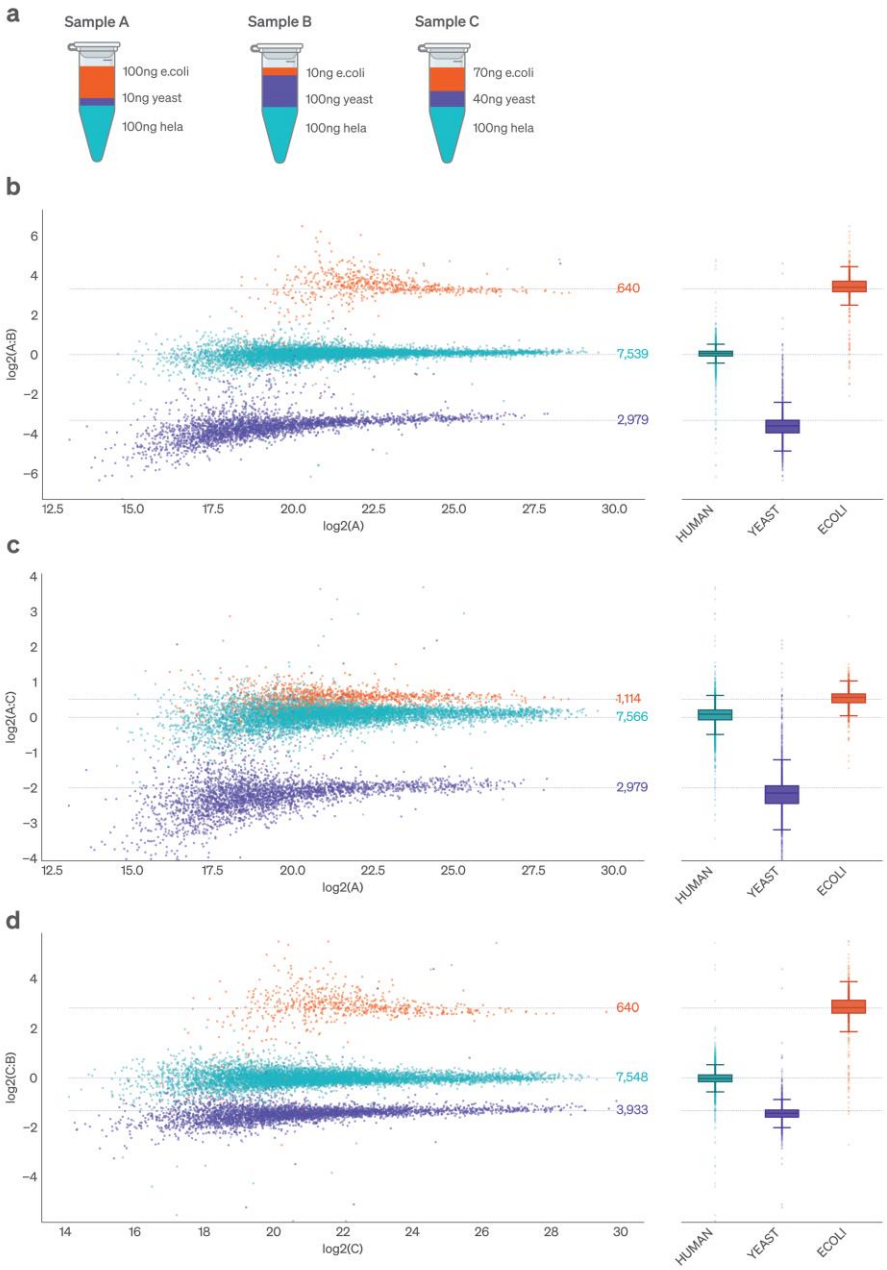
Extended Data Fig. 8 | FDR benchmarking using Arabidopsis entrapments. Target Mouse and Yeast libraries we're spiked in with increasing amounts of known false positive Arabidopsis precursors as provided by Lou et al.[33] **a-d**, Number of global known false positive Arabidopsis proteins as a fraction of all identified proteins is shown as entrapment FDR. Search results are shown for increasing amounts of entrapment precursors, relative to the target library. **a**, Benchmarking data acquired on timsTOF, entrapment FDR calculated on the protein group level. **b**, Benchmarking data acquired on QE-HF, entrapment FDR calculated on the protein group level. **c**, Benchmarking data acquired on timsTOF, entrapment FDR calculated on the precursor level. **d**, Benchmarking data acquired on QE-HF, entrapment FDR calculated on the precursor level.



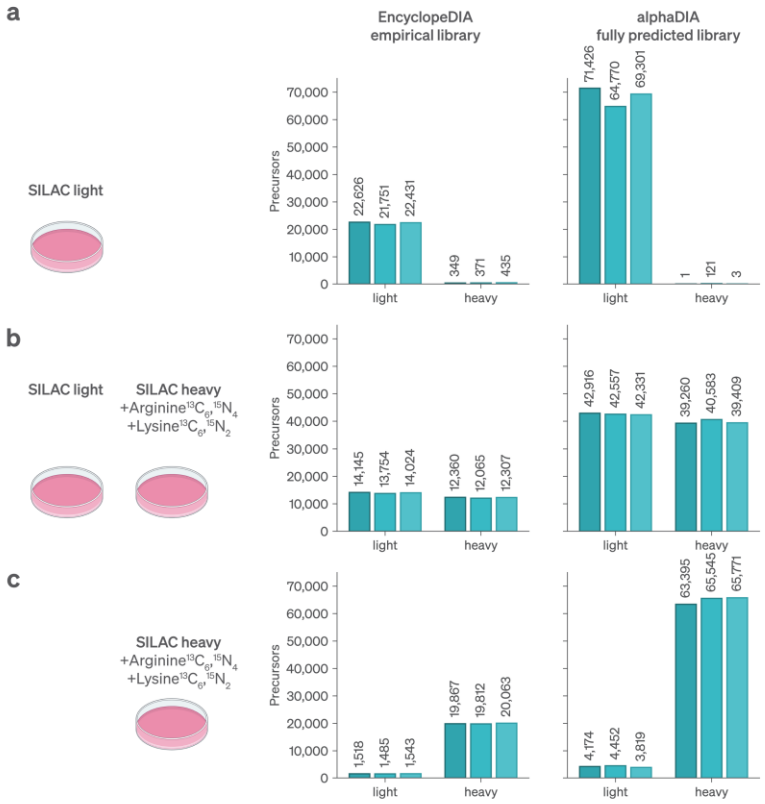
Extended Data Fig. 9 | Quantitative accuracy for ratios in the benchmarking dataset. a Ratios were calculated as described in the original study. The absolute error between the expected and observed ratio is shown for different search engines.



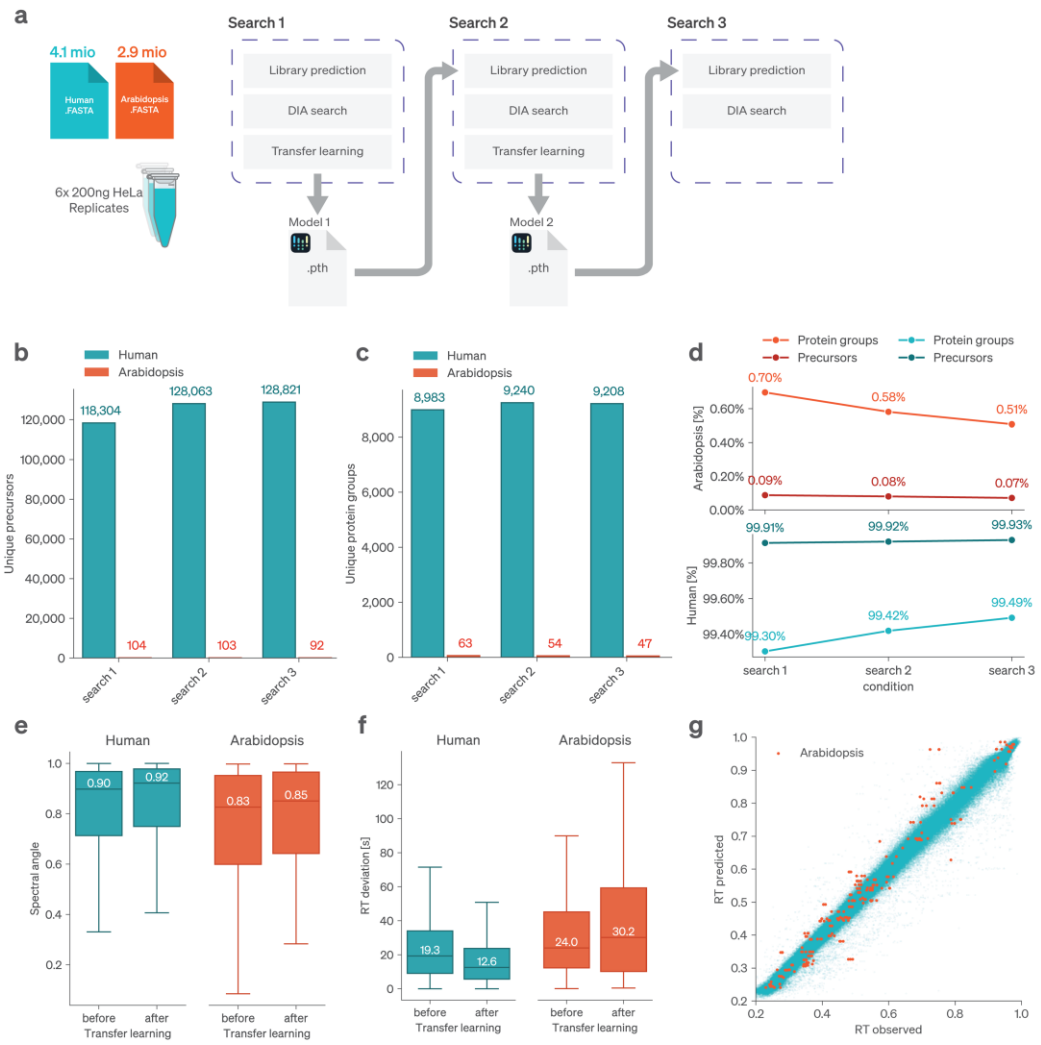
Extended Data Fig. 10 | Comparison of identifications for fully predicted library search across search engines. a, Data completeness of precursor identifications across replicates. **b**, Data completeness of modified peptide identifications across replicates. **c**, Data completeness of protein identifications across runs. **d-e** Peptides were mapped back to the human reference proteome to enable comparison independent of grouping. All peptides matching to multiple proteins were discarded. **d**, Venn diagram comparing the peptides identified by the different search engines. **e,** Venn diagram comparing the proteins identified by different search engines.



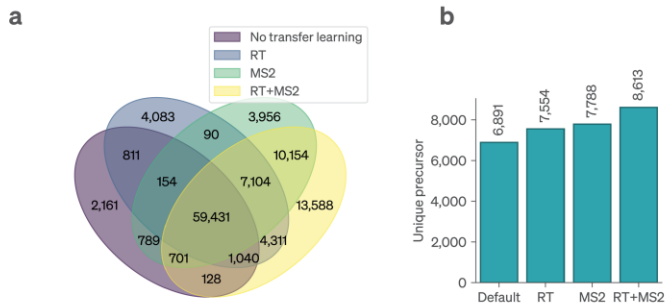
Extended Data Fig. 11 | Quantitative accuracy benchmark using mixed species proteomes on the Orbitrap Astral. a, Five replicates of three samples were prepared with Yeast, E.coli and human proteomes mixed in defined ratios. **b**, Comparison of median protein group intensities at 1% FDR between sample A and B. **c**, Comparison of median protein group intensities at 1% FDR between sample A and B. **d**, Comparison of median protein group intensities at 1% FDR between sample C and B.



Extended Data Fig. 12 | Validation of identification in SILAC labeled samples. SILAC data is from a method optimization study by the Garcia group that was originally analyzed by EncyclopeDIA and an empirical library[39]. This is compared to a fully alphaPeptDeep predicted library and database search by AlphaDIA. Triplicates results from the original paper are plotted in the left-hand panels and the AlphaDIA results on the same data in the right-hand panels. **a,** Percentage of false identifications in the heavy channel are median of 1.6% with EncyclopeDIA and 0.0043% with alphaDIA, which identified a threefold more precursors. **b,** For the combined sample, the heavy to light ratios are similar (46.7% heavy in EncyclopeDIA to 48.1% heavy in alphaDIA). **c**, After extended incorporation both analyses found similar percentage of light peptides (7.1% light in EncyclopeDIA vs 6.0% light in alphaDIA).



Extended Data Fig. 13 | Entrapment validation of end-to-end transfer learning across for iterations. a, Overview of the validation workflow. A Human and Arabidopsis fasta file digest was used for fully predicted library search. All identified precursors at 1% FDR were subsequently used for end-to-end transfer learning, including false positive Arabidopsis identifications. This process was repeated twice, using the transfer learned deep-learning model for library prediction. **b**, Total unique identified precursors across six replicates. Precursors mapping to both species, including leucine and isoleucine pairs were removed. **c**, Total unique identified protein groups. **d**, Entrapment FDR given as the percentage of false positive Arabidopsis identifications. **e**, MS2 spectral angle for precursors before and after transfer learning. Median spectral angle is shown for each plot. **f**, Retention time deviation in seconds before and after transfer learning. The median retention time deviation is shown. **g**, Predicted vs observed retention time following transfer learning. False positive Arabidopsis identifications are highlighted in red.



Extended Data Fig. 14 | Comparison of identification with transfer learning of dimethylation. a, Venn diagram showing the overlap of precursor identifications before and after transfer learning. **b**, Total number of unique protein groups identified across replicates after different stages of transfer learning.

3.2 Applications of Orbitrap Astral technology for spatial proteomics

As shown before, the Orbitrap Astral, as well as other highly sensitive TOF detector instruments such as the timsTOF Ultra/SCP, have pushed the boundaries of sensitivity, acquisition speed, and identification. This has shown to be particularly advantageous for low input applications and is broadening the possibilities for applications such as Deep Visual Proteomics (DVP) and single cell proteomics.[215,217,395,426,427] While previous DVP studies relied on the classification and laser-microdissection-based extraction of 700-1000 cell shapes to achieve sufficient depth, the sensitivity of the Orbitrap Astral MS allows for great proteomics depth at much lower sample input.[394,395,414] Especially when paired with an optimized and tailored acquisition strategy, something I have been focusing on during my PhD. This knowledge and experience served as a building stone for multiple DVP projects focusing on personalized medicine (**Article 4**), the evaluation of phenotypic shifts after xenotransplantation (**Article 5**) and single cell DVP (scDVP) in the context of alpha-1-antitrypsin deficiency (**Article 6**).

Article 4: Deep Visual Proteomics reveals DNA replication stress as a hallmark of Signet Ring Cell Carcinoma

Pre-print published online: bioRxiv (2024), doi: 10.1101/2024.08.07.606985, in revision at Precision Oncology

Sonja Kabatnik[1], Xiang Zheng[1,2], Georgios Pappas[1], **Sophia Steigerwald[3]**, Matthew P Padula[4], Matthias Mann[1,3]*

[1]NNF Center for Protein Research, Faculty of Health Science, University of Copenhagen, Copenhagen, Denmark

[2]Department of Biomedicine, Aarhus University, Denmark

[3]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany

[4]School of Life Sciences and Proteomics Core Facility, Faculty of Science, University of Technology Sydney,Ultimo, Australia

**Corresponding author*

Signet ring cell carcinoma (SRCC) is a rare and highly aggressive form of adenocarcinoma. SRCC is defined by the formation of a mucin filled vacuole, which leads to nuclei dislocation to the periphery and gives SR cells their characteristic signet ring morphology. It most commonly originates in the glandular cells of the stomach, but can also arise from other tissues such as the gall- or urinary bladder.[428,429] In comparison to other gastrointestinal cancers it has a poor prognosis, largely due to late diagnosis

and limited treatment options.[430,431] Due to its rarity, little is known about the mechanisms of this malignant cancer.

In this study, the first author Sonja Kabatnik had the unique opportunity to use DVP to investigate the proteome of the primary tumor and metastasized tissues of a single SRCC patient and utilize the gained information to make a tailored treatment recommendation. After optimizing a universal staining strategy and training a segmentation model, Sonja dissected 500 cell shapes, equating to ~50 SR cells, from the bladder (primary tumor side), the prostate, the seminal vesicles and a lymph node as well as non-cancerous epithelial prostate cells as a control. Using an input-optimized MS acquisition method on the Orbitrap Astral MS, which I advised on, a median of > 6,500 proteins could be identified per sample. While these included a number of tissue specific proteins, we could establish a core proteome of 4,825 proteins across all four tissue types. The initial analysis showed a clear clustering of samples based on tissue type and identified the disease status, healthy control vs. SR cells, as the primary driver of separation in a principal component analysis. This separation was primarily driven by known markers for prostate cancer, proteins related to epithelial-mesenchymal transition and classic SRCC markers, such as carcinoembryonic antigen-related cell adhesion molecule (CEACAM) and mucin (MUC) proteins. Among these, CEACAM5 and CEACAM6, as well as MUC1, 2, and 13, showed the most differential abundance between cancerous and epithelial controls. Further analysis indicated an upregulation of proteins associated with DNA replication, DNA damage response (DDR) and ataxia-telangiectasia mutated and Rad3-related (ATR) signaling, as well as defective mismatch repair (MMR). Together, this hints towards replication stress as a signature of SRCC. Moreover, proteomic and histological analysis indicated high levels of immune-related proteins, including programmed cell death ligand protein 1 (PD-L1), and infiltration of PD-1-positive cytotoxic T cells. This points towards tumor immunogenicity and suggests immunotherapy, especially PD-1 or PD-L1 inhibitors, as potential treatment options. In line with this, treatment with pembrolizumab, a PD-L1 inhibitor, was administered to the patient and showed a positive treatment response and effectively halted tumor progression. Overall, this highlights the potential of MS-based proteomics or DVP in particular, for precision oncology.

Contribution:

Co-authorship. This study was conceptualized by Sonja Kabatnik, Xiang Zheng, Matthew Padua and Matthias Mann. Sonja Kabatnik conducted the study. I advised on the MS acquisition strategy and gave feedback on data visualization. Alongside the other co-authors, I contributed to revising and editing the manuscript

Deep Visual Proteomics reveals DNA replication stress as a hallmark of Signet Ring Cell Carcinoma

Sonja Kabatnik[1], Xiang Zheng[1,2], Georgios Pappas[1], Sophia Steigerwald[3], Matthew P Padula[4], Matthias Mann[1,3]

[1]NNF Center for Protein Research, Faculty of Health Science, University of Copenhagen, Copenhagen, Denmark.
[2]Department of Biomedicine, Aarhus University, Denmark,
[3]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany.
[4]School of Life Sciences and Proteomics Core Facility, Faculty of Science, University of Technology Sydney, Ultimo, Australia.

Running title: Replication stress in signet ring cell carcinoma cells.

Abstract

Signet Ring Cell Carcinoma (SRCC) is a rare and highly malignant form of adenocarcinoma with increasing incidence and poor prognosis due to late diagnosis and limited treatment options. We employed Deep Visual Proteomics (DVP), which combines AI directed cell segmentation and classification with laser microdissection and ultra-high sensitivity mass spectrometry, for cell-type specific proteomic analysis of SRCC across the bladder, prostate, liver, and lymph nodes of a single patient. DVP identified significant alterations in DNA damage response (DDR) proteins, particularly within the ATR and mismatch repair (MMR) pathways, indicating replication stress as a crucial factor in SRCC mutagenicity. Additionally, we observed substantial enrichment of immune-related proteins, reflecting high levels of cytotoxic T lymphocyte infiltration and elevated PD-1 expression. These findings suggest that pembrolizumab immunotherapy may be more effective than conventional chemotherapy for this patient. Our results provide novel insights into the proteomic landscape of SRCC, identifying potential targets and open up for personalized therapeutic strategies in managing SRCC.

Introduction

Signet Ring (SR) cell carcinoma (SRCC) is a rare and highly aggressive type of adenocarcinoma that can occur in multiple organs. While the stomach is the most common primary tumor site, SRCC has also been reported in the prostate, breast, lung, and bladder [1]. Regardless of origin it typically metastasizes rapidly to distal sites [2,3]. Incidences of gastric SRCC have persistently increased over the last few decades [4,5].

If SRCC occurs from cells other than stomach glandular cells this may make disease classification in the effected organ more difficult [6–8]. However, there is one pathological feature that characterizes SR cells as such: a high concentration of intercellular mucin that builds up in large vacuoles, pushing the nucleus to the periphery of the cell and giving it the distinctive shape of a signet ring [9].

Despite clinical advances in gastric cancer classification, grading and treatment, the SR cell carcinoma subtype remains a substantial clinical burden [9,10]. Due to its rarity and a propensity for late symptom onset, SRCC patients are often diagnosed at an advanced stage, limiting treatment options and therapeutic efficacy [11,12]. Surgical resection followed by postoperative chemotherapy and radiotherapy are the main management options for advanced disease[13]. However, these treatments have limited impact on overall survival and can have numerous negative effects that worsen patient wellbeing [12]. The rarity of SRCC and the substantial knowledge gap regarding its fundamental biology and underlying signaling pathways thus combine to limit personalized therapeutic strategies for this distinct cancer subtype.

Investigations into SRCC biology have primarily revolved around this cancer's inherently increased proliferation rate, characterized by aberrations of the RAS/RAF/MAPK[14], HER2 or Wnt/β-catenin [15] signaling pathways and mutation of the E-cadherin gene CDH1 [16]. Microsatellite instability and strong lymphocyte infiltration have also been linked with colorectal SRCC, clinicopathological signatures typically rather associated with colorectal cancer than specifically with SRCC [17,18]. It is also kown that in colorectal SRCC, the SMAD complex triggers the epithelial-mesenchymal transition (EMT) in response to transforming growth factor (TGF)-β signaling, which accounts for the distinctive change of epithelial cell junctions and polarity in SRCC of the colon[19,20].

So far, most research on SRCC has been limited to clinical observations, histological classifications [21,22], and obtaining genomic sequencing data specific to occurrences in affected organs [20,23], predominantly the colon. We reasoned that global molecular analyses at the protein level could contribute to elucidating the broader biological context and distinctive pathogenic mechanisms of SRCC. The spatial proteomics field has made significant strides in recent years, and is potentially able to address the above challenge [24,25]. In particular our

group has developed the Deep Visual Proteomics (DVP) technology which combines high-resolution image acquisition with machine learning-guided segmentation and classification, followed by single-cell type enriched high-sensitivity mass spectrometry (MS)-based proteomics[26].

In this study, we took a precision oncology approach by using DVP to examine SRCC in four different organs—the bladder, prostate, liver, and lymph node—within a single patient. We reasoned that this spatial context would allow us to explore proteome differences and similarities of SR cells across tissues, offering valuable insights into tumor origin, potential mechanisms of metastasis and to make treatement recommendations.

Results

Patient Disease Background and Interventions

The patient was diagnosed with SRCC , with the bladder identified as the primary site of origin, following the removal of a suspicious mass on the bladder wall that was revealed by magnetic resonance imaging (MRI). Hematoxylin and eosin (H&E) staining of the mass revealed the typical signet ring morphology, and the patient was subjected to a radical cystectomy that removed the bladder (B.), prostate (P.), seminal vesicles (S.V.) and 14 lymph nodes (L.N.) (Figure 1A, B). Post-surgery pathology of these organs showed cells with signet ring morphology in all organs and nine out of fourteen lymph nodes. To enhance the therapeutic options for the patient, a genomic analysis was performed and a molecular tumor board report was filed, noting a microsatellite instability of only 0.8%, an ATRX (alpha thalassemia/mental retardation syndrome X-linked) frameshift mutation, MYCL and RICTOR (Rapamycin-insensitive companion of mTOR) amplification and KDM6A (Lysine-specific demethylase 6A) biallelic loss. The patient underwent chemotherapy with a combination of oxaliplatin, which was discontinued after four months due to the onset of continuous neuropathy, and capecitabine, likewise discontinued after seven months, before being monitored by quarterly computed tomography (CT) scans (Figure 1C). Twelve months after the cessation of chemotherapy, CT scan revealed suspicious enlargement of several lower abdominal lymph nodes. After further evaluation through a positron emission tomography (PET) scan, an accessible lymph node was removed by ultrasound guided biopsy, in which pathology confirmed the presence of cells of signet ring morphology. The patient received a combination of immunotherapy with pembrolizumab, which is ongoing, and chemotherapy with carboplatin, which was again stopped after four months due to side effects (Figure 1C). The tissues used in this study were obtained prior to any treatment.

A simple stain allows robust segmentation and classification for the DVP workflow

For spatial proteomics we sectioned formalin-fixed, paraffin-embedded (FFPE) tissue blocks of all four organs (bladder, prostate, seminal vesicle and one lymph node) at three μm thickness using a microtome and mounted the tissue sections on polyethylene naphthalate (PEN) membrane-coated microscopy glass slides (Figure 2A). Tissues were stained with DAPI for nuclear visualization. A crucial step in DVP is delination of the cell plasma membrand for subsequent laser microdissection. For our samples, we found that staining by wheat germ agglutinin (WGA), a lectin that binds to specific carbohydrates in the plasma membrane, was sufficient for this purpose (Figure 2B). In comparison to other staining methods, such as cytokeratin 1 (CK1) or the conventional H&E, WGA staining proved superior in terms of

efficiency and simplicity. Continuing with the DVP pipeline, we imaged the tissue slides with a standard immunoflourecent microscope (Zeiss Axio) and processed images with the Biology Image Analysis Software (BIAS) [26] (Figure 2A, B). For cell segmentation we fine-tuned a pre-trained model in BIAS. We trained a machine learning model for cell classification, which involved manual annotation of more than 1000 SRCC and lymphocytes from each organ to capture morphological diversity, ensuring accurate classification across tissue types (Figure 2B). Prediction accuracy of SR cells was 95% based on 10-fold cross validation and indendent validation by a pathologist.

Shapes were subsequently exported to a second microscope for semi-automated laser microdissection (Leica LMD7). In total, we dissected 500 cell shapes per organ, corresponding to approximately 50 SR cells, in triplicates. Collected cell shapes were lysed and enzymatically digested for subsequent MS-based proteomics (Figure 2D). Peptides were separated by to the Evosep One chromatography system [27], coupled to the Orbitrap Astral™ mass spectrometer [24]. This was followed by protein identification and quantification using the DIA-NN software [28] (Figure 2E, see Methods).

Proteomic analysis identifies organ-specific SRCC and DDR protein signatures

Analyzing MS data from the equivalent of 50 SR cells in all four organs, and including non-cancerous epithelial prostate cells as controls, we quantified a median of 6,638 different proteins (Figure 3A), with a excellent coefficient of variation (CV) of approximately 11% across the tissues (Figure 3C). A total of 4,648 proteins were present across all triplicates and organs and 7,157 in at least 70% of samples of each organ indicating high completeness of our data set (Figure 3C). Across the four organs, we identified 4,825 proteins as a common core proteome (Figure 3E). As expected, proteins uniquely present in each organ mirror specific organ functions, such as semenogelin-2 (SEMG2) in the seminal vesicle which is responsible for gel matrix formation for spermatozoa[29] (Figure 3F).

Principle component analysis (PCA) clearly clustered samples originating from the same tissue, but also the cancerous SR cell away from the control (Figure 3E). Likewise, SR cells from the lymph node, prostate, and bladder were clearly distinct from SR cells of seminal vesicles (Figure 3E). Well known markers for prostate cancer and proteins involved in EMT including dipeptidyl peptidase 4 (DPP4), transglutaminase 4 (TGM4), keratin 7 (KRT7), acid phosphatase 3 (ACP3), kallikrein-related peptidase 3 (KLK3) and solute carrier family 45 member 4 (SLC45A4) [30–34] were among the proteins driving the separation between SRCC and epithelial control in our PCA along the load component 1 (Figure 3F). Proteins that are instead enriched in the SR cells compared to the epethilial control cells include

carcinoembryonic antigen-related cell adhesion molecule 5 (CEACAM5) and CEACAM6, mucins (MUC2, MUC5B), and calcium-activated cloride channel regulator 1 (CLCA1), classical markers for SRCC (Figure 3F). Fatty acid binding protein 4 (FABP4) and glycerol-3-phosphate dehydrogenase 1 (GPD1) separate the SR cells from the seminal vesicle from the other organs through component 2, likely due to tissue-specific differences in cellular proteomes, function, due to interactions between SR cells and their tumor environment (Figure 3F). Thus, DVP recapitulated expected or recently described physiological patterns while adding novel molecular players.

In the prostate, there was a clear and significant enrichment of MUC and CEACAM proteins between the epithelial control and SR cells (Figure 3G). In contrast, we observed minimal differences in the levels of prostate and prostate cancer-associated proteins, including KLKB1, KLK2, KLK3, APC3, and SLC45A4, conventional adenocarcinoma of the prostate (Figure 3H).

To control for SRCC-specific protein patterns and to investigate proteins with the most significant differential changes, we focused on two well-known protein families strongly associated with SRCC, mucins and CEACAMs.

MUC1, MUC2, and MUC13 showed the strongest – up to ten-fold - and most consistent enrichments in SR cells across all organs compared to the epithelial control cells of the prostate (Figure 3H). MUC1 and MUC2 are already well known to be overexpressed in gastric cancers, however, MUC13, a transmembrane mucin might play an yet unknown role in cell signaling and eptithelial barrier protection. MUC4, MUC5AC, MUC5B, and MUC12 had significant but fluctuating fold-changes between organs. SR cells in the seminal vesicles exhibited protein levels similar to those of non-cancerous control cells in the prostate. Mucin-like 1 protein (MUCL)1 has structural similarities and glycosylation patterns to classical mucins, but interestingly its expression profile was not significantly changed across all tissues analyzed, demonstrating that changes and overexpression in SR cells are specific to classical mucins.

Regarding the CEACAM family, CEACAM5 and CEACAM6 expression increased up to ten-fold between cancerous and epithelial controls, with the sole exception of CEACAM6 in the SR cells of the seminal vesicle. CEACAM1 and CEACAM21, who have different functions and structures, remained uniform across the different tissues supporting the notion that they are not directly involved in [35].

We next asked if the proteins highly enriched in prostate SR cells could point us to any therapeutically relevant pathways. Indeed, the top ones in terms of fold-change and statistical significance in a Gene Ontology (GO) enrichment analysis were all related to DNA replication and DNA damage response (DDR), including ‘nucleotide excision repair (NER)’, ‘base

excision repair (BER)' and 'mismatch repair (MMR)' (Figure 3I). We additionally found that the enrichment of MMR pathways is universal to all SR-positive tissues. The majority of the constituent proteins were upregulated, however, a number of prominent replication proteins (RPAs) were substantially downregulated (Figure 3J).

Signet ring cells exhibit multiple DDR pathway deficiencies across organs

Following up on our observation that proteins of the DDR showed abundance changes between prostate SR cells and epithelial cells, we next investigated tissue-specific protein changes by correlating the fold-changes between them (Figure 4A). Comparing two tissues at a time, observed that Ly6/PLAUR domain-containing protein 8 (LYPD8) and UDP-glucuronosyltransferase 2B17 (UGT2B17) showed similar patterns to the above mentioned CEACAM5 and CEACAM6 proteins. LYPD8 is also involved in epethilial cell junction integrity, pointing to a dysregulation of cell-cell adhesion, as well as potential deficiencies in tissue protection. UGT2B17 is involved in the metabolism of steroid hormones and xenobiotics, which can alter the tumor microenvironment.

S100 calcium binding protein P (S100P), MUC2, and CLCA1 also had similar expression patterns across the tissue (Figure 4A), in line with CLCA1 (Calcium-activated chloride channel regulator 1) affecting mucin secretion through Ca^{2+} signalling and its possible implications in cancer pathophysiology[36]. KLK3 and TGM4, well-known prostate-specific markers, consistently exhibit a negative or zero fold change between tissues and non-cancerous control cells (Figure 4A). Thus signet ring cells may arise due to different molecular mechanisms distinct from those of conventional prostate adenocarcinoma and metastases.

To globally examine protein patterns prevalent across all SR cells and contrast them with epithelial cells as a control, we performed unsupervised hierarchical clustering on the 1,560 ANOVA significant proteins, which revealed two prominent clusters, those upregulated or downregulated with respect to control (upper, red cluster and lower, blue cluster in Figure 4B). We performed GO term enrichment analysis on the upregulated cluster using Reactome, NetPath, and Biological Processes, which highlighted diverse pathways active in SRCC cells. These included Wnt, leptin, epidermal growth factor (EGF) receptor and transforming growth factor β (TGFβ) receptor pathways, all well-known for their roles in various carcinomas including stomach, colorectal and SRCC, (Figure 4C). Apart from these, the most prominent pathways were again associated with DNA replication and DDR (Figure 4C).

Next, by comparing the SR cells to the epithelial control cells, we ran a gene set enrichment analysis (GSEA) on proteins which showed a significant enrichment following pairwise proteomic comparison. Remarkably, 7 of the top 10 pathways are part of DDR, namely

'Activation of [the] pre-replicative complex', 'activation of ATR (ataxia-telangiectasia mutated and Rad3-Related), a pathway triggered upon perturbations affecting DNA replication dynamics characterized as replication stress (RS)', 'PCNA-dependent long patch base excision repair (LP-BER)', 'gap-filling DNA repair synthesis and ligation in global-genome nucleotide excision repair (GG-NER)', and 'DNA strand elongation' (Figure 4D, E).

To validate our proteomic results regarding ATR signaling activation, we stained all SRCC-positive tissues for phospho-ATR (pATR), the activated form of the protein kinase which phosphorylates downstream key proteins involved in DDR[37–40]. Our staining results confirmed the presence of pATR across our tissue samples, with the highest positivity observed in the seminal vesicle tissue (Figure 4F). We confirmed that the seminal vesicle is particularly highly positive for pATR whereas the bladder, prostate, and lymph node also display pATR signals, but to a lesser extent (Figure 4F).

Pathways implicated in metabolic processes such as 'glycogen metabolism; and signaling mechanisms such as the 'Ca^{2+} pathway' and G-protein beta:gamma signaling' are negatively enriched (Figure 4D, E). Downregulation of these pathways in SR cells likely reflects metabolic reprogramming of cancer cells, alterations in calcium signaling to support uncontrolled growth and survival, and specific adaptations of SRCC to facilitate mucin production and secretion.

Given the observations of significant changes in protein abundances related to DDR pathways and the ATR signaling axis in SR cell-positive tissues compared to epithelial control cells, we further investigated proteins involved in stalled replication fork (RF) protection and repair of complex DNA lesions formed in case of replication fork collapse, a key part of the cellular response to DDR. These included proteins of the Fanconi Anemia (FA) pathway specifically the FA group D2 protein (FANCD2) and its interactor [41], Fanconi Anemia complementation group I (FANCI) [42]. Additional mediators of the same process including DNA unwinding RecQ like helicase 5 (RECQL5), Werner syndrome helicase (WRN), and helicase-like transcription factor (HLTF) all displaying a similar positive fold change (Figure 4G). Our data provides strong indications of an ongoing RS and of the subsequent response of the SR cells to maintain their genomic stability by upregulating various RF protection mechanisms.

Upon persistent RS and prolonged RF stalling, replisome structure is impaired and RFs collapse, leading to the emergence of single-end double strand breaks (seDSBs), the most deleterious form of DNA lesions. Cells then trigger the highly error-prone break induced replication pathway (BIR) to deal with this threat [43,44]. GSEA on our proteomic data showed a significant enrichment of this mutagenic pathway (Figure 4F). Moreover, DNA polymerase delta subunit POLD3, an essential subunit of DNA polymerase delta upon BIR, together with POLD2 and DNA polymerase epsilon (POLE) show a positive fold change enrichment

comparing SR cells of the prostate, the seminal vesicle, the lymph node, and the bladder to the epithelial control (Figure 4G).

APOBEC3s, members of the Apolipoprotein B mRNA-editing enzyme catalytic polypeptides (APOBECs) superfamily, exhibit overexpression across various cancer types, notably bladder [45–47] and prostate cancer [48,49]. The induced hyper-mutations of long stretches of single strand DNA (ssDNA) formed during BIR (with APOBEC3A and APOBEC3B being the major mutators) through deamination, foster genome instability in cancer cells, a phenomenon referred to as “kataegis”. However, proteins of the APOBEC3 family of enzymes were markedly reduced in abundance in SR cells of every tissue (Figure 4G), possibly as a protective feedback mechanism to mitigate the mutational burden and maintain genomic stability [50–53].

Collectively our analysis of the proteome changes of SR cells from the bladder, identified as the primary tumor site, as well as from metastatic sites, namely the prostate, seminal vesicle, and lymph node revealed consistent patterns of a severe dysregulation of multiple DNA repair mechanisms, with a potential negative impact on genome integrity.

Enrichment of Complement System and PD-1 Signaling Proteins in Signet Ring Cells result in a cytotoxic T lymphocyte infiltration

DDR genes’ mutations and expression profiles have been recently associated with alterations of immune regulatory gene expression and CD8+ T cell infiltration in the tumor microenvironment, serving as a predictive marker of immune checkpoint blockade (ICB) therapy efficiency [54,55]. We therefore hypothesized that our unique protein signatures could indicate a higher immunogenicity and a greater mutational burden of the SRCC. The Reactome-curated ‘Complement system’ pathway displayed a positive fold change across all tissues, with the most marked increases seen in the C1q subcomponent subunits A (C1QA), B (C1QB), and C (C1QC) (Figure 5A). A similar expression pattern was observed in immunoglobulins and proteins involved in the programmed cell death protein 1 (PD-1) signaling pathway (Figure 5A).

To confirm our hypotheses derived from our proteomic analyses, which pointed to tumour immunogenicity and DNA damage response pathways, we immunostained for PD-1 and CD8-positive cytotoxic T cells in bladder tissue (primary tumor site), and in lymph nodes (metastasis) (Figure 5B). These tissues exhibited stubstantial or moderate infiltration of PD-1+ cytotoxic T cells, respectively (Figure 5B). We also observed a pronounced upregulation of programmed cell death ligand protein 1 (PD-L1) on SR cells of the bladder (Figure 5C). This

suggests that immunotherapy, particularly PD-1/PD-L1 inhibitors, could be a promising therapeutic approach for targeting these tumors.

In line with our findings, the PD-1 inhibitor pembrolizumab had indeed been recommended and administered as a therapy following recurrence rather than chemotherapy. Our results indicate that the ladder, would have been unlikely to be effectivel while having the usual adverse effects. Initiated in 2022, the pembrolizumab ICB therapy on our patient has successfully halted tumor progression, with MRI scans conducted quarterly confirming tumor stasis.

Based on our results we propose a model is which the DNA damage repair mechnisms and the replication stress response takes center stage in SRCCs (Figure 5 D): These SR cells hyper-activate the epidermal growth factor receptor (EGFR) pathway with subsequent hyper-proliferation. Increased DNA replication combined with defective MMR then results in numerous unrepaired post-replication DNA lesions across the genome. The repair of these lesions relies on the cells' excision repair mechanisms, including base excision repair (BER) and nucleotide excision repair (NER), which we observed to be upregulated at the protein level. The abundance of such lesions, along with the increased rate of DNA replication, are major driving forces behind replication stress, leading to the activation of the ATR signaling pathway. Proteomics indicates SRCC cells respond to this stress by upregulating proteins mediating stalled replication fork protection and collapsed replication forks repair, striving to maintain their genome integrity.

Discussion

In this study, we employed Deep Visual Proteomics (DVP) to investigate the proteome landscape of Signet Ring Cell Carcinoma (SRCC) across primary and metastatic sites from a single patient. Our analysis of approximately 50 SR cells per organ yielded up to 7,700 proteins, providing unprecedented insights into the tumorigenic properties and potential signaling pathways of SRCC.

We identified both shared and organ-specific protein patterns in SR cells, with a clear distinction from normal epithelial control cells. Key drivers of this difference include mucins, CLCA1, CEACAM5, and CEACAM6. Mucins, particularly MUC1, MUC2, and MUC13, showed significant enrichment in SR cells across all organs, directly contributing to the characteristic signet ring morphology[56]. CLCA1, closely linked to mucin production, can significantly alter the tumor microenvironment, affecting cell adhesion and migration [36,57]. The upregulation of CEACAM5 and CEACAM6, immunoglobulin-related glycoproteins and adhesion molecules, is notable. While CEACAMs are known to facilitate cellular connection and are frequently elevated in various cancers [58–60], their specific role in SRCC has not been previously emphasized. Their overexpression may contribute to the distinctive morphology and aggressive behavior of SRCC through promotion of invasion and metastasis[61–64].

Our data revealed significant alterations in DNA damage response (DDR) pathways across SR cells in different organs. We observed changes in excision repair mechanisms, including DNA mismatch repair (MMR), base excision repair (BER), and nucleotide excision repair (NER). The upregulation of most MMR proteins, coupled with the downregulation of MLH1, suggests a defective MMR pathway, consistent with previous studies linking microsatellite instability to colorectal SRCC [17,18].

A key finding of our study is the upregulation of the ATR signaling axis, indicating ongoing replication stress – a recognized hallmark of cancer driving genome instability[65]. Our proposed model suggests that replication fork stalling and collapse result from an increasing load of post-replicative lesions combined with increased proliferation and DNA replication rates. In response to this stress, the ATR signaling pathway is activated, triggering mediators of stalled replication fork protection, and collapsed replication fork repair and restart (Figure 4G, 4F, 5D)[66]. Single-ended double-strand breaks, the most deleterious form of DNA lesions formed upon replication fork collapse, are addressed by the break-induced replication (BIR) pathway[67,68]. We demonstrated that BIR is upregulated in SR cells across all four organs examined. The error-prone nature of BIR has been associated with high mutation rates, gross chromosomal rearrangements (GCRs), and loss of heterozygosity, further fostering genomic instability[50,69].

In line with this model, SR cells exhibited a considerable increase in the abundance of poly (ADP-ribose) polymerase (PARP), a key player in DNA repair, and a decreased abundance of APOBEC (apolipoprotein B mRNA editing catalytic polypeptide-like) enzymes compared to adjacent non-tumorigenic epithelial cells. This protein profile suggests a complex interplay between DNA damage accumulation and repair mechanisms in SRCC.

The activation of these DNA damage response and repair pathways likely contributes to the high mutation rate and genomic instability observed in SRCC. This genomic instability, particularly the disruptions in the MMR pathway, is linked to microsatellite instability, which in turn can lead to increased tumor immunogenicity. These findings provide a mechanistic explanation for the observed enrichment of immune-related protein signatures in our proteomics data, and the potential efficacy of immunotherapy in SRCC[17,18], which we could confirm by immunofluorescence (IF) imaging, revealing strong cytotoxic T lymphocyte infiltration and PD-1 expression[22]. Moreover, we observed alterations related to the complement cascade pathway do occur in SRCC tissues as previously reported[70]

Our results provide a rationale for the observed clinical response to pembrolizumab immunotherapy in this patient[55,71–73], despite initial sequencing results showing only 0.8% unstable microsatellite sites. This highlights the potential of proteomic analysis in guiding treatment decisions, especially in cases where genomic data alone may not fully capture the tumor's biology. The identification of replication stress as a central feature of SRCC opens new avenues for targeted therapies. Our findings suggest that targeting the ATR pathway or exploiting vulnerabilities in DNA repair mechanisms could be promising strategies. Additionally, the overexpression of CEACAMs points to potential targets for antibody-drug conjugates or other targeted therapies.

This study demonstrates the power of spatial proteomics in uncovering the molecular intricacies of rare cancers like SRCC. By providing a comprehensive view of the proteome across different organs, we've identified common features of SR cells that transcend their tissue of origin, as well as organ-specific adaptations. This approach offers valuable insights into tumor biology that may not be apparent from genomic or transcriptomic analyses alone.

In conclusion, our DVP-based analysis of SRCC reveals a complex interplay of DNA damage response, replication stress, and immune signaling pathways. These findings not only deepen our understanding of SRCC biology but also suggest potential therapeutic strategies. The success of pembrolizumab in this case, explained retrospectively by our proteomic data, underscores the potential of precision oncology approaches guided by comprehensive molecular profiling.

Future studies should aim to validate these findings in larger cohorts of SRCC patients and explore the therapeutic potential of targeting the pathways identified here. Moreover, integrating proteomic data with genomic and transcriptomic profiles could provide an even more comprehensive understanding of SRCC biology, potentially leading to improved diagnostic and therapeutic strategies for this aggressive cancer subtype.

Acknowledgements

The authors thank L. Drici (NNF CPR Proteomics Program) for technical assistance. We acknowledge Richard Denis Maxime De Mets from the Core Facility of Integrated Microscopy for microscopy support. We also thank S. Adams, A. Mund, F.H. Post, L. Niu, JJ. Wang and E. Krismer for engaging and productive discussions.

Funding

This work is supported financially by the Novo Nordisk Foundation (grant NNF14CC0001) and the Max Planck Society. Additionally, S. Kabatnik was supported by the Novo Nordisk Foundation grant NNF20SA0035590.

CRediT Author contributions

S. Kabatnik, (Conceptualization: Equal; Investigation: Equal; Formal analysis: Lead; Data Curation: Lead; Visualization: Lead; Validation: Equal; Writing – Original Draft Preparation: Lead; Project administration: Equal).

X. Zheng, PhD (Conceptualization: Equal; Investigation: Equal, Visualization: Supporting, Data curation: Supporting; Visualization: Supporting; Validation: Equal; Writing – Original Draft Preparation: Supporting, Writing – Review & Editing: Equal).

G. Pappas, PhD (Conceptualization: Supporting; Writing – Writing & Editing: Equal)

S. Steigerwald (Data curation: Supporting; Original Draft Preparation: Supporting).

M. Padula, PhD (Conceptualization: Equal; Resources: Lead; Supervision: Supporting)

M. Mann, PhD (Conceptualization: Equal; Supervision: Lead; Resources: Lead; Project administration: Equal; Funding acquisition: Lead; Writing – Writing & Editing: Lead).

Disclosure and competing interests statement

M. M. is an indirect investor in Evosep Biosystems.

Data availability

The proteomics raw data and quantified files were submitted to the ProteomeXchange Consortium through the PRIDE partner repository (https://www.ebi.ac.uk/pride/) with the identifier PXD053079. Image data will be provided upon request by contacting Sonja Kabatnik at sonja.kabatnik@cpr.ku.dk.

Materials and Methods

Study design and ethical permission

This is a case study. All experiments were performed on a single individual patient who provided us with FFPE blocks from four organs with SRCC presence: bladder, lymph node, prostate, and seminal vesicle. After consultation with the Nepean Blue Mountains Local Health District, they concluded that 'there is no need for formal application to the Human Research Ethics Committee' (HREC). The patient provided full consent as a subject of study (HREC study reference: UTS ETH22-7236), including the provision that the proteomic analysis of signet ring adenocarcinoma will be not followed up by any clinical intervention, and there is 'no risk to privacy or confidentiality'. Thus, the letter and communication with the Nepean Blue Mountains Local Health District acts as 'evidence of waiver of the need for HREC approval'.

Immunohistochemistry and high-resolution microscopy

A detailed protocol for FFPE tissue mounting and staining on membrane PEN slides 1.0 (Zeiss, 415190-9041-000) is provided in the original Deep Visual Proteomics (DVP) article[26]. The tissue sections were initially subjected to deparaffinization and hydration through three cycles involving xylene and decreasing ethanol concentrations from 99.6% to 70%. For Wheat Germ Agglutinin (WGA) labeling, sections on membrane PEN slides were incubated with WGA staining solution (Biotium, 29023; diluted 1:1000) in a light-protected environment at 37°C for 10 min. For pan-cytokeratin (CK), CD8, PD1, PDL1 and pATR staining, antigen retrieval was achieved by immersing the tissue sections on glass slides in EDTA buffer (Sigma, E1161; pH 8.5) at 90°C for 30 min. Following this, the tissue sections were blocked with TBS protein-free blocking buffer (LI-COR, 927-80000) for 20 min at room temperature. For CD8/PD1/CK triple staining, the sections underwent overnight incubation at 4°C with anti-CD8 antibody (Abcam, ab17147; 1:100), followed by slide washing and subsequent incubation with Alexa Fluor® 647 goat anti-mouse antibody (Invitrogen, A-21235; 1:1000) for one hour at room temperature. After rinsing, the slides were further incubated overnight at 4°C with anti-PD1 antibody (Miltenyi Biotec, 130-117-384; 1:100) and anti-CK antibody (Invitrogen, 53-9003-82; 1:500). For PDL1/CK double staining, slides were incubated with anti-PDL1 antibody (Invitrogen, 12-5983-42; 1:100) and anti-CK antibody (Invitrogen, 53-9003-82; 1:500) overnight at 4°C. For pATR staining, slides were incubated with anti-pATR antibody (GeneTex, GTX128145; 1:500) overnight at 4°C, followed by slide washing and subsequent incubation with Alexa Fluor® 647 donkey anti-rabbit antibody (Invitrogen, A-31573; 1:1000) for one hour at room temperature. Finally, we used DAPI (Abcam, ab228529; 1:1000) for nuclear counterstaining for 5 min at room temperature and the slides were mounted with Anti-

Fade Fluorescence Mounting Medium (Abcam, ab104135) before examination under an AxioScan7 microscope (Zeiss, for WGA, CK, CD8, PD1 and PDL1 imaging) or PANNORAMIC 250 Flash III (3Dhistech, for pATR imaging).

Tumor regions were identified using CK staining, WGA staining, or simply by including the auto-fluorescent signal of mucin.

Cell segmentation and classification

Microscopy images were imported into BIAS (Biology Image Analysis Software, single-cell-technologies.com), for machine learning-based cell segmentation, classification, and subsequent single-shape export for semi-automated laser microdissection. For SR cell segmentation, we utilized a pre-trained deep neuronal network on our IF WGA-stained tissues. Detection confidence was set to 60% and the contour confidence to 20%. Cell shapes with a larger area than 1000 μm^2 were excluded. To accurately classify SR cells, we trained a BIAS-integrated multilayer perceptron (MLP) feedforward neural network on manually identified SR cells across all four tissues. We set the weight scale and the momentum parameter to 0.01, and the number of iterations to 10,000. Subsequently, reference points were set, and SR cell contours were exported for semi-automated laser microdissection[26].

Laser microdissection

After aligning the reference points using the LMD7 (Leica) microscope, we imported the shape contours to facilitate semi-automated laser microdissection, which was conducted with the following parameters: laser power at 34, aperture set to one, cutting speed at 28, the middle pulse count to tree, final pulse to one, head current at 47 percent, pulse frequency at 2,600 Hz, and an offset of 180. For each type of organ tissue, SR cell shapes were excised in triplicates, and collected into 384-well plates, deliberately omitting the outermost rows and columns. After microdissection, we spun down the plate at 1,000 g for 10 min, and the dissected cell shapes were preserved by freezing at –20°C for later processing.

MS sample preparation

The entire MS sample preparation protocol was adapted from the original DVP paper [26]. After protein digestion, samples were vacuum dried, resuspended in 20 μL Evosep buffer A (0.1% formic acid v/v) and directly loaded on Evotips (https://www.evosep.com/).

LC-MS

Subsequently after Evotip loading, our low input samples were analyzed on our Orbitrap Astral mass spectrometer (Thermo Fisher Scientific) connected to the EvoSep One chromatography system (https://www.evosep.com/). We utilized a commercial analytical column (Aurora Elite TS, IonOpticks) and an EASY-Spray™ source to run our samples with the 40 Samples Per Day ('40 SPD') method (31-min gradient). All samples we recorded in DIA (data independent acquisition) mode. The Orbitrap analyzer of the mass spectrometer was utilized for full MS1 analyses with a resolution setting of 240,000 within a full scan range of 380 – 980 m/z. The automatic gain control (AGC) for the full MS1 was adjusted to 500%. For the acquisition of our low-input FFPE DVP samples, we set the MS/MS scan isolation window to 3 Th (200 windows), the ion injection time (IIT) to 5 ms, and the MS/MS scanning range to cover 150–2000 m/z. Selected ions were fragmented by higher-energy collisional dissociation (HCD)[74]. at a normalized collision energy (NCE) of 25%.

MS data analysis

Raw files were first converted to the mzML file format using the MSConvert software (https://proteowizard.sourceforge.io/) from Proteowizard, keeping the default parameters and selecting 'Peak Picking' as filter. Afterwards, mzML files were quantified in DIA-NN[28] (version 1.8.1) using the FASTA (2023, UP000005640_9606, with 20,594 gene entries) from the UniProt database and a direct-DIA approach. The enzyme specificity was set to 'Trypsin/P' with a maximum of two missed cleavages. Parameters for post-translational modifications were set to including N-terminal methionine excision, methionine oxidation and N-terminal acetylation were all activated, and a maximum of two variable modifications were allowed. Precursor FDR was set to 1%, and both mass and MS1 accuracy were set to 15 ppm. 'Use isotopologues', 'heuristic protein inference', 'no shared spectra' and 'match between runs' (MBR) were enabled. Protein inference was set to 'genes' and the neural network classifier run in 'single-pass mode'. We chose the 'robust LC (high precision)' as quantification method and a retention time-dependent cross-run normalization strategy. SRCC samples, microdissected across all four organs', were searched together.

Bioinformatic analysis

After quantification in DIA-NN, the protein group matrix was imported into Perseus[75], and samples were annotated according to the organ of origin (bladder, lymph node, prostate, and seminal vesicle). Proteins with 70% of quantitative values present 'in at least one group' were then kept for imputation of missing values based on their normal distribution (width=0.3;

downshift = 1.5). Further, all statistical tests were corrected for multiple hypothesis testing, applying a permutation-based false discovery rate (FDR) cut off either 5% or 1%.

Gene Set Enrichment Analysis (GSEA) was conducted using Python (version 3.9.7) and the GSEApy package (documentation: https://github.com/zqfang/GSEApy, version 1.0.4).

For the purpose of data visualization, our analyses were performed using the Python programming language (version 3.9.7), and essential libraries such as NumPy (version 1.20.3), Pandas (version 1.3.4), Matplotlib (version 3.4.3), and Seaborn (version 0.12.2). Additionally, the ShinyGo web tool (documentation: http://bioinformatics.sdstate.edu/go/), version 0.77, was used to perform gene ontology (GO) term enrichment analysis.

References

1. Benesch, M. G. K. & Mathieson, A. Epidemiology of Signet Ring Cell Adenocarcinomas. *Cancers (Basel)* **12**, (2020).
2. He, C.-L., Chen, P., Xia, B.-L., Xiao, Q. & Cai, F.-L. Breast metastasis of gastric signet-ring cell carcinoma: a case report and literature review. *World J Surg Oncol* **13**, 120 (2015).
3. Kwon, K.-J. *et al.* Clinicopathological characteristics and prognosis of signet ring cell carcinoma of the stomach. *Gastric Cancer* **17**, 43–53 (2014).
4. Henson, D. E., Dittus, C., Younes, M., Nguyen, H. & Albores-Saavedra, J. Differential trends in the intestinal and diffuse types of gastric carcinoma in the United States, 1973-2000: increase in the signet ring cell type. *Arch Pathol Lab Med* **128**, 765–70 (2004).
5. Pernot, S. *et al.* Signet-ring cell carcinoma of the stomach: Impact on prognosis and specific therapeutic challenge. *World J Gastroenterol* **21**, 11428–38 (2015).
6. LAUREN, P. THE TWO HISTOLOGICAL MAIN TYPES OF GASTRIC CARCINOMA: DIFFUSE AND SO-CALLED INTESTINAL-TYPE CARCINOMA. AN ATTEMPT AT A HISTO-CLINICAL CLASSIFICATION. *Acta Pathol Microbiol Scand* **64**, 31–49 (1965).
7. Ming, S. C. Gastric carcinoma. A pathobiological classification. *Cancer* **39**, 2475–85 (1977).
8. Patel, M. I. *et al.* Seventh edition (2010) of the AJCC/UICC staging system for gastric adenocarcinoma: is there room for improvement? *Ann Surg Oncol* **20**, 1631–8 (2013).
9. Nagtegaal, I. D. *et al.* The 2019 WHO classification of tumours of the digestive system. *Histopathology* **76**, 182–188 (2020).
10. Li, Y., Zhu, Z., Ma, F., Xue, L. & Tian, Y. Gastric Signet Ring Cell Carcinoma: Current Management and Future Challenges. *Cancer Manag Res* **12**, 7973–7981 (2020).
11. Voron, T. *et al.* Is signet-ring cell carcinoma a specific entity among gastric cancers? *Gastric Cancer* **19**, 1027–1040 (2016).
12. Van Cutsem, E., Sagaert, X., Topal, B., Haustermans, K. & Prenen, H. Gastric cancer. *Lancet* **388**, 2654–2664 (2016).
13. Van Cutsem, E. *et al.* The diagnosis and management of gastric cancer: expert discussion and recommendations from the 12th ESMO/World Congress on Gastrointestinal Cancer, Barcelona, 2010. *Annals of Oncology* **22**, v1–v9 (2011).
14. An, Y. *et al.* Clinicopathological and Molecular Characteristics of Colorectal Signet Ring Cell Carcinoma: A Review. *Pathol Oncol Res* **27**, 1609859 (2021).
15. Lei, Z.-N. *et al.* Signaling pathways and therapeutic interventions in gastric cancer. *Signal Transduct Target Ther* **7**, 358 (2022).
16. Aitchison, A. *et al.* CDH1 gene mutation in early-onset, colorectal signet-ring cell carcinoma. *Pathol Res Pract* **216**, 152912 (2020).
17. Kakar, S. & Smyrk, T. C. Signet ring cell carcinoma of the colorectum: correlations between microsatellite instability, clinicopathologic features and survival. *Modern Pathology* **18**, 244–249 (2005).
18. Boland, C. R. & Goel, A. Microsatellite instability in colorectal cancer. *Gastroenterology* **138**, 2073-2087.e3 (2010).
19. Lamouille, S., Xu, J. & Derynck, R. Molecular mechanisms of epithelial–mesenchymal transition. *Nat Rev Mol Cell Biol* **15**, 178–196 (2014).
20. Nam, J.-Y. *et al.* Molecular Characterization of Colorectal Signet-Ring Cell Carcinoma Using Whole-Exome and RNA Sequencing. *Transl Oncol* **11**, 836–844 (2018).

21. Wang, Y. *et al.* Early onset, development and histological features of gastric signet-ring cell carcinoma. *Front Oncol* **13**, (2023).
22. Teng, Q.-L. *et al.* Immunohistochemical analysis of PD-L1 and tumor-infiltrating immune cells expression in the tumor microenvironment of primary signet ring cell carcinoma of the prostate. *Asian J Androl* **24**, 525 (2022).
23. Korphaisarn, K. *et al.* Signet ring cell colorectal cancer: genomic insights into a rare subpopulation of colorectal adenocarcinoma. *Br J Cancer* **121**, 505–510 (2019).
24. Stewart, H. I. *et al.* Parallelized Acquisition of Orbitrap and Astral Analyzers Enables High-Throughput Quantitative Analysis. *Anal Chem* **95**, 15656–15664 (2023).
25. Mund, A., Brunner, A.-D. & Mann, M. Unbiased spatial proteomics with single-cell resolution in tissues. *Mol Cell* **82**, 2335–2349 (2022).
26. Mund, A. *et al.* Deep Visual Proteomics defines single-cell identity and heterogeneity. *Nat Biotechnol* (2022) doi:10.1038/s41587-022-01302-5.
27. Bache, N. *et al.* A Novel LC System Embeds Analytes in Pre-formed Gradients for Rapid, Ultra-robust Proteomics. *Molecular & Cellular Proteomics* **17**, 2284–2296 (2018).
28. 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. *Nat Methods* **17**, 41–44 (2020).
29. Malm, J., Hellman, J., Magnusson, H., Laurell, C. B. & Lilja, H. Isolation and characterization of the major gel proteins in human semen, semenogelin I and semenogelin II. *Eur J Biochem* **238**, 48–53 (1996).
30. Yang, F. *et al.* Inhibition of Dipeptidyl Peptidase-4 Accelerates Epithelial-Mesenchymal Transition and Breast Cancer Metastasis via the CXCL12/CXCR4/mTOR Axis. *Cancer Res* **79**, 735–746 (2019).
31. Farashi, S., Kryza, T., Clements, J. & Batra, J. Post-GWAS in prostate cancer: from genetic association to biological contribution. *Nat Rev Cancer* **19**, 46–59 (2019).
32. Chuang, T.-D. *et al.* Human prostatic acid phosphatase, an authentic tyrosine phosphatase, dephosphorylates ErbB-2 and regulates prostate cancer cell growth. *J Biol Chem* **285**, 23598–606 (2010).
33. An, Q. *et al.* KRT7 promotes epithelial-mesenchymal transition in ovarian cancer via the TGF-β/Smad2/3 signaling pathway. *Oncol Rep* **45**, 481–492 (2020).
34. Chen, W. *et al.* SLC45A4 promotes glycolysis and prevents AMPK/ULK1-induced autophagy in TP53 mutant pancreatic ductal adenocarcinoma. *J Gene Med* **23**, e3364 (2021).
35. Han, Z.-W. *et al.* The old CEACAMs find their new role in tumor immunotherapy. *Invest New Drugs* **38**, 1888–1898 (2020).
36. HU, D., ANSARI, D., BAUDEN, M., ZHOU, Q. & ANDERSSON, R. The Emerging Role of Calcium-activated Chloride Channel Regulator 1 in Cancer. *Anticancer Res* **39**, 1661–1666 (2019).
37. Maréchal, A. & Zou, L. DNA damage sensing by the ATM and ATR kinases. *Cold Spring Harb Perspect Biol* **5**, (2013).
38. Zhou, B.-B. S. & Elledge, S. J. The DNA damage response: putting checkpoints in perspective. *Nature* **408**, 433–439 (2000).
39. Matsuoka, S. *et al.* ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science* **316**, 1160–6 (2007).

40. Smolka, M. B., Albuquerque, C. P., Chen, S. & Zhou, H. Proteome-wide identification of in vivo targets of DNA damage checkpoint kinases. *Proc Natl Acad Sci U S A* **104**, 10364–9 (2007).
41. Zhu, J. *et al.* FANCD2 influences replication fork processes and genome stability in response to clustered DSBs. *Cell Cycle* **14**, 1809–22 (2015).
42. Zhan, S. *et al.* Focal Point of Fanconi Anemia Signaling. *Int J Mol Sci* **22**, (2021).
43. Patel, J. A. *et al.* Replisome dysfunction upon inducible TIMELESS degradation synergizes with ATR inhibition to trigger replication catastrophe. *Nucleic Acids Res* **51**, 6246–6263 (2023).
44. Casas-Delucchi, C. S., Daza-Martin, M., Williams, S. L. & Coster, G. The mechanism of replication stalling and recovery within repetitive DNA. *Nat Commun* **13**, 3953 (2022).
45. Caswell, D. & Swanton, C. Distinct Mutagenic Activity of APOBEC3G Cytidine Deaminase Identified in Bladder Cancer. *Cancer Res* **83**, 487–488 (2023).
46. Shi, R. *et al.* APOBEC-mediated mutagenesis is a favorable predictor of prognosis and immunotherapy for bladder cancer patients: evidence from pan-cancer analysis and multiple databases. *Theranostics* **12**, 4181–4199 (2022).
47. Glaser, A. P. *et al.* APOBEC-mediated mutagenesis in urothelial carcinoma is associated with improved survival, mutations in DNA damage response genes, and immune response. *Oncotarget* **9**, 4537–4548 (2018).
48. Gerhauser, C. *et al.* Molecular Evolution of Early-Onset Prostate Cancer Identifies Molecular Risk Markers and Clinical Trajectories. *Cancer Cell* **34**, 996-1011.e8 (2018).
49. Li, X. *et al.* Loss of SYNCRIP unleashes APOBEC-driven mutagenesis, tumor heterogeneity, and AR-targeted therapy resistance in prostate cancer. *Cancer Cell* **41**, 1427-1449.e12 (2023).
50. Kockler, Z. W., Osia, B., Lee, R., Musmaker, K. & Malkova, A. Repair of DNA Breaks by Break-Induced Replication. *Annu Rev Biochem* **90**, 165–191 (2021).
51. Petljak, M. *et al.* Mechanisms of APOBEC3 mutagenesis in human cancer cells. *Nature* **607**, 799–807 (2022).
52. Scully, R., Panday, A., Elango, R. & Willis, N. A. DNA double-strand break repair-pathway choice in somatic mammalian cells. *Nat Rev Mol Cell Biol* **20**, 698–714 (2019).
53. Zou, J., Wang, C., Ma, X., Wang, E. & Peng, G. APOBEC3B, a molecular driver of mutagenesis in human cancers. *Cell Biosci* **7**, 29 (2017).
54. Vidotto, T., Nersesian, S., Graham, C., Siemens, D. R. & Koti, M. DNA damage repair gene mutations and their association with tumor immune regulatory gene expression in muscle invasive bladder cancer subtypes. *J Immunother Cancer* **7**, 148 (2019).
55. Teo, M. Y. *et al.* Alterations in DNA Damage Response and Repair Genes as Potential Marker of Clinical Benefit From PD-1/PD-L1 Blockade in Advanced Urothelial Cancers. *J Clin Oncol* **36**, 1685–1694 (2018).
56. Kerckhoffs, K. G. P. *et al.* Mucin expression in gastric- and gastro-oesophageal signet-ring cell cancer: results from a comprehensive literature review and a large cohort study of Caucasian and Asian gastric cancer. *Gastric Cancer* **23**, 765–779 (2020).
57. Liu, C.-L. & Shi, G.-P. Calcium-activated chloride channel regulator 1 (CLCA1): More than a regulator of chloride transport and mucus production. *World Allergy Organization Journal* **12**, 100077 (2019).
58. Blumenthal, R. D., Leon, E., Hansen, H. J. & Goldenberg, D. M. Expression patterns of CEACAM5 and CEACAM6 in primary and metastatic cancers. *BMC Cancer* **7**, 2 (2007).

59. Thompson, J. A., Grunert, F. & Zimmermann, W. Carcinoembryonic antigen gene family: Molecular biology and clinical perspectives. *J Clin Lab Anal* **5**, 344–366 (1991).
60. Shively, J. E. & Beatty, J. D. Cea-related antigens: Molecular biology and clinical significance. *Crit Rev Oncol Hematol* **2**, 355–399 (1985).
61. Kuespert, K., Pils, S. & Hauck, C. R. CEACAMs: their role in physiology and pathophysiology. *Curr Opin Cell Biol* **18**, 565–571 (2006).
62. Powell, E. *et al.* A functional genomic screen in vivo identifies CEACAM5 as a clinically relevant driver of breast cancer metastasis. *NPJ Breast Cancer* **4**, 9 (2018).
63. Zang, M. *et al.* CEACAM6 Promotes Gastric Cancer Invasion and Metastasis by Inducing Epithelial-Mesenchymal Transition via PI3K/AKT Signaling Pathway. *PLoS One* **9**, e112908 (2014).
64. Zhang, Y. *et al.* CEACAM6 promotes tumor migration, invasion, and metastasis in gastric cancer. *Acta Biochim Biophys Sin (Shanghai)* **46**, 283–290 (2014).
65. Macheret, M. & Halazonetis, T. D. DNA replication stress as a hallmark of cancer. *Annu Rev Pathol* **10**, 425–48 (2015).
66. Saldivar, J. C., Cortez, D. & Cimprich, K. A. The essential kinase ATR: ensuring faithful duplication of a challenging genome. *Nat Rev Mol Cell Biol* **18**, 622–636 (2017).
67. Sedletska, Y., Radicella, J. P. & Sage, E. Replication fork collapse is a major cause of the high mutation frequency at three-base lesion clusters. *Nucleic Acids Res* **41**, 9339–9348 (2013).
68. Zhang, F. & Gong, Z. Regulation of DNA double-strand break repair pathway choice: a new focus on 53BP1. *J Zhejiang Univ Sci B* **22**, 38–46 (2021).
69. Sakofsky, C. J. *et al.* Break-induced replication is a source of mutation clusters underlying kataegis. *Cell Rep* **7**, 1640–1648 (2014).
70. Fan, Y. *et al.* Proteomic Profiling of Gastric Signet Ring Cell Carcinoma Tissues Reveals Characteristic Changes of the Complement Cascade Pathway. *Mol Cell Proteomics* **20**, 100068 (2021).
71. Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nat Rev Cancer* **12**, 252–64 (2012).
72. Kavun, A. *et al.* Microsatellite Instability: A Review of Molecular Epidemiology and Implications for Immune Checkpoint Inhibitor Therapy. *Cancers (Basel)* **15**, 2288 (2023).
73. Luchini, C. *et al.* ESMO recommendations on microsatellite instability testing for immunotherapy in cancer, and its relationship with PD-1/PD-L1 expression and tumour mutational burden: a systematic review-based approach. *Ann Oncol* **30**, 1232–1243 (2019).
74. Olsen, J. V *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nat Methods* **4**, 709–712 (2007).
75. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* **13**, 731–740 (2016).

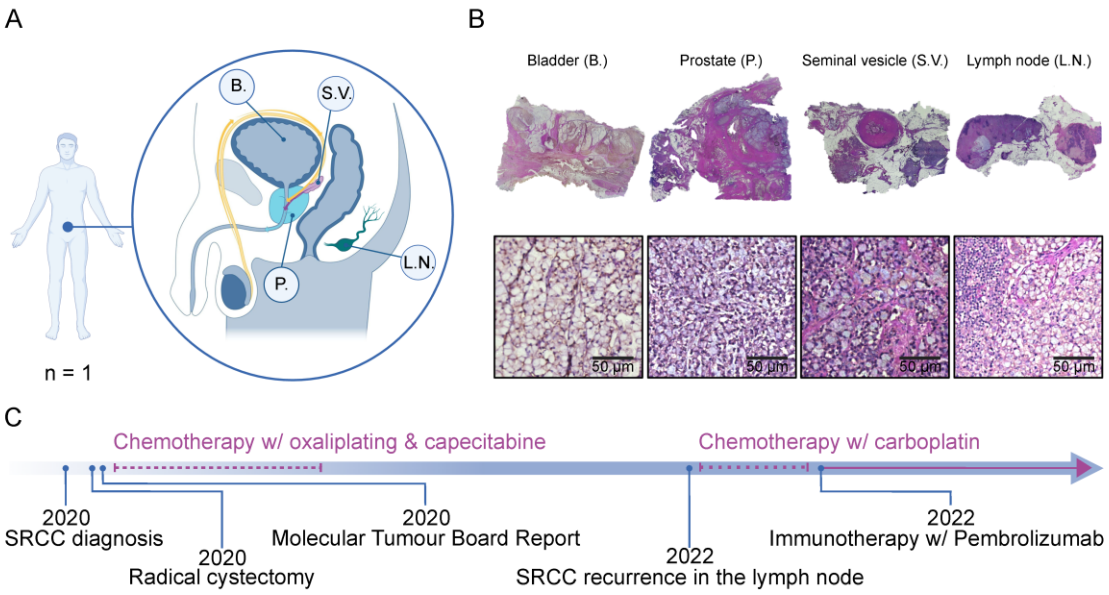


Figure 1 Signet ring cell carcinoma samples and timeline of medical interventions.

A Sample overview of signet ring cell carcinoma (SRCC)-positive tissues including the bladder (B.), the seminal vesicle (S.V.), one lymph node (L.N.) and the prostate (P.). Image was adapted from tulsaprocedure.com and modified. **B** Images of hematoxylin and eosin (H&E) stained SRCC formalin-fixed, paraffin-embedded (FFPE) tissues. **C** Chronological timeline of medical interventions. Illustrated with BioRender.

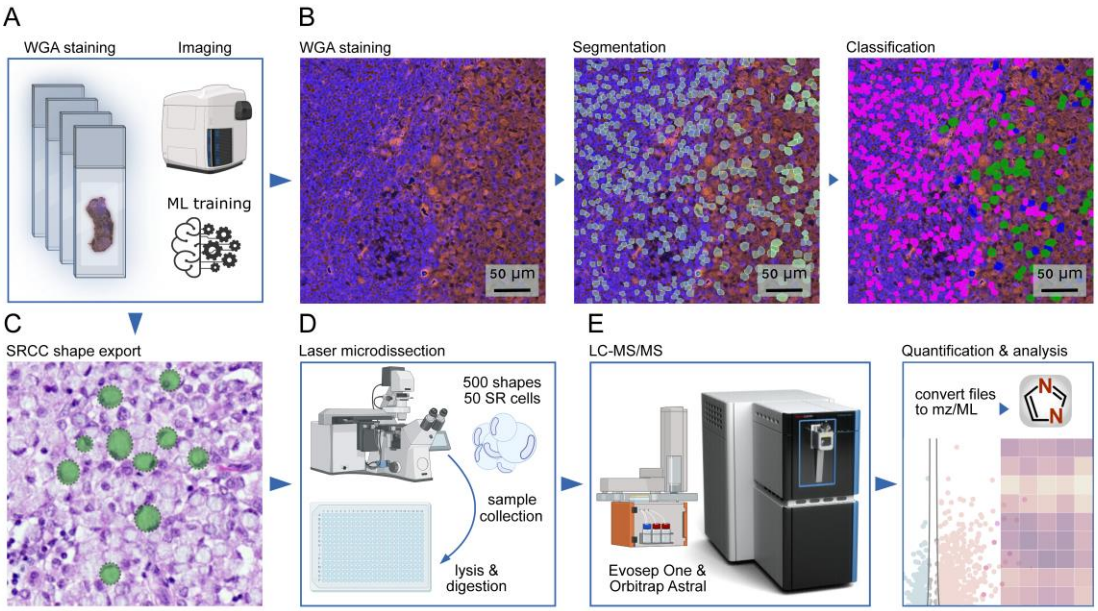


Figure 2 Deep Visual Proteomics workflow on WGA-stained tissues.

A Cell-type specific tissue preparation for the Deep Visual Proteomics (DVP) spatial proteomics pipeline, starting with FFPE tissue sectioning, mounting, staining and image acquisition. **B** Representative images of WGA-stained lymph node tissue, showing one raw, one segmented and one classified image (lymphocytes in pink, SR cells in green and segmentation artifacts in blue). **C** Export mask of classified SR cells. **D** Illustration of the semi-automated laser microdissection sample collection and processing, followed by **E** the liquid chromatography-mass spectrometry (LC-MS) setup. Illustrated with BioRender.

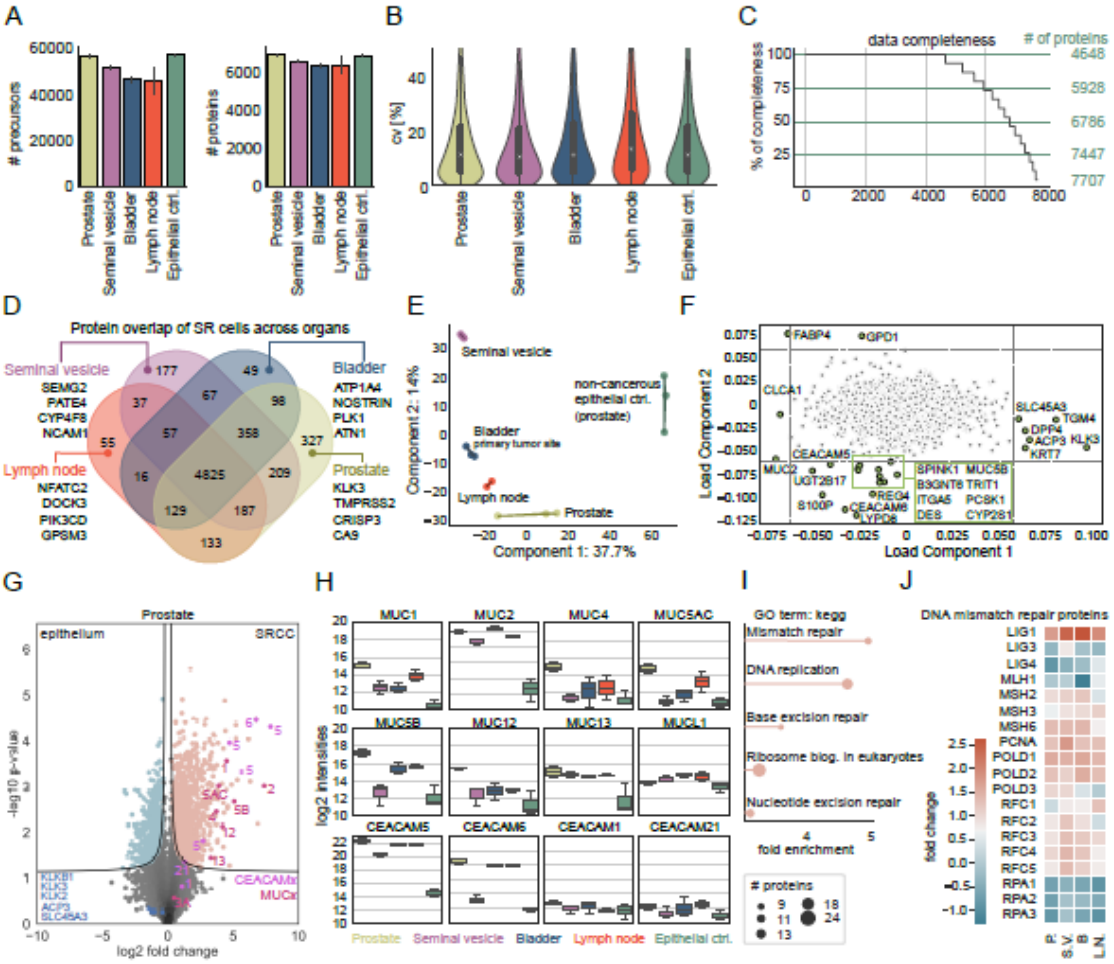


Figure 3 Proteomic depth and signatures of signet ring cells across tissues.

A Number of precursors and proteins across all tissues. **B** Coefficients of variation (cv). **C** Data completeness and highlighted cutoffs at 100, 75, 50 and 25%. Proteins were ranked in descending order based on the number of valid values present across organs and triplicates. **D** Overlap of SR cell proteomes across tissues, highlighting organ-specific proteins for the seminal vesicle, bladder, lymph node, and prostate. **E** Principal component analysis (PCA) of SR proteins across tissues. **F** Loading plot of the PCA, highlighting outlier proteins. **G** Pairwise proteomic comparison of the non-cancerous epithelial control (Prostate ctrl.) cells to the SR cells of the prostate (two-sided t-test, FDR <0.01, s_0 = 0.1). **H** Log2 normalized protein intensities of the mucin (MUC) and the carcinoembryonic antigen-related cell adhesion molecule (CEACAM) family members. **I** Gene Ontology (GO) term enrichment analysis using KEGG pathways of proteins significantly upregulated in the previous pairwise proteomic comparison. **J** Heatmap showing fold changes in DNA mismatch repair proteins between SR cells of the prostate, seminal vesicle, bladder, and lymph node to epithelial prostate cells as control.

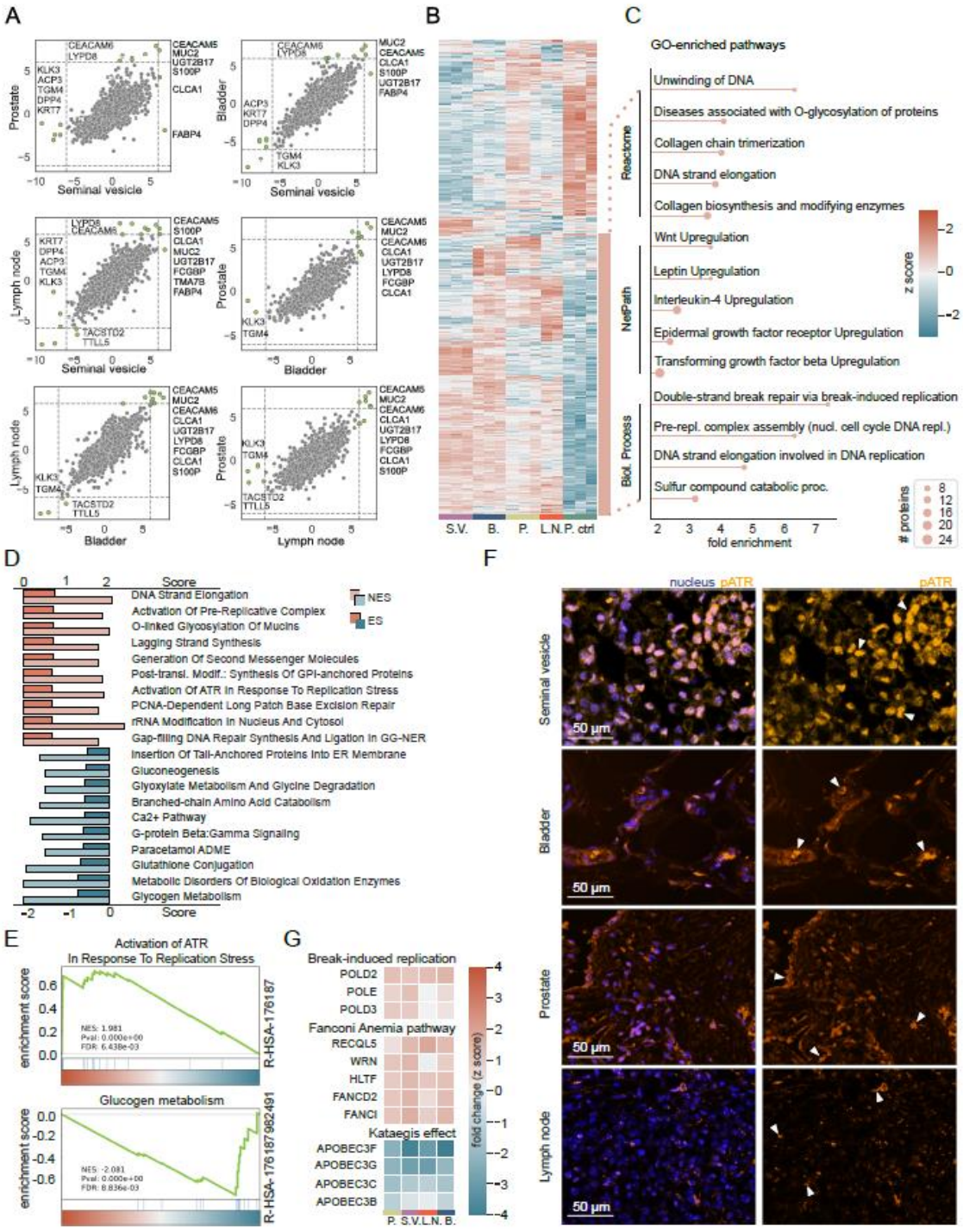


Figure 4 Proteomic profiling of signet ring cells in the context of DNA damage.

A Inter-organ fold change correlation plots, with emphasis on significant protein variations highlighted in green (cutoffs at ±6 fold-change). **B** Unsupervised hierarchical clustering of ANOVA significant proteins (premutation-based FDR <0.01, $s_0 = 0.1$). **C** GO term enrichment analysis of the bottom, upregulated cluster (in orange), highlighting the top five enriched pathways within Reactome, NetPath, and Biological Process. **D** Gene Set Enrichment Analysis (GSEA) of significantly positively and negatively enriched proteins after a pairwise proteomic comparison of SR cells to the epithelial cells of the prostate (two-sided t-test, FDR <0.01, $s_0 = 0.1$). Top ten pathways, sorted in a descending sequence according to their enrichment score (ES), with the corresponding normalized enrichment score (NES). **E** Two representative GSEA graphs, showing one positively and one negatively enriched pathway. **F** Representative images of SRCC-positive regions of the seminal vesicle, bladder, prostate and lymph node, stained for pATR and DAPI (nucleus). The auto-fluorescence signal of the mucus was initially used to identify SRCC-positive tumor regions. The scale bar is at 50 μm, and white arrows indicate strong pATR accumulation with the nuclei. **G** Heatmaps showing fold changes of proteins involved in 'break-induced replication', the 'Fanconi Anemia pathway' and the 'Kataegis effect'. SR cells of all four tissues (prostate, bladder, lymph node and the seminal vesicle) were compared to the epithelial cells of the prostate.

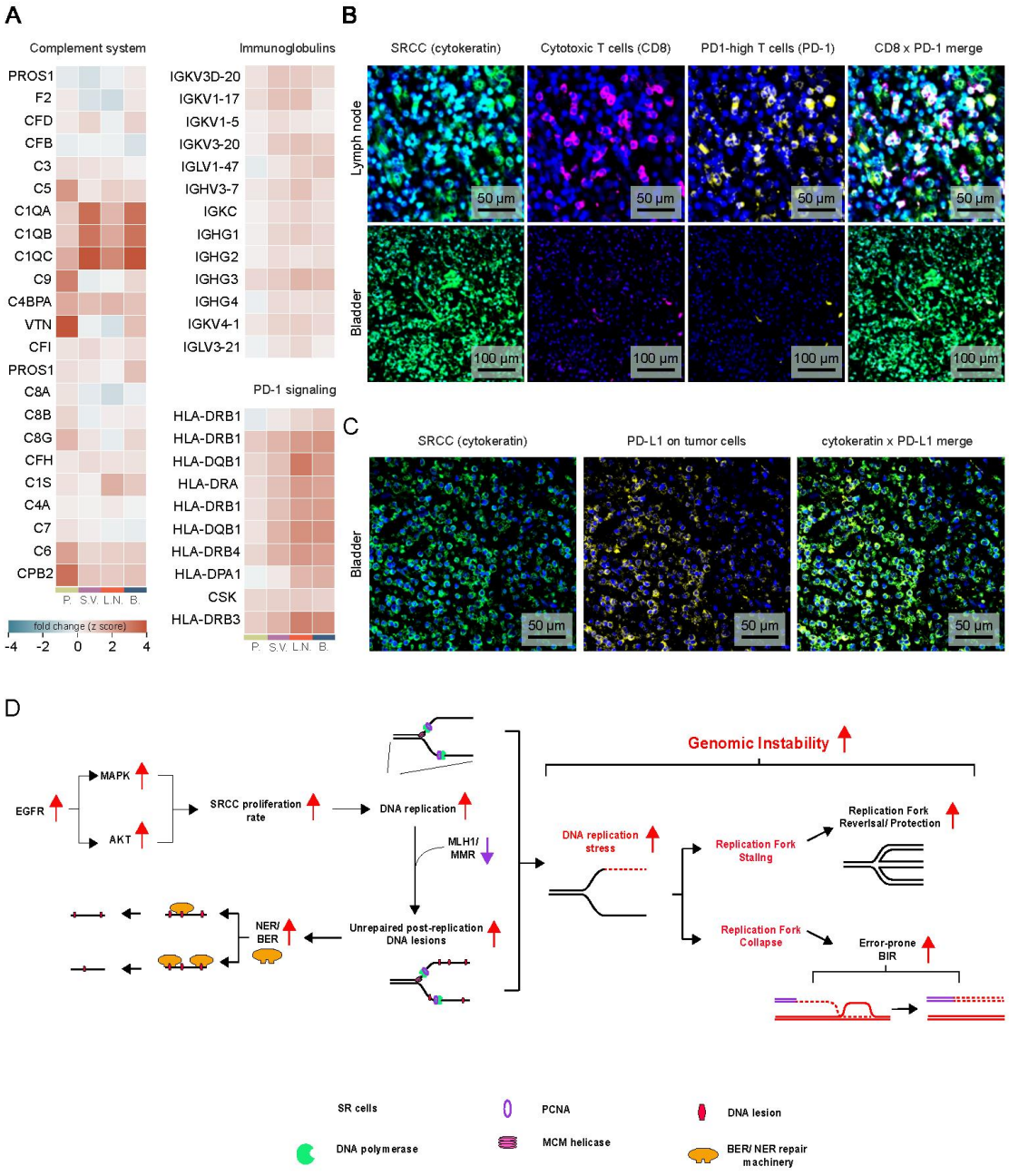


Figure 5 SRCC shows immunogenicity and cytotoxic T cell infiltration.

A Fold-changes of proteins involved in GO 'Complement system' pathway, 'immunoglobulins' and 'PD-1 signaling'. SR cells of all four tissues (prostate, bladder, lymph node and the seminal vesicle) were compared to the epithelial cells of the prostate. **B** Representative images of immunofluorescent-stained lymph node tissue and bladder for SR cells (cytokeratin, green), cytotoxic T cells (CD8, pink), the programmed death protein 1 (PD-1, yellow) and the nucleus (DAPI). **C** Representative image of the bladder tissue with SR cells (cytokeratin, green) and the programmed death protein ligand 1 (PD-L1, yellow). **D** Proposed model of SRCC DNA damage repair mechanisms and replication stress response.

Article 5: Deep Visual Proteomics advances human colon organoid models by revealing a switch to an *in vivo*-like phenotype upon xenotransplantation

Pre-print published online: bioRxiv (2024), doi: 10.1101/2024.05.13.593888, Submitted to Cell

Frederik Post[1,6], Annika Hausmann[2,6]*, Sonja Kabatnik[1], **Sophia Steigerwald[3]**, Alexandra Brand[2], Ditte L. Clement[2], Jonathan Skov[2], Theresa L. Boye[4], Toshiro Sato[5], Casper Steenholdt[4], Andreas Mund[1], Ole H. Nielsen[4], Kim B. Jensen[2,7]*, Matthias Mann[1,3,7]*

[1]*Proteomics Program, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark*
[2]*reNEW, Novo Nordisk Foundation Center for Stem Cell Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark*
[3]*Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany*
[4]*Department of Gastroenterology, Herlev Hospital, University of Copenhagen, Herlev, Denmark*
[5]*Department of Integrated Medicine and Biochemistry, Keio University School of Medicine*
[6]*equal contribution*
[7]*equal contribution*

Intestinal epithelial cells (IECs), organized in crypt-villus units with a stem cell niche in the crypt bottom, provide the intestinal mucosa's first line of defense against harmful luminal components and pathogens.[432] To maintain intestinal homeostasis and tissue integrity, the intestinal epithelial lining is renewed every 3-5 days.[433] Dysregulation of the gut homeostasis or intestinal epithelium, for instance caused by chemotherapy, food allergies or overuse of alcohol or aspirin, predispose to the development of inflammatory bowel disease and are hallmark symptoms of ulcerative colitis or Crohn's disease.[434–436] The study of IECs and epithelial maintenance, therefore, is vital for understanding gut health as well as preventing and treating these conditions.

Here the main authors Frederik Post and Annika Hausmann, aimed to evaluate the suitability of human colon organoid models to study human IECs by applying an optimized DVP workflow. To first establish a proteomic ground truth of the human colon mucosa, epithelial, goblet, immune cells, and fibroblasts were isolated from the upper crypt and crypt bottom. Spatially separating the crypt sections using our DVP pipeline, circumventing the otherwise challenging identification of intestinal stem cells with antibodies. Using the Orbitrap Astral MS, an unprecedented depth of almost 9,000 proteins across all cell types with a median of 6,780 proteins per sample could be achieved and revealed a number of differentially regulated proteins across the analyzed cell types and crypt localizations. While most samples consisted of about 500 dissected shapes, even samples of rare cell types, such as stem cells of xenotransplanted

organoids, yielded more about 5000 protein groups from fewer than 50 shapes. The proteomic analysis of the *in vitro* organoids revealed high overlap with the *in vivo* atlas, indicating the preservation of key features of the crypt bottom and the upper crypt. Despite the big overlap, the *in vitro* organoids showed high levels of proliferation and lacked functional signatures of the healthy human mucosa, such as secretion pathways. Further analysis revealed that this proliferative state was primarily driven by WNT pathway activation and could be shifted closer to more differentiated, functional states by reducing WNR supplementation in the culturing medium. These adjusted culture conditions improved the reliability of *in vitro* organoid models for studying IECs or conducting drug screenings.

Interestingly, these proliferative signatures were also reverted to a more *in vivo*-like state upon xenotransplantation of the organoids into the murine colon. Particularly, cells isolated from the upper crypt showed upregulation of CA1 and MUC17, proteins involved in ion transport and mucosal barrier formation respectively. This demonstrates that while organoids are already a powerful tool to study human IECs, the mucosal microenvironment is important to recapitulate functional characteristics and highlights the use of xenotransplantation to enhance organoid models.

Contribution

Co-authorship. The study was conceptualized by Frederik Post, Annika Hausmann, Kim. B Jensen and Matthias Mann. Frederik Post and Annika Hausmann conducted the experiments. I shared my established Orbitrap Astral methods and advised on the MS acquisition strategy, enabling the quantification of almost 9000 protein groups across all samples. Alongside the other co-authors, I contributed to revising and editing the manuscript.

.

Deep Visual Proteomics advances human colon organoid models by revealing a switch to an *in vivo*-like phenotype upon xenotransplantation

Frederik Post[1,6], Annika Hausmann[2,6*], Sonja Kabatnik[1], Sophia Steigerwald[3], Alexandra Brand[2], Ditte L. Clement[2], Jonathan Skov[2], Theresa L. Boye[4], Toshiro Sato[5], Casper Steenholdt[4], Andreas Mund[1], Ole H. Nielsen[4], Kim B. Jensen[2, 7*], Matthias Mann[1,3,7*]

[1] Proteomics Program, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark

[2] reNEW, Novo Nordisk Foundation Center for Stem Cell Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark

[3] Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany

[4] Department of Gastroenterology, Herlev Hospital, University of Copenhagen, Herlev, Denmark

[5] Department of Integrated Medicine and Biochemistry, Keio University School of Medicine

[6] equal contribution

[7] equal contribution

* correspondence: annika.hausmann@sund.ku.dk, kim.jensen@sund.ku.dk, mmann@biochem.mpg.de

Abstract

Intestinal epithelial damage predisposes to chronic disorders like inflammatory bowel disease. The organoid model allows cultivation, expansion and analysis of primary intestinal epithelial cells and has been instrumental in studying epithelial behavior in homeostasis and disease. Recent advances in organoid transplantation allow studying human epithelial cell behavior within the intestinal tissue context. However, it remained unclear how organoid transplantation into the colon affects epithelial phenotypes, which is key to assessing the model's suitability to study human epithelial cells. We employed Deep Visual Proteomics, integrating AI-guided cell classification, laser microdissection, and an improved proteomics pipeline to study the human colon. This created an in-depth cell type-resolved proteomics resource of human intestinal epithelial cells within human tissue, *in vitro* organoids, and the murine colon post-xenotransplantation. Our findings reveal that *in vitro* conditions induce a proliferative organoid phenotype, which was reversible upon transplantation and adjustment of organoid culturing conditions.

Introduction

The intestinal epithelium forms an integral barrier between the intestinal lumen, filled with microbiota and dietary components, and the lamina propria containing immune cells and fibroblasts. Continuous proliferation of epithelial stem cells located within the epithelial crypts ensures constant replenishment of intestinal epithelial cells (IECs). As stem cell progeny move towards the crypt top, they cease to divide and differentiate terminally, establishing a heterogeneous continuum along the crypt axis. These terminally differentiated IECs include absorptive colonocytes, mucus producing goblet cells, and hormone secreting enteroendocrine cells, which all perform key functions in intestinal physiology[1].

Epithelial maintenance is key for human health and requires tight molecular regulation balancing cell proliferation, differentiation and death. Murine models have provided substantial mechanistic insights into these intricate relations. There are, however, clear differences between the human and mouse, e.g. unique cell types identified in the human intestine[2], highlighting the need for human models. Addressing mechanistic questions in humans *in vivo* is challenging, and organoids[3–5] have emerged as an important model system to culture primary human cells and allow experimental manipulation. Human intestinal organoids have provided insights into e.g. cell fate choices, with applications in molecular medicine, drug testing and cellular therapies[1,6–9]. Conventional organoid culture features epithelial cells, but lacks other cell types present in the intestinal mucosa, such as immune cells and fibroblasts[1]. To address this limitation, orthotopic transplantation models have recently been developed[10,11]. They enable the transplantation of wild-type or genetically engineered mouse or human organoids into the murine colon to mechanistically dissect epithelial phenotypes within the mucosal microenvironment, which was previously only possible in mouse models. Furthermore, autologous transplantation of organoids into patients with impaired IEC phenotypes has great therapeutic potential in regenerative medicine, e.g. for inflammatory bowel disease (IBD) and short bowel syndrome[6,8]. This tractable xenotransplantation system enables the assessment of human IEC phenotypes in the mucosal microenvironment[11,12], but we still only have limited knowledge on how well human IECs transplanted into the murine colon recapitulate human IECs *in vivo*.

Fully leveraging the potential of human organoids requires in-depth characterization and validation of organoid models[1,13], which necessitates an accurate reference data set of their *in vivo* counterpart. Such a resource could guide future evaluation of disease-related changes, cellular and disease markers, and improvement of *in vitro* model systems. An accurate assessment of cellular phenotypes should account for their spatial context, especially in delicately organized tissues like the colon mucosa. Spatial transcriptomics and fluorescent *in situ* hybridization (FISH)-based techniques have provided valuable insights into the cellular heterogeneity of the colon[14,15]. These approaches, however, require pre-defined target panels and are biased by current knowledge. Single cell RNA-sequencing (scRNAseq), facilitates in-depth characterization of cellular phenotypes[16–18], but lacks spatial information. Typically, it also requires cellular dissociation

and long enrichment protocols, which in itself can impact epithelial phenotypes[19]. In the context of organoids, scRNAseq has been used to assess cellular composition[5,20], but in-depth phenotypic benchmarking including direct comparison to the *in vivo* counterparts remains limited, especially for the human colon.

Recent studies suggest that deep and sensitive proteomics provides more robust readouts for cellular states than transcriptomes, while directly pinpointing functional consequences of perturbation-induced changes[21,22]. The sensitivity of proteomics has advanced massively in the last decades from the quantification of a few thousand proteins from milligrams of input material in the beginning of the millennium to comparable numbers from single cells to date[22–24]. However, so far none of these methods have reached substantially complete coverage of cell type-specific proteomes. To address this, we here substantially further develop our Deep Visual Proteomics (DVP)[25] pipeline, which employs high-resolution fluorescence imaging, AI-guided cell segmentation and classification, single-cell isolation by laser capture microdissection, and high-sensitivity proteomics. To date, the conventional DVP pipeline generally yielded up to 5,000 proteins by combining a few hundred contours of single cell contours of the same type[25]. Our improved workflow using low flow gradients and the novel Orbitrap Astral analyzer[26], improved proteome coverage substantially, from even fewer contours. This allowed us to build a spatial proteome atlas of the human colon mucosa with unprecedented cell type-specific proteome depth. Importantly, the increased depth of protein quantifications at decreased input amounts enabled us to robustly and accurately benchmark human colon organoids grown *in vitro* and transplanted into the murine colon.

Our findings reveal that despite a robust correlation between *in vitro* and *in vivo* proteomes, IECs grown as organoids *in vitro* display high proliferation and low functional signatures. Strikingly, this is reverted upon xenotransplantation, rendering xenotransplanted human IECs a valuable tool to dissect human IEC phenotypes and illustrating that organoids retain their ability to reform colonic epithelium. Combined with iterative, proteomics guided improvements in organoid cell culture conditions this is a promising approach in regenerative medicine.

Results

DVP enables in-depth spatial proteomic profiling of cellular populations in the human colon

The assessment of human organoid models requires the determination of the *status quo* of the human colon mucosa. We made use of DVP (Fig. 1A) to generate a high sensitivity spatial proteome atlas of the human colon mucosa and analyze organoid models. In total, we analyzed 11 human colon tissue sections, 15 sections of organoids *in vitro*, and 50 sections of transplanted organoids.

The analysis of the human colon mucosa included different populations of colonic epithelial cells (EPCAM$^+$) and their microenvironment (lamina propria fibroblasts (PDGFRA$^+$), immune cells (CD45$^+$)) (Fig. 1B). Intestinal stem cells can be identified by *LGR5* expression[27], but it has proven difficult to generate antibodies for reliable detection of LGR5. Alternative strategies for isolating human intestinal stem cells have been developed based on expression of EPHB2[28,29], PTK7[30] and OLFM4[31], however, it remained challenging to detect epithelial stem cells in the human colon mucosa. We capitalized on the DVP technology to address this pertinent problem, enabling us to separate the epithelial crypt bottoms (enriched for stem cells, hereafter referred to as “crypt bottom”) from the upper part of the crypt (hereafter referred to as “upper crypt”) (Fig. 1C) based on spatial context. We used cellpose to segment high-resolution images for cell detection[32]. The resulting cell shapes and marker staining intensity were used to classify epithelial, goblet, immune cells and fibroblasts from the crypt bottom and upper crypt region using the biological image analysis software (BIAS) resulting in contours (one contour ≈ one cell in a 5 µm tissue section) (Fig. 1C). Technological limitations concerning availability of material and reliance on cellular markers for in-depth analysis of specific cellular subpopulations have so far hindered the characterization of functional states and phenotypes of human colonic epithelial cell subpopulations at protein levels. To address this, we isolated ~500 contours per population by laser capture microdissection, lysed the collected contours, digested the proteins and performed proteome acquisition on the Evosep One liquid chromatography system coupled to an Orbitrap Astral mass spectrometer (Experimental Methods). With this approach, we achieved unprecedented sensitivity of cell populations directly isolated from fresh-frozen tissue, featuring 8,865 unique proteins across all cell populations and a median of 6,780 unique proteins per sample with a throughput of 40 samples per day (Fig. 1D-E, S1A-B). The limited sample amount from transplanted organoids restricted us to collecting a maximum of 100 contours from transplanted stem cells and ~200 contours of transplanted epithelial cells in the upper crypt. Remarkably, the quantification of these samples still yielded ~5,000 or ~7,000 proteins, respectively (Fig. S1A).

Downstream principal component analysis (PCA) of the resulting data revealed that the samples from the human colon mucosa separated into two main clusters according to epithelium and lamina propria (immune cells and fibroblasts) along PC1, and further distributed according to the position along the crypt axis (bottom or top) along PC2 (Fig. 1F). To assess the reliability of identification and isolation of the different cell populations, we next assessed the abundance of previously described cellular markers for the isolated subpopulations in our sample set (Fig. 1C, G, S1C) and identified high expression of keratin (KRT)20 in upper crypt epithelial cells, Ephrin-type B receptor (EPHB)2 in crypt bottom epithelial cells, mucin (MUC)2 in goblet cells, thymocyte antigen (THY)1 in fibroblasts, as well as cluster of differentiation (CD)3E and human leukocyte antigen (HLA)-DRA in immune cells, thereby validating our human colonic mucosa proteome atlas. Interestingly, within the epithelial and lamina propria clusters, sample location along the crypt axis (upper crypt/bottom) rather than cell type drove their distribution (Fig. 1F). Differential activity of

WNT and BMP signaling along the crypt axis regulate cellular organization, proliferation and differentiation within the intestinal epithelium, suggesting that these pathways might partially drive observed differences. The protein transgelin (TAGLN) was associated with the crypt bottom compartment irrespective of the cell type (Fig. S1D). In line with high WNT activity around the epithelial stem cell niche in the crypt bottom, TAGLN$^+$ stromal cells have been identified as WNT producers[33]. The protein Zinc Finger ZZ-Type And EF-Hand Domain Containing (ZZEF)1, on the other hand, was enriched in the upper crypt compartment (Fig. S1D). ZZEF1 acts as a transcriptional regulator in cooperation with Krueppel-like factor (KLF)6 and KLF9[34] which regulate IEC proliferation[35] and absorption[36], and might be modulated by the intestinal microbiota[37], indicating a potential involvement in the integration of environmental stimuli into epithelial phenotypes. The interplay between luminal inputs and intrinsic regulation of mucosal gradients along the crypt axis and their molecular basis warrants further investigation.

In summary, we successfully generated a proteome atlas of the human colon mucosa in unprecedented depth with our DVP approach, which reveals differentially regulated protein levels along the crypt axis across cell types.

DVP analysis reveals a robust correlation between human IECs *in vivo* and grown as organoids

For an in-depth characterization of human colon organoids at proteome level, we adapted the DVP pipeline described above to organoids. The accurate and sensitive assessment of functional cellular states at proteome level within a spatial context in combination with the *in vivo* proteome atlas as reference data set enables the benchmarking of model systems for human IECs (Fig. 1). Here we made use of a genetically engineered human colon cell organoid line, expressing the fluorescent reporter TdTomato under the control of the LGR5 promoter[12] to identify epithelial stem cells (Fig. 2A). This allowed us to use the DVP workflow described above to identify, isolate and analyze human colonic stem cells (LGR5-TdTomato$^+$ cells, hereafter referred to as “stem cells”), LGR5-TdTomato$^-$ cells (hereafter referred to as “LGR5$^-$ cells”), and goblet cells to generate a proteome atlas of human IECs grown as organoids *in vitro*. It should be noted that the half-life of the reporter protein might be longer than LGR5, thus TdTomato$^+$ cells could contain a fraction of cells which have recently exited the stem cell state (e.g. transit amplifying progenitors). In the PCA, samples clustered according to different epithelial populations (Fig. 2B) with PC1 separating stem cells from the remaining IECs and PC5 separating goblet cells from stem cells and LGR5$^-$ IECs. Expectedly, KRT20 was enriched in the LGR5$^-$ cells (Fig. 2C), the stem cell marker EPHB2 in LGR5$^+$ stem cells and MUC2 in goblet cells (Fig, 2C).

A comparison of significantly changed proteins in stem versus LGR5$^-$ cells measured *in vitro*, and those measured *in vivo* in the crypt bottom versus upper crypt respectively, showed a robust correlation between the lower and upper crypt compartments *in vivo* and *in vitro* (Pearson coefficient 0.77, Fig. 2D), which is in a similar range to the correlation of transcriptomes of murine

small intestinal IECs *in vitro* and *in vivo* [38,39]. Notably, ~70% (crypt bottom) or ~60% (upper crypt) of significantly enriched proteins in the respective populations *in vivo* were shared with organoids grown *in vitro* (Fig. 2E). Among these, we identified a number of described markers associated with the analyzed populations, indicating that key aspects of crypt bottom and upper crypt epithelial cells are preserved in *in vitro* culture. The higher number of proteins identified as differentially abundant *in vitro* is likely due to more homogenous populations isolated from *in vitro* than *in vivo* conditions (e.g., LGR5$^+$ cells/crypt bottom), which allow for a more robust comparison.

To conclude, with our DVP approach we successfully benchmark human colon organoids to IECs *in vivo,* revealing a robust preservation of key compartment-associated features in organoids and highlighting their applicability as a model system for human colon IECs *in vitro*.

Orthotopic transplantation reverts organoid phenotypes to an *in vivo*-like state

The transplantation of human organoids into the murine colon emerges as a novel model to dissect human IEC phenotypes and behavior within the mucosal environment[10,11], but our current knowledge on how well human IECs transplanted into the murine colon recapitulate human IECs *in vivo* is limited to the assessment of selected markers for epithelial subpopulations[10,11]. To address this, we transplanted the genetically engineered human reporter organoids (Fig. 2A) into the murine colon (Fig. 3A). Consistent with previous reports, the cultured cells integrated into the murine colon mucosa and recapitulated the organotypic crypt structure featuring LGR5-TdTomato$^+$ cells at the crypt bottom (Fig. 3B)[10–12]. For a comprehensive, unbiased assessment of epithelial phenotypes upon transplantation, we performed DVP analysis on the transplanted cells, focusing on stem (LGR5-TdTomato$^+$, hereafter referred to as “stem cells”) and remaining cells (LGR5-TdTomato$^-$, hereafter referred to as “LGR5$^-$ cells”). Transplant size varies between mice and sometimes comprises only a few crypts. In protocols that require tissue dissociation (e.g. for scRNAseq), it can be challenging to efficiently recover these relatively rare cells. Furthermore, they often include lengthy enrichment steps such as cell sorting, which impacts IEC phenotypes[19]. For our DVP approach instead, we localized the transplants during sectioning, which enabled us to efficiently isolate transplanted IECs directly from their mucosal microenvironment. Strikingly, our DVP analysis revealed that transplanted organoids clustered with the *in vivo* IECs rather than organoid samples (Fig. 3C). This is particularly remarkable given that all organoid samples derive from the same organoid line (i.e., the same donor), while the IECs *in vivo* derive from three different donors, indicating that the phenotypic shift across conditions is stronger than interindividual differences.

To gauge the biological magnitude of this shift, we included the lamina propria cells (fibroblasts, immune cells) isolated from the colon mucosa *in vivo* as outlier groups into the PCA (Fig. S3A). Surprisingly, despite the robust correlation between IECs *in vivo* and *in vitro* observed above, the distance between IECs grown as organoids *in vitro* and *in vivo* was very similar to the distance along PC1 between lamina propria cells and IECs *in vivo*, which are different cell types. A major

driver for this differential clustering were components of the mucosal immunoglobulin A (IgA) (Fig. S3B), an important adaptive immune component of the mucosal barrier, which is secreted into the intestinal mucosa by B cells and subsequently transported into the intestinal lumen by IECs[40]. This indicates that the mucosal microenvironment has a significant impact on cellular proteomes across cell types, which should be considered when translating findings from organoid studies to *in vivo* phenotypes.

Collectively, the DVP analysis of orthotopically transplanted human colon organoids into the murine colon demonstrates that the cellular environment strongly impacts on IEC proteome profiles, pushing organoid phenotypes towards their *in vivo* counterparts.

To assess the cellular features driving phenotypic differences between IECs *in vitro* and within the mucosa, we performed a Kruskal-Wallis test across all epithelial samples. Hierarchical clustering of significantly changed proteins confirmed a separation of IECs grown *in vitro* from those isolated from the mucosa (*in vivo*, transplant) (Fig. 3D). Protein abundance patterns among these samples yielded eight clusters. Pathway analysis for the proteins within each cluster (Fig. 3E) revealed that signatures high in transcription (Cluster 4), translation (Cluster 1, 3), and proliferation (Cluster 5) characterized organoids cultured *in vitro* (partially shared with crypt bottom *in vivo* & transplanted stem cells), whereas *in vivo* and transplanted IECs were characterized by signatures associated with mucosal barrier function[41] (e.g. complement activation, Cluster 8), functional features of mature IECs (e.g. ion transport, secretion, Cluster 7), and oxidative phosphorylation (Cluster 6). A direct comparison between IECs *in vivo* and *in vitro* confirmed these observations (Fig. S3C-H). The increased proliferative features *in vitro* were also evident as a specific enrichment of proteins involved in proliferation in IECs *in vitro*, which was decreased upon transplantation to levels similar to *in vivo* (Fig. 3F)[42]. To identify markers associated with upper crypt IEC phenotypes *in vivo*, we next assessed the PC loadings to identify proteins that drive the separation between IECs *in vivo* and *in vitro* (Fig. S3B). Here, carbonic anhydrase (CA)1 and MUC17 were amongst the highest scoring proteins. CA1 mediates ion transport, which is key for the regulation of water absorption in the intestine[43]. MUC17 is a membrane mucin forming the glycocalyx, an important barrier against bacterial attachment to the mucosa, which is compromised in IBD[44] (Fig. S3I). In summary, components of two aspects of functional IECs *in vivo*, ion transport and barrier function, are underrepresented in IECs grown *in vitro* under the conditions tested here.

Altogether, our DVP approach revealed that the *in vitro* culturing conditions used here induce a high proliferation, low functional profile of IECs *in vitro*, and that these characteristics are reversible upon transplantation into the colon mucosa. This underscores the value of transplanted organoids as a system for the molecular dissection of epithelial phenotypes in a more *in vivo*-like setting, and highlights their applicability in regenerative medicine, e.g., for approaches to replenish impaired epithelium.

Integrated DVP analysis identifies a human stem cell signature

The use of fluorescent reporters has enabled studies of intestinal epithelial stem cells in mice *in vivo* and in genetically engineered human organoids *in vitro* but it has so far been difficult to specifically isolate and analyze human stem cells *in vivo* due to the lack of antibody-stainable stem cell markers. Our study design uniquely allowed the collective in-depth proteome analysis of LGR5-TdTomato^{+} human stem cells *in vitro* and upon xenotransplantation in comparison to stem cell-enriched human IECs *in vivo*. The comparisons across these datasets enabled us to identify a shared protein profile enriched in stem cells *in vitro* and upon transplantation, and crypt bottom cells *in vivo*, which were downregulated in upper crypt cells *in vivo*. This human stem cell proteome signature includes 48 proteins (Fig. 3G) and as expected, contains a number of proteins associated with cell proliferation. The assessment of the expression patterns of these proteins via the Human Protein Atlas[45] confirmed their localization at the crypt bottom *in vivo* (Fig. S4A-B). Notably, while all identified proteins localized within the stem cell niche, their abundance towards the crypt's upper part varied (Fig. S4A-B). Based on this, we postulate that markers with a relatively confined expression such as EPHB3, meiotic recombination 11 (MRE11) and minichromosome maintenance complex component 2 (MCM2) could be suitable markers for a strongly stem cell-enriched IEC population. In comparison to previously published markers for stem cell enrichment in the human colon such as PTK7, EPHB2 and OLFM4 [28,30,31], expression of these markers is more restricted to the crypt bottom (Fig. S4B). EPHB3 is a receptor tyrosine kinase involved in regulation of stem cell positioning along the crypt axis and regulates mitogenic activity in cooperation with WNT[29,46]. As an antibody-stainable surface protein, we expect it to be a valuable marker for the enrichment of human stem cells, e.g. in cell sorting, which would address a major technical gap. MRE11[47] and MCM2[48] regulate DNA double-strand break repair and DNA replication, respectively. Other markers such as PCNA, MCM3, MCM4 likely include transit amplifying populations as well, in line with their roles in cell division[49–51].

With this, our DVP approach has enabled the identification of EPHB3 as a potential novel surface marker for strong enrichment of stem cells, together with MRE11 and MCM2 as additional, antibody-stainable markers.

WNT withdrawal induces upregulation of *in vivo* IEC markers

The protocols for expansion of IECs as organoids have been optimized for growth at the expense of differentiation. This is achieved via activation of the WNT pathway (supplementation of signals activating the canonical WNT pathway – WNT surrogate and R-spondin1), which is active in the crypt bottom compartment *in vivo*, and inhibition of BMP signaling (supplementation of Noggin), which is active in the upper crypt compartment *in vivo*[3,5,52]. We hypothesized that these conditions could be drivers of the observed *in vitro* characteristics shaped by high proliferation and lower functional features when compared to the *in vivo* and transplanted IECs. In line with this,

both stem cells and LGR5^{-} cells *in vitro* were enriched for active WNT signaling[53] when compared to their *in vivo* counterparts (Fig. S5A-B).

To address the impact of WNT and BMP signaling on epithelial phenotypes, we cultured organoids *in vitro* under conventional (+WNT, Noggin, RSPO (WNR)) or differentiation (-WNR +/- BMP) conditions[12,54]. We observed a clear shift in organoid proteome profiles upon withdrawal of WNR while the addition of BMP only had a minor additional effect (Fig 4A). As hypothesized, WNR withdrawal led to a decrease in WNT activation (Fig. S5C-D). It furthermore induced a downregulation of stem cell- and proliferation-associated proteins such as SOX9, MKI67, MCM2 and PCNA (Fig. 4B-C). This was also evident at a more global level when we assessed expression of proteins assigned to the proliferation signature[42] and our stem cell signature identified above (Fig. 4D). At the same time, WNR withdrawal coincided with an upregulation of markers of mature IECs, such as KRT20, as well as CA1 and MUC17, which we identified in the analysis above as strongly associated with IECs *in vivo* (Fig. 4E). Similarly, the oxidative phosphorylation signature, which was enriched *in vivo* compared to organoids (Fig. S3F) was increased upon WNR withdrawal, indicating that IEC metabolic function is in part driven by IEC maturation state (Fig. 4F). Importantly, immunostaining of MUC17 in organoids upon WNR withdrawal revealed increased abundance of MUC17 at the apical surface, suggesting glycocalyx formation under these conditions. (Fig. 4G). Altogether, this indicates, as suggested previously, that withdrawal of WNR indeed drives organoids towards a more *in vivo*, upper crypt-like phenotype[55].

Discussion

We here employ DVP to generate an in-depth proteome atlas of the human colon mucosa, which we use to benchmark human colon organoids grown *in vitro* and upon orthotopic xenotransplantation. We originally developed DVP as a spatial proteomics technology that enabled the acquisition of the proteome of about 10 samples per day, quantifying up to 5,000 proteins from input material equivalent to 100 – 200 cells[25]. In our improved workflow, which includes coupling the Evosep One liquid chromatography system to the Orbitrap Astral analyzer, throughput is increased to 40 samples per day. Remarkably, total proteome acquisition time for this in depth, functional organoid study encompassing 136 samples was only 88 hours. Despite faster acquisition, we increased the proteome depth to a total of 8,865 unique proteins. This setup also enabled the quantification of ~5,000 proteins from as little as 100 transplanted stem cell contours, corresponding to only 20 intact cells. The increased proteome depth was essential to enable conducting this study since it enabled us to identify low abundant proteins such as SOX9 or LGR5 from cells dispersed over several slides.

Based on this improved DVP pipeline, the benchmarking of human colon organoids reveals a robust correlation of IECs grown *in vitro* and *in vivo*. Nevertheless, IECs grown *in vitro* display high proliferation and altered functional and metabolic signatures compared to *in vivo*, which has important implications for the use of organoids as models to dissect epithelial phenotypes. We

show that these features are driven by organoid culture conditions and are largely reverted upon organoid transplantation into the murine mucosa, as well as, in part, by altering organoid culturing conditions (WNR withdrawal). Altogether, our study validates the applicability of orthotopically xenotransplanted organoids as tools to mechanistically dissect human IEC phenotypes in an *in vivo*-like setting and highlights their potential to accurately replenish the intestinal epithelium in a regenerative medicine approaches.

Human organoid models are instrumental for assessing key biological questions in a human context. The premise that the organoid model truly recapitulates *in vivo* phenotypes, and an awareness of its limitations, is crucial for the translatability of *in vitro* results to *in vivo* applications. A key gap currently limiting the exploitation of the full potential of human organoids in biomedical research is the characterization and validation of organoids as accurate models for human biology[1,7,13]. An in-depth characterization of native IEC states within their *in vivo* environment is essential to establish a reference for benchmarking of human-like model systems. We have here tackled this issue, using our DVP approach to generate an in-depth proteome atlas of the homeostatic human colon, which serves as an important reference for future studies assessing e.g. disease-associated changes in the human colon. Notably, the DVP setup does not require fresh tissue dissociation and enrichment of living cells, which reduces the impact of lengthy isolation protocols on cellular phenotypes and thereby enabled us to assess the proteomes of mucosal cell types in their native state. We successfully identified and differentiated the isolated mucosal cell populations. Interestingly, aside from cell type-specific protein abundance patterns, we observed location-skewed protein abundance along the mucosal crypt axis. A similar zonation has been reported previously for murine small intestinal epithelial cells at transcriptome level[14], and it is well known that differences in e.g. WNT and BMP signaling along the crypt axis regulate epithelial phenotypes[1]. We here address this comprehensively across the different cell types in the mucosa at the protein level and identify the proteins ZZEF1 and TAGLN, which associate with the upper or crypt bottom compartment across the analyzed cell types, respectively. In the future, it will be interesting to study this protein regulation along the crypt axis in further detail and to dissect how e.g. WNT and BMP signaling gradients, as well as luminal cues such as microbiota shape protein abundance and cellular identity. This will shed light on the regulatory pathways maintaining tissue structures which are key for intestinal homeostasis and abrogated for example in the context of colorectal cancer[56,57].

Our human colon proteome atlas further enabled us to benchmark widely used *in vitro* [3–5] and emerging organoid transplantation models[10] for human IECs. Importantly, while we detect robust proteome correlation between IECs grown *in vitro* and their *in vivo* counterparts, which mirrors previous reports on transcriptome level in murine small intestine, we observe a striking phenotypic switch of organoids upon transplantation into the mucosa, rendering them *in vivo*-like. A major difference between organoids grown *in vitro* and transplanted into the mouse colon is a reduction in the proliferation signature, comparable to *in vivo* IECs, upon reintroduction into the mucosa. In

addition to the high proliferation state, organoids grown *in vitro* display lower functional features (e.g. ion transport), as well as a different metabolic signature characterized by lower oxidative phosphorylation. In the murine small intestine, oxidative phosphorylation has been linked to the regulation of stem cell identity and differentiation into Paneth cells[58]. We here find that proteins associated with oxidative phosphorylation are, at least in part, differentially regulated depending on epithelial maturation state. It remains to be shown whether this correlates with actual changes in metabolism between epithelial subpopulations, and whether/how epithelial differentiation and metabolism are linked in colonic IECs[59]. We further make use of our dataset to identify CA1 (ion transport/water homeostasis[43]) and MUC17 (glycocalyx in the brush border of differentiated IECs/barrier function[44]) as markers for human upper crypt IECs *in vivo*.

We show that high proliferation and low functional features observed in IECs grown *in vitro* are driven by the culture conditions (high WNT, low BMP signaling), rather than an intrinsic cellular feature selected for during culture, and that this state, including abundance of CA1 and MUC17, can be partially reverted by adjustments in culturing conditions (-WNR +BMP). Notably, recent advances in organoid-on-a-chip models using hydrogels which recapitulate the mucosal crypt structure and molecular gradients, feature similar IEC shifts to a more *in vivo*-like phenotype at transcriptome level[60]. These findings have important implications for the use of organoids to study IEC functions *in vitro*, especially when focusing on the role of upper crypt IECs, e.g. in host-microbe interactions.

The phenotypic reversion of organoids transplanted into the murine colon to a more *in vivo*-like phenotype highlights a remarkable homology between mouse and human stem cell niche factors. A more detailed analysis of the differences between transplanted organoids and IECs *in vivo* will reveal which molecular pathways drive the difference we observed between these two populations. One key aspect aside from the limited compatibility of mouse and human growth factor signaling could be the fact that we used immunocompromised mice for the xenotransplantation to prevent rejection. Future studies comparing human to murine organoids transplanted into the murine colon will be able to dissect the impact of species-specificity and the presence of immune cells on transplanted epithelial cells.

Finally, we capitalize on the unprecedented possibility to characterize human LGR5^{+} stem cells in the colon mucosa to identify a human stem cell proteome signature, which reveals EPHB3, MRE11 and MCM2 as antibody-stainable markers for the enrichment of human colonic stem cells *in vivo*. Notably, expression of these markers is more strongly restricted to the crypt bottom *in vivo* compared to previously published markers for the enrichment of human stem cells (EPHB2, PTK7, OLFM4). As EPHB3 is a surface protein, we expect that this marker will be of great value for the community to identify and isolate stem cell-enriched IECs for future studies of human intestinal stem cells. Furthermore, this showcases the strength of DVP to i) efficiently isolate rare cells from tissue in their native state and to ii) use proteome data to directly identify antibody-stainable markers. In addition, it serves as a proof-of-principle for the specific isolation and analysis of

genetically modified xenotransplanted human IECs from the murine colon and lays the base for future mechanistic studies, e.g. in the context of tissue damage and repair, and host-microbe interactions.

We here advanced the DVP pipeline, demonstrating that DVP is a uniquely well-suited methodology for the faithful in-depth analysis of functional cellular phenotypes in a densely packed tissue like the colon mucosa. Limited sensitivity has so far been a major difficulty for the use of proteomics to dissect dynamic tissue processes, especially in the context of tightly regulated responses such as inflammation (i.e., low abundant, spatially restricted proteins). An additional limitation has been the ability to isolate cells in a near to native state, in the absence of alterations by tissue handling including single cell isolation. The DVP protocol we use here tackles these hurdles, enabling higher throughput and requiring less input material than the original method, and preserving spatial context while reducing the impact of isolation protocols on cellular phenotypes. These technological advancements are promising regarding the expansion of DVP for the acquisition of proteomes of single cells[61]. This opens exciting perspectives for the use of DVP to study dynamic tissue processes such as inflammation, even from rare patient material.

Taken together, the presented data has important implications for the selection of *in vitro* organoid systems to study specific aspects of epithelial cell biology. The phenotypic reversion of organoids transplanted into the murine colon to a more *in vivo*-like phenotype highlights the impressive homology between mouse and human stem cell niche factors, underlines the suitability of the murine (orthotopic transplantation) model for studies of epithelial-niche interactions with a translational perspective and opens exciting possibilities for the use of organoid transplantation in regenerative medicine.

Methods

Human colon mucosa samples

All individuals included in this study were attending the Department of Gastroenterology, Herlev Hospital, University of Copenhagen, Denmark, for the Danish National Screening Program for Colorectal Cancer or evaluated for various gastrointestinal symptoms but were included only if all subsequent examinations were normal. The exclusion criteria included age below 18 or over 80 years; impaired cognitive functions, e.g., dementia; pregnant or lactation women; ongoing treatment with anticoagulation, and patients unable to understand Danish language. The study was approved by the Scientific Ethics Committee of the Capital Region of Denmark (reg. no. H-21038375). All individuals were informed of the study both orally and in writing, in compliance with the Declaration of Helsinki and the guidelines of the Danish National Scientific Ethics Committee. Written informed consent was obtained prior to inclusion.

For those individuals included, human colon mucosa samples (cancer-associated bowel resection or biopsies (healthy individuals undergoing cancer screening)) were immediately transferred to 4% PFA (Sigma) upon sampling and fixed at 4 °C for 2-10 days, depending on

sample size. Samples were then washed in PBS and transferred to 30% sucrose/PBS and dehydrated for 2-10 days at 4 °C. Next, samples were embedded in OCT, frozen on dry ice and stored at -80C until further analysis.

Human colon organoid culture

Human colon organoids were cultured as previously described[5]. Briefly, upon single cell dissociation, 3,000 – 4,000 single cells were seeded in 30 µL Matrigel domes and maintained in advanced DMEM/F-12, supplemented with penicillin-streptomycin, 10 mM HEPES, 2 mM GlutaMAX, 100 ng/mL recombinant mouse Noggin, 1x B27, 500 nM A83-01, 1% NGS-WNT, 1 mg/mL recombinant human R-spondin-1, 100 ng/mL recombinant human IGF, 50 ng/mL recombinant human FGF2, 1 mM N-Acetylcysteine and 10 nM recombinant human Gastrin. For WNR withdrawal, organoids were cultured in conventional medium until d7. Organoids were then reseeded in fresh Matrigel domes (no splitting) and maintained until d10 in advanced DMEM/F-12, supplemented with penicillin-streptomycin (Penstrep), 10 mM HEPES, 2 mM GlutaMAX, 1x B27, 500 nM A83-01, 100 ng/mL recombinant human IGF, 50 ng/mL recombinant human FGF2, 1 mM N-Acetylcysteine and 10 nM recombinant human Gastrin in the presence or absence of BMP4 (10 ng/ml). Organoids were split every 7d for maintenance. Organoids were harvested at d10 for the analyses presented in this study. Human colon organoids from healthy individuals have been used for this study. The LGR5-TdTomato reporter organoid line has been described before[12]. To introduce a constitutive GFP reporter to the cells for easier localization of the transplant, eight wells (i.e. eight 30 µL Matrigel domes) of organoids were mechanically disrupted, washed and resuspended ~600 µL media supplemented with Y-27632 (10 uM). Lenti virus was added to the cells to transduce them with a plasmid expressing GFP under the SFFV promotor[62]. The cells were incubated for 4 h at 37 °C, washed three times in DMEM medium and subsequently seeded into four Matrigel domes (30 µL). After three days of culture, transduced cells were selected by addition of 2 ug/ml Puromycin to the media. Cells were passaged twice, tested according to FELASA standards (IDEXX), and subsequently used for transplantation.

For cryosamples, 500 µL ice cold cell recovery solution was added to each well. Matrigel domes were carefully scraped off with a cut open P1000 pipet tip and transferred to 5 ml cell recovery solution (R&D systems) on ice. After 30 min, the supernatant was removed, organoids were resuspended in 4% PFA and fixed for 1h at ambient temperature. Subsequently, organoids were washed three times in 5 ml PBS (if necessary, organoids were spun down for 2 min at 100g), embedded in OCT (Tissue Tek) in cryomolds, and frozen on dry ice. Samples were stored at -80 °C until further analysis.

For bulk proteome analysis, organoids were harvested as previously described[63]. Briefly, 1 ml ice cold 0.1% BSA/PBS was added to each well and matrigel domes were broken up by pipetting 10 times with a P1000 pipet. Organoids from four wells were pooled per sample in a tube containing 3 ml 0.1% BSA/PBS. Cells were pelleted by centrifugation (5 min, 300 x *g*, 4 °C), supernatant was

491 removed and cells were resuspended in 1 ml 0.1% BSA/PBS and pelleted again. Upon removal of
492 the supernatant, cells were resuspended in 200 µL 0.1% BSA/PBS and transferred to a 1.5 ml
493 Eppendorf tube (pre-coated with 0.1% BSA/PBS) and kept on ice until further processing.
494
495 **Orthotopic xenotransplantation**
496 NOD.Cg-Prkdc^{scid} Il2rg^{tm1Sug}/JicTac (NOG) mice were used for transplantation assays. All
497 animal procedures were approved by the Danish Animal Inspectorate (license number 2018-15-
498 0201-01569 to Kim B. Jensen).
499 In preparation of the transplantation, organoids were grown as described above until d5-6 in 6-
500 well plates containing nine Matrigel domes per well. 3 ml ice cold cell recovery solution was added
501 to each well. Matrigel domes were carefully scraped off with a cut open P1000 pipet tip and
502 transferred to 5 ml cell recovery solution (R&D systems) on ice for 20 min. Cells were subsequently
503 pelleted for 3 min at 300 x *g*, washed once in PBS and resuspended in 200 µL of 5% Matrigel/PBS
504 per mouse. Right before transplantation, organoids were dissociated by pipetting 20x with a pre-
505 wet P1000 pipette.
506 Transplantation was performed as described previously[11], with slight modifications. Mice were
507 anesthetized with 2% isoflurane before the procedure. The colon content was flushed with PBS
508 and an electric interdental brush, soaked in prewarmed 0.5 M EDTA, was used to brush crypts off
509 on one side of the colon. The organoids suspension was subsequently infused into the conditioned
510 colon. Glue (Histo-acryl, B. Braun) was added to the anal verge and left for 3h to avoid the ejection
511 of the organoid suspension and thereby enhance the engraftment of the infused material. Mice
512 were monitored daily. Transplanted samples were isolated six weeks after transplantation. For
513 cryosectioning, the colon was isolated, cut open and placed under a fluorescent microscope (Evos)
514 to locate GFP$^+$ transplanted cells. The colon area containing the transplant was subsequently cut
515 out, fixed in 4% PFA at 4 °C over night, dehydrated in 30% sucrose/PBS over night at 4 °C and
516 then embedded in OCT and frozen on dry ice. Samples were kept at -80 °C until further analysis.
517
518 **Cryosectioning, immunofluorescent staining and imaging for DVP**
519 2-mm-thick polyethylene naphthalate membrane slides (Zeiss) were pretreated by ultraviolet
520 ionization for 3 h. Without delay, slides were consecutively washed for 5 min each in 350 ml
521 acetone and 7 ml VECTABOND reagent to 350 ml with acetone, and then washed in ultrapure
522 water for 30 s before drying in a gentle nitrogen air flow. The slides were treated with a dilution of
523 7 mL Vectabond in 350 mL acetone for 5 minutes without prior washing in acetone or subsequent
524 washing in water. Afterwards, the slides were dried in an incubator at 30 °C for 3 hours.
525 Frozen samples in OCT were cut with a Leica cryostat in 5 um sections. Samples were
526 subsequently dried for 1h at ambient temperature, rehydrated with 500 µL PBS for 1 min and
527 permeabilized with 300 ul PBS/0.5% TritonX-100. Tissue sections were blocked in 200 µL
528 PBS/donkey serum for 30 min at room temperature and subsequently incubated with the primary

antibody mix in blocking buffer overnight at 4C. The next day, samples were washed three times with 500 µL PBS and incubated for 40 min at ambient temperature with the secondary antibody mix in PBS. Upon washing three times with PBS, samples were mounted using anti-fade fluorescence mounting medium (abcam). Samples were subsequently imaged as described below and, if necessary subjected to a second round of staining. For this, samples were bleached using bleaching buffer (24 mM NaOH and 4.5% H2O2) for 10 min at room temperature, washed with PBS and stained as above.

Antibodies and staining reagents used in this study: CD45-BV421 (30-F11, Biolegend, 1:100), Lrig1 (R&D Systems AF3688, 1:50), PDGFR (EPR22059-270, abcam, 1:100), UEA-Atto550 (Atto-Tec, 1:500), EPCAM-APC (EBA1, BD Biosciences, 1:50), EPCAM-APC (G8.8, Fisher Scientific,1:50), ECAD (ECCD2, Thermo Fisher, 1:200), CD45 (HI30, Stem cell, 1:200), DAPI (Sigma), MUC17 (Merck HPA031634, 1:200), CA1 (EPR5193, abcam, 1:200), Pan-Laminin-AF647 (Novus Biologicals NB300-144AF647, 1:100).

The samples were imaged on a Zeiss Axioscan 7 microscope slide scanner at a magnification of 20×, with three z-layers with intervals of 2.5 mm. Human colon tissues were imaged in two secutive rounds. For the first round, the acquisition settings were 4 ms illumination time and 1.49% 385 nm laser for DAPI, 20 ms illumination time and 100% 475 nm laser for AF488, and 300 ms illumination time and 100% 735 nm laser for AF750. For the second round, the acquisition settings were 4 ms illumination time and 1.49% 385 nm laser for DAPI, 15 ms illumination time and 100% 475 nm laser for AF488, 60 ms illumination time and 100% 567 nm laser for AF568, and 20 ms illumination time and 100% 630 nm laser for AF647. For *in vitro* organoids, the acquisition settings were 2 ms illumination time and 1.1% 385 nm laser for DAPI, 2.2 ms illumination time and 100% 475 nm laser for FITC, 30 ms illumination time and 100% 567 nm laser for Rhoda, and 8 ms illumination time and 100% 630nm laser for AF647. Transplanted organoids were imaged in two staining rounds. The first round was imaged with an illumination time of 1.2ms and 1.5% 385 nm laser for DAPI, 3 ms illumination time and 100% 475 nm laser for Af488, 80 ms illumination time and 100% 567 nm laser for tdTomato, 20 ms illumination time and 100% 630 nm laser for Af647, and 100 ms illumination time and 100% 735 nm laser for Af750.

Image Analysis

Corresponding images of the two imaging rounds were cropped and subsequently concatenated in imagej. Afterwards, the images were registered using the RigidBody transformation in HyperStackReg on the GFP and tdTomato channel in the transplanted organoids and DAPI in the *in vivo* human colon, and all channels were merged.

Images were split into tiles using the Biological Image Analysis Software (BIAS , Single-Cell Technologies Ltd.) and each tile was segmented in Napari using the cellpose cytosolic algorithm in the serialcellpose plugin. Images were not treated as RGB, batch size was set to 3, flow threshold was set to 3, cell probability threshold was set to –4, diameter was set to 30, the magenta

channel was set as channel to segment, and the yellow channel was used as a helper channel. Image analysis was continued in in BIAS by filtering shapes for a minimum size of 50 µm^2 and a maximum size of 2000 µm^2. Features of segmented cells were extracted and classified using a multi-layer perceptron classifier with default settings. For human colon tissue, the bottom part of crypts was manually annotated using the region feature to distinguish stem cells and differentiated epithelium. Contours of cells were sorted using the "Greedy" setting and coordinates of the contours were exported.

Laser Microdissection

Contours were imported at 63× magnification, and laser microdissection performed with the LMD7 (Leica) in a semi-automated manner at the following settings: power 46, aperture 1, speed 40, middle pulse count 4, final pulse 8, head current 46-50%, and pulse frequency 2,600. Contours were sorted into a low-binding 384-well plate (Eppendorf 0030129547). 500 contours were collected per sample except for immune cells surrounding upper crypt of which 700 contours were collected. Due to limited sample amount in the transplanted organoids, 200 contours were collected for differentiated cells and about 100 contours were collected for stem cells. An overview of collected biological replicates and technical replicates per cell population can be found in the supplementary data (Table S1). Contours were rinsed to the bottom of the well by filling the wells up with 40 mL acetonitrile, vortexing for 10 seconds, and centrifuging at 2000 x *g* at ambient temperature for 5 min. A SpeedVac was used to evaporate the acetonitrile at 60 °C for 20 min or until achieving complete dryness and the contours were stored at 4 °C.

DVP proteome sample preparation and acquisition

Lysis was performed in 4 mL of 0.01 % n-dodecyl-beta-maltoside in 60 mM triethyl ammonium bicarbonate (TEAB, pH 8.5, Sigma) at 95 °C in a PCR cycler with a lid temperature of 110 °C for 1 h. 1 mL of 60% acetonitrile in 60 mM TEAB was added and lysis continued at 75 °C for 1 h. Proteins were first digested with 4 ng LysC at 37 °C for 3 h and subsequently digested overnight using 6 ng trypsin at 37 °C. The digestion was terminated by adding 1.5 mL 5 % TFA. Samples were dried in a SpeedVac at 60 °C for 40 min and stored at -80 °C.

C-18 tips (Evotip Pure, EvoSep) were washed with 100 µL of buffer B (0.1% formic acid in acetonitrile), activated for 1 min in 1-propanol, and washed once with 20 µL buffer A (0.1% formic acid). Samples were resuspended in 20 mL buffer A on a thermoshaker at room temperature at 700 x *g* for 15 min. Peptides were loaded on the C-18 tips, washed with 20 mL buffer A, and then toped up with 100 mL buffer A. All centrifugation steps were performed at 700 x *g* for 1 min, except peptide loading at 800 x *g* for 1 min.

Samples were measured with the Evosep One LC system (EvoSep) coupled to an Orbitrap Astral mass spectrometer (Thermo Fisher). Peptides were separated on an Aurora Elite column (15 cm x 75 mm ID with 1.7 mm media, IonOpticks) at 40 °C running the Whisper40 gradient. The

mobile phases were 0.1% formic acid in liquid chromatography (LC)–MS-grade water (buffer A) and 0.1% formic acid in acetonitrile (buffer B). For samples consisting of 500 contours, the Orbitrap Astral MS was operated at a full MS resolution of 240,000 with a full scan range of 380 − 980 m/z. The AGC target was set to 500% for full scans and fragment ion scans. Fragment ion scans were recorded with a maximum injection time of 5 ms and with 300 windows of 2 Th scanning from 150 − 2000 m/z. Fragmentation of precursor ions took place using HCD with 25% NCE. Samples consisting of 200 contours (stem cells from transplanted organoids) were acquired using a full maximum injection time of 100 ms for MS1. Fragment ion scans were recorded with a maximum injection time of 14 ms (MS2), an AGC target of 800 %, and with 75 windows of 8 Th scanning from 150 − 2000 m/z.

DVP raw MS data analysis

Raw files were converted to mzML using MSconvert and analyzed in DIA-NN 1.8.1 using an in-silico DIA-NN predicted spectral library (101370 protein isoforms, 177027 protein groups and 7821224 precursors in 3872218 elution groups)[64]. A human proteome reference database, including isoform information and the tdTomato fluorophore sequence, was used to generate the library and search the raw files (Uniprot March 2023). Following configuration was set for the search: N-terminal methionine excision was enabled, digest was performed at K* and R*, maximum number of missed cleavages was set to 2, maximum number of variable modifications was set to 2, oxidation of methionine was considered as variable, acetylation of the N-terminus was considered as variable, Protein inference = “Genes”, Neural network classifier = “Single-pass mode”, Quantification strategy = “Robust LC(high precision)”, Cross-run normalization = “RT-dependent”, Library Generation = “Smart profiling”, and Speed and RAM usage = “Optimal results”. Mass accuracy and MS1 accuracy were set to 15. “Use isotopologues”, “No shared spectra”, “Heuristic protein inference” and “MBR” were activated.

DVP data analysis

Data analysis was mostly performed in Perseus and AlphaPeptStats [65,66]. Python and R were used to conduct further analyses and visualize the data. The first technical replicate of the second biological replicate of fibroblasts at the bottom of crypts (fib_top_02_01) was removed due to the quantification of less than 2000 proteins. Raw data was imported into Perseus, and proteins filtered for 80 % data completeness within samples of the same cell type and same location in the human tissue. Missing values were replaced from a normal distribution with a width of 0.3 and a down shift of 1.3. Data was normalized by aligning the median intensity of all samples. Median intensities of each sample were determined, and the median of these median intensities was divided by the median of each sample. The resulting factor was multiplied with each intensity of the sample. Differential abundance analyses for volcano plots and enrichment analyses were performed in Perseus. Kruskal-Wallis tests were performed in Perseus with Benjamini-Hochberg FDR correction

and a threshold of 0.01. GSEAs were performed using the GSEApy (v 1.0.6) package against the GO_Biological_Process_2023 dataset[67,68].

Bulk Proteome sample preparation and acquisition

200 mL 60 mM TEAB lysis buffer was added to the washed and pelleted organoids. Samples were lysed at 95 °C shaking at 800 rpm for 30 min. Afterwards the lysate was sonicated at 4 °C in 30 s intervals for 10 min. 18 mL ACN was added to bring the lysis buffer to a final concentration of 12.5 % ACN and lysis continued at 95 °C shaking at 800 rpm for another 30 min. Debris was pelleted at 4 °C at 20,000 x g for 10 min and supernatants transferred to fresh tubes. Protein concentration of supernatants was determined using nanodrop and 200 mg were used for further processing. Lys-C and trypsin were added at a protein to enzyme ratio of 50:1. Digestion took place at 37 °C shaking at 800 rpm overnight. Peptides were lyophilized using a SpeedVac at 60 °C for 1 hour. Peptides were resuspended in 200 mL Evosep buffer A (0.1 % formic acid) and 60 mL corresponding to 60 mg were loaded in triplicates on 3 layers of SDB-RPS membranes. About 10 ng were loaded on Evotips Pure.

Samples were measured with the Evosep One LC system (EvoSep) coupled to an Orbitrap Astral mass spectrometer (Thermo Fisher). Peptides were separated on an Aurora Elite column (15 cm x 75 mm ID with 1.7 mm media, IonOpticks) at 40 °C running the Whisper40 gradient. The mobile phases were 0.1% formic acid in liquid chromatography (LC)–MS-grade water (buffer A) and 0.1% formic acid in acetonitrile (buffer B). The Orbitrap Astral MS was operated at a full MS resolution of 240,000 with a full scan range of 380 − 980 m/z and a maximum injection time of 100 ms. The AGC target was set to 500% for full scans and fragment ion scans. Fragment ion scans were recorded with a maximum injection time of 5 ms and with 300 windows of 2 Th scanning from 150 − 2000 m/z. Fragmentation of precursor ions took place using HCD with 25% NCE.

Bulk proteome raw MS data analysis

Raw files were converted to mzML using MSconvert and analyzed together with the DVP samples in DIA-NN 1.8.1 using an in-silico DIA-NN predicted spectral library (101370 protein isoforms, 177027 protein groups and 7821224 precursors in 3872218 elution groups)[64,69]. A human proteome reference database, including isoform information and the tdTomato fluorophore sequence, was used to generate the library and search the raw files (Uniprot March 2023). Following configuration was set for the search: N-terminal methionine excision was enabled, digest was performed at K* and R*, maximum number of missed cleavages was set to 2, maximum number of variable modifications was set to 2, oxidation of methionine was considered as variable, acetylation of the N-terminus was considered as variable, Protein inference = “Genes”, Neural network classifier = “Single-pass mode”, Quantification strategy = “Robust LC(high precision)”, Cross-run normalization = “RT-dependent”, Library Generation = “Smart profiling”, and Speed and

RAM usage = “Optimal results”. Mass accuracy and MS1 accuracy were set to 15. “Use isotopologues”, “No shared spectra”, “Heuristic protein inference” and “MBR” were activated.

Bulk proteome data analysis

Data analysis was mostly performed in Perseus and AlphaPeptStats. Python and R were used to conduct further analyses and visualize the data. Raw data was imported into Perseus, and proteins filtered for 80 % data completeness within samples of the same cell type and same location in the human tissue. Missing values were replaced from a normal distribution with a width of 0.3 and a down shift of 1.3. Differential abundance analyses for volcano plots and enrichment analyses were performed in Perseus and visualized in python and R. GSEAs were performed using the GSEApy (v 1.0.6) package against the GO_Biological_Process_2023 dataset.

Author contributions

Conceptualization: FP, AH, AM, KJB, MM. Experimentation: FP, AH, SK, SS, AB, DLC, JS. Reagents and material: TLB, TS, CS, OHN. Writing – original draft: FP, AH, KJB, MM. Writing – review and editing: all authors.

Acknowledgements

The authors thank Daniela Mayer and Hjalte L. Larsen for critical reading of the manuscript, Kira Petzold for preparing pretreated membrane slides, Xiang Zheng for providing staining protocols, and the Mann and Jensen groups for fruitful discussions. We acknowledge support by the CPR/reNEW imaging facility as well as Core Facility for Microscopy. This work was funded by grants from the Novo Nordisk Foundation (NNF14CC0001, NNF15CC0001 to MM and NNF18OC0034066, NNF20OC0064376 to KBJ)) and the Independent Research Fund Denmark (0134-00111B) to KBJ. Additionally, SK and FP were supported by the Novo Nordisk Foundation grant NNF20SA0035590 and NNF0069780. AH acknowledges funding by EMBO (ALTF 179-2021) and the European Crohn´s and Colitis Organization (PROP-1495). The Novo Nordisk Foundation Center for Stem Cell Medicine was supported by a Novo Nordisk Foundation grant (NNF21CC0073729). MM is funded by the Max Planck Society for the Advancement of Science.

Conflicts of interest

CS lectures for MSD and Janssen-Cilag and received a research grant from Takeda. MM is an indirect shareholder in Evosep.

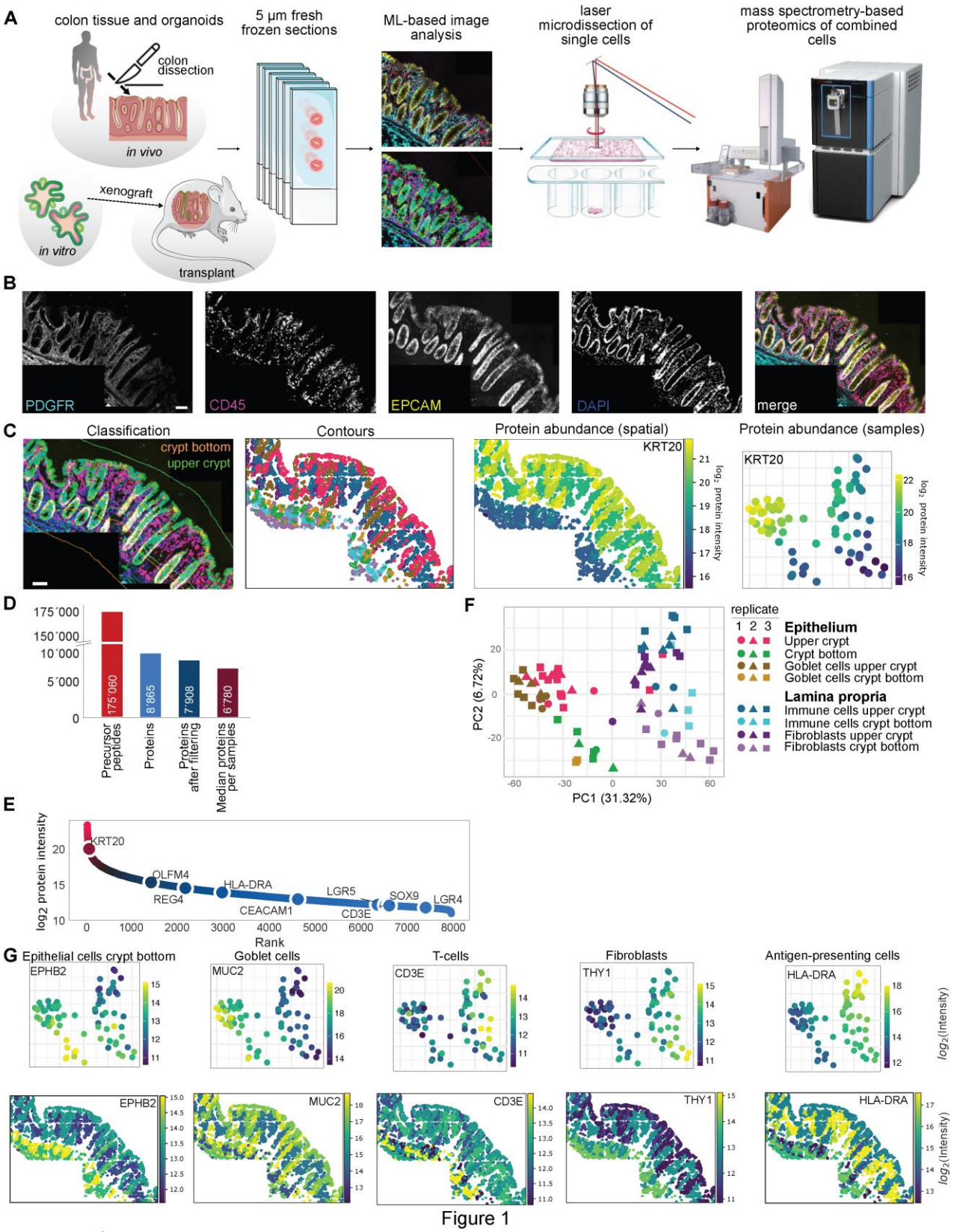

A
colon tissue and organoids
5 µm fresh frozen sections
ML-based image analysis
laser microdissection of single cells
mass spectrometry-based proteomics of combined cells
colon dissection
in vivo
xenograft
in vitro
transplant
B
PDGFR
CD45
EPCAM
DAPI
merge
C
Classification
crypt bottom
upper crypt
Contours
Protein abundance (spatial)
KRT20
log2 protein intensity
Protein abundance (samples)
D
175´000
150´000
10´000
5´000
0
175´060
8´865
7´908
6´780
Precursor peptides
Proteins
Proteins after filtering
Median proteins per samples
F
replicate
1 2 3
Epithelium
Upper crypt
Crypt bottom
Goblet cells upper crypt
Goblet cells crypt bottom
Lamina propria
Immune cells upper crypt
Immune cells crypt bottom
Fibroblasts upper crypt
Fibroblasts crypt bottom
PC2 (6.72%)
PC1 (31.32%)
E
log2 protein intensity
KRT20
OLFM4
REG4
HLA-DRA
CEACAM1
LGR5
CD3E
SOX9
LGR4
Rank
G
Epithelial cells crypt bottom
Goblet cells
T-cells
Fibroblasts
Antigen-presenting cells
EPHB2
MUC2
CD3E
THY1
HLA-DRA
log2(Intensity)

Figure 1

Figure 1: DVP analysis faithfully assesses cellular heterogeneity in the human colon
A Study design for the validation of organoids *in vitro* and organoid transplantation using Deep Visual Proteomics. **B** Immunofluorescence image of the human colon mucosa stained for fibroblasts (PDGFR), immune cells (CD45) and epithelial cells (EPCAM). **C** Crypt bottom and upper crypts were defined by a manually drawn line. Single cells were segmented and classified, contours exported, microdissected, and analyzed. This analysis reveals protein abundance across the colon mucosa and cell populations, as exemplified here for KRT20, a marker of differentiated epithelial cells. **D** Protein and precursor peptide identifications across all samples. **E** Median dynamic range of identified proteins across all samples after imputation and normalization. **F** PCA plot of samples isolated from the colon mucosa (three donors) as indicated by classification in C. **G** Protein abundance and spatial distribution of previously described cell type markers for different subpopulations in the human colon.

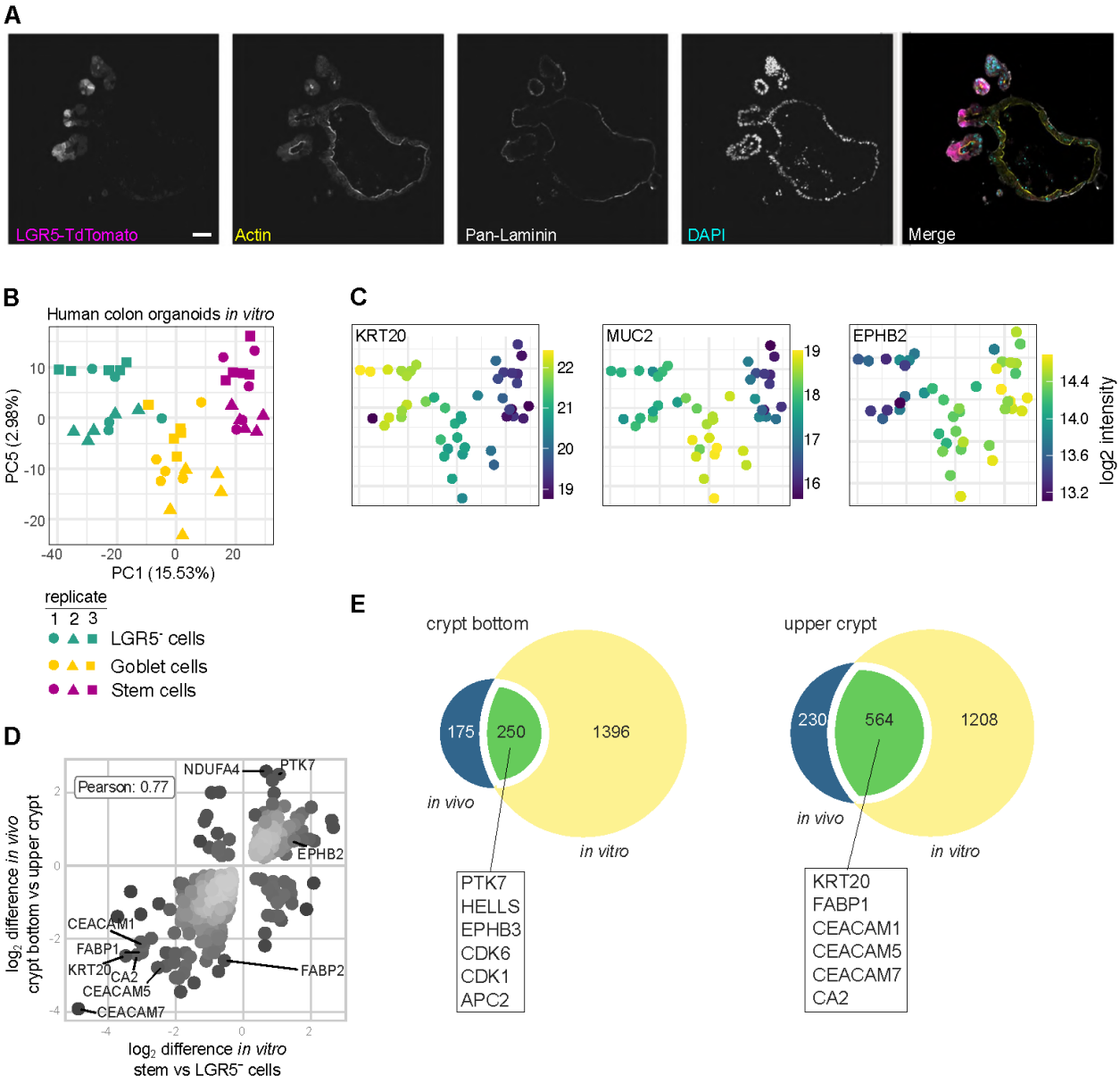

A
LGR5-TdTomato
Actin
Pan-Laminin
DAPI
Merge
B
Human colon organoids in vitro
PC5 (2.98%)
PC1 (15.53%)
replicate
1 2 3
LGR5- cells
Goblet cells
Stem cells
C
KRT20
MUC2
EPHB2
log2 intensity
D
Pearson: 0.77
NDUFA4
PTK7
EPHB2
CEACAM1
FABP1
KRT20
CA2
CEACAM5
CEACAM7
FABP2
log2 difference in vivo crypt bottom vs upper crypt
log2 difference in vitro stem vs LGR5- cells
E
crypt bottom
175
250
1396
in vivo
in vitro
PTK7
HELLS
EPHB3
CDK6
CDK1
APC2
upper crypt
230
564
1208
in vivo
in vitro
KRT20
FABP1
CEACAM1
CEACAM5
CEACAM7
CA2

Figure 2

Figure 2: DVP analysis reveals a robust correlation between human IECs *in vivo* and grown as organoids

A Immunofluorescence image of a human colon organoid genetically engineered to express TdTomato under an LGR5 reporter for the identification of LGR5^{+} epithelial stem cells. **B** PCA plot of samples isolated from human colon organoids (three biological replicates (one organoid line, three separate passages), five technical replicates). **C** Abundance of previously described markers for different epithelial subpopulations (Krt20 – differentiated epithelial cells, MUC2 – goblet cells, EPHB" – stem cells). **D** Correlation plot of protein intensities of significantly changed proteins in epithelial cells located in the crypt bottom vs upper crypt *in vitro* and *in vivo*. **E** Venn Diagram of significantly changed proteins in epithelial cells in crypt bottom vs upper crypt *in vitro* and *in vivo*. Lines indicate selected overlapping proteins between *in vitro* and *in vivo* crypts.

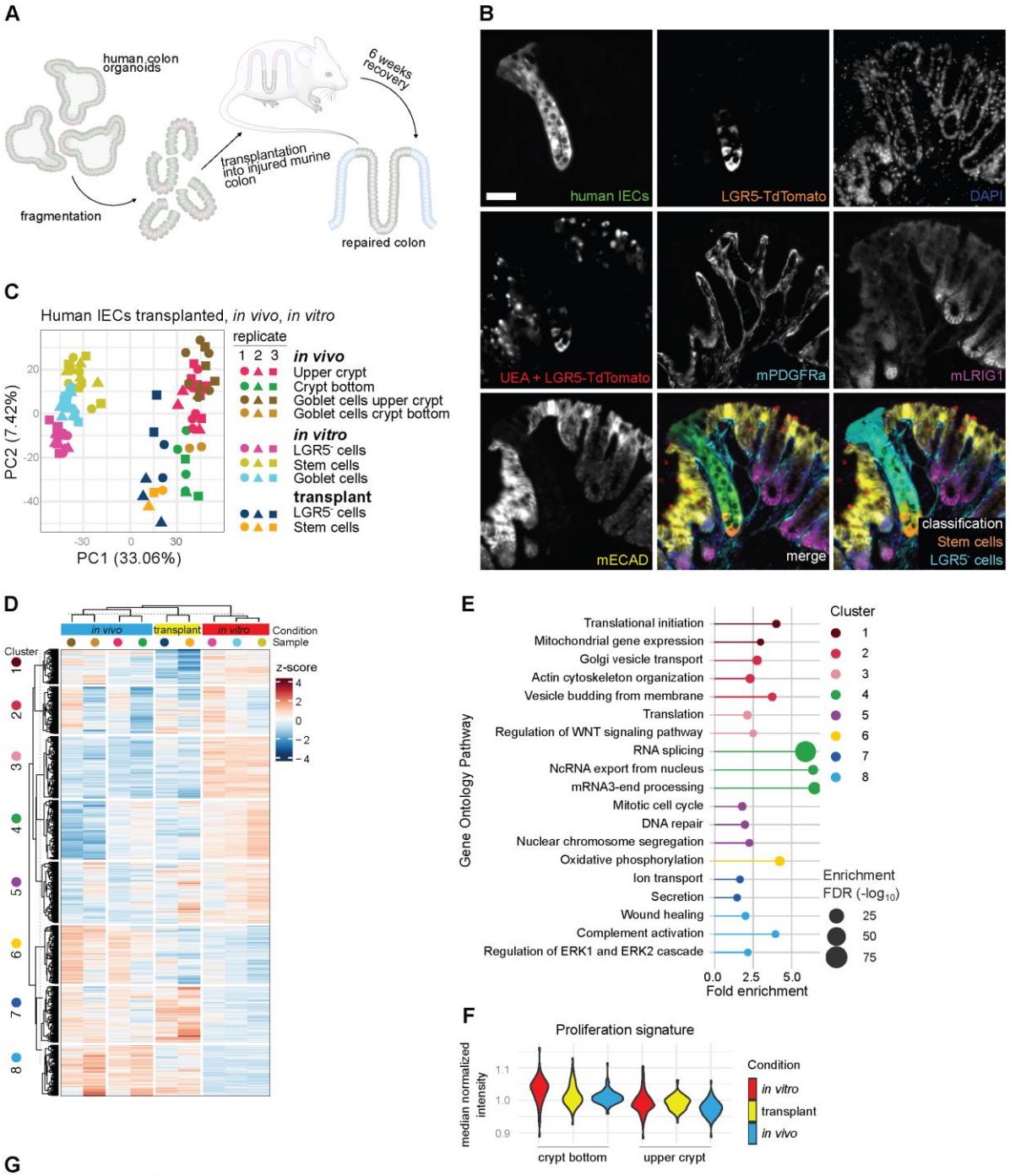

A
human colon organoids
6 weeks recovery
transplantation into injured murine colon
fragmentation
repaired colon
B
human IECs
LGR5-TdTomato
DAPI
UEA + LGR5-TdTomato
mPDGFRa
mLRIG1
mECAD
merge
classification
Stem cells
LGR5- cells
C
Human IECs transplanted, in vivo, in vitro
replicate
1 2 3
in vivo
Upper crypt
Crypt bottom
Goblet cells upper crypt
Goblet cells crypt bottom
in vitro
LGR5- cells
Stem cells
Goblet cells
transplant
LGR5- cells
Stem cells
PC2 (7.42%)
PC1 (33.06%)
D
in vivo
transplant
in vitro
Condition
Sample
Cluster
z-score
E
Translational initiation
Mitochondrial gene expression
Golgi vesicle transport
Actin cytoskeleton organization
Vesicle budding from membrane
Translation
Regulation of WNT signaling pathway
RNA splicing
NcRNA export from nucleus
mRNA3-end processing
Mitotic cell cycle
DNA repair
Nuclear chromosome segregation
Oxidative phosphorylation
Ion transport
Secretion
Wound healing
Complement activation
Regulation of ERK1 and ERK2 cascade
Gene Ontology Pathway
Cluster
Enrichment FDR (-log10)
Fold enrichment
F
Proliferation signature
median normalized intensity
Condition
in vitro
transplant
in vivo
crypt bottom
upper crypt
G
Stem cell signature
Stem cells in vitro
Crypt bottom cells in vivo
Stem cells transplant
Upper crypt in vivo
XRCC1 XPO5 WNK2 UTP14A TRIM28 TP53BP1 TMOD1 SNTB1 SLC19A1 PTK7 POLR1G POLA1 PCNA OLFM4 NLE1 NEK9 MSH2 MRE11 MCM7 MCM6 MCM4 MCM3 MCM2 MAP3K20 MAOB LRIG1 LGR5 KIFC1 HELLS HAUS GRHLD2 FRMD5 EPHB4 EPHB3 DDX56 DCP1B DCAF13 DAG1 CHGB CHDH CHAMP1 DCKN1C CDK6 CDC42EP1 CDC20 C8orf34 C1orf198 ALDH1B1
z-score

Figure 3

Figure 3: Human colon organoids transplanted into the murine colon recapitulate human colonocytes *in vivo*

A Workflow for orthotopic transplantation of organoids into the murine colon. **B** Immunofluorescence image human colon organoids (Fig. 2) transplanted into the murine colon (transplant). (GFP: human IECs. LGR5: stem cells (human). mECAD: epithelial cells (mouse). mPDGFR: fibroblasts (mouse). mLRIG1: crypt bottom compartment (mouse). UEA: mucus (goblet cells). **C** PCA plot of human colonocytes transplanted into the murine colon (one organoid line, three mice, one to three technical replicates), *in vitro* (organoids) and *in vivo* (human colon). **D** Heatmap of significantly changed proteins between organoids *in vitro*, transplanted organoids, and epithelial cells *in vivo*. **E** Gene ontology pathway enrichments of clustered proteins based on the heatmap in 3D. **F** Normalized protein intensities *in vitro*, in transplant, and *in vivo* of proteins that are associated with a proliferation signature in epithelial cells in the crypt bottom and the upper crypt. **G** Human stem cell proteome signature.

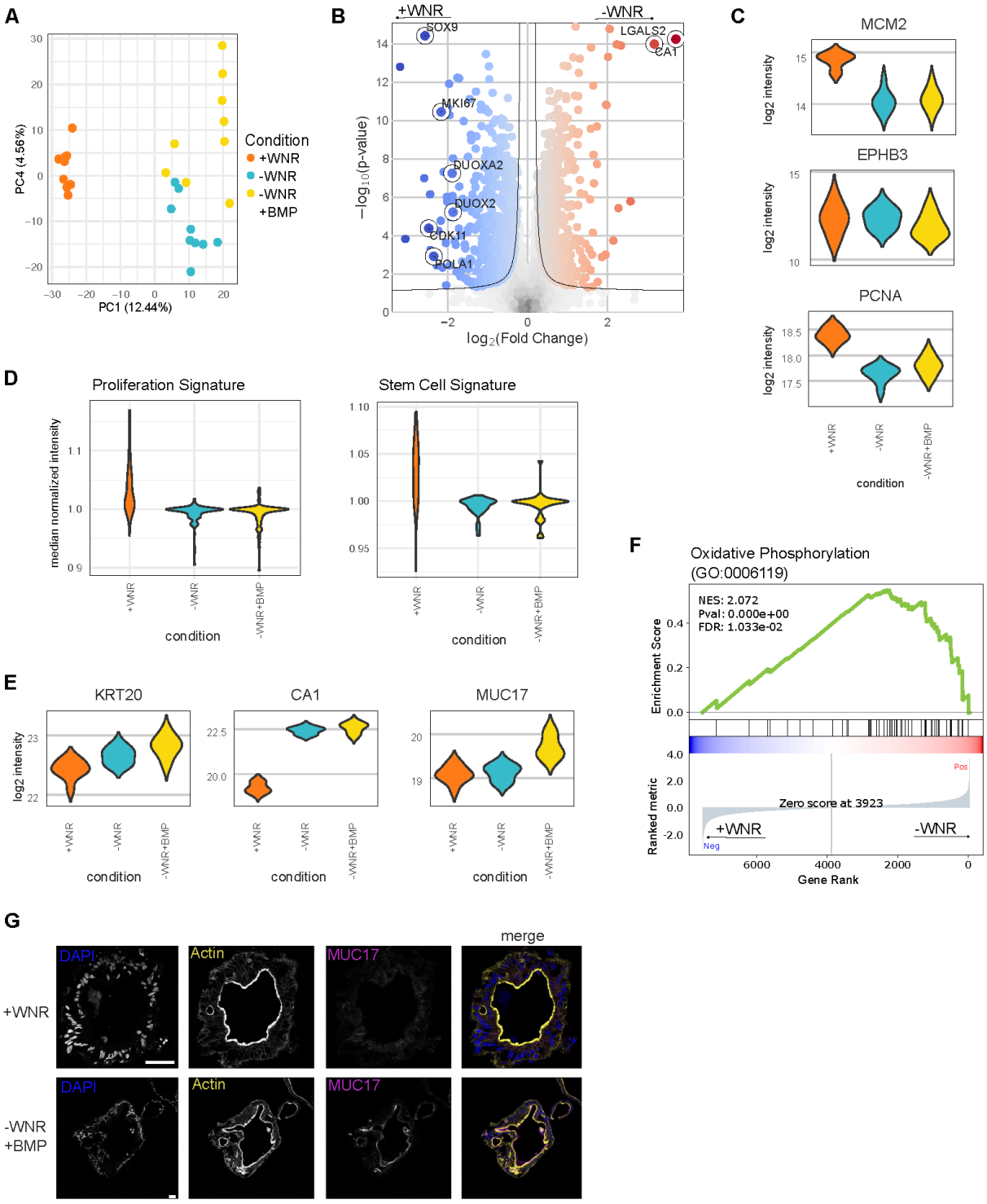

A
PC4 (4.56%)
PC1 (12.44%)
Condition
+WNR
-WNR
-WNR
+BMP
B
+WNR
-WNR
SOX9
LGALS2
CA1
MKI67
DUOXA2
DUOX2
CDK11
POLA1
−log10(p-value)
log2(Fold Change)
C
MCM2
EPHB3
PCNA
log2 intensity
+WNR
-WNR
-WNR+BMP
condition
D
Proliferation Signature
Stem Cell Signature
median normalized intensity
condition
E
KRT20
CA1
MUC17
log2 intensity
condition
F
Oxidative Phosphorylation
(GO:0006119)
NES: 2.072
Pval: 0.000e+00
FDR: 1.033e-02
Enrichment Score
Ranked metric
Zero score at 3923
+WNR
-WNR
Pos
Neg
Gene Rank
G
merge
+WNR
-WNR
+BMP
DAPI
Actin
MUC17

Figure 4

Figure 4: WNR withdrawal in colon organoids cultured *in vitro* induces upregulation of *in vivo* IEC markers.

A PCA of organoids cultured with WNR (WNT3a (W), Noggin (N), R-spondin-3 (R))(+WNR), without WNR (-WNR), and with BMP (Bone Morphogenetic Protein) but without WNR (-WNR +BMP). **B** Volcano plot of organoids cultured with WNR and without WNR. **C** Decrease of stem cell markers of colonic epithelial cells by withdrawal of WNR and addition of BMP. **D** Median normalized intensity of a proliferation signature[41] and stem cell signature in +WNR, -WNR, and -WNR +BMP. **E** Increase of differentiation markers of colonic epithelial cells by withdrawal of WNR and addition of BMP. **F** Fluorescence microscopy showing the increase of MUC17 in colon organoids upon withdrawal of WNR and addition of BMP. **G** Gene Set Enrichment Analysis showing an increase of the oxidative phosphorylation Gene Ontology pathway in -WNR vs +WNR.

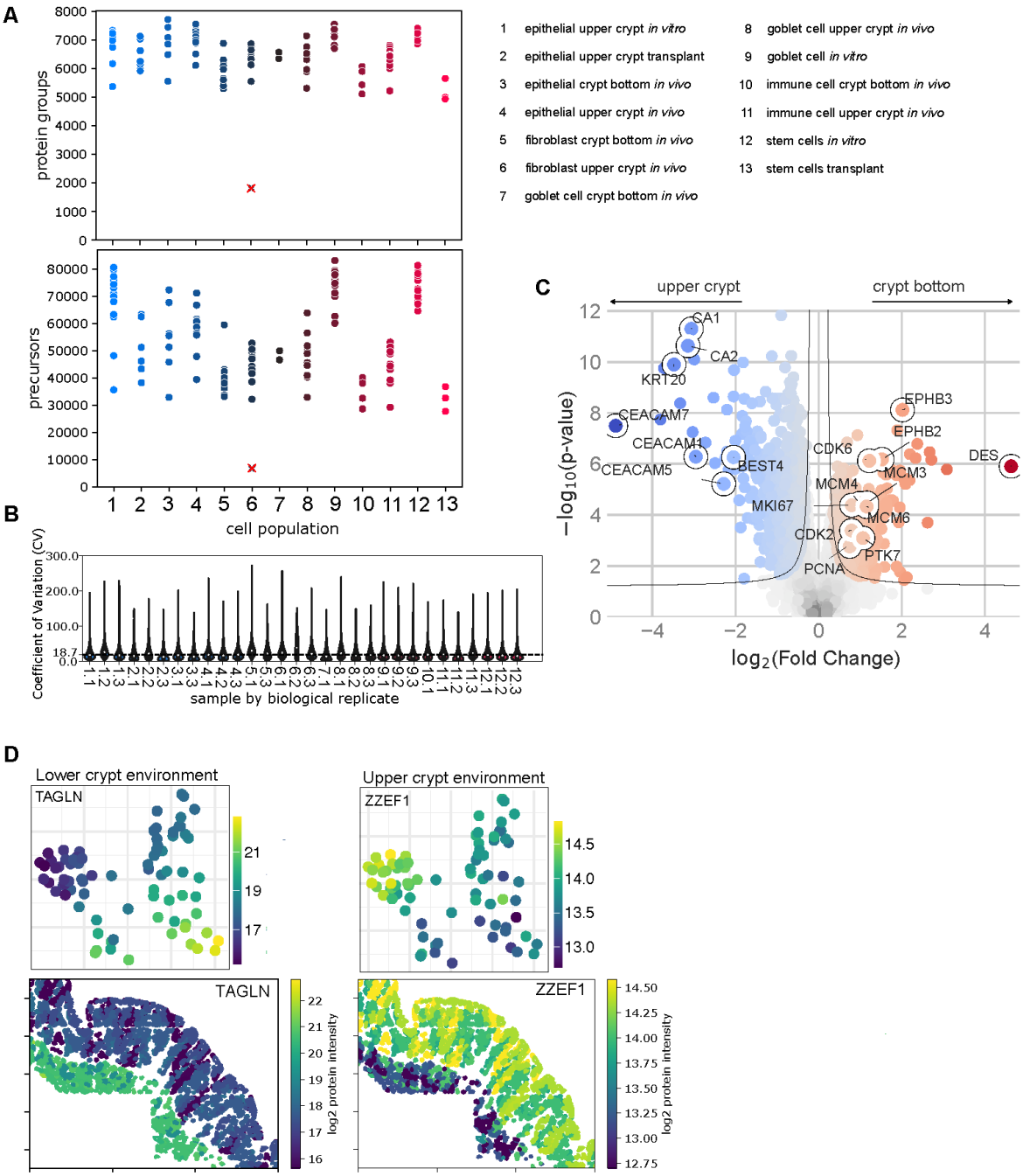

A
protein groups
precursors
cell population
1 epithelial upper crypt *in vitro*
2 epithelial upper crypt transplant
3 epithelial crypt bottom *in vivo*
4 epithelial upper crypt *in vivo*
5 fibroblast crypt bottom *in vivo*
6 fibroblast upper crypt *in vivo*
7 goblet cell crypt bottom *in vivo*
8 goblet cell upper crypt *in vivo*
9 goblet cell *in vitro*
10 immune cell crypt bottom *in vivo*
11 immune cell upper crypt *in vivo*
12 stem cells *in vitro*
13 stem cells transplant
B
Coefficient of Variation (CV)
sample by biological replicate
C
upper crypt
crypt bottom
CA1
CA2
KRT20
CEACAM7
CEACAM1
CEACAM5
BEST4
MKI67
EPHB3
EPHB2
CDK6
DES
MCM3
MCM4
MCM6
CDK2
PTK7
PCNA
$-\log_{10}$(p-value)
$\log_2$(Fold Change)
D
Lower crypt environment
Upper crypt environment
TAGLN
ZZEF1
log2 protein intensity

Figure S1

Figure S1

A Number of identified proteins and precursors per sample. **B** Coefficient of variation of technical replicates. **C** Volcano plot comparing epithelial cells from the crypt bottom and upper crypt *in vivo*. **D** Protein abundance and spatial distribution of TAGLN and ZZEF1, which are differentially abundant in the crypt bottom versus upper crypt region in the colon mucosa across different cell types.

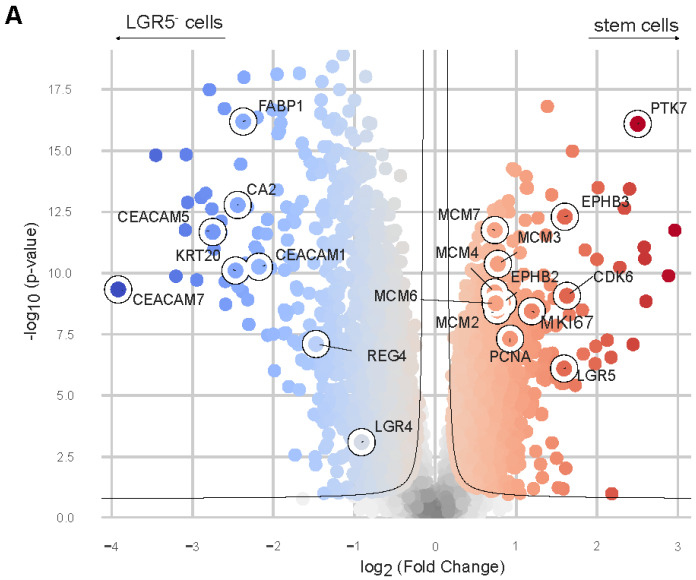

A
LGR5- cells
stem cells
FABP1
PTK7
CA2
CEACAM5
KRT20
CEACAM1
CEACAM7
MCM7
MCM3
MCM4
EPHB3
EPHB2
CDK6
MCM6
MCM2
MKI67
PCNA
REG4
LGR5
LGR4
-log10 (p-value)
log2 (Fold Change)

Figure S2

Figure S2

A Volcano plot of stem cells (LGR5-TdTomato^{+}) and LGR5-TdTomato^{-} cells in organoids *in vitro*.

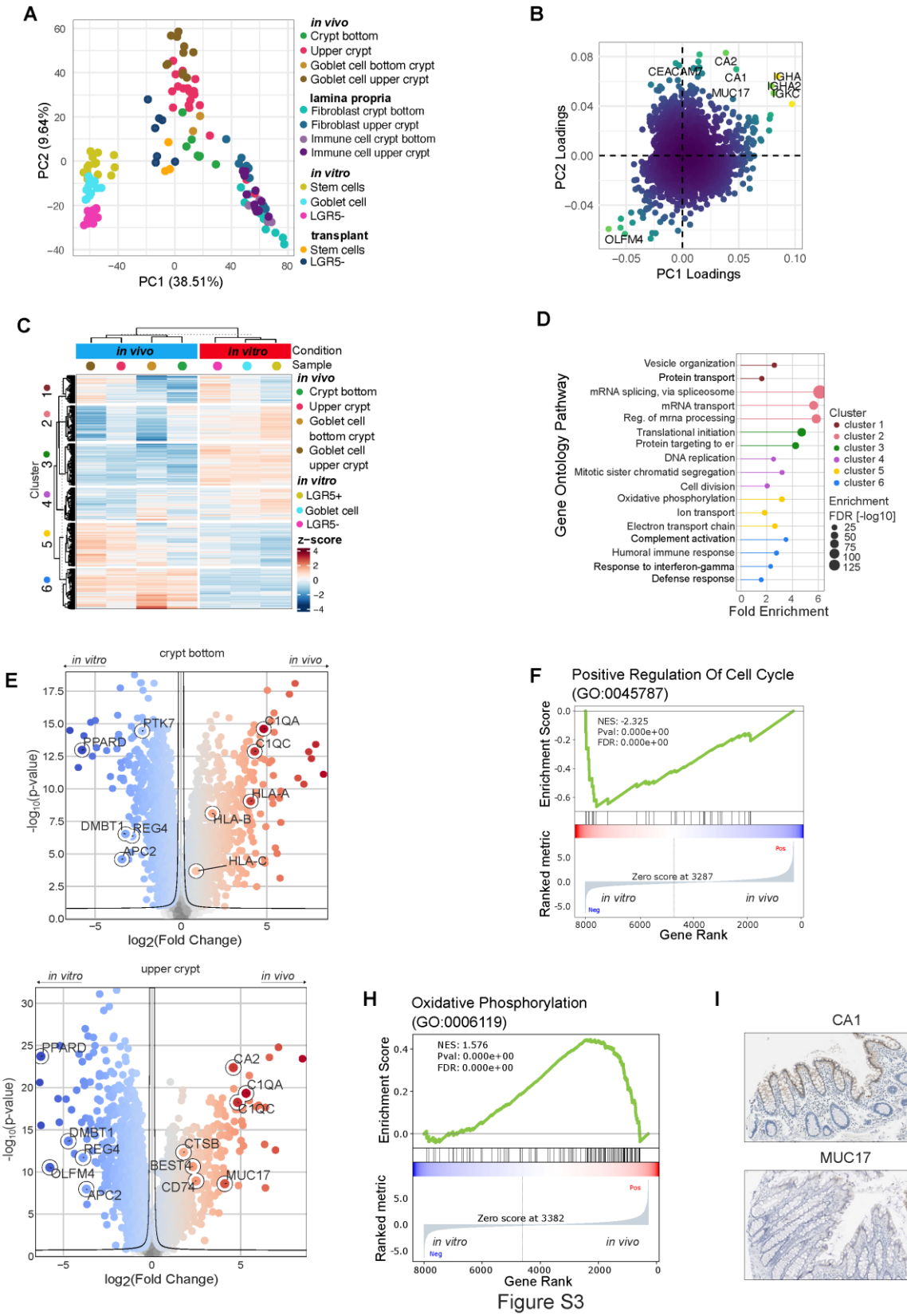
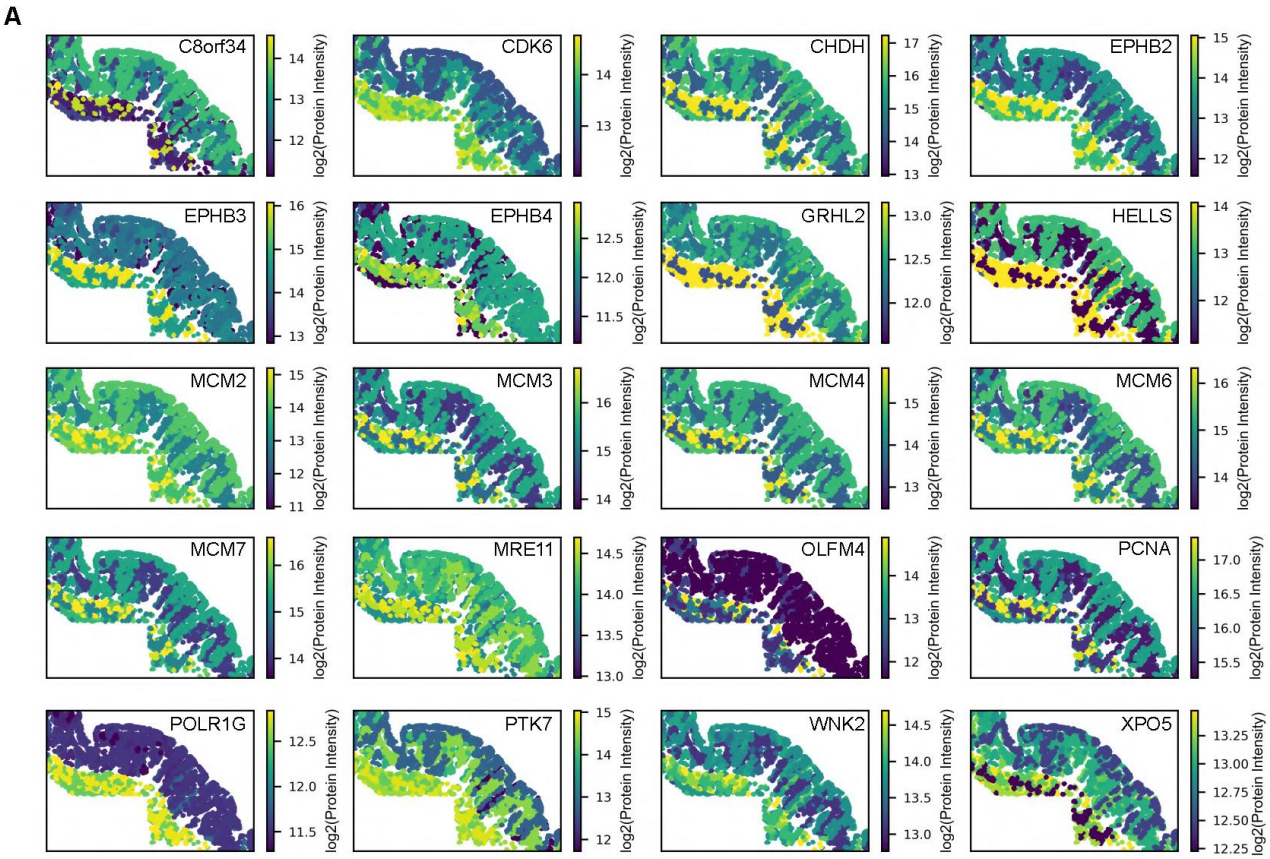

A
PC2 (9.64%)
PC1 (38.51%)
in vivo
Crypt bottom
Upper crypt
Goblet cell bottom crypt
Goblet cell upper crypt
lamina propria
Fibroblast crypt bottom
Fibroblast upper crypt
Immune cell crypt bottom
Immune cell upper crypt
in vitro
Stem cells
Goblet cell
LGR5-
transplant
Stem cells
LGR5-
B
PC2 Loadings
PC1 Loadings
CEACAM7
CA2
CA1
MUC17
IGHA
IGHA2
IGKC
OLFM4
C
in vivo
in vitro
Condition
Sample
Cluster
in vivo
Crypt bottom
Upper crypt
Goblet cell bottom crypt
Goblet cell upper crypt
in vitro
LGR5+
Goblet cell
LGR5-
z-score
D
Gene Ontology Pathway
Vesicle organization
Protein transport
mRNA splicing, via spliceosome
mRNA transport
Reg. of mrna processing
Translational initiation
Protein targeting to er
DNA replication
Mitotic sister chromatid segregation
Cell division
Oxidative phosphorylation
Ion transport
Electron transport chain
Complement activation
Humoral immune response
Response to interferon-gamma
Defense response
Fold Enrichment
Cluster
cluster 1
cluster 2
cluster 3
cluster 4
cluster 5
cluster 6
Enrichment FDR [-log10]
E
crypt bottom
in vitro
in vivo
-log10(p-value)
log2(Fold Change)
PTK7
PPARD
DMBT1
REG4
APC2
C1QA
C1QC
HLA-A
HLA-B
HLA-C
F
Positive Regulation Of Cell Cycle (GO:0045787)
NES: -2.325
Pval: 0.000e+00
FDR: 0.000e+00
Enrichment Score
Ranked metric
Zero score at 3287
in vitro
in vivo
Gene Rank
G
upper crypt
in vitro
in vivo
-log10(p-value)
log2(Fold Change)
PPARD
DMBT1
REG4
OLFM4
APC2
CA2
C1QA
C1QC
CTSB
BEST4
CD74
MUC17
H
Oxidative Phosphorylation (GO:0006119)
NES: 1.576
Pval: 0.000e+00
FDR: 0.000e+00
Enrichment Score
Ranked metric
Zero score at 3382
in vitro
in vivo
Gene Rank
I
CA1
MUC17

Figure S3

Figure S3

A PCA of the top 3000 most varying proteins across samples in *in vitro*, transplant, and *in vivo*. **B** Loadings describing proteins driving the PCA in S3A. CA1, CA2, MUC17 and CEACAM7 are strongly associated with a crypt top *in vivo* colonocyte phenotype. **C** Heatmap of significantly changed proteins between epithelial cells *in vitro* and *in vivo*. **D** Pathway enrichments of proteins in clusters of Fig S3C. **E** Volcano plot of stem cells *in vitro* vs crypt bottom epithelial cells *in vivo*. **F** Gene Set Enrichment Analysis (GSEA) of the Gene Ontology term "positive regulation of cell cycle" on protein differences of S3E. **G** Volcano plot of epithelial cells in the upper crypt *in vitro* vs *in vivo*. **H** GSEA of the Gene Ontology term "oxidative phosphorylation" on protein differences of S3G. **I** Staining for CA1 and MUC17 in the human colon mucosa from the Human Protein Atlas[43].


A
C8orf34
CDK6
CHDH
EPHB2
EPHB3
EPHB4
GRHL2
HELLS
MCM2
MCM3
MCM4
MCM6
MCM7
MRE11
OLFM4
PCNA
POLR1G
PTK7
WNK2
XPO5
log2(Protein Intensity)

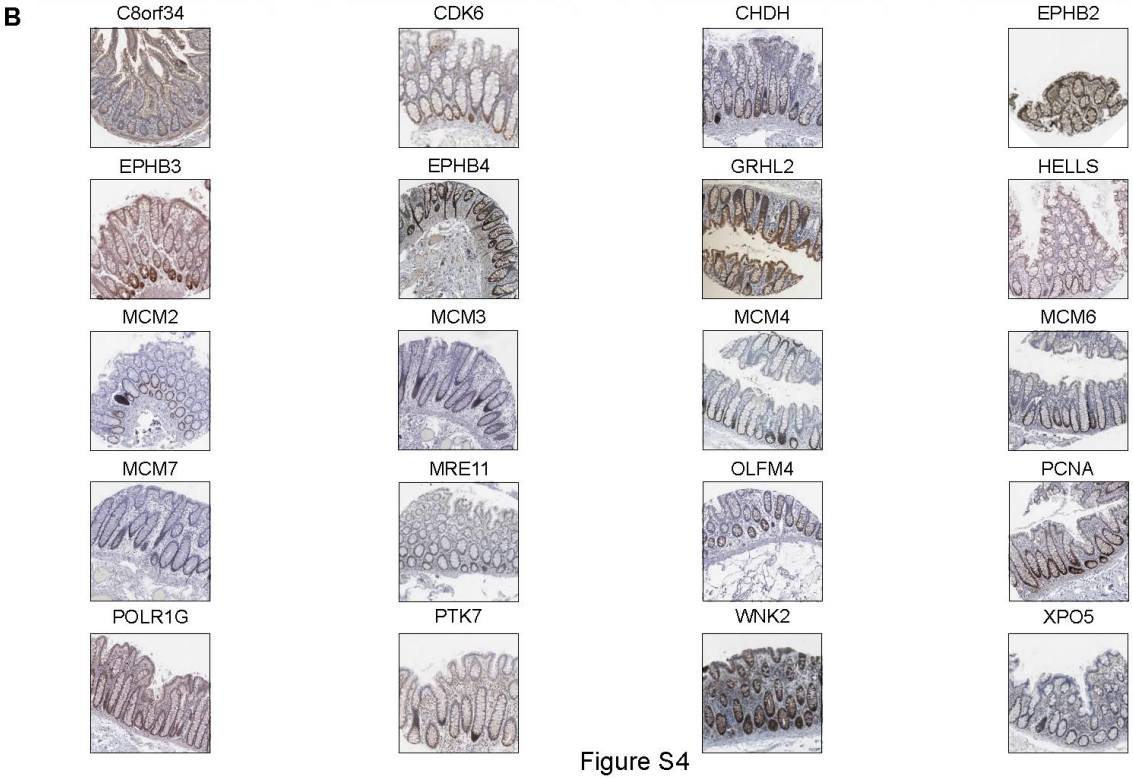

B
C8orf34
CDK6
CHDH
EPHB2
EPHB3
EPHB4
GRHL2
HELLS
MCM2
MCM3
MCM4
MCM6
MCM7
MRE11
OLFM4
PCNA
POLR1G
PTK7
WNK2
XPO5

Figure S4

Figure S4

A Spatial distribution of proteins that were identified in the colon stem cell signature. **B** Immunohistochemistry staining from the Human Protein Atlas[43] in human colon of proteins that were identified as potential colon stem cell markers.

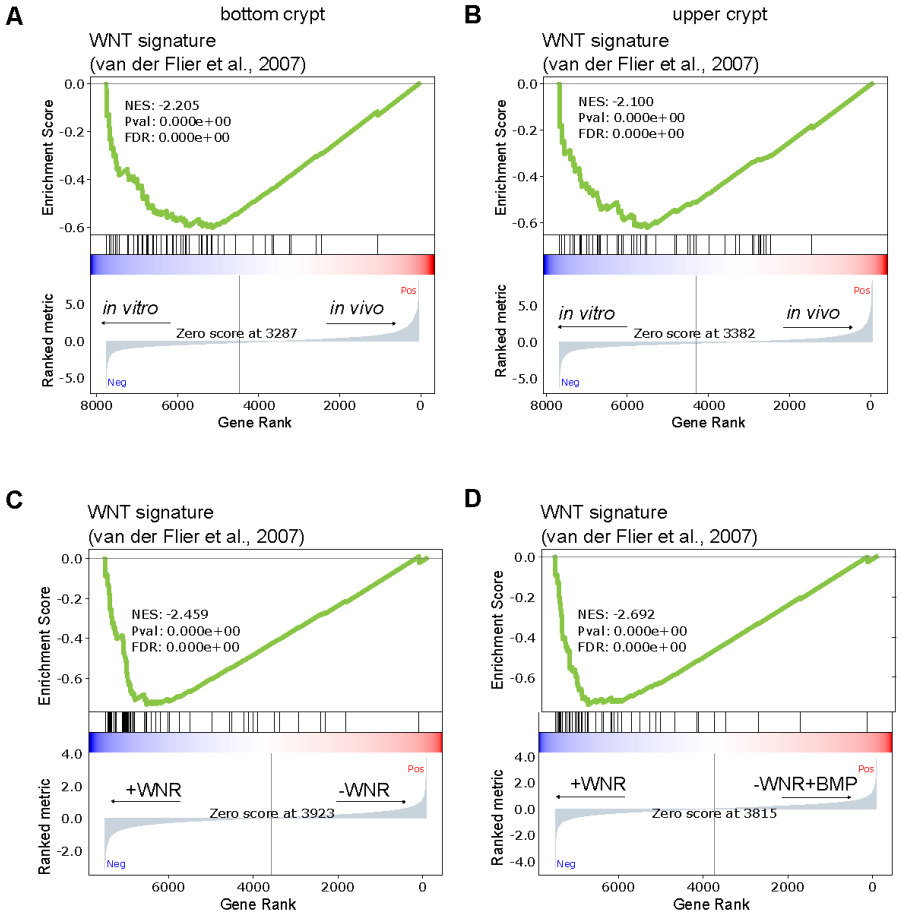

A
bottom crypt
WNT signature
(van der Flier et al., 2007)
NES: -2.205
Pval: 0.000e+00
FDR: 0.000e+00
Enrichment Score
Ranked metric
in vitro
in vivo
Zero score at 3287
Pos
Neg
Gene Rank
B
upper crypt
WNT signature
(van der Flier et al., 2007)
NES: -2.100
Pval: 0.000e+00
FDR: 0.000e+00
Enrichment Score
Ranked metric
in vitro
in vivo
Zero score at 3382
Pos
Neg
Gene Rank
C
WNT signature
(van der Flier et al., 2007)
NES: -2.459
Pval: 0.000e+00
FDR: 0.000e+00
Enrichment Score
Ranked metric
+WNR
-WNR
Zero score at 3923
Pos
Neg
Gene Rank
D
WNT signature
(van der Flier et al., 2007)
NES: -2.692
Pval: 0.000e+00
FDR: 0.000e+00
Enrichment Score
Ranked metric
+WNR
-WNR+BMP
Zero score at 3815
Pos
Neg
Gene Rank

Figure S5

Figure S5

A Gene set enrichment analysis of WNT signature proteins[51] in crypt bottom IECs *in vivo* versus stem cells *in vitro*, **B** in upper crypt IECs *in vivo* versus LGR5-TdTomato$^-$ cells *in vitro*, **C** organoids grown under -WNR versus +WNR conditions and **D** organoids grown under -WNR+BMP and +WNR conditions.

References

1. Hausmann, A., Steenholdt, C., Nielsen, O. H. & Jensen, K. B. Immune cell-derived signals governing epithelial phenotypes in homeostasis and inflammation. *Trends in Molecular Medicine* **0**, (2024).
2. Burclaff, J. *et al.* A Proximal-to-Distal Survey of Healthy Adult Human Small Intestine and Colon Epithelium by Single-Cell Transcriptomics. *Cellular and Molecular Gastroenterology and Hepatology* **13**, 1554–1589 (2022).
3. Sato, T. *et al.* Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. *Nature* **459**, 262–265 (2009).
4. Sato, T. *et al.* Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. *Gastroenterology* **141**, 1762–1772 (2011).
5. Fujii, M. *et al.* Human Intestinal Organoids Maintain Self-Renewal Capacity and Cellular Diversity in Niche-Inspired Culture Condition. *Cell Stem Cell* **23**, 787-793.e6 (2018).
6. Clevers, H. *et al.* Tissue-Engineering the Intestine: The Trials before the Trials. *Cell Stem Cell* **24**, 855–859 (2019).
7. Jensen, K. B. & Little, M. H. Organoids are not organs: Sources of variation and misinformation in organoid biology. *Stem Cell Reports* **18**, 1255–1270 (2023).
8. Sugimoto, S. *et al.* An organoid-based organ-repurposing approach to treat short bowel syndrome. *Nature* **592**, 99–104 (2021).
9. Driehuis, E., Kretzschmar, K. & Clevers, H. Establishment of patient-derived cancer organoids for drug-screening applications. *Nat Protoc* **15**, 3380–3409 (2020).

738 10. Watanabe, S. *et al.* Transplantation of intestinal organoids into a mouse model of
739 colitis. *Nat Protoc* **17**, 649–671 (2022).
740 11. Sugimoto, S. *et al.* Reconstruction of the Human Colon Epithelium In Vivo. *Cell*
741 *Stem Cell* **22**, 171-176.e5 (2018).
742 12. Ishikawa, K. *et al.* Identification of Quiescent LGR5+ Stem Cells in the Human
743 Colon. *Gastroenterology* **163**, 1391-1406.e24 (2022).
744 13. Bock, C. *et al.* The Organoid Cell Atlas. *Nat Biotechnol* **39**, 13–17 (2021).
745 14. Moor, A. E. *et al.* Spatial Reconstruction of Single Enterocytes Uncovers Broad
746 Zonation along the Intestinal Villus Axis. *Cell* **175**, 1156-1167.e15 (2018).
747 15. Parigi, S. M. *et al.* The spatial transcriptomic landscape of the healing mouse
748 intestine following damage. *Nat Commun* **13**, 828 (2022).
749 16. Haber, A. L. *et al.* A single-cell survey of the small intestinal epithelium. *Nature* **551**,
750 333–339 (2017).
751 17. Smillie, C. S. *et al.* Intra- and Inter-cellular Rewiring of the Human Colon during
752 Ulcerative Colitis. *Cell* **178**, 714-730.e22 (2019).
753 18. Fawkner-Corbett, D. *et al.* Spatiotemporal analysis of human intestinal
754 development at single-cell resolution. *Cell* **184**, 810-826.e23 (2021).
755 19. Machado, L. *et al.* Tissue damage induces a conserved stress response that initiates
756 quiescent muscle stem cell activation. *Cell Stem Cell* **28**, 1125-1135.e7 (2021).
757 20. Bues, J. *et al.* Deterministic scRNA-seq captures variation in intestinal crypt and
758 organoid composition. *Nat Methods* **19**, 323–330 (2022).
759 21. Hausser, J., Mayo, A., Keren, L. & Alon, U. Central dogma rates and the trade-off
760 between precision and economy in gene expression. *Nat Commun* **10**, 68 (2019).

761 22. Brunner, A.-D. *et al.* Ultra-high sensitivity mass spectrometry quantifies single-cell
762 proteome changes upon perturbation. *Mol Syst Biol* **18**, e10798 (2022).
763 23. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure
764 and function. *Nature* **537**, 347–355 (2016).
765 24. Ye, Z. *et al.* One-Tip enables comprehensive proteome coverage in minimal cells and
766 single zygotes. *Nat Commun* **15**, 2474 (2024).
767 25. Mund, A. *et al.* Deep Visual Proteomics defines single-cell identity and
768 heterogeneity. *Nat Biotechnol* **40**, 1231–1240 (2022).
769 26. Stewart, H. I. *et al.* Parallelized Acquisition of Orbitrap and Astral Analyzers Enables
770 High-Throughput Quantitative Analysis. *Anal. Chem.* **95**, 15656–15664 (2023).
771 27. Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker
772 gene Lgr5. *Nature* **449**, 1003–1007 (2007).
773 28. Jung, P. *et al.* Isolation and in vitro expansion of human colonic stem cells. *Nat Med*
774 **17**, 1225–1227 (2011).
775 29. Batlle, E. *et al.* Beta-catenin and TCF mediate cell positioning in the intestinal
776 epithelium by controlling the expression of EphB/ephrinB. *Cell* **111**, 251–263 (2002).
777 30. Jung, P. *et al.* Isolation of Human Colon Stem Cells Using Surface Expression of
778 PTK7. *Stem Cell Reports* **5**, 979–987 (2015).
779 31. van der Flier, L. G., Haegebarth, A., Stange, D. E., van de Wetering, M. & Clevers, H.
780 OLFM4 is a robust marker for stem cells in human intestine and marks a subset of
781 colorectal cancer cells. *Gastroenterology* **137**, 15–17 (2009).
782 32. Pachitariu, M. & Stringer, C. Cellpose 2.0: how to train your own model. *Nat Methods*
783 **19**, 1634–1641 (2022).

784 33. Das, S. *et al.* Colonic healing requires Wnt produced by epithelium as well as Tagln+
785 and Acta2+ stromal cells. *Development* **149**, dev199587 (2022).
786 34. Yu, Y., Tencer, A., Xuan, H., Kutateladze, T. G. & Shi, X. ZZEF1 is a Histone Reader and
787 Transcriptional Coregulator of Krüppel-Like Factors. *J Mol Biol* **433**, 166722 (2021).
788 35. Cheung, P. *et al.* Regenerative Reprogramming of the Intestinal Stem Cell State via
789 Hippo Signaling Suppresses Metastatic Colorectal Cancer. *Cell Stem Cell* **27**, 590-
790 604.e9 (2020).
791 36. Liu, S. *et al.* Transcription factor Klf9 controls bile acid reabsorption and
792 enterohepatic circulation in mice via promoting intestinal Asbt expression. *Acta*
793 *Pharmacol Sin* **43**, 2362–2372 (2022).
794 37. Kostopoulos, I. *et al.* A Continuous Battle for Host-Derived Glycans Between a
795 Mucus Specialist and a Glycan Generalist in vitro and in vivo. *Front. Microbiol.* **12**,
796 (2021).
797 38. Lu, J. *et al.* Characterization of an in vitro 3D intestinal organoid model by using
798 massive RNAseq-based transcriptome profiling. *Sci Rep* **11**, 16668 (2021).
799 39. Pikkupeura, L. M. *et al.* Transcriptional and epigenomic profiling identifies YAP
800 signaling as a key regulator of intestinal epithelium maturation. *Sci Adv* **9**, eadf9460
801 (2023).
802 40. Lycke, N. Y. & Bemark, M. The regulation of gut mucosal IgA B-cell responses: recent
803 developments. *Mucosal Immunol* **10**, 1361–1374 (2017).
804 41. Hausmann, A. & Hardt, W.-D. The Interplay between Salmonella enterica Serovar
805 Typhimurium and the Intestinal Mucosa during Oral Infection. *Microbiol Spectr* **7**,
806 (2019).

807 42. Merlos-Suárez, A. *et al.* The intestinal stem cell signature identifies colorectal
808 cancer stem cells and predicts disease relapse. *Cell Stem Cell* **8**, 511–524 (2011).
809 43. Sowden, J., Leigh, S., Talbot, I., Delhanty, J. & Edwards, Y. Expression from the
810 proximal promoter of the carbonic anhydrase 1 gene as a marker for differentiation
811 in colon epithelia. *Differentiation* **53**, 67–74 (1993).
812 44. Layunta, E. *et al.* MUC17 is an essential small intestinal glycocalyx component that
813 is disrupted in Crohn's disease. *bioRxiv* 2024.02.08.578867 (2024)
814 doi:10.1101/2024.02.08.578867.
815 45. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science*
816 **347**, 1260419 (2015).
817 46. Holmberg, J. *et al.* EphB Receptors Coordinate Migration and Proliferation in the
818 Intestinal Stem Cell Niche. *Cell* **125**, 1151–1163 (2006).
819 47. Stracker, T. H. & Petrini, J. H. J. The MRE11 complex: starting from the ends. *Nat Rev*
820 *Mol Cell Biol* **12**, 90–103 (2011).
821 48. Sun, Y., Cheng, Z. & Liu, S. MCM2 in human cancer: functions, mechanisms, and
822 clinical significance. *Molecular Medicine* **28**, 128 (2022).
823 49. Strzalka, W. & Ziemienowicz, A. Proliferating cell nuclear antigen (PCNA): a key
824 factor in DNA replication and cell cycle regulation. *Ann Bot* **107**, 1127–1140 (2011).
825 50. Madine, M. A., Khoo, C. Y., Mills, A. D. & Laskey, R. A. MCM3 complex required for
826 cell cycle regulation of DNA replication in vertebrate cells. *Nature* **375**, 421–424
827 (1995).
828 51. Komamura-Kohno, Y. *et al.* Site-specific phosphorylation of MCM4 during the cell
829 cycle in mammalian cells. *FEBS J* **273**, 1224–1239 (2006).

830 52. Beumer, J. *et al.* BMP gradient along the intestinal villus axis controls zonated
831 enterocyte and goblet cell states. *Cell Reports* **38**, 110438 (2022).
832 53. Van der Flier, L. G. *et al.* The Intestinal Wnt/TCF Signature. *Gastroenterology* **132**,
833 628–632 (2007).
834 54. Wilson, S. S. *et al.* Optimized Culture Conditions for Improved Growth and
835 Functional Differentiation of Mouse and Human Colon Organoids. *Front Immunol*
836 **11**, 547102 (2021).
837 55. Pleguezuelos-Manzano, C. *et al.* Establishment and Culture of Human Intestinal
838 Organoids Derived from Adult Stem Cells. *Curr Protoc Immunol* **130**, e106 (2020).
839 56. Heino, S. *et al.* Lef1 restricts ectopic crypt formation and tumor cell growth in
840 intestinal adenomas. *Sci Adv* **7**, eabj0512 (2021).
841 57. Mulholland, E. J. *et al.* GREMLIN1 disrupts intestinal epithelial-mesenchymal
842 crosstalk to induce a wnt-dependent ectopic stem cell niche via stromal
843 remodelling. Preprint at https://doi.org/10.1101/2024.04.28.591245 (2024).
844 58. Rodríguez-Colman, M. J. *et al.* Interplay between metabolic identities in the
845 intestinal crypt supports stem cell function. *Nature* **543**, 424–427 (2017).
846 59. Rath, E. & Haller, D. Intestinal epithelial cell metabolism at the interface of microbial
847 dysbiosis and tissue injury. *Mucosal Immunol* **15**, 595–604 (2022).
848 60. Mitrofanova, O., Broguiere, N., Nikolaev, M. & Lutolf, M. P. Bioengineered human
849 colon organoids with in vivo-like complexity and function. 2023.10.05.560991
850 Preprint at https://doi.org/10.1101/2023.10.05.560991 (2023).
851 61. Rosenberger, F. A. *et al.* Spatial single-cell mass spectrometry defines zonation of
852 the hepatocyte proteome. *Nat Methods* **20**, 1530–1536 (2023).

853 62. Hansen, S. L. *et al.* An organoid-based CRISPR-Cas9 screen for regulators of
854 intestinal epithelial maturation and cell fate. *Sci Adv* **9**, eadg4055.
855 63. Hausmann, A. *et al.* Germ-free and microbiota-associated mice yield small
856 intestinal epithelial organoids with equivalent and robust transcriptome/proteome
857 expression phenotypes. *Cell. Microbiol.* e13191 (2020) doi:10.1111/cmi.13191.
858 64. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural
859 networks and interference correction enable deep proteome coverage in high
860 throughput. *Nat Methods* **17**, 41–44 (2020).
861 65. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of
862 (prote)omics data. *Nat Methods* **13**, 731–740 (2016).
863 66. Krismer, E., Bludau, I., Strauss, M. T. & Mann, M. AlphaPeptStats: an open-source
864 Python package for automated and scalable statistical analysis of mass
865 spectrometry-based proteomics. *Bioinformatics* **39**, btad461 (2023).
866 67. Fang, Z., Liu, X. & Peltz, G. GSEApy: a comprehensive package for performing gene
867 set enrichment analysis in Python. *Bioinformatics* **39**, btac757 (2023).
868 68. Thomas, P. D. *et al.* PANTHER: Making genome-scale phylogenetics accessible to
869 all. *Protein Science* **31**, 8–22 (2022).
870 69. Adusumilli, R. & Mallick, P. Data Conversion with ProteoWizard msConvert. in
871 *Proteomics: Methods and Protocols* (eds. Comai, L., Katz, J. E. & Mallick, P.) 339–
872 368 (Springer, New York, NY, 2017). doi:10.1007/978-1-4939-6747-6_23.
873

Article 6: The proteomic landscape of proteotoxic stress in fibrogenic liver disease

Pre-print published online: bioRxiv (2024), doi: 10.1101/2024.11.01.621457, In revision at Nature

Florian A. Rosenberger[1]*, Sophia C. Mädler[1,15], Katrine Holtz Thorhauge[2,3,15], **Sophia Steigerwald[1,15],** Malin Fromme[4], Mikhail Lebedev[1], Caroline A. M. Weiss[1], Marc Oeller[1], Maria Wahle[1], Maximilian Zwiebel[1], Niklas A. Schmacke[5], Sönke Detlefsen[3,6], Peter Boor[7], Joseph Kaserman[8,9], Andrew Wilson[8,9], Ondřej Fabián[10,11], Soňa Fraňková[12], Aleksander A. Krag[2,3,13], Pavel Strnad[4], Matthias Mann[1,14]*

[1]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany
[2]Centre for Liver Research, Department of Gastroenterology and Hepatology, Odense, Denmark
[3]Department of Clinical Research, Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark
[4]Department of Internal Medicine III and IZKF, Gastroenterology, University Hospital Aachen, Germany
[5]Gene Center and Department of Biochemistry, Ludwig-Maximilians-Universität München
[6]Department of Pathology, Odense University Hospital, Odense, Denmark
[7]Institute of Pathology, University Hospital Aachen, RWTH Aachen University, Aachen, Germany
[8]Center for Regenerative Medicine of Boston University and Boston Medical Center, Boston, MA
[9]The Pulmonary Center and Department of Medicine, Boston University School of Medicine, Boston, MA, United States
[10]Clinical and Transplant Pathology Centre, Institute for Clinical and Experimental Medicine, Prague, Czech Republic
[11]Department of Pathology and Molecular Medicine, 3rd Faculty of Medicine, Charles University and Thomayer Hospital, Prague, Czech Republic
[12]Department of Hepatogastroenterology, Institute for Clinical and Experimental Medicine, Prague, Czech Republic
[13]Danish Institute of Advanced Study (DIAS), University of Southern Denmark, Odense, Denmark
[14]NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark
[15]These authors contributed equally
**Corresponding authors*

Alpha-1 antitrypsin deficiency (AATD) is a fibrogenic liver disease caused by mutations in the *SERPINA1* gene. This causes misfolding and accumulation of alpha-1 antitrypsin (AAT) in hepatocytes and ultimately leads to liver cirrhosis and can negatively impact lung function. Most severe ATTD cases can be attributed to a homozygous Z-variant, which has a prevalence of 1:2,000.[437,438] The mechanisms driving the disease progression and outcome heterogeneity, however, are largely unknown and treatment options remain unexplored.

In this study, we characterized hepatocyte responses to proteotoxic stress in AATD using DVP, machine learning and AI-guided image-based cell phenotyping. In a first line of evidence, we evaluated proteomic differences of cells with low, medium and high AAT aggregate load. While this confirmed known AATD signatures, it also enabled a pseudo-time analysis of disease progression. This identified a prominent peroxisomal

biogenesis response and unfolded protein response (UPR) as early and late hepatocyte responses to AAT accumulation, respectively. Interestingly, the distribution of AAT aggregate positive cells shows a distinct spatial component, with clear separation of areas with AAT+ and AAT- cells, and even occurrence of single AAT+ cells. To map the spatial proteomes of these cells and regions, we utilized the previously described single cell DVP workflow.[411] Aiming to improve the proteomic depth of the workflow and use it on formalin-fixed paraffin-embedded tissue sections, I optimized a variable window DIA method based for the acquisition of these single hepatocyte shapes on the Orbitrap Astral MS. With this, we achieved an unprecedented depth of up to 3,600 proteins in the equivalent of one-third to one-half of a hepatocyte, a 50% increase compared to the previously achieved depth on isolated hepatocytes from frozen tissue sections. Single-cell analysis of the AAT+ and AAT- border regions indicated that proteotoxic stress is cell-intrinsic and not propagated between neighboring cells. Correlating the earlier protein markers for early and late proteotoxic response with the border regions showed that late response markers, such as DNAJB11, remained unchanged in two out of three tissue samples. Moreover, in one sample we detected upregulation of an apoptotic marker in AAT+ border cells, which correlated with the observed aggregate morphology. Building on this, we integrated image featurization to isolate cells with different aggregate morphologies and identified globular aggregate morphology as a terminal cellular feature prior to cell death in AATD. Aggregating the results of the different spatial approaches, hundreds of dysregulated proteins could be identified, which offers novel candidates for treatment of AATD.

Contribution:

Co-authorship and shared second author. This study was conceptualized by Florian Rosenberger and Matthias Mann. I was the study lead for the single-cell DVP section of the manuscript. I selected regions of interest, processed the scDVP samples, and developed and optimized a tailored MS method for the acquisition of single hepatocyte shapes based on the expected precursor distribution. I supervised data quality control and performed initial biological analyses. Furthermore, for the first and last part of this study, I performed initial experiments to advise on the Orbitrap Astral acquisition for the DVP samples. I wrote the MS method section for the scDVP acquisition and contributed to revising and editing the manuscript alongside the other co-authors.

The proteomic landscape of proteotoxic stress in a fibrogenic liver disease

Florian A. Rosenberger[1*], Sophia C. Mädler[1,15], Katrine Holtz Thorhauge[2,3,15], Sophia Steigerwald[1,15], Malin Fromme[4], Mikhail Lebedev[1], Caroline A. M. Weiss[1], Marc Oeller[1], Maria Wahle[1], Maximilian Zwiebel[1], Niklas A. Schmacke[5], Sönke Detlefsen[3,6], Peter Boor[7], Joseph Kaserman[8,9], Andrew Wilson[8,9], Ondřej Fabián[10,11], Soňa Fraňková[12], Aleksander A. Krag[2,3,13], Pavel Strnad[4], Matthias Mann[1,14*]

Affiliations

[1]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany

[2]Centre for Liver Research, Department of Gastroenterology and Hepatology, Odense, Denmark

[3]Department of Clinical Research, Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark

[4]Department of Internal Medicine III and IZKF, Gastroenterology, University Hospital Aachen, Germany

[5]Gene Center and Department of Biochemistry, Ludwig-Maximilians-Universität München

[6]Department of Pathology, Odense University Hospital, Odense, Denmark

[7]Institute of Pathology, University Hospital Aachen, RWTH Aachen University, Aachen, Germany

[8]Center for Regenerative Medicine of Boston University and Boston Medical Center, Boston, MA

[9]The Pulmonary Center and Department of Medicine, Boston University School of Medicine, Boston, MA, United States

[10]Clinical and Transplant Pathology Centre, Institute for Clinical and Experimental Medicine, Prague, Czech Republic

[11]Department of Pathology and Molecular Medicine, 3rd Faculty of Medicine, Charles University and Thomayer Hospital, Prague, Czech Republic

[12]Department of Hepatogastroenterology, Institute for Clinical and Experimental Medicine, Prague, Czech Republic

[13]Danish Institute of Advanced Study (DIAS), University of Southern Denmark, Odense, Denmark

[14]NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark

[15]These authors contributed equally.

[*]Corresponding authors: rosenberger@biochem.mpg.de, mmann@biochem.mpg.de.

ABSTRACT

Protein misfolding diseases, including alpha-1 antitrypsin deficiency (AATD), pose significant health challenges, with their cellular progression still poorly understood [1–3]. We utilize spatial proteomics by mass spectrometry and machine learning to map AATD in human liver tissue. Combining Deep Visual Proteomics (DVP) with single-cell analysis[4,5], we probe intact patient biopsies to resolve molecular events during hepatocyte stress in pseudo-time across fibrosis stages. We achieve unprecedented proteome depth of up to 3,800 proteins from a third of a single cell in formalin-fixed, paraffin-embedded (FFPE) tissue. This dataset revealed a potentially clinically actionable peroxisomal upregulation that precedes the canonical unfolded protein response. Our single-cell proteomics data show alpha-1 antitrypsin accumulation is largely cell-intrinsic, with minimal stress propagation between hepatocytes. We integrated proteomic data with AI-guided image-based phenotyping across multiple disease stages, revealing a terminal hepatocyte state characterized by globular protein aggregates and distinct proteomic signatures, notably including elevated TNFSF10/TRAIL expression. This phenotype may represent a critical disease progression stage. Our study offers novel insights into AATD pathogenesis and introduces a powerful methodology for high-resolution, in situ proteomic analysis of complex tissues. This approach holds potential to unravel molecular mechanisms in various protein misfolding disorders, setting a new standard for understanding disease progression at the single-cell level in human tissue.

MAIN TEXT

Spatial omics technologies are revolutionizing our ability to deconvolute molecular events at single-cell resolution within a tissue context. While much focus has been placed on spatial genomics and transcriptomics, recent advances in multiplexed imaging and proteomics are beginning to shed light on the functional proteomic layer. Mass spectrometry-based proteomics has made significant strides towards biologically informative single-cell analysis, now enabling quantification of up to 5,000 proteins in cultured cells [6,7]. In the tissue context, we have recently introduced Deep Visual Proteomics (DVP), which integrates staining, AI-guided cell segmentation and classification, laser microdissection of single-cell shapes, and high-sensitivity mass spectrometry [4,5]. DVP excels in digital pathology applications with pronounced spatial and visual components, providing simultaneous and deep proteomic characterization at the level of thousands of proteins.

We reasoned that these emerging technologies would be ideally suited to elucidate molecular events during the progressive worsening of proteotoxicity as it unfolds in patients.

Proteotoxicity, characterized by the accumulation of misfolded and aggregated proteins leading to cell damage, is a hallmark of many diseases, including neurodegenerative pathologies such as Alzheimer's and Parkinson's disease [8–10]. The underlying cause of proteotoxicity is a disruption in protein homeostasis, resulting in an imbalance between protein synthesis, folding, and clearance mechanisms [3].

To investigate proteotoxicity in a clinically relevant context, we focused on a disorder with unmet clinical need that exemplifies the challenges of protein misfolding and aggregation in a vital organ. The fibrogenic liver disease alpha-1 antitrypsin deficiency (AATD), is a genetic disorder caused by autosomal, co-dominant mutations in the *SERPINA1* gene resulting in misfolding and accumulation of alpha-1 antitrypsin (AAT) in hepatocytes. Most severe AATD cases are caused by a homozygous Z-variant (Pi*ZZ genotype) with a peak incidence of 1:2,000 in individuals of European descent [1,2,11,12]. Current hypotheses suggest that the severity of liver damage correlates with the amount of accumulated AAT [13–18]. However, the mechanisms driving fibrogenesis or hepatocyte survival versus death remain unclear, leaving potentially druggable targets unexplored.

To address this challenge, we curated a cohort of formalin-fixed paraffin-embedded (FFPE) biopsies and liver explants from patients homozygous for the pathogenic Z-variant (Pi*ZZ), encompassing all fibrosis stages (n = 35, Extended Data Fig. 1a, Supplementary Table S1). Despite the same underlying disease-causing mutation at a similar median age (57.3 ± SD 9.9 years) and BMI (25.4 ± SD 4.0), the fibrosis stages varied drastically, indicating unexplored molecular resilience or risk profiles.

Proteomic mapping of hepatocyte responses to proteotoxic stress

To elucidate the molecular basis of the observed clinical heterogeneity in AATD patients, we implemented a comprehensive proteomic mapping approach to characterize hepatocyte responses to proteotoxic stress. We first laser microdissected 3 µm thick FFPE sections from patient biopsies and analyzed them with mass spectrometry following our DVP workflow. After staining for cell outlines and AAT, we segmented and stratified cells into low, moderate, and high aggregate load groups based on their microscopy images (Fig. 1a and 1b). The proteome of 100 shapes, equivalent to the volume of 10–15 complete hepatocytes, was then acquired on the recently introduced Orbitrap Astral mass spectrometer, yielding a high-quality dataset with a mean proteomic depth exceeding 5,000 proteins per sample (Extended Data Fig. 1b and 1c, Supplementary Table S1). We observed a striking 32-fold difference in AAT levels

between low and high-load cells. The AAT load was captured on the second principal component, preceded only by the fibrosis stage on the first component (Extended Data Fig. 1d to 1f). Given the sparsity of AAT+ cells in biopsy material, this validated our laser microdissection approach as it allowed the biological phenotype to emerge more clearly. Biopsies with a low fibrosis stage exhibited lower AAT baseline loading compared to high fibrosis stages on both proteomics and imaging data, while the maximum load remained fairly equal across all stages (Extended Data Fig. 1g). The proteomes of the three load classes differed markedly (16.2% significant hits at < 5% FDR, paired two-sided t-test; Fig. 1c). Alongside AAT, several known markers of AATD liver pathology were highly enriched in aggregate-positive cells, such as a 1.8-fold increased ER chaperone HSPA5 and a 2.8-fold increased ER-Golgi cargo receptor LMAN1 (Fig. 1d) [19–21].

Among the most dysregulated hits, we identified other secretory proteins, including many SERPINs, coagulation, and complement factors (Fig. 1c, Extended Data Fig. 1i). This corroborates the notion of ineffective processing and crowding in the ER space, with

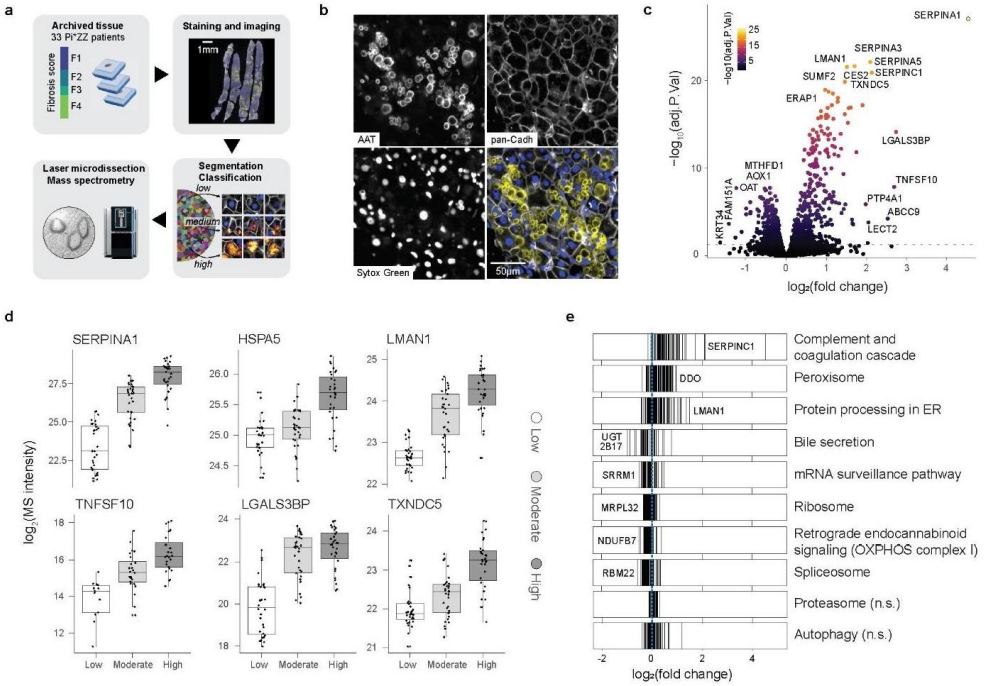


Fig. 1: Proteomic mapping of hepatocyte stress response. a, Overview of the Deep Visual Proteomics workflow. Fibrosis stages are Kleiner scores. **b,** Immunofluorescence staining of alpha-1 antitrypsin (AAT), the cell outline marker pan-cadherin (pan-Cadh), nucleus (SytoxGreen), and three-color overlay. **c,** Proteomic changes in high versus moderate versus low AAT-accumulating cells. Enriched in high on the right side. Top significant and top changed hits are named (paired two-sided t test with load class as covariable, multiple testing corrected, n = 95 at 100 shapes per sample). **d,** MS intensity of selected proteins across three classes. One dot is one sample from a patient (n = 32). **e,** Significantly (FDR < 0.05) enriched KEGG pathways after Gene Set Enrichment Analysis. Each line is a member of the pathway. n.s. not significant.

117 pathological implications due to the systemic deficiency of multiple plasma proteins [16].
118 Galectin-3 binding protein LGALS3BP and the apoptotic inducer TNFSF10/TRAIL had the
119 most pronounced positive changes (Fig. 1c and 1d). LGALS3BP is a hepatocyte-produced
120 protein targeted for secretion that is elevated in plasma from patients with liver disease [22].
121 Reports describing the immune-modulatory activity of LGALS3BP could explain the
122 involvement of immune cells in AATD liver pathology [13,23,24].
123 Pathway enrichment analysis showed a strong elevation of proteins related to the three branches
124 of unfolded protein response (UPR) mediated through ATF6, PERK and IRE1 along with a
125 general upregulation of chaperones, accompanied by a reduction of the transcription and
126 translation machinery. This occurred at the expense of physiological functions such as bile
127 secretion (Fig. 1e). Strikingly, many responses converged into a protective response to reactive
128 oxygen species (ROS) with upregulation of thioredoxins and glutaredoxins, including an
129 atypical increase in the peroxisomal compartment and reduction of mitochondrial complex I
130 (Fig. 1d, Extended Data Fig. 1h to 1m). Proteasomal and autophagy proteins remained largely
131 unchanged, and neither did we detect disturbances of calcium homeostasis (Fig. 1d, Extended
132 Data Fig. 1n).

133 **Early and late-stage responses to proteotoxic stress**

134 Our experimental design, encompassing three aggregate load classes, should allow us to
135 resolve the step-wise progression of molecular events. To determine the sequence in which
136 molecular responses occur during AAT build-up, we first correlated AAT with other protein
137 levels to identify 'followers' that tightly track AAT levels. Proteins of the endoplasmic
138 reticulum were among the top ten hits, many destined for secretion (Fig. 2a, Extended Data
139 Fig. 2a and 2b). This included many structurally similar SERPINs, and the tight tracking of
140 AAT levels suggests that these proteins accumulate in tandem with AAT rather than being co-
141 regulated.

142 We then categorized proteins into early and late responders to proteotoxic stress caused by
143 AAT accumulation (Fig. 2b, Supplementary Table S2). We observed the most consistent
144 relation with AAT load among co-elevated proteins, with the majority (77%) manifesting as
145 late responders and only a smaller fraction as early responders. The immune-modulatory
146 marker LGALS3BP, was most prominent among early responders, followed by the ER cargo
147 receptor MCFD2 together with its co-binder LMAN1 (Fig. 2c). Intriguingly, a strong
148 peroxisomal biogenesis response emerged early on, characterized by the peroxisomal
149 proliferation factor PEX11B and other membrane-integral proteins, along with lipid

metabolism and superoxide detoxifying proteins (Fig. 2d and 2e, Extended Data Fig. 2c and 3). In contrast, most proteins of the core machinery of the unfolded protein response appeared later during AAT build-up, despite visual protein accumulation at earlier stages (Fig. 2d, Extended Data Fig. 2d and 2e). The crosstalk between UPR and peroxisomal activity remains poorly understood, yet lipid metabolism, cholesterol metabolism, and ROS detoxification intersect both pathways. Together, the data indicate a dominant increase of the endoplasmic reticulum oxidoreductase 1 alpha (ERO1A), a major peroxide producer (Extended Data Fig. 2b).

We then analyzed samples at various fibrosis stages, revealing major dysregulations with increasing fibrosis stage in proteotoxicity-responsive pathways (Fig. 2f, Extended Data Fig. 4). Notably, this included the peroxisomal response, which showed a gradually prolonged onset time relative to AAT load (Fig. 2g). Importantly, peroxisomal chaperones or chaperone-like proteins remained unaltered, suggesting that peroxisomes are unlikely to contribute to the clearance of unfolded proteins (Extended Data Fig. 2c).

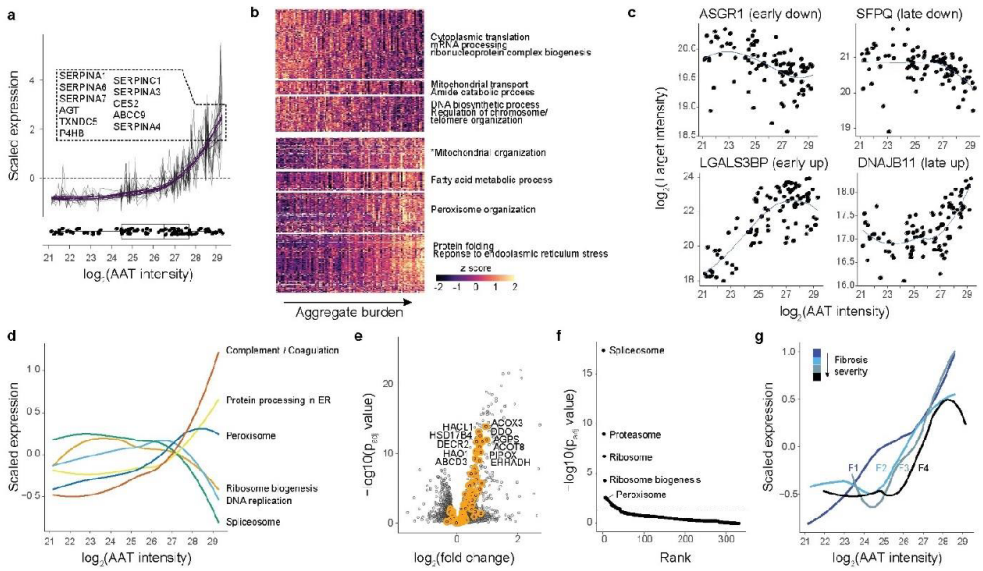


Fig. 2: Early and late responses to proteotoxic stress. a, Expression profile of the top-ten proteins correlating with AAT. All DVP sample are plotted, and values belonging to the same protein are on one line. Purple, polynomial fit (third order). Boxplot, distribution of AAT expression values along the x axis. **b**, Clustering into early and late responding genes to proteotoxic stress, order on x axis by AAT levels. The y axis was broken into seven groups to achieve good coverage of all response types. Significant KEGG term per box are shown, *not significant. **c**, Pseudo-time expression of top early and late responders by directionality. **d,** Cumulative changes of indicated KEGG pathways expressed as z scores. **e,** Changes of proteins levels across three AAT bins, highlighting peroxisomal proteins. Top significant and top changed hits are named (paired two-sided t test with load class as covariable, multiple testing corrected, n = 95). **f,** Top differential functional categories between F1 and F4 fibrotic samples during early AAT accumulation (log2(AAT intensity) < 25; two-sided Wilcoxon test, multiple testing corrected). **g,** Cumulative expression of peroxisomal proteins across four fibrosis stages.

Single-cell mapping in intact tissue

The accumulation of AAT in intact tissue exhibits a pronounced spatial component. Prior work has demonstrated that AAT accumulates unequally along the zonation gradient from portal to central vein axis in AATD-patients with then Pi*ZZ genotype [13,25,26]. Yet, sharp borders and the absence of gradual changes between neighboring AAT+ and AAT- cells, as well as single positive cells, indicate a more complex picture (Fig. 3a). To map the spatial proteome in these regions, we built upon our previous single-cell DVP workflow [5] and isolated single shapes from selected regions in 10 µm thick FFPE sections (equivalent to one-third to one-half of a complete hepatocyte) from three F1-stage biopsies. We quantified the proteome of these 'shapes' one at a time, allowing us to map back the proteome information onto the tissue with preserved single-cell spatial resolution (Fig. 3a).

In this way, we quantified the proteome of 132 single shapes in three biopsies at a median depth of 2,735 proteins, and reaching up to 3,600 proteins in some cells (Fig. 3b, Supplementary Table S3). The laser capturing proved highly efficient (9.9% dropout rate) and precise, as

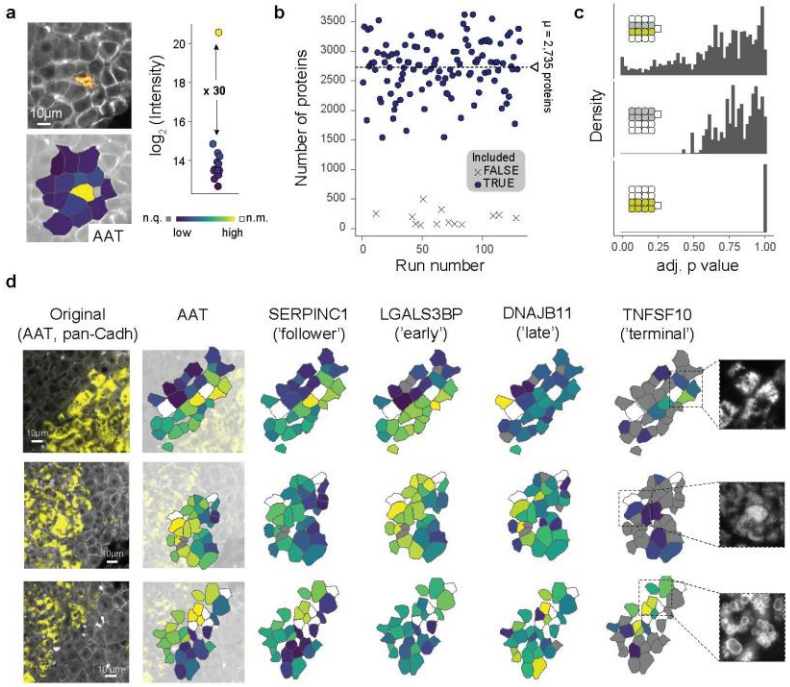


Figure 3, Mapping intact tissue at single cell level. a, Enrichment efficiency of the workflow as shown by isolating adjacent cells from FFPE tissue. Proteome quantification of AAT mapped back onto tissue. Boxplot shows AAT expression enrichment. **b,** Number of proteins detected per single shape across all 132 runs. **c,** Distribution of p values when comparing single cells at a border (top, n = 68), direct AAT- neighbours (middle, n = 69) and direct AAT+ neighbours (bottom, n = 49; two-sided unpaired t test after multiple testing correction). **d,** Mapping of proteomic information onto the original microscopic image. Cut-out images show AAT staining only. Gray, protein not quantified (n.q.); white, shape not captured and measured (n.m.) (N = 3, n = 132).

178 evidenced by the complete separation of adjacent AAT+ and AAT- cells (Fig. 3a, Extended
179 Data Fig. 5a to 5d). Upon comparing AAT+ and AAT- cells at border regions, we identified
180 similar proteotoxic stress markers as before (Extended Data Fig. 5e to 5g). Interestingly, cells
181 of the first or second row within a border region and within their respective AAT class
182 displayed very similar proteomes (Fig. 3c). Consistent with this, the AAT-accumulation
183 markers LGALS3BP and ERO1A were markedly different between AAT+ and AAT- cells, but
184 not among first and second-order neighbors. Consequently, the data supports an absence of
185 dedicated stress propagation between neighboring cells, suggesting that proteotoxic stress is a
186 cell-intrinsic response.
187 AAT accumulation has been previously characterized as a peri-portal event [27]. However, our
188 data indicate only partial or no dependence of AAT accumulation on zonation, as evidenced
189 by a drastic change in the expression levels of the portal marker ASS1 at borders, but not HAL
190 and ARG1, or the central markers ADH1 and CYP2E1. Notably, we observed a marked loss
191 of subunits of oxidative phosphorylation in AAT+ cells, including complex IV subunits (mt-
192 CO2, COX5B, COX6C, and others), a signal that was largely undetectable when comparing
193 bulk samples of three groups (Extended Data Fig. 5c, 5h). Importantly, we did not observe any
194 zonation effect in single AAT+ cells compared to AAT- direct neighbors (Extended Data Fig.
195 5i).
196 Upon mapping early- and late-responder markers back onto tissue, we found the expected
197 pattern at border regions for SERPINC1 and LGALS3BP, which mirrored AAT levels early
198 on. The late marker DNAJB11 remained unchanged in two of the three samples, indicating that
199 we captured the accumulation event at an early to medium stage (Fig. 3d). However, we
200 detected upregulation of the apoptotic inducer TNFSF10 in the border cells in one sample.
201 Further inspection revealed that the aggregate morphology was markedly different, with a
202 globular phenotype in contrast to amorphous AAT accumulation in the other two samples.
203 Differential expression analysis highlighted intracellular sequestration of iron (FTH1, FTL),
204 the apoptotic marker TNFSF10, and MBL2, as well as several enzymes related to detoxification
205 functions.
206 **Globular aggregates mark apoptotic cells**
207 Motivated by this observation, we enhanced our DVP workflow to connect cellular phenotypes
208 with proteomic data acquisition. We obtained liver resection samples containing thousands of
209 cells with various AAT aggregate morphologies on one slide. After staining and confocal
210 imaging of 3 µm thick sections of four biological and five technical samples, we segmented

cells and transformed the AAT channel signal within cell boundaries into 2048 features representing AAT morphology using the ConvNeXt convolutional neural network [28]. We projected these representations into a two-dimensional space using UMAP and determined 50 equally distributed center points across the image information layer, from which selected the 50 closest cells. These were isolated by laser microdissection and measured by MS, resulting in 250 morphology classes representing a total of 12,500 cells (Fig. 4a).

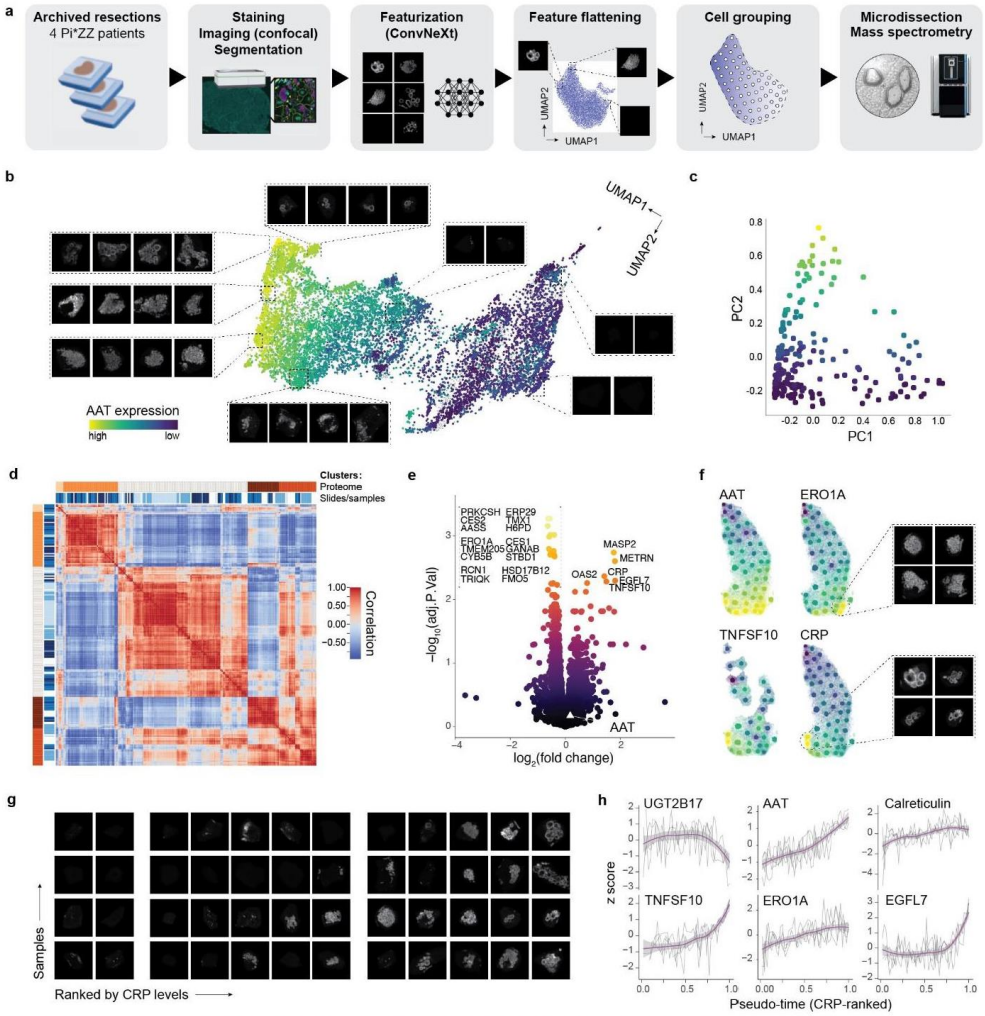


Fig. 4: The proteome of cells with various aggregate morphologies (n = 4). a, Overview of the CNN-DVP pipeline. **b,** Projection of all laser microdissected cells (12,500) and representative AAT images in indicated areas. Color scheme refers to AAT expression level (proteomic). **c,** Proteomic data of 209s samples (after filtering) reduced by PCA. **d,** Proteomic sample correlation heatmap, indicating proteome clusters based on k means clustering (5 groups manually chosen) and samples slides. **e,** Comparison of proteomes from cells with globular versus amorphous aggregates after selecting for similar AAT levels (AAT indicated as white triangle). Up in globular on the right, top hits annotated (paired two-sided t test after multiple testing correction). **f,** Projection of proteomics data onto image-based UMAP space of one representative sample, with representative images of indicated clusters. **g,** Pseudo time-sorted images of all four biological replicates. Groups mark inflection points of CRP. **h,** Expression levels of indicated proteins in CRP-ranked pseudo-time. Each line is one sample, smoothing curve in purple with 95%-confidence interval in grey.

Employing UMAP to project the representation of these micro-dissected cells into a 2D space validated that the utilized CNN could indeed stratify cells by aggregate morphologies, with aggregate-devoid cells clustering on one end and globular and amorphous morphologies located at the opposite side and clearly separated from one another (Fig. 4b). We achieved a median proteomic depth of 5,970 proteins from the equivalent of 5 to 10 complete hepatocytes (Extended Data Fig. 6a, Supplementary Table S4). The main drivers of our proteomic data were dynamic changes in keratins and AAT levels on principal components 1 and 2, respectively (Fig. 4c, Extended Data Fig. 6b to 6d). When grouping samples by proteome into clusters, patient samples were equally distributed across proteomic clusters without apparent genotypic or technical biases (Fig. 4d). As an inverse proof-of-principle, we successfully mapped the proteomic clusters back onto the UMAP image space with clear dimensional separation (Extended Data Fig. 6e). Consistently, samples of one proteome cluster also exhibited the shortest distances to one another on a proteomic UMAP and t-SNE plot (Extended Data Fig. 6f and 6g).

To better understand the molecular responses underlying morphology types, we comparatively analyzed samples with clear globular versus amorphous aggregates (Fig. 4e). Contrary to expectation, markers that typically follow AAT levels, like CES2 and ERO1A, were decreased in globular types. Conversely, the apoptotic inducer TNFSF10 and the inflammatory marker C-reactive protein (CRP) were positively enriched, indicating this to be a terminal phenotype preceding intrinsic or extrinsic apoptosis. We then mapped levels of marker proteins back onto the UMAP-derived image space. Intriguingly, ERO1A and TNFSF10 were localized in two distinct cell populations (Fig. 4f). While ERO1A, indicative of an ongoing UPR response, was highly enriched in amorphous aggregate types, TNFSF10 was mostly present in cells with globular aggregates alongside innate immune system activators. In line with this, Gene Set Enrichment Analysis further identified processes related to cell death as upregulated in globular types (Extended Data Fig. 6h).

Given a rather linear response rate of CRP across the image UMAP space (Fig. 4f), we then sorted all samples in pseudo-time by CRP expression levels. Across all four biological samples, we observed the emergence and disappearance of small corpuscular aggregates despite retained CRP signal. This was followed by a fulminant amorphous aggregation prior to condensation into globular aggregates as a terminal feature before cell death and clearance (Fig. 4g). In addition to TNFSF10, we identified EGF-like domain-containing protein 7 (EGFL7) as a viable marker of this stage that appeared late in the AATD phenotype. Notably, EGFL7 is also

upregulated in hepatocellular carcinoma, and high expression levels are associated with poor prognosis [29]. However, a potential link between globular phenotypes and HCC incidence in AATD remains unexplored. This terminal phenotype was further characterized by a stagnating or even declining unfolded protein response in late stages, as evidenced by Calreticulin and ERO1A levels, while reclining levels of proteins such as UGT2B17 suggest the termination of physiological functions in this hepatocyte subtype (Fig. 4h).

DISCUSSION

We present a pseudo-time resolved proteome of individual hepatocytes undergoing proteotoxic stress due to AAT aggregation. Our findings, derived from FFPE biopsies and resections from patients, provide novel insights into the progression and hepatic manifestation in AAT deficiency. While there are several model systems in the field, including murine models [30] and patient-derived induced pluripotent stem cells (iPSCs) [31], our approach uniquely captures responses to proteotoxic stress directly in patients via human tissue specimens representing the full disease spectrum (stages F1-F4). Notably, our data reveal that existing Pi*ZZ models do not accurately recapitulate the UPR, which manifests as a late but fulminant mode of action in our patient-derived samples [1,32]. This discrepancy extends to the globular phenotype, which we now identify as the terminal cellular feature preceding cell death [14]. Our approach strikingly underlines the power of harnessing patient cohorts and tissues. As many potentially druggable targets and pathways are intrinsically more difficult to validate when appropriate model systems are not in place, this inverts the traditional biomedical discovery cycle.

We here developed a single-cell proteomics approach to generate high-resolution maps of adjacent hepatocytes in intact tissue, leveraging recent advancements in ultra-low input mass spectrometry [6,7,33]. Building upon our previous work mapping zonation profiles in frozen mouse liver sections at single-cell resolution [5], we now quantify 50% more proteins and apply single-cell Deep Visual Proteomics (scDVP) to formalin-fixed tissue. This compatibility with FFPE tissue specimens, the gold standard in diagnostic pathology, expands access to cohorts of virtually any origin, age, and size [34], broadening the potential applications of this technology.

Our findings indicate that cells without aggregates are not directly affected or triggered by seeding-like mechanisms from adjacent aggregate-bearing cells. However, the presence of large patches of positive cells implies a propagation mechanism. Given the extensive metabolic perturbations observed, including alterations in fatty acid metabolism and detoxification pathways, AAT aggregate formation in one cell may lead to changes in the metabolic microenvironment, thereby inducing stress and proteostatic imbalance in adjacent cells. This

hypothesis aligns with other reports in the AATD field and similar mechanisms have been proposed in the context of neurodegenerative proteotoxic disorders where, however, it remains subject of ongoing debate [35,36].

We present an integration of image featurization and DVP that enables characterization of the entire proteomic and phenotypic lifecycle of stressed hepatocytes in a proteotoxic and fibrogenic liver disease. This methodology establishes a robust framework for dissecting complex cellular processes in situ across a spectrum of proteotoxic diseases. This strategy, an example of digital pathology with quantitative and very deep proteomic readout, yielded exceptionally deep proteomes of 6,000 quantified proteins, sufficient to inter most of the functional proteome of a given cell type. Importantly, our datasets are large enough to generate robust models capable of predicting the proteome of a cell based solely on its phenotype. This advancement paves the way for whole-slide proteomics in the future, representing a leap forward in our ability to comprehensively analyze tissue types at exceptional molecular and spatial resolution by mass spectrometry.

The methods developed here recapitulate known disease progression markers while identifying hundreds of additional dysregulated proteins. The present study is necessarily limited in functional follow-ups, yet these novel candidates clearly offer a valuable resource for biological and clinical validation. Of particular clinical relevance, we uncover an early upregulation of the peroxisomal compartment in samples from patients with low-grade liver fibrosis. This response is significantly delayed in high-grade fibrotic samples, suggesting a potential window for therapeutic intervention. PPAR-α agonists, such as fibrates, which increase peroxisome load in the liver, may be promising candidates for treating patients with late-diagnosed advanced liver fibrosis due to AATD. Given their well-established safety profiles, we suggest that these drugs could be repurposed for AATD, potentially transforming the treatment landscape of this proteotoxic disorder.

METHODS

Clinical cohorts and sample preparation

Patient biopsies and explant samples were obtained at two different sites, Odense University Hospital (OUH, Denmark) and Aachen RWTH University Hospital (UKA, Germany). The sample origin is indicated in Supplementary Table 1. Following ethical guidelines, the clinical data provided here is de-identified by only reporting sample type, fibrosis score, and site of origin.

OUH patient recruitment – Patients were recruited through the Danish patient organization (Alfa-1 Denmark) and clinical departments for liver and lung diseases as part of a cohort study. The cohort was designed to investigate liver health among non-pregnant adults (minimum age 18 years) diagnosed with AATD of any genotype and carrier status. This specific study includes 16 individuals diagnosed with Pi*ZZ who consented to undergo the procedure. The study was approved by the Danish Ethical Committee (S-2016987), and participants gave informed consent prior to enrollment. Participants without a history of liver transplant or decompensated cirrhosis were offered a percutaneous liver biopsy. The patients underwent liver core needle biopsies at Odense University Hospital (OUH) between 2017 and 2021. Liver core needle biopsies were taken during this period, stored in 4% formalin, and embedded in paraffin. For the assessment of fibrosis stage, FFPE blocks were cut on a microtome into 3µm thin sections and mounted on FLEX IHC slides (Dako, Glostrup, Denmark). Tissue sections were deparaffinized with xylene, rehydrated in serial dilutions of ethanol, and stained with Sirius Red. A certified hepatopathologist (S.D.) assessed the Kleiner fibrosis stage (0-4) according to the Pathology Committee of the NASH Clinical Research Network (NAS-CRN).

UKA patient recruitment – The recruitment of patients is described in detail in reference [37]. Of this cohort, the present study includes 19 individuals diagnosed with Pi*ZZ, of whom 14 underwent liver core needle biopsies due to medical indication and five received a liver transplantation due to end-stage liver disease. Samples were stored in 4% formalin and embedded in paraffin. Fibrosis stage was assessed after trichrome staining of 5µm thin sections by a certified hepatopathologist. Blocks were stored at room temperature. Ethical approval was provided by the institutional review board of Aachen University (EK 173/15). All participants provided written informed consent and were treated following the ethical guidelines of the Helsinki Declaration (Hong Kong Amendment) as well as Good Clinical Practice (European guidelines).

Staining

Two micrometer PEN membrane slides (MicroDissect GmbH) were exposed to UV light (254 nm) for one hour and then coated with Vectabond (Vector Laboratories; SP-1800-7) according to the manufacturer's protocol. Three (DVP, ML) or ten (scDVP) micrometer thin FFPE sections were mounted onto these slides and dried at 37°C overnight. Slides were stored at 4°C until further processing, upon which slides were baked at 55°C for 40 minutes, and then deparaffinized and rehydrated (xylene 2 x 2 min, 100% EtOH 2 x 1 min, 90% EtOH 2 x 1 min, 75% EtOH 2 x 1 min, 30% EtOH 2 x 1 min, ddH2O 2 x 1 min). Slides were transferred to prewarmed glycerol-supplemented antigen retrieval buffer (DAKO pH 9 S2367 + 10% Glycerol) at 88°C for 20 minutes, followed by a 20-minute cooldown at room temperature (RT 22°C). After washing in water, sections were blocked with 5% BSA in PBS for one hour, followed by an overnight incubation with primary antibodies in 1% BSA/PBS at 4°C in a humid staining chamber (1:200 mouse IgG1 monoclonal AAT 2C1, Hycult HM2289; 1:200 rabbit recombinant anti-pan cadherin [EPR1792Y], Abcam ab51034). After three washes in PBS for two minutes each, secondary antibodies (1:400 goat anti-mouse IgG1, Invitrogen A21127; 1:400 goat anti-rabbit AF647, Invitrogen A21245) in 1% BSA/PBS were applied for 90 minutes, followed by two 2-minute washes in PBS, 15 minutes in SYTOX™ Green (1:40,000 in PBS, Invitrogen S 7020), and three final 2-minute washes in PBS. Excess liquid was removed and samples were coverslipped using SlowFade Diamond Antifade Mountant (Invitrogen, S36963).

Imaging

Widefield Imaging – For DVP and scDVP experiments (Figures 1–3), sections were imaged using a Zeiss Axioscan 7. For all excitation wavelengths (504 nm, 577 nm, 653 nm), 50% light source intensity was used. The illumination time was specified on one section and applied to all consecutive samples within one experimental group. Three z-stacks at an interval of 2 µm were recorded with a Plan-Apochromat 20x/0.8 M27 objective and an Axiocam 712 camera at 14-bit, with a binning of 1 and a tile overlap of 10%, resulting in a scaling of 0.173 µm x 0.173 µm. Multiscene images were then split into single scenes, z-stacks combined into a single plane using extended depth of focus (variance method, standard settings), and stitched on the pan cadherin channel using the proprietary Zeiss Zen Imaging software.

Confocal Imaging – For experiments with downstream ML applications (Figure 4), sections were imaged on an PerkinElmer OperaPhenix high-content microscope, controlled with Harmony v4.9 software, at 40× magnification and 0.75 numerical aperture, with a binning of

1 and a per tile overlap of 10%. Only one z-plane was recorded, which was manually specified for each slide and channel. The three channels were imaged consecutively after deactivation of simultaneous recording to avoid any leakage between channels.

Cell selection (BIAS)

Images were imported as .czi files into the Biological Image Analysis Software (BIAS) using the packaged import tool [4]. Within BIAS, images were then retiled to 1024x1024 pixels with an overlap of 10%, and empty tiles were excluded from further analyses. Cell outlines were identified based on anti-pan cadherin stains using Cellpose 2.0 with the default cyto2 model [38]. Masks were imported into BIAS, and duplicates, as well as cells touching the borders of a tile (0.1% on each side), were removed. Further filtering was applied to retain cells with a minimum size of 3000 pixels, enriching for the hepatocyte population. For classification based on low, medium, and high aggregate load, the cell populations were divided per sample into five classes using a multilayer perceptron (MLP) with the following parameters: weight scale 0.01, momentum 0.01, maximum iterations 10,000, epsilon 0.0005, and 5 neurons in the hidden layer. Classification was based on the AAT (alpha-1 antitrypsin) maximum, median, and mean intensity within the cell outline mask. No human feedback was provided during this process. The low class was attributed to the cells with the lowest normalized mean intensity, medium to the third highest, and high to the highest normalized mean intensity; the other two intermediate classes were dropped. Reference points were selected based on prominent nuclear and histological features. One hundred cells were randomly picked for excision.

For single shape experiments, three characteristic low-fibrosis samples (all F1) and regions were selected that presented with a clear border-like phenotype (i.e., a row of AAT+ cells in direct neighborhood to AAT- cells) or with single AAT+ cells surrounded by AAT- cells. The cells were selected manually in BIAS, starting from the innermost cell and moving spiral-like to the outermost cell, thus avoiding cross-contamination of consecutively cut material.

Single-cell image generation

Images were flat-field corrected during image acquisition using the Perkin Elmer Harmony software (v4.9). Stitching of the flat-field corrected image tiles was performed using SPARCStools (https://github.com/MannLabs/SPARCStools). The stitched tile positions were calculated using the anti-pan cadherin stains imaged in the Alexa647 channel as a reference and then transferred to the other image channels. During stitching, the tile overlap was set to 0.1, the filter sigma parameter to 1, and the max shift parameter to 50.

405 The stitched images were then further processed in the python library SPARCSpy
406 (https://github.com/MannLabs/SPARCSpy). Cell outlines were identified based on the 7X
407 downsampled anti-pan cadherin stains using Cellpose 2.0 with the pretrained "cyto" model [38].
408 Segmentation was performed in a tiled mode with a 100px overlap. After resolving the cell
409 outlines from overlapping regions, the resulting segmentation mask was upscaled to the
410 original input dimensions during which the edges of the masks were smoothened by applying
411 an erosion and dilation operation with a kernel size of 7.

412 Then, the generated segmentation mask was used to extract single-cell image datasets with a
413 size of 280px x 280px. During extraction, the same single-cell image masks are used to obtain
414 the pixel information from each channel for each cell. The resulting single-cell images were
415 then rescaled to the [0, 1] range while preserving relative signal intensities. The resulting
416 single-cell image datasets were filtered to only contain cells from within manually annotated
417 regions in the tissue section containing hepatocytes but not fibrotic tissue.

418 **Cell selection (CNN)**

419 The filtered single-cell image datasets produced by SPARCSpy were further filtered to remove
420 any cells that fell outside the 5 to 97.5% size percentile. Representations of the remaining cells
421 were generated by featurization using the natural image-pretrained ConvNext model [28]. For
422 this, the single-cell images depicting the Alpha-1 channel were rescaled to the expected image
423 dimensions of Npx x Npx and triplicated to generate a pseudo rgb image. Inference was then
424 performed using the huggingface transformers package v. 4.26 [39].

425 The resulting 2048 image features were projected into a two-dimensional space using the
426 UMAP algorithm [40].The UMAP dimensions were calculated on the basis of the first 50
427 principal components and the 15 nearest neighbours. Using the spectral clustering algorithm
428 from scikit-learn [41], the resulting UMAP space was split into 50 clusters. The geometric centre
429 of each cluster was calculated and the 50 cells with the smallest Euclidean distance to the
430 cluster centre were selected for laser microdissection.

431 Contour outlines of the selected cells were generated in SPARCSpy using the py-lmd package
432 [42], whereby the cell outlines were dilated with a kernel size of 3 and a smoothing filter of 25
433 was applied. Furthermore, the number of points defining each shape were compressed by a
434 factor of 30 to improve LMD cutting performance. The cutting path, i.e. which cell is cut after
435 one another, was optimized using the Hilbert algorithm
436 (https://github.com/galtay/hilbertcurve).

Laser microdissection

After aligning the reference points, contour outlines were imported, and shapes were cut using the LMD7 (Leica) laser microdissection system in a semi-automated mode with the following settings: power 45, aperture 1, speed 40, middle pulse count 1, final pulse 0, head current 42-50%, pulse frequency 2,982, and offset 190. The microscope was operated with the LMD beta 10 software, calibrated for the gravitational stage shift into 384-well plates (Eppendorf 0030129547), leaving the outermost rows and columns empty. To prevent sorting errors, a 'wind shield' plate was placed on top of the sample stage. Plates were then sealed, centrifuged at 1,000 g for 5 minutes, and subsequently frozen at −20°C for further processing.

Peptide preparation and Evotip loading

Peptides were prepared as previously described using a BRAVO pipetting robot (Agilent) as per reference [43]. Briefly, 384-well plates were thawed, and shapes (both combined and individual) were rinsed from the walls into the bottom of the well with 28µL of 100% acetonitrile (ACN). The wells were completely dried in a SpeedVac at 45°C, followed by the addition of 6µL of 60mM triethylammonium bicarbonate (TEAB, Supelco 18597) (pH 8.5) supplemented with 0.013% n-Dodecyl-beta-D-maltoside (DDM, Sigma-Aldrich D5172). Plates were sealed and incubated at 95 °C for one hour. After adjusting to 10% ACN, samples were incubated again at 75 °C for one hour. Subsequently, 6ng and 4ng of trypsin and Lys-C protease, respectively, in 1 µL of 60 mM TEAB buffer were added to each sample, and proteins were digested for 16 hours at 37 °C. The reaction was quenched by adding trifluoroacetic acid (TFA) to a final concentration of 1%. Peptide samples were then frozen at -20 °C.

For loading, new Evotips were first soaked in 1-propanol for one minute, then rinsed twice with 50 µL of buffer B (ACN with 0.1% formic acid). After another 1-propanol soaking step for three minutes, the tips were equilibrated with two washes of 50 µL buffer A (0.1% formic acid). Samples were loaded into 70 µL of pre-loaded buffer A. Following one additional buffer A wash, the peptide-containing C18 disk was overlaid with 150µL buffer A and briefly centrifuged through the disk. All centrifugation steps were performed at 700g for one minute. The final tips were stored in buffer A for a maximum of four days prior to LC-MS.

LC-MS data acquisition

The peptide samples were analyzed using an Evosep One liquid chromatography (LC) system (Evosep) coupled to an Orbitrap Astral mass spectrometer (Thermo Fisher Scientific). Peptides were eluted from the Evotips with up to 35% acetonitrile (ACN) and separated using an Evosep

low-flow "Whisper" gradient for DVP samples, or an experimental Evosep "Whisper Zoom" gradient for single shapes and DVP-ML samples, with a throughput of 40 samples per day (SPD) on an Aurora Elite TS column of 15 cm length, 75 µm internal diameter (i.d.), packed with 1.7 µm C18 beads (IonOpticks). The column temperature was maintained at 50°C using a column heater (IonOpticks).

The Orbitrap Astral mass spectrometer was equipped with a FAIMS Pro interface and an EASY-Spray source (both Thermo Fisher Scientific). A FAIMS compensation voltage of −40V and a total carrier gas flow of 3.5 L/min were used. An electrospray voltage of 1900V was applied for ionization, and the RF level was set to 40. Orbitrap MS1 spectra were acquired from 380 to 980 m/z at a resolution of 240,000 (at m/z 200) with a normalized automated gain control (AGC) target of 500% and a maximum injection time of 100 ms.

For the Astral MS/MS scans in data-independent acquisition (DIA) mode, we experimentally determined the optimal methods across the precursor selection range of 380-980 m/z: (a) For DVP samples, a window width of 5 Th, a maximum injection time of 10 ms, and a normalized AGC target of 800% were used. (b) For DVP-ML samples, a window width of 6 Th, a maximum injection time of 13 ms, and a normalized AGC target of 500% were applied. (c) For single shapes and other DIA scans, the window width was optimized based on precursor density across the selection range of 380-980 m/z. A total of 45 variable-width DIA windows (see supplementary table 3) were acquired with a maximum injection time of 28 ms and an AGC target of 800%. The isolated ions were fragmented using higher-energy collisional dissociation (HCD) with 25% normalized collision energy.

Detailed method descriptions are provided in a default format with each supplementary data table.

Spectral searches and normalization

The raw files were searched together with match-between run in library-free mode within each experimental group with DIA-NN v1.8.1 [44]. A FASTA file containing only canonical sequences was obtained from Uniprot (20,404 entries, downloaded on 2023-01-02), and the disease-causing amino acid was manually changed (E342K). We allowed a missed cleavage rate of up to 1, and set mass accuracy to 8, MS1 accuracy to 4, and the scan window to 6. Proteins were inferred based on genes, and the neural network classifier was set to 'single-pass mode'. For DVP and DVP-ML samples, precursor intensities in the 'report.tsv' file were then normalized using the directLFQ GUI at standard settings including a minimum number of non-

nan ion intensities required to derive a protein intensity of one [45]. The single shape data was additionally median normalized to a set of proteins quantified across all samples (621 proteins quantified in 100% of included samples; see Supplementary Table S3), thereby correcting for the dependence of protein numbers on shape size [5].

Data analysis and statistics

Data was analyzed using R version 4.4.1. The directLFQ output file 'pg_matrix.tsv' was utilized for all subsequent data analysis, including the reported protein counts. Samples were included if the number of protein groups exceeded the mean minus (a) 1.5 standard deviations for DVP and single shape samples, resulting in 1.0% (1/96) and 10.6% (14/132) dropouts, respectively; and (b) 0.5 standard deviations for DVP-ML samples, resulting in 16.4% (41/250) dropouts. This lower cutoff was selected after manual inspection of the data distribution. Although some samples were collected in technical duplicates per patient biopsy, only the first replicate was used for statistical analyses and all reported measurements were taken from distinct samples. Coefficients of variation were calculated on non-transformed intensity values. For principal component analysis (PCA), the R package PCAtools 2.16.0 was used on a complete data matrix, removing the lower 10% of variables based on variance. Statistical analyses were performed assuming normality using the limma package version 3.60.3 with two-sided moderated t-tests and "fdr" as a multiple testing correction method. A per-patient statistical pairing was applied for DVP and single shape experiments. Intensity and fold changes are reported as log2-transformed values unless indicated otherwise. Gene Set Enrichment Analysis (GSEA) was conducted using WebGestalt 2024 against the indicated databases, with a false discovery rate (FDR) of < 0.05 considered significant [46]. Interaction networks were calculated with STRING database at standard settings [47]. The timing of responses ranked by the absolute difference between B values of limma's moderated t test comparing three AAT load groups: low to moderate, and moderate to high. Only proteins that were significant in either or both comparisons were considered. Differential pathway expression across fibrosis stages was calculated by fitting a linear model through log2-transformed intensity values of individual proteins in samples with log2(AAT)-intensity < 25, and the slopes of proteins in a particular pathway were compared between F1 and F4 samples by a two-sided Wilcoxon rank test without assumption of normality. Indicated p values are corrected for multiple testing using the 'fdr' method. Spatial data was mapped using the 'simple features' package.

Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [48] partner repository with the dataset identifier PXD054440 (Username: reviewer_pxd054440@ebi.ac.uk, Password: R14c41PdHVK0).

REFERENCES

1. Greene, C. M. *et al.* α1-Antitrypsin deficiency. *Nat Rev Dis Primers* **2**, 16051 (2016).
2. Strnad, P., McElvaney, N. G. & Lomas, D. A. Alpha1-Antitrypsin Deficiency. *N Engl J Med* **382**, 1443–1455 (2020).
3. Hipp, M. S., Park, S.-H. & Hartl, F. U. Proteostasis impairment in protein-misfolding and -aggregation diseases. *Trends Cell Biol* **24**, 506–514 (2014).
4. Mund, A. *et al.* Deep Visual Proteomics defines single-cell identity and heterogeneity. *Nat Biotechnol* **40**, 1231–1240 (2022).
5. Rosenberger, F. A. *et al.* Spatial single-cell mass spectrometry defines zonation of the hepatocyte proteome. *Nat Methods* **20**, 1530–1536 (2023).
6. Petrosius, V. *et al.* Evaluating the capabilities of the Astral mass analyzer for single-cell proteomics. 2023.06.06.543943 Preprint at https://doi.org/10.1101/2023.06.06.543943 (2023).
7. Guzman, U. H. *et al.* Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition. *Nat Biotechnol* 1–12 (2024) doi:10.1038/s41587-023-02099-7.
8. Chiti, F. & Dobson, C. M. Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade. *Annual Review of Biochemistry* **86**, 27–68 (2017).
9. Selkoe, D. J. & Hardy, J. The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Molecular Medicine* **8**, 595–608 (2016).
10. Goedert, M., Jakes, R. & Spillantini, M. G. The Synucleinopathies: Twenty Years On. *Journal of Parkinson's Disease* **7**, S51–S69 (2017).
11. Lomas, D. A., Evans, D. L., Finch, J. T. & Carrell, R. W. The mechanism of Z alpha 1-antitrypsin accumulation in the liver. *Nature* **357**, 605–607 (1992).
12. Brantly, M., Nukiwa, T. & Crystal, R. G. Molecular basis of alpha-1-antitrypsin deficiency. *The American Journal of Medicine* **84**, 13–31 (1988).
13. Clark, V. C. *et al.* Clinical and histologic features of adults with alpha-1 antitrypsin deficiency in a non-cirrhotic cohort. *J Hepatol* **69**, 1357–1364 (2018).

14. Lindblad, D., Blomenkamp, K. & Teckman, J. Alpha-1-antitrypsin mutant Z protein content in individual hepatocytes correlates with cell death in a mouse model. *Hepatology* **46**, 1228–1235 (2007).
15. Rudnick, D. A. *et al.* Analyses of hepatocellular proliferation in a mouse model of alpha-1-antitrypsin deficiency. *Hepatology* **39**, 1048–1055 (2004).
16. Chambers, J. E. *et al.* Z-α1-antitrypsin polymers impose molecular filtration in the endoplasmic reticulum after undergoing phase transition to a solid state. *Science Advances* **8**, eabm2094 (2022).
17. Segeritz, C.-P. *et al.* hiPSC hepatocyte model demonstrates the role of unfolded protein response and inflammatory networks in α1-antitrypsin deficiency. *J Hepatol* **69**, 851–860 (2018).
18. Fromme, M., Schneider, C. V., Trautwein, C., Brunetti-Pierri, N. & Strnad, P. Alpha-1 antitrypsin deficiency: A re-surfacing adult liver disorder. *J Hepatol* **76**, 946–958 (2022).
19. Zhang, Y. *et al.* LMAN1-MCFD2 complex is a cargo receptor for the ER-Golgi transport of α1-antitrypsin. *Biochem J* **479**, 839–855 (2022).
20. Schmidt, B. Z. & Perlmutter, D. H. Grp78, Grp94, and Grp170 interact with alpha1-antitrypsin mutants that are retained in the endoplasmic reticulum. *Am J Physiol Gastrointest Liver Physiol* **289**, G444-455 (2005).
21. Werder, R. B. *et al.* Adenine base editing reduces misfolded protein accumulation and toxicity in alpha-1 antitrypsin deficient patient iPSC-hepatocytes. *Mol Ther* **29**, 3219–3229 (2021).
22. Niu, L. *et al.* Noninvasive proteomic biomarkers for alcohol-related liver disease. *Nat Med* **28**, 1277–1287 (2022).
23. Cho, S.-H. *et al.* Lgals3bp suppresses colon inflammation and tumorigenesis through the downregulation of TAK1-NF-κB signaling. *Cell Death Discov.* **7**, 1–13 (2021).
24. Khodayari, N. *et al.* Characterization of hepatic inflammatory changes in a C57BL/6J mouse model of alpha1-antitrypsin deficiency. *Am J Physiol Gastrointest Liver Physiol* **323**, G594–G608 (2022).
25. Porat-Shliom, N. Compartmentalization, cooperation, and communication: The 3Cs of Hepatocyte zonation. *Curr Opin Cell Biol* **86**, 102292 (2024).
26. Piccolo, P. *et al.* Down-regulation of hepatocyte nuclear factor-4α and defective zonation in livers expressing mutant Z α1-antitrypsin. *Hepatology* **66**, 124 (2017).
27. Crowther, D. C. *et al.* Practical genetics: alpha-1-antitrypsin deficiency and the serpinopathies. *Eur J Hum Genet* **12**, 167–172 (2004).

28. Liu, Z. *et al.* A ConvNet for the 2020s. Preprint at https://doi.org/10.48550/arXiv.2201.03545 (2022).
29. Yang, C. *et al.* Increased expression of epidermal growth factor-like domain-containing protein 7 is predictive of poor prognosis in patients with hepatocellular carcinoma. *J Cancer Res Ther* **14**, 867–872 (2018).
30. Carlson, J. A. *et al.* Accumulation of PiZ alpha 1-antitrypsin causes liver damage in transgenic mice. *J Clin Invest* **83**, 1183–1190 (1989).
31. Yusa, K. *et al.* Targeted gene correction of α1-antitrypsin deficiency in induced pluripotent stem cells. *Nature* **478**, 391–394 (2011).
32. Hidvegi, T., Schmidt, B. Z., Hale, P. & Perlmutter, D. H. Accumulation of mutant alpha1-antitrypsin Z in the endoplasmic reticulum activates caspases-4 and -12, NFkappaB, and BAP31 but not the unfolded protein response. *J Biol Chem* **280**, 39002–39015 (2005).
33. Rosenberger, F. A., Thielert, M. & Mann, M. Making single-cell proteomics biologically relevant. *Nat Methods* **20**, 320–323 (2023).
34. Coscia, F. *et al.* A streamlined mass spectrometry–based proteomics workflow for large-scale FFPE tissue analysis. *The Journal of Pathology* **251**, 100–112 (2020).
35. Henrich, M. T. *et al.* Determinants of seeding and spreading of α-synuclein pathology in the brain. *Science Advances* **6**, eabc2487 (2020).
36. Bassil, F. *et al.* Amyloid-Beta (Aβ) Plaques Promote Seeding and Spreading of Alpha-Synuclein and Tau in a Mouse Model of Lewy Body Disorders with Aβ Pathology. *Neuron* **105**, 260-275.e6 (2020).
37. Schneider, C. V. *et al.* Liver Phenotypes of European Adults Heterozygous or Homozygous for Pi∗Z Variant of AAT (Pi∗MZ vs Pi∗ZZ genotype) and Noncarriers. *Gastroenterology* **159**, 534-548.e11 (2020).
38. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat Methods* **18**, 100–106 (2021).
39. Wolf, T. *et al.* HuggingFace's Transformers: State-of-the-art Natural Language Processing. Preprint at https://doi.org/10.48550/arXiv.1910.03771 (2020).
40. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at https://doi.org/10.48550/arXiv.1802.03426 (2020).
41. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

42. Schmacke, N. A. *et al. SPARCS, a Platform for Genome-Scale CRISPR Screening for Spatial Cellular Phenotypes*. http://biorxiv.org/lookup/doi/10.1101/2023.06.01.542416 (2023) doi:10.1101/2023.06.01.542416.
43. Thielert, M., Weiss, C. A. M., Mann, M. & Rosenberger, F. A. Spatial Proteomics of Single Hepatocytes with Multiplexed Data-Independent Acquisition (mDIA). in *Mass Spectrometry Based Single Cell Proteomics* (eds. Vegvari, A., Teppo, J. & Zubarev, R. A.) 97–113 (Springer US, New York, NY, 2024). doi:10.1007/978-1-0716-3934-4_9.
44. 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. *Nat Methods* **17**, 41–44 (2020).
45. Ammar, C., Schessner, J. P., Willems, S., Michaelis, A. C. & Mann, M. Accurate Label-Free Quantification by directLFQ to Compare Unlimited Numbers of Proteomes. *Molecular & Cellular Proteomics* **22**, (2023).
46. Elizarraras, J. M. *et al.* WebGestalt 2024: faster gene set analysis and new support for metabolomics and multi-omics. *Nucleic Acids Research* **52**, W415–W421 (2024).
47. 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 (2023).
48. Perez-Riverol, Y. *et al.* The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Research* **50**, D543–D552 (2022).
49. Rath, S. *et al.* MitoCarta3.0: an updated mitochondrial proteome now with sub-organelle localization and pathway annotations. *Nucleic Acids Research* **49**, D1541–D1547 (2021).

ACKNOWLEDGEMENTS

We thank our colleagues at the Department of Proteomics and Signal Transduction at the Max Planck Institute of Biochemistry as well as our colleagues at the Center for Proteome Research in Copenhagen for their input and support. We are particularly grateful for the technical assistance of Dirk Wischnewski, and for input from Thierry Nordmann, Marvin Thielert and Vincenth Brennsteiner. We thank the Computing Centre and the Imaging Facility of the MPI of Biochemistry for their support and resources. F.A.R. is an EMBO postdoctoral fellow (ALTF 399-2021). S.C.M. is a PhD fellow of the Boehringer Ingelheim Fonds. This study has been supported by the Horizon-2020 under the MICROB-PREDICT program (M.M., A.K., no. 825694); by the Max Planck Society for Advancement of Science (M.M.); by a grant from the

668 Alpha-1 Foundation (F.A.R.) and Alfa-1 Denmark (A.K.); by the Deutsche
669 Forschungsgemeinschaft DFG through SFB 1382 (P.S., ID 403224013); P.S. holds a
670 Heisenberg professorship (STR1095/6-1).

671 **CONTRIBUTIONS**

672 Conceptualization: F.A.R., K.H.T., S.C.M., P.S. and M.M. Project teams were led by S.C.M.
673 (image analysis and machine learning), K.H.T. (clinical data), and S.S. (single-cell analysis).
674 Methodology: F.A.R., S.C.M., S.S. Software: S.C.M., M.L., N.A.S. Validation: F.A.R.,
675 C.A.M.W., M.O., M.W., M.Z., J.K. Formal Analysis: F.A.R., S.C.M., M.L. Investigation:
676 F.A.R., S.C.M., K.H.T., S.S., M.L., C.A.M.W., M.W., M.Z., J.K. Resources: K.H.T., M.F.,
677 S.D., P.B., A.W., O.F., S.F., A.K., P.S., M.M. Data Curation: F.A.R., S.C.M. Writing –
678 Original Draft: F.A.R., M.M. Writing – Review & Editing: all authors. Visualization: F.A.R.,
679 S.C.M., M.L., Supervision: F.A.R, P.S., M.M.

680 **COMPETING INTEREST STATEMENT**

681 MM is an indirect investor in Evosep. The authors declare no other competing interests.

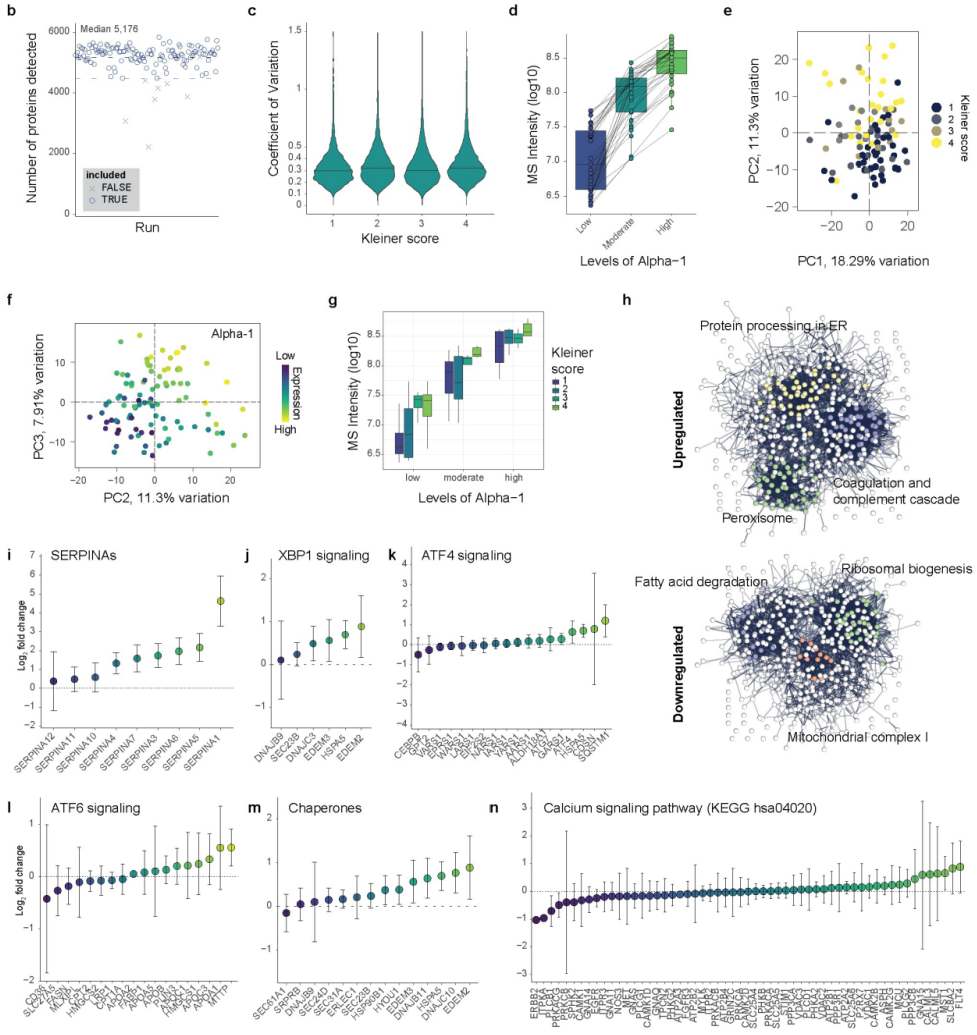
682 **CORRESPONDING AUTHORS**

683 Correspondence to Florian A. Rosenberger (rosenberger@biochem.mpg.de) or Matthias
684 Mann (mmann@biochem.mpg.de).

EXTENDED DATA FIGURES

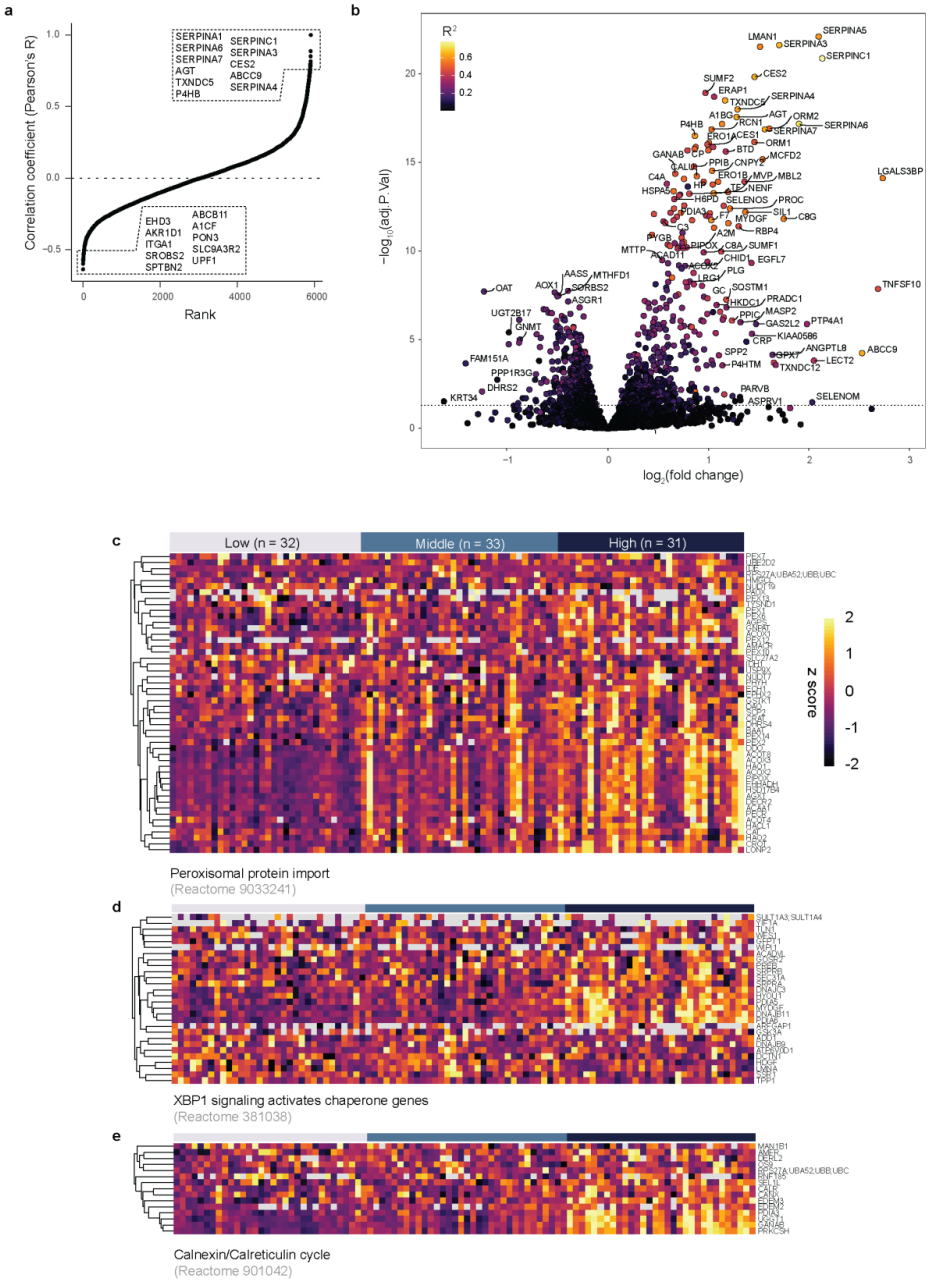
a

Kleiner score	F1	F2	F3	F4	Resections
Genotype			PI*ZZ (100%)		
Female sex	45.5% (5/11)	0% (0/7)	33.3% (1/3)	66.6% (4/6)	100% (3/3, 2 NA)
Age at biopsy (y)	56.9 (±11.5)	49.2 (±9.88)	63.0 (±8.5)	55.6 (±4.0)	63.6 (±5.7)
Alcohol	18.1% (2/11)	28.6% (2/7, 2 NA)	0% (0/3)	33.3% (2/6)	0% (4/6, 1 NA)
Diabetes			2.9% (1/35)		
BMI	24.0 (±3.0)	25.5 (±2.8)	24.2 (±2.3)	26.2 (±1.6)	33.5 (±7.5)



Extended Data Fig. 1, Proteomic mapping of hepatocyte stress response. a, Summary of clinical metadata expressed in number of patients, or percentages with absolute numbers in brackets. Mean ± SD is reported. **b,** Number of proteins detected across all runs prior to exclusion of technical replicates (n = 134). Upper dotted line: median number of protein

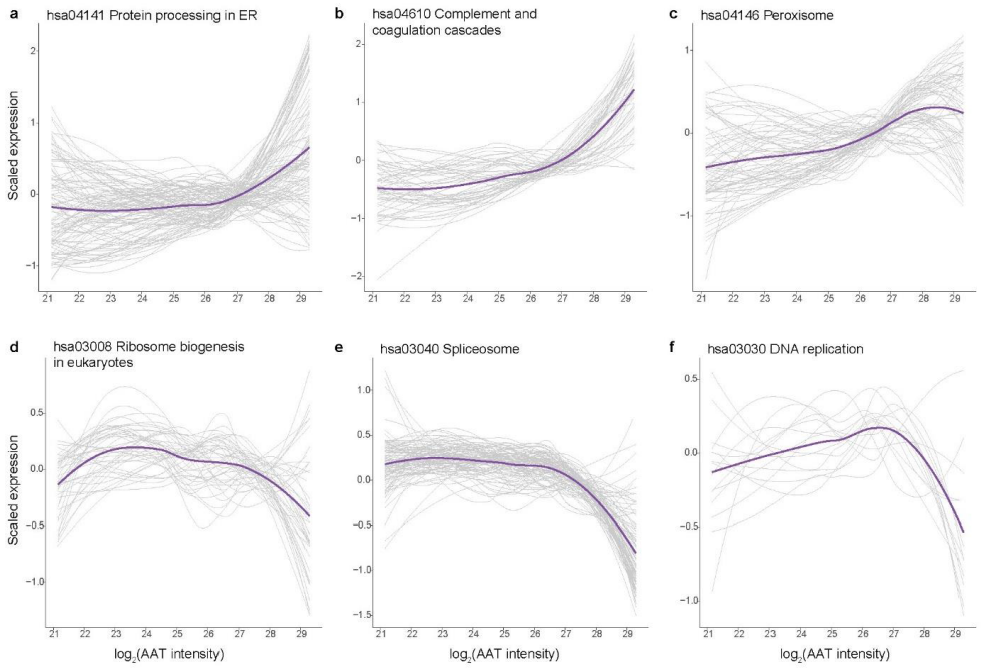
691 groups. Lower dotted line: Median – 1.5 SD. Samples below were excluded and are marked as
692 a cross. **c,** Coefficient of variation across fibrosis stages. **d,** MS intensity of alpha-1 antitrypsin
693 in the three distinctly microdissected cell classes. **e,** Principal component analysis with
694 principal components 1 and 2 color by fibrosis stage, and **f,** with principal component 2 and 3
695 colored by alpha-1 antitrypsin level. Each dot is one sample (n = 95). **g,** Levels of alpha-1
696 antitrypsin by fibrosis stage across the three microdissected cell classes (n = 32 patients). h,
697 STRING interaction network of significantly (FDR < 0.05) upregulated (top) or downregulated
698 proteins in cells (see Fig. 1c). **i – n,** levels of selected proteins in indicated pathways in cells
699 with compared to without aggregates. Circles indicate mean, bars are SD across patient samples
700 (n = 32). The proteins in i to m were manually selected, n is retrieved from KEGG.



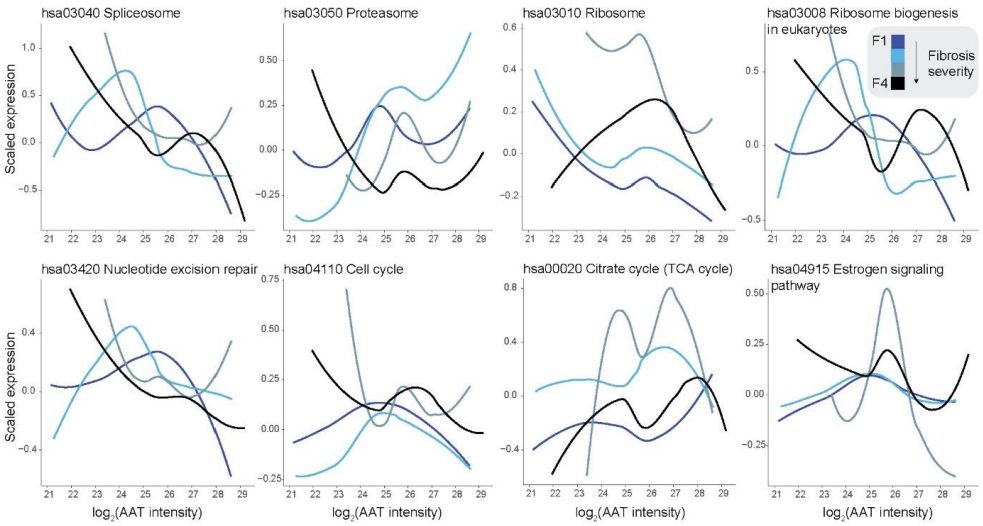
701

702 **Extended Data Fig. 2, Early and late responses to proteotoxic stress. a,** Pearson's R
703 correlation coefficient of each detected protein with alpha-1 antitrypsin levels calculated per
704 MS sample. Top and low-10 protein names are indicated in boxes. b, Proteomic changes in
705 high versus moderate versus low AAT-accumulating cells colored by their R^2 value against
706 alpha-1 antitrypsin expression. Enriched in high on the right side. Top significant and top

changed hits are named (paired two-sided t test with load class as covariable, multiple testing corrected, n = 95 at 100 shapes per sample). **c,** Expression levels of indicated proteins colored by z score (assuming normality) across all samples split by load class and related to peroxisomal protein import, **d,** XBP1 signaling and **e,** the Calnexin/Calreticulin cycle. Database IDs given below each graph (n = 95 in 32 patients).

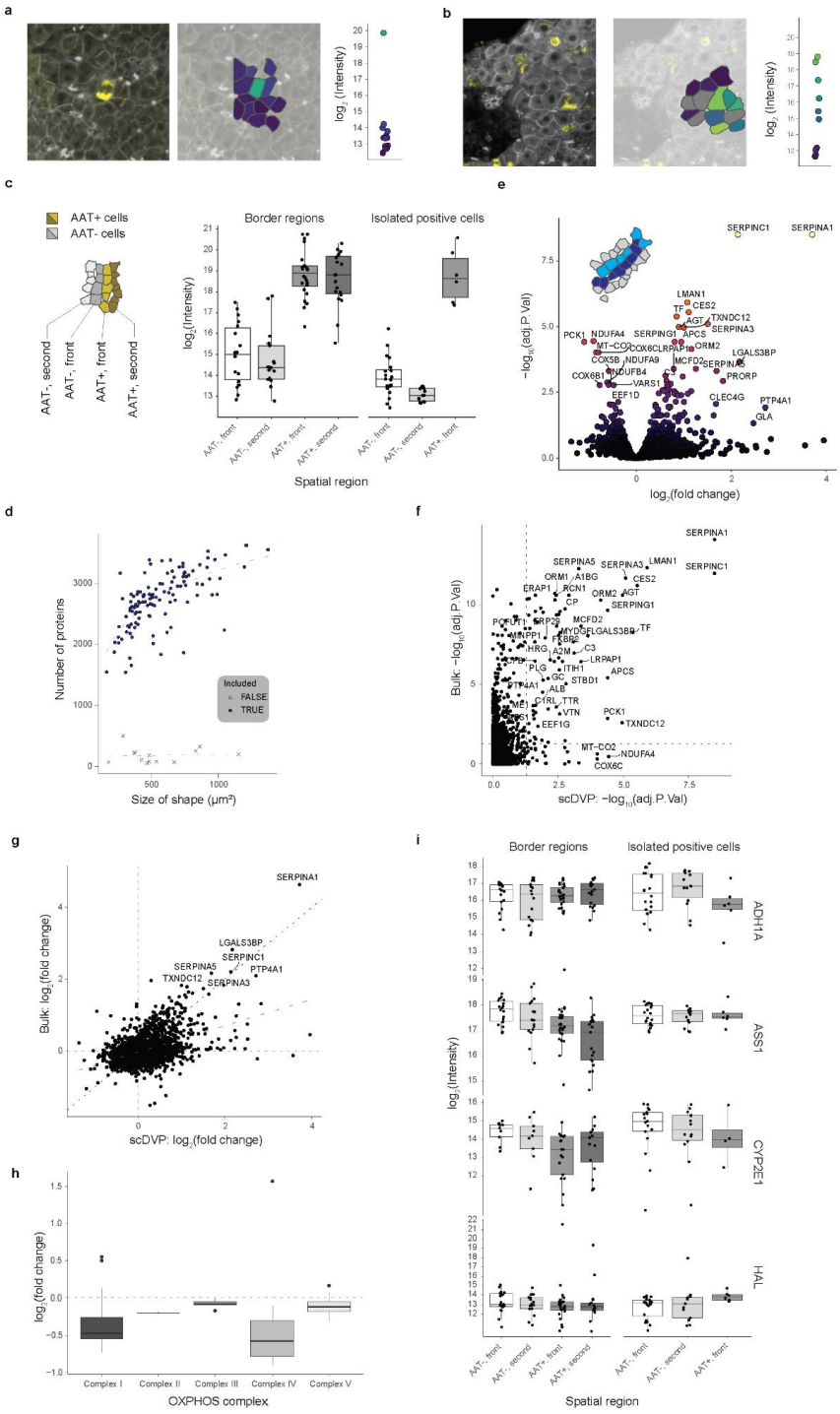


Extended Data Fig. 3, Changes of functional pathways. a-f, Scaled intensity (z scored) of all detected proteins in indicated KEGG pathways against AAT intensity. 'hsa00000' are KEGG identifiers. Purple line is the local regression (span 0.75, degree 2).



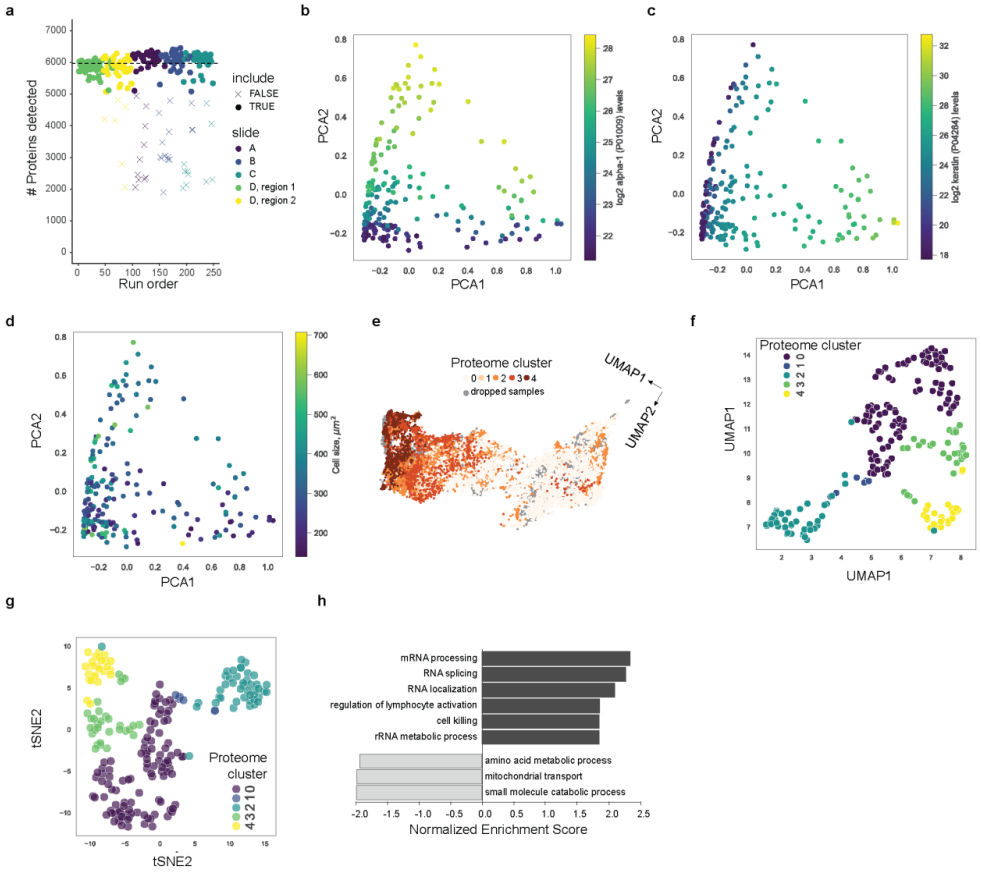
716

717 **Extended Data Fig. 4, Impact of fibrosis on functional pathways in relation to AAT load.**
718 Scaled intensity (z scored per fibrosis group) of all detected proteins in indicated KEGG
719 pathways against AAT intensity. 'hsa00000' are KEGG identifiers. Purple line is the local
720 regression (span 0.75, degree 2). Legend for all panels on top right.


a
b
c
AAT+ cells
AAT- cells
AAT-, second
AAT-, front
AAT+, front
AAT+, second
Border regions
Isolated positive cells
log2(Intensity)
Spatial region
e
SERPINC1
SERPINA1
−log10(adj.P.Val)
log2(fold change)
d
Number of proteins
Size of shape (µm²)
Included
FALSE
TRUE
f
Bulk: −log10(adj.P.Val)
scDVP: −log10(adj.P.Val)
g
Bulk: log2(fold change)
scDVP: log2(fold change)
h
log2(fold change)
Complex I
Complex II
Complex III
Complex IV
Complex V
OXPHOS complex
i
ADH1A
ASS1
CYP2E1
HAL

721

722 **Extended Data Fig. 5, The single-cell proteome. a and b,** Color-coded AAT expression in
723 regions with single-positive cells. AAT expression levels of all indicated shapes are also shown
724 in the dot graph on the right of each spatial mapping. **c,** Expression of AAT in indicated regions
725 determined by immunofluorescence signal across all included samples (n = 118). **d,** Number
726 of proteins detected in relation to the cut shape area. Excluded samples are indicated with a
727 cross. **e,** Statistical comparison of AAT+ and AAT- cells at the three regions classified as
728 'borders' (paired two-sided t test, multiple testing corrected, 30 AAT+ cells and 38 AAT-
729 cells). **f,** Comparison of adjusted p-values and **g,** log2(fold changes) of AAT+ and AAT- single
730 shape comparisons on the x axis versus cells along the accumulation gradient (refer to Fig. 1
731 and 2) on the y axis. Statistics as in e, and Fig. 1c. **h,** Relative expression levels of subunits of
732 the oxidative phosphorylation system (OXPHOS) in AAT+ versus AAT- single shapes.
733 Proteins are retrieved from Mitocarta 3.0 [49]. **i,** Expression of protein indicated on the right in
734 respective spatial region. Periportal markers: ASS1 and HAL; pericentral markers: ALDH1A1
735 and CYP2E1. The boxes are first and third quartiles, the thick line is the median, whiskers are
736 ±1.5 interquartile range and outliers are indicated as individual points.



Extended Data Fig. 6, The proteome of cells with various aggregate morphologies. a, Number of protein groups detected per sample. Each dot is one sample, the horizontal line indicates the mean across all included samples (n = 209 included, n = 41 excluded and marked with a cross). Exclusion criteria were that the number of detected proteins was smaller than mean minus 0.5 SD. **b,** Principal component analysis of all included samples with AAT, **c,** KRT1 expression levels, or **d,** shape size color coded (n = 209). **e,** Annotation of the proteome cluster in Fig. 4d onto the image space UMAP. Dropped samples are in grey (n = 12,500). **f,** Representation of individual samples color coded by proteome cluster in a proteomic UMAP, or **g,** tSNE space (n = 209). **h,** Gene Set Enrichment Analysis (GO: Biological Process noRedundant) of globular versus amorphous aggregate types.

4. Discussion and Outlook

The increasing sensitivity and speed of MS instrumentation in the last years, has driven a great wave of impactful publications. Since its introduction in June 2023, already more than 100 papers using the Orbitrap Astral MS were published, of which over 40 are peer-reviewed. These cover a wide range of applications, from full proteomes over PTM analysis to microbiomes and additionally enable unprecedented proteome depths of >5000 proteins from single cells.[168,215,217,326,328,426,439,440]

Presenting at the instrument release, I could showcase initial results for our DVP workflow. In a titration experiment of epithelial cells from patients with high grade serous ovarian cancer, we could identify almost 2,000 protein groups from as little of 10 cell shapes and more than 5,000 protein groups from 100 shapes. A depth that previously required the analysis of 500-700 shapes.[395] Additionally, we were able to identify the primary ovarian cancer biomarker, CA-125, in as little as 25 cell shapes. With these promising results in hand, we focused our effort on these projects in the last year. DVP presents a unique opportunity to preserve the spatial aspect of cell type-specific proteomes in the context of intact tissue. In contrast, other methodologies, such as macrodissection or cell sorting, can only preserve the spatial or cell type resolution respectively. DVP, therefore, is of particular interest when studying specific cell types in cases of distinct spatial characteristics, such as the crypt-villus architecture in the intestinal mucosa (Article 5) or to spatially differentiate cancerous and non-cancerous cells (Article 4). With more advanced MS technology, the number of required cells per cell-type further decreases and broadens the applicability of DVP to the study of more rare cell types or where total cell amount is limited, as is the case in organoid models, for instance.

For research questions that require higher spatial resolution, scDVP offers a more fine-grained analysis of the spatial proteome of single cells in intact tissue. First applied to fresh frozen tissue sections and used to study the spatial organization of hepatocytes in the central to portal vein axis, we could extend the workflow to FFPE (Article 6) to study liver sections of patients with AATD. As clinical tissue samples are commonly archived as FFPE tissue and often available as parts of biobanks, this greatly extends the number of sample cohorts that can be studied using scDVP. Further, we increased the achievable proteome depth from a mean of 1700 proteins to 2800 proteins by a combination of technological advances and optimal method design, and were able to

pick up biologically and clinically relevant proteomic changes. With this more sensitive and optimized set-up, scDVP studies of smaller cell types become increasingly more feasible further extending the possible use cases. Similar to AATD, it would be valuable to apply scDVP to other clinical conditions featuring protein misfolding and aggregation in individual cells, such as the neurodegenerative diseases Parkinson and Alzheimer.[441] This is currently being investigated in our group. Apart from the disease context scDVP could give insights into developmental and regenerative processes that require a spatial single cell resolution. Altogether, the technological improvements in the DVP and scDVP studies presented in this thesis showcase the potential of DVP, especially in spatially resolved clinical proteomics and highlight use cases in precision oncology or personalized medicine. They, however, also spotlight the importance of highly sensitive mass spectrometers, such as the Orbitrap Astral MS, and tailored method design to achieve high proteomic depth and quantitative accuracy.

While new MS instruments have greatly improved on the previously achievable proteomic depth, there is still room for further improvements. The fast scanning speeds of modern MS analyzers enable us to reduce DIA windows to an almost DDA-like width, considerably reducing the spectral complexity of each DIA window and in turn increasing identification.[326] This, however, means we are only ever analyzing a small fraction of the total ion population. In contrast, an ideal mass spectrometer or acquisition strategy would utilize all entering ions for subsequent analysis. Over the years, multiple approaches, both technical and methodological, have been proposed to improve on this. One such methodological strategy is BoxCar, which increases ion utilization, total injection time for MS full scans, and with it dynamic range and sensitvity.[442] While DIA acquisition shifted the focus towards MS2 spectra, recent experimental data reemphasizes the importance of high quality MS1 data, especially for low input samples and for improved quantification.[61] Depending of the mass analyzer used, this, however, often requires long transient times, which might not be feasible especially for high-throughput applications. In these cases, the ability to acquire MS1 and MS2 spectra in parallel, as instruments with more than one mass analyzer can do, is particularly advantageous. BoxCar-like acquisition strategies for MS1 or potentially MS2 level in combination with tribrid instruments or the Orbitrap Astral MS could help increase ion utilization while maintaining high proteomic depth. Ultimately, optimal ion usage on the Orbitrap Astral MS will require a technical solution similar to the trapped ion mobility spectrometry (TIMS) device on Bruker instruments, or Sciex's Zeno Trap technology.

While discussing the impressive performance of novel mass spectrometers and potential ideas to further improve their capabilities, one should, however, note that not every group has the financial means to upgrade to the newest instrument releases. For this reason, I think it is important to also extend the functionality of existing mass spectrometers through hardware or software add-ons and increasingly more refined acquisition strategies. On the side of Thermo Fisher Scientific instrumentation for instance, the unification of the MS front-end design between tribrid and hybrid MS instruments enabled the use of the FAIMS ion mobility device for the hybrid MS instruments. While this can extend the time between cleaning cycles, making the instruments more robust, it can also greatly improve the performance for low-input applications by removing background ions. On the same line, ΦSDM (Article 1), potentially as a commercially available upgrade, could increase the performance and functionality of existing Orbitrap mass spectrometers in groups that cannot afford to exchange their MS instruments with the newest generation.

Even if we are utilizing our mass spectrometers to the best of their abilities, all of this is diminished without analysis or post-processing software that makes optimal use of the acquired data. In line with this, the introduction of AlphaDIA (Article 3) provides a great framework for the search of DIA data, particularly for potentially “noisier” TOF data. The aggregation of evidence across multiple dimensions allows the confident identification of peptides and precursors even at low fragment intensities. As novel analyzers promise single ion detection, this will be of particular importance to retain low FDR and high identification confidence. Moreover, AlphaDIA’s flexible processing algorithm combined with alphaRaw’s efficient raw data handling promised high adaptability to novel and complex scan modes, including synchro-PASEF.[300] The integration of AlphaPeptDeep, for prediction of spectral libraries, and directLFQ provide an end-to-end solution for raw data analysis.[126,131] The former also highlight the use of deep learning for the prediction of peptide properties, training of highly tailored models, including HLA peptides and PTMs, and generation of *in-silico* libraries. AlphaDIA, as well as other software solutions of the alphaX universe, are built with modern and open-source tools like Python and PyTorch and openly provided to the community on GitHub. This stand in contrast to other commonly used DIA analysis software, whose “inner workings” more often than not are “black boxes”. With this, AlphaDIA sets an example for transparent, open science that performs on par or better than other popular DIA search platforms, particularly for TOF analyzers, such as the Orbitrap Astral MS. Here, AlphaDIA was able to identify 9,500 proteins groups from a 21 min run, outperforming the other analysis tools.

While in this case proper FDR control ensures high confidence in the identified peptides and proteins, one should always prioritize reproducible, high-quality datasets over a sole focus on who gets the highest numbers. As such, there is much to say about the “numbers game” in proteomics. On the side of MS instrumentation continuously or sometimes drastically improving instrument parameters upkeep the commercial competition between MS vendors. This promotes innovation in order to stay competitive, driving the field forwards. Just in the last years this enables, almost routine identification of full proteomes, deeper plasma proteomes, and covers the single cell proteome at a biologically and potentially clinically relevant depth, goals that the community was working towards for a long time.[369,443,444] However, we have also seen that purely focusing on achieving the highest numbers possible, through any means necessary, might be accompanied with higher false identifications, unreproducible results, and, in translation to clinical proteomics, can lead to the misidentification of biomarkers. Examples for this can be found in the early days of plasma proteomics, were achieved depths and identified biomarkers were, in hindsight, associated with cohort batch effects or lack of sample quality. Consequently, this decreased trust that MS-based plasma proteomics could aid in the efforts to identify disease biomarkers.[369,445,446] With a revival of the plasma proteomics field in the last years, a greater focus was placed on achieving translatable data, including proposed improvements to cohort design and awareness of sample quality biases.[344,362,447] The latter, revealed that a great number of previously identified plasma biomarkers can be attributed to sample processing artefacts, such as erythrocyte and platelet contaminations. The proposed contamination marker panel provides a useful tool to evaluate cohort quality and increases confidence in potential protein markers, such as HPR in our bed rest study (Article 2). As a result of these efforts, first examples of promising MS-based marker panels for diagnosis have set the stage for the MS-based proteomic approaches in the clinic.[17,394,448]

The translation to clinical application, however, will require further validation, the establishment of easy-to-use MS-based assays and MS systems that can be maintained and operated by non-expert users.[361,449] Steps in the right direction are the recent advances in MS systems focusing on targeted proteomics, such as the Thermo Stellar MS, which improves on the dated triple-quad technology and allows rapid and highly sensitive PRM and MS3 targeting.[342,343] This enables the targeting of thousands of peptides in a single run and can adapt target lists that were previously generated using discovery DIA on high-resolution mass spectrometers, such as the Orbitrap Astral. Implementation of an auto-calibration source for easier maintenance, additionally makes this instrument more user-friendly. As such it could provide a solution for establishing

targeted MS-based assays for a variety of disease marker panels, as highlighted by the development of a targeted assay for the previously proposed alcohol-related liver disease biomarkers.[17,342]

In summary, the recent improvements and innovations in MS technology have greatly and positively impacted the proteomics field and will go hand-in-hand with advances in data analysis, such as AlphaDIA and applications of machine learning and artificial intelligence in proteomics. In my thesis, I highlighted the performance of novel MS instrumentation, namely the Orbitrap Astral MS, and the importance of tailored acquisition strategies. The application to clinical proteomics with a focus on spatial proteomics and biomarker discovery, showcased the great potential and adaptability of our previously introduced DVP workflow. Altogether, I am sure there are exciting times and great discoveries ahead and I for one am looking forward to what the future and my continued journey in MS technology will bring.

5. References

1. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
2. Mattick, J. S. & Makunin, I. V. Non-coding RNA. *Hum. Mol. Genet.* **15**, R17–R29 (2006).
3. Palazzo, A. F. & Gregory, T. R. The Case for Junk DNA. *PLoS Genet.* **10**, e1004351 (2014).
4. Author, N. G. *Understanding Our Genetic Inheritance: The US Human Genome Project, The First Five Years FY 1991--1995*. DOE/ER-0452P, 6958032 http://www.osti.gov/servlets/purl/6958032/ (1990) doi:10.2172/6958032.
5. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
6. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
7. Rhie, A. *et al.* The complete sequence of a human Y chromosome. *Nature* **621**, 344–354 (2023).
8. Omenn, G. S. *et al.* The 2023 Report on the Proteome from the HUPO Human Proteome Project. *J. Proteome Res.* **23**, 532–549 (2024).
9. Ponomarenko, E. A. *et al.* The Size of the Human Proteome: The Width and Depth. *Int. J. Anal. Chem.* **2016**, 1–6 (2016).
10. Aebersold, R. *et al.* How many human proteoforms are there? *Nat. Chem. Biol.* **14**, 206–214 (2018).
11. Tyers, M. & Mann, M. From genomics to proteomics. *Nature* **422**, 193–197 (2003).
12. Wilkins, M. R. *et al.* From Proteins to Proteomes: Large Scale Protein Identification by Two-Dimensional Electrophoresis and Arnino Acid Analysis. *Nat. Biotechnol.* **14**, 61–65 (1996).
13. Carbonara, K., Andonovski, M. & Coorssen, J. R. Proteomes Are of Proteoforms: Embracing the Complexity. *Proteomes* **9**, 38 (2021).
14. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
15. Kobayashi, E. *et al.* Biomarkers for Screening, Diagnosis, and Monitoring of Ovarian Cancer. *Cancer Epidemiol. Biomarkers Prev.* **21**, 1902–1912 (2012).
16. Åkesson, J. *et al.* Proteomics reveal biomarkers for diagnosis, disease activity and long-term disability outcomes in multiple sclerosis. *Nat. Commun.* **14**, 6903 (2023).
17. Niu, L. *et al.* Noninvasive proteomic biomarkers for alcohol-related liver disease. *Nat. Med.* **28**, 1277–1287 (2022).

18. Santos, R. *et al.* A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.* **16**, 19–34 (2017).
19. Ebrahimi, S. B. & Samanta, D. Engineering protein-based therapeutics through structural and chemical design. *Nat. Commun.* **14**, 2411 (2023).
20. Müller, T. D., Blüher, M., Tschöp, M. H. & DiMarchi, R. D. Anti-obesity drug discovery: advances and challenges. *Nat. Rev. Drug Discov.* **21**, 201–223 (2022).
21. Garvey, W. T. *et al.* Two-year effects of semaglutide in adults with overweight or obesity: the STEP 5 trial. *Nat. Med.* **28**, 2083–2091 (2022).
22. Sinitcyn, P. *et al.* Global detection of human variants and isoforms by deep proteome sequencing. *Nat. Biotechnol.* **41**, 1776–1786 (2023).
23. Wright, B. W., Yi, Z., Weissman, J. S. & Chen, J. The dark proteome: translation from noncanonical open reading frames. *Trends Cell Biol.* **32**, 243–258 (2022).
24. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
25. Domon, B. & Aebersold, R. Mass Spectrometry and Protein Analysis. *Science* **312**, 212–217 (2006).
26. Catherman, A. D., Skinner, O. S. & Kelleher, N. L. Top Down proteomics: Facts and perspectives. *Biochem. Biophys. Res. Commun.* **445**, 683–693 (2014).
27. Lermyte, F., Tsybin, Y. O., O'Connor, P. B. & Loo, J. A. Top or Middle? Up or Down? Toward a Standard Lexicon for Protein Top-Down and Allied Mass Spectrometry Approaches. *J. Am. Soc. Mass Spectrom.* **30**, 1149–1157 (2019).
28. Sidoli, S. & Garcia, B. A. Middle-down proteomics: a still unexploited resource for chromatin biology. *Expert Rev. Proteomics* **14**, 617–626 (2017).
29. Zhang, Y., Fonslow, B. R., Shan, B., Baek, M.-C. & Yates, J. R. Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.* **113**, 2343–2394 (2013).
30. Roberts, D. S. *et al.* Top-down proteomics. *Nat. Rev. Methods Primer* **4**, 38 (2024).
31. Toby, T. K., Fornelli, L. & Kelleher, N. L. Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annu. Rev. Anal. Chem.* **9**, 499–519 (2016).
32. Hildonen, S., Halvorsen, T. G. & Reubsaet, L. Why less is more when generating tryptic peptides in bottom-up proteomics. *PROTEOMICS* **14**, 2031–2041 (2014).
33. Olsen, J. V., Ong, S.-E. & Mann, M. Trypsin Cleaves Exclusively C-terminal to Arginine and Lysine Residues. *Mol. Cell. Proteomics* **3**, 608–614 (2004).
34. Cristobal, A. *et al.* Toward an Optimized Workflow for Middle-Down Proteomics. *Anal. Chem.* **89**, 3318–3325 (2017).

35. Pandeswari, P. B. & Sabareesh, V. Middle-down approach: a choice to sequence and characterize proteins/proteomes by mass spectrometry. *RSC Adv.* **9**, 313–344 (2018).
36. Gillet, L. C., Leitner, A. & Aebersold, R. Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. *Annu. Rev. Anal. Chem.* **9**, 449–472 (2016).
37. 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).
38. 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. *Nat. Methods* **14**, 513–520 (2017).
39. Bruderer, R. *et al.* Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. *Mol. Cell. Proteomics MCP* **16**, 2296–2309 (2017).
40. 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. *Nat. Methods* **17**, 41–44 (2020).
41. Hein, M. Y., Sharma, K., Cox, J. & Mann, M. Chapter 1 - Proteomic Analysis of Cellular Systems. in *Handbook of Systems Biology* (eds. Walhout, A. J. M., Vidal, M. & Dekker, J.) 3–25 (Academic Press, San Diego, 2013). doi:10.1016/B978-0-12-385944-0.00001-0.
42. Means, G. E. & Feeney, R. E. Reductive alkylation of amino groups in proteins. *Biochemistry* **7**, 2192–2201 (1968).
43. Ren, Y. *et al.* Evaluation and minimization of over-alkylation in proteomic sample preparation. *Int. J. Mass Spectrom.* **481**, 116919 (2022).
44. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **6**, 359–362 (2009).
45. Saveliev, S. *et al.* Trypsin/Lys-C protease mix for enhanced protein mass spectrometry analysis. *Nat. Methods* **10**, i–ii (2013).
46. Sinha, A. & Mann, M. A beginner's guide to mass spectrometry–based proteomics. *The Biochemist* **42**, 64–69 (2020).
47. Wiśniewski, J. R. Filter-Aided Sample Preparation. in *Methods in Enzymology* vol. 585 15–27 (Elsevier, 2017).
48. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11**, 319–324 (2014).

49. Hughes, C. S. *et al.* Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nat. Protoc.* **14**, 68–85 (2019).
50. Batth, T. S. *et al.* Protein Aggregation Capture on Microparticles Enables Multipurpose Proteomics Sample Preparation*. *Mol. Cell. Proteomics* **18**, 1027a–11035 (2019).
51. Kulak, N. A., Geyer, P. E. & Mann, M. Loss-less Nano-fractionator for High Sensitivity, High Coverage Proteomics. *Mol. Cell. Proteomics* **16**, 694–705 (2017).
52. Batth, T. S., Francavilla, C. & Olsen, J. V. Off-Line High-pH Reversed-Phase Fractionation for In-Depth Phosphoproteomics. *J. Proteome Res.* **13**, 6176–6186 (2014).
53. Huang, J., Wang, F., Ye, M. & Zou, H. Enrichment and separation techniques for large-scale proteomics analysis of the protein post-translational modifications. *J. Chromatogr. A* **1372**, 1–17 (2014).
54. Sharma, K. *et al.* Ultradeep Human Phosphoproteome Reveals a Distinct Regulatory Nature of Tyr and Ser/Thr-Based Signaling. *Cell Rep.* **8**, 1583–1594 (2014).
55. Cantin, G. & Yatesiii, J. Strategies for shotgun identification of post-translational modifications by mass spectrometry. *J. Chromatogr. A* **1053**, 7–14 (2004).
56. Horváth, C., Melander, W. & Molnár, I. Solvophobic interactions in liquid chromatography with nonpolar stationary phases. *J. Chromatogr. A* **125**, 129–156 (1976).
57. Kovalchuk, S. I., Jensen, O. N. & Rogowska-Wrzesinska, A. FlashPack: Fast and Simple Preparation of Ultrahigh-performance Capillary Columns for LC-MS*[S]. *Mol. Cell. Proteomics* **18**, 383–390 (2019).
58. Müller-Reif, J. B. *et al.* A New Parallel High-Pressure Packing System Enables Rapid Multiplexed Production of Capillary Columns. *Mol. Cell. Proteomics* **20**, 100082 (2021).
59. Müller, J. B. *et al.* The proteome landscape of the kingdoms of life. *Nature* **582**, 592–596 (2020).
60. Stadlmann, J. *et al.* Improved Sensitivity in Low-Input Proteomics Using Micropillar Array-Based Chromatography. *Anal. Chem.* **91**, 14203–14207 (2019).
61. Petrosius, V. *et al.* Exploration of cell state heterogeneity using single-cell proteomics through sensitivity-tailored data-independent acquisition. *Nat. Commun.* **14**, 5910 (2023).
62. Desmet, G. *et al.* Separation efficiency kinetics of capillary flow micro-pillar array columns for liquid chromatography. *J. Chromatogr. A* **1626**, 461279 (2020).

63. Futagami, S. *et al.* Study of peak capacities generated by a porous layered radially elongated pillar array column coupled to a nano-LC system. *Analyst* **144**, 1809–1817 (2019).
64. Bache, N. *et al.* A Novel LC System Embeds Analytes in Pre-formed Gradients for Rapid, Ultra-robust Proteomics*. *Mol. Cell. Proteomics* **17**, 2284–2296 (2018).
65. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. https://www.science.org/doi/10.1126/science.2675315 doi:10.1126/science.2675315.
66. Li, K.-Y., Tu, H. & Ray, A. K. Charge Limits on Droplets during Evaporation. *Langmuir* **21**, 3786–3794 (2005).
67. Banerjee, S. & Mazumdar, S. Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte. *Int. J. Anal. Chem.* **2012**, 282574 (2012).
68. Wilm, M. S. & Mann, M. Electrospray and Taylor-Cone theory, Dole's beam of macromolecules at last? *Int. J. Mass Spectrom. Ion Process.* **136**, 167–180 (1994).
69. Wilm, M. & Mann, M. Analytical Properties of the Nanoelectrospray Ion Source. *Anal. Chem.* **68**, 1–8 (1996).
70. Chen, S., Zeng, J., Zhang, Z., Xu, B. & Zhang, B. Recent advancements in nanoelectrospray ionization interface and coupled devices. *J. Chromatogr. Open* **2**, 100064 (2022).
71. Liu, H. *et al.* The Prediction of Peptide Charge States for Electrospray Ionization in Mass Spectrometry. *Procedia Environ. Sci.* **8**, 483–491 (2011).
72. Wong, S. F., Meng, C. K. & Fenn, J. B. Multiple charging in electrospray ionization of poly(ethylene glycols). *J. Phys. Chem.* **92**, 546–550 (1988).
73. Fernandez de la Mora, J. Electrospray ionization of large multiply charged species proceeds via Dole's charged residue mechanism. *Anal. Chim. Acta* **406**, 93–104 (2000).
74. Biniossek, M. L. & Schilling, O. Enhanced identification of peptides lacking basic residues by LC-ESI-MS/MS analysis of singly charged peptides. *PROTEOMICS* **12**, 1303–1309 (2012).
75. Wahle, M. *et al.* IMBAS-MS Discovers Organ-Specific HLA Peptide Patterns in Plasma. *Mol. Cell. Proteomics* **23**, 100689 (2024).
76. Wilhelm, M. *et al.* Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics. *Nat. Commun.* **12**, 3346 (2021).
77. Dupree, E. J. *et al.* A Critical Review of Bottom-Up Proteomics: The Good, the Bad, and the Future of This Field. *Proteomes* **8**, 14 (2020).

78. 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).
79. Haag, A. M. Mass Analyzers and Mass Spectrometers. in *Modern Proteomics – Sample Preparation, Analysis and Practical Applications* (eds. Mirzaei, H. & Carrasco, M.) 157–169 (Springer International Publishing, Cham, 2016). doi:10.1007/978-3-319-41448-5_7.
80. Li, C. *et al.* Towards Higher Sensitivity of Mass Spectrometry: A Perspective From the Mass Analyzers. *Front. Chem.* **9**, (2021).
81. Makarov, A. Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. *Anal. Chem.* **72**, 1156–1162 (2000).
82. Douglas, D. J., Frank, A. J. & Mao, D. Linear ion traps in mass spectrometry. *Mass Spectrom. Rev.* **24**, 1–29 (2005).
83. Guilhaus, M., Selby, D. & Mlynski, V. Orthogonal acceleration time-of-flight mass spectrometry. *Mass Spectrom. Rev.* **19**, 65–107 (2000).
84. Bogdanov, B. & Smith, R. D. Proteomics by FTICR mass spectrometry: Top down and bottom up. *Mass Spectrom. Rev.* **24**, 168–200 (2005).
85. Meier, F. *et al.* Online Parallel Accumulation–Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer*. *Mol. Cell. Proteomics* **17**, 2534–2545 (2018).
86. Glish, G. L. & Burinsky, D. J. Hybrid mass spectrometers for tandem mass spectrometry. *J. Am. Soc. Mass Spectrom.* **19**, 161–172 (2008).
87. Roepstorff, P. & Fohlman, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom.* **11**, 601 (1984).
88. Wysocki, V. H., Resing, K. A., Zhang, Q. & Cheng, G. Mass spectrometry of peptides and proteins. *Methods* **35**, 211–222 (2005).
89. Good, D. M., Wirtala, M., McAlister, G. C. & Coon, J. J. Performance Characteristics of Electron Transfer Dissociation Mass Spectrometry*. *Mol. Cell. Proteomics* **6**, 1942–1951 (2007).
90. Brodbelt, J. S., Morrison, L. J. & Santos, I. Ultraviolet Photodissociation Mass Spectrometry for Analysis of Biological Molecules. *Chem. Rev.* **120**, 3328 (2019).
91. Kim, M.-S. & Pandey, A. Electron Transfer Dissociation Mass Spectrometry in Proteomics. *Proteomics* **12**, 530 (2012).
92. Geromanos, S. J. *et al.* The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS. *PROTEOMICS* **9**, 1683–1695 (2009).

93. Michalski, A., Cox, J. & Mann, M. More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority is Inaccessible to Data-Dependent LC−MS/MS. *J. Proteome Res.* **10**, 1785–1793 (2011).
94. Meier, F. *et al.* diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nat. Methods* **17**, 1229–1236 (2020).
95. 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*. *Mol. Cell. Proteomics* **11**, O111.016717 (2012).
96. Ludwig, C. *et al.* Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* **14**, e8126 (2018).
97. Lou, R. & Shui, W. Acquisition and Analysis of DIA-Based Proteomic Data: A Comprehensive Survey in 2023. *Mol. Cell. Proteomics MCP* **23**, 100712 (2024).
98. Bruderer, R. *et al.* Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell. Proteomics MCP* **14**, 1400–1410 (2015).
99. Doerr, A. DIA mass spectrometry. *Nat. Methods* **12**, 35–35 (2015).
100. Sinitcyn, P. *et al.* MaxDIA enables library-based and library-free data-independent acquisition proteomics. *Nat. Biotechnol.* **39**, 1563–1573 (2021).
101. Frejno, M. *et al.* Unifying the analysis of bottom-up proteomics data with CHIMERYS. 2024.05.27.596040 Preprint at https://doi.org/10.1101/2024.05.27.596040 (2024).
102. Röst, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* **32**, 219–223 (2014).
103. Venable, J. D., Dong, M.-Q., Wohlschlegel, J., Dillin, A. & Yates, J. R. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods* **1**, 39–45 (2004).
104. Yu, F. *et al.* Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. *Nat. Commun.* **14**, 4154 (2023).
105. Lange, V., Picotti, P., Domon, B. & Aebersold, R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* **4**, 222 (2008).
106. Gallien, S., Duriez, E., Demeure, K. & Domon, B. Selectivity of LC-MS/MS analysis: Implication for proteomics experiments. *J. Proteomics* **81**, 148–158 (2013).
107. Gallien, S., Bourmaud, A., Kim, S. Y. & Domon, B. Technical considerations for large-scale parallel reaction monitoring analysis. *J. Proteomics* **100**, 147–159 (2014).

108. Peterson, A. C., Russell, J. D., Bailey, D. J., Westphall, M. S. & Coon, J. J. Parallel Reaction Monitoring for High Resolution and High Mass Accuracy Quantitative, Targeted Proteomics *. *Mol. Cell. Proteomics* **11**, 1475–1488 (2012).
109. Whiteaker, J. R. *et al.* Targeted Mass Spectrometry Enables Quantification of Novel Pharmacodynamic Biomarkers of ATM Kinase Inhibition. *Cancers* **13**, 3843 (2021).
110. Remes, P. M., Yip, P. & MacCoss, M. J. Highly Multiplex Targeted Proteomics Enabled by Real-Time Chromatographic Alignment. *Anal. Chem.* **92**, 11809–11817 (2020).
111. Nesvizhskii, A. I., Vitek, O. & Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **4**, 787–797 (2007).
112. Standing, K. G. Peptide and protein *de novo* sequencing by mass spectrometry. *Curr. Opin. Struct. Biol.* **13**, 595–601 (2003).
113. Tran, N. H., Zhang, X., Xin, L., Shan, B. & Li, M. De novo peptide sequencing by deep learning. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 8247–8252 (2017).
114. Taylor, J. A. & Johnson, R. S. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom. RCM* **11**, 1067–1075 (1997).
115. Seidler, J., Zinn, N., Boehm, M. E. & Lehmann, W. D. De novo sequencing of peptides by MS/MS. *PROTEOMICS* **10**, 634–649 (2010).
116. Muth, T. & Renard, B. Y. Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Brief. Bioinform.* **19**, 954–970 (2018).
117. Strauss, M. T. *et al.* AlphaPept: a modern and open framework for MS-based proteomics. *Nat. Commun.* **15**, 2168 (2024).
118. Cox, J. *et al.* Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
119. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
120. Perkins, D. N., Pappin, D. J. C., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *ELECTROPHORESIS* **20**, 3551–3567 (1999).
121. Tsou, C.-C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **12**, 258–264 (2015).

122. Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* **73**, 2092–2123 (2010).
123. Bernhardt, O. *et al.* Spectronaut: a fast and efficient algorithm for MRM-like processing of data independent acquisition (SWATH-MS) data. in (2014).
124. Navarro, P. *et al.* A multicenter study benchmarks software tools for label-free proteome quantification. *Nat. Biotechnol.* **34**, 1130–1136 (2016).
125. Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518 (2019).
126. Zeng, W.-F. *et al.* AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics. *Nat. Commun.* **13**, 7238 (2022).
127. Bouwmeester, R., Gabriels, R., Hulstaert, N., Martens, L. & Degroeve, S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nat. Methods* **18**, 1363–1369 (2021).
128. Isaksson, M., Karlsson, C., Laurell, T., Kirkeby, A. & Heusel, M. MSLibrarian: Optimized Predicted Spectral Libraries for Data-Independent Acquisition Proteomics. *J. Proteome Res.* **21**, 535–546 (2022).
129. Yang, Y. *et al.* In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nat. Commun.* **11**, 146 (2020).
130. Wallmann, G. *et al.* AlphaDIA enables End-to-End Transfer Learning for Feature-Free Proteomics. 2024.05.28.596182 Preprint at https://doi.org/10.1101/2024.05.28.596182 (2024).
131. 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. Proteomics* **22**, (2023).
132. Prieto, G. & Vázquez, J. Calculation of False Discovery Rate for Peptide and Protein Identification. in *Mass Spectrometry Data Analysis in Proteomics* (ed. Matthiesen, R.) 145–159 (Springer, New York, NY, 2020). doi:10.1007/978-1-4939-9744-2_6.
133. 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).
134. Asara, J. M., Christofk, H. R., Freimark, L. M. & Cantley, L. C. A label-free quantification method by MS/MS TIC compared to SILAC and spectral counting in a proteomics screen. *PROTEOMICS* **8**, 994–999 (2008).

135. Cox, J. *et al.* Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ*. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).
136. Nahnsen, S., Bielow, C., Reinert, K. & Kohlbacher, O. Tools for Label-free Peptide Quantification *. *Mol. Cell. Proteomics* **12**, 549–556 (2013).
137. Thompson, A. *et al.* Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **75**, 1895–1904 (2003).
138. Ross, P. L. *et al.* Multiplexed Protein Quantitation in *Saccharomyces cerevisiae* Using Amine-reactive Isobaric Tagging Reagents*. *Mol. Cell. Proteomics* **3**, 1154–1169 (2004).
139. Virreira Winter, S. *et al.* EASI-tag enables accurate multiplexed and interference-free MS2-based proteome quantification. *Nat. Methods* **15**, 527–530 (2018).
140. Wiese, S., Reidegeld, K. A., Meyer, H. E. & Warscheid, B. Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research. *PROTEOMICS* **7**, 1004–1004 (2007).
141. Sivanich, M. K., Gu, T.-J., Tabang, D. N. & Li, L. Recent advances in isobaric labeling and applications in quantitative proteomics. *PROTEOMICS* **22**, 2100256 (2022).
142. Wang, Z. *et al.* 27-Plex Tandem Mass Tag Mass Spectrometry for Profiling Brain Proteome in Alzheimer's Disease. *Anal. Chem.* **92**, 7162–7170 (2020).
143. Zuniga, N. R. *et al.* Achieving a 35-Plex Tandem Mass Tag Reagent Set through Deuterium Incorporation. *J. Proteome Res.* (2024) doi:10.1021/acs.jproteome.4c00668.
144. Jiang, H. & English, A. M. Quantitative Analysis of the Yeast Proteome by Incorporation of Isotopically Labeled Leucine. *J. Proteome Res.* **1**, 345–350 (2002).
145. Ong, S.-E. *et al.* Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics*. *Mol. Cell. Proteomics* **1**, 376–386 (2002).
146. Kang, U.-B., Yeom, J., Kim, H. & Lee, C. Quantitative Analysis of mTRAQ-Labeled Proteome Using Full MS Scans. *J. Proteome Res.* **9**, 3750–3758 (2010).
147. DeSouza, L. V. *et al.* Multiple Reaction Monitoring of mTRAQ-Labeled Peptides Enables Absolute Quantification of Endogenous Levels of a Potential Cancer Marker in Cancerous and Normal Endometrial Tissues. *J. Proteome Res.* **7**, 3525–3534 (2008).

148. Boersema, P. J., Raijmakers, R., Lemeer, S., Mohammed, S. & Heck, A. J. R. Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat. Protoc.* **4**, 484–494 (2009).
149. Hsu, J.-L., Huang, S.-Y., Chow, N.-H. & Chen, S.-H. Stable-isotope dimethyl labeling for quantitative proteomics. *Anal. Chem.* **75**, 6843–6852 (2003).
150. Derks, J. *et al.* Increasing the throughput of sensitive proteomics by plexDIA. *Nat. Biotechnol.* **41**, 50–59 (2023).
151. Pino, L. K., Baeza, J., Lauman, R., Schilling, B. & Garcia, B. A. Improved SILAC Quantification with Data-Independent Acquisition to Investigate Bortezomib-Induced Protein Degradation. *J. Proteome Res.* **20**, 1918–1927 (2021).
152. Haynes, S. E., Majmudar, J. D. & Martin, B. R. DIA-SIFT: A Precursor and Product Ion Filter for Accurate Stable Isotope Data-Independent Acquisition Proteomics. *Anal. Chem.* **90**, 8722–8726 (2018).
153. Ctortecka, C. *et al.* Comparative Proteome Signatures of Trace Samples by Multiplexed Data-Independent Acquisition. *Mol. Cell. Proteomics* **21**, 100177 (2022).
154. Thielert, M. *et al.* Robust dimethyl-based multiplex-DIA doubles single-cell proteome depth via a reference channel. *Mol. Syst. Biol.* **19**, e11503 (2023).
155. Tian, X., de Vries, M. P., Permentier, H. P. & Bischoff, R. A Versatile Isobaric Tag Enables Proteome Quantification in Data-Dependent and Data-Independent Acquisition Modes. *Anal. Chem.* **92**, 16149–16157 (2020).
156. 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.* **19**, 2915–2938 (2024).
157. Choi, M. *et al.* MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* **30**, 2524–2526 (2014).
158. 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).
159. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
160. Gerault, M.-A., Camoin, L. & Granjeaud, S. DIAgui: a Shiny application to process the output from DIA-NN. *Bioinforma. Adv.* **4**, vbae001 (2024).
161. Schessner, J. P., Voytik, E. & Bludau, I. A practical guide to interpreting and generating bottom-up proteomics data visualizations. *PROTEOMICS* **22**, 2100103 (2022).

162. Bittremieux, W., Valkenborg, D., Martens, L. & Laukens, K. Computational quality control tools for mass spectrometry proteomics. *PROTEOMICS* **17**, 1600159 (2017).
163. Voytik, E. *et al.* AlphaViz: Visualization and validation of critical proteomics data directly at the raw data level. 2022.07.12.499676 Preprint at https://doi.org/10.1101/2022.07.12.499676 (2022).
164. MannLabs/alpharaw. Mann Labs (2024).
165. Steigerwald, S. *et al.* Full Mass Range ΦSDM Orbitrap Mass Spectrometry for DIA Proteome Analysis. *Mol. Cell. Proteomics* **23**, 100713 (2024).
166. XIX. Further experiments on positive rays. *Philos. Mag. Ser. 1* **24**, 209–253 (1912).
167. Thomson, J. J. *Rays of Positive Electricity and Their Application to Chemical Analyses*. (Longmans, Green and Company, 1913).
168. Serrano, L. R. *et al.* The One Hour Human Proteome. *Mol. Cell. Proteomics* **23**, 100760 (2024).
169. Paul, W. & Steinwedel, H. Notizen: Ein neues Massenspektrometer ohne Magnetfeld. *Z. Für Naturforschung A* **8**, 448–450 (1953).
170. Makarov, A. A. Mass spectrometer. (1999).
171. Brunnée, C. 50 Years of MAT in Bremen. *Rapid Commun. Mass Spectrom.* **11**, 694–707 (1997).
172. Gal, J.-F. A History of European Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **24**, (2013).
173. Habfast, K. Karleugen Habfast. *Phys. Bl.* **56**, 22–22 (2000).
174. Orbitrap Against All Odds. *The Analytical Scientist* https://theanalyticalscientist.com/fields-applications/orbitrap-against-all-odds (2013).
175. Makarov, A. Orbitrap journey: taming the ion rings. *Nat. Commun.* **10**, 3743 (2019).
176. Makarov, A. *et al.* Performance Evaluation of a Hybrid Linear Ion Trap/Orbitrap Mass Spectrometer. *Anal. Chem.* **78**, 2113–2120 (2006).
177. Kingdon, K. H. A Method for the Neutralization of Electron Space Charge by Positive Ionization at Very Low Gas Pressures. *Phys. Rev.* **21**, 408–418 (1923).
178. Knight, R. D. Storage of ions from laser-produced plasmas. *Appl. Phys. Lett.* **38**, 221–223 (1981).
179. Gall, L. N.; Golikov, Y. K.; Aleksandrov, M. L.; Pechalina, Y. E.; Holin, N. A. USSR Inventor's Certificate 1247973.
180. Hu, Q. *et al.* The Orbitrap: a new mass spectrometer. *J. Mass Spectrom.* **40**, 430–443 (2005).

181. Perry, R. H., Cooks, R. G. & Noll, R. J. Orbitrap mass spectrometry: instrumentation, ion motion and applications. *Mass Spectrom. Rev.* **27**, 661–699 (2008).
182. Scigelova, M. & Makarov, A. Orbitrap mass analyzer--overview and applications in proteomics. *Proteomics* **6 Suppl 2**, 16–21 (2006).
183. Amster, I. J. Fourier Transform Mass Spectrometry. *J. Mass Spectrom.* **31**, 1325–1337 (1996).
184. Scigelova, M., Hornshaw, M., Giannakopulos, A. & Makarov, A. Fourier Transform Mass Spectrometry. *Mol. Cell. Proteomics* **10**, M111.009431 (2011).
185. Makarov, A., Denisov, E. & Lange, O. Performance evaluation of a high-field orbitrap mass analyzer. *J. Am. Soc. Mass Spectrom.* **20**, 1391–1396 (2009).
186. Kelstrup, C. D. *et al.* Performance Evaluation of the Q Exactive HF-X for Shotgun Proteomics. *J. Proteome Res.* **17**, 727–738 (2018).
187. Scheltema, R. A. *et al.* The Q Exactive HF, a Benchtop Mass Spectrometer with a Pre-filter, High-performance Quadrupole and an Ultra-high-field Orbitrap Analyzer. *Mol. Cell. Proteomics* **13**, 3698–3708 (2014).
188. Lange, O., Damoc, E., Wieghaus, A. & Makarov, A. Enhanced Fourier transform for Orbitrap mass spectrometry. *Int. J. Mass Spectrom.* **369**, 16–22 (2014).
189. Guevremont, R. High-field asymmetric waveform ion mobility spectrometry: A new tool for mass spectrometry. *J. Chromatogr. A* **1058**, 3–19 (2004).
190. Purves, R. W. & Guevremont, R. Electrospray ionization high-field asymmetric waveform ion mobility spectrometry-mass spectrometry. *Anal. Chem.* **71**, 2346–2357 (1999).
191. Lapthorn, C., Pullen, F. & Chowdhry, B. Z. Ion mobility spectrometry-mass spectrometry (IMS-MS) of small molecules: Separating and assigning structures to ions. *Mass Spectrom. Rev.* **32**, 43–71 (2013).
192. Swearingen, K. E. & Moritz, R. L. High-field asymmetric waveform ion mobility spectrometry for mass spectrometry-based proteomics. *Expert Rev. Proteomics* **9**, 505–517 (2012).
193. Barnett, D. A., Belford, M., Dunyach, J.-J. & Purves, R. W. Characterization of a Temperature-Controlled FAIMS System. *J. Am. Soc. Mass Spectrom.* **18**, 1653–1663 (2007).
194. Pfammatter, S. *et al.* A Novel Differential Ion Mobility Device Expands the Depth of Proteome Coverage and the Sensitivity of Multiplex Proteomic Measurements*. *Mol. Cell. Proteomics* **17**, 2051–2067 (2018).

195. Wang, Q., Fang, F., Wang, Q. & Sun, L. Capillary zone electrophoresis-high field asymmetric ion mobility spectrometry-tandem mass spectrometry for top-down characterization of histone proteoforms. *PROTEOMICS* **24**, 2200389 (2024).
196. Shvartsburg, A. A. *Differential Ion Mobility Spectrometry: Nonlinear Ion Transport and Fundamentals of FAIMS*. (CRC Press, Boca Raton, 2008). doi:10.1201/9781420051070.
197. Barnett, D. A., Ells, B., Guevremont, R. & Purves, R. W. Application of ESI-FAIMS-MS to the analysis of tryptic peptides. *J. Am. Soc. Mass Spectrom.* **13**, 1282–1291 (2002).
198. Hebert, A. S. *et al.* Comprehensive Single-Shot Proteomics with FAIMS on a Hybrid Orbitrap Mass Spectrometer. *Anal. Chem.* **90**, 9529–9537 (2018).
199. Winter, D. L., Wilkins, M. R. & Donald, W. A. Differential Ion Mobility–Mass Spectrometry for Detailed Analysis of the Proteome. *Trends Biotechnol.* **37**, 198–213 (2019).
200. McKetney, J. *et al.* Deep Learning Predicts Non-Normal Peptide FAIMS Mobility Distributions Directly from Sequence. 2024.09.11.612538 Preprint at https://doi.org/10.1101/2024.09.11.612538 (2024).
201. Bekker-Jensen, D. B. *et al.* A Compact Quadrupole-Orbitrap Mass Spectrometer with FAIMS Interface Improves Proteome Coverage in Short LC Gradients. *Mol. Cell. Proteomics* **19**, 716–729 (2020).
202. Adoni, K. R., Cunningham, D. L., Heath, J. K. & Leney, A. C. FAIMS Enhances the Detection of PTM Crosstalk Sites. *J. Proteome Res.* **21**, 930–939 (2022).
203. Sweet, S. *et al.* The addition of FAIMS increases targeted proteomics sensitivity from FFPE tumor biopsies. *Sci. Rep.* **12**, 13876 (2022).
204. Trujillo, E. A. *et al.* Rapid Targeted Quantitation of Protein Overexpression with Direct Infusion Shotgun Proteome Analysis (DISPA-PRM). *Anal. Chem.* **94**, 1965–1973 (2022).
205. Sinn, L. R., Giese, S. H., Stuiver, M. & Rappsilber, J. Leveraging Parameter Dependencies in High-Field Asymmetric Waveform Ion-Mobility Spectrometry and Size Exclusion Chromatography for Proteome-wide Cross-Linking Mass Spectrometry. *Anal. Chem.* **94**, 4627–4634 (2022).
206. Hale, O. J., Illes-Toth, E., Mize, T. H. & Cooper, H. J. High-Field Asymmetric Waveform Ion Mobility Spectrometry and Native Mass Spectrometry: Analysis of Intact Protein Assemblies and Protein Complexes. *Anal. Chem.* **92**, 6811–6816 (2020).

207. Deng, W. *et al.* High-Field Asymmetric Waveform Ion Mobility Spectrometry Interface Enhances Parallel Reaction Monitoring on an Orbitrap Mass Spectrometer. *Anal. Chem.* **94**, 15939–15947 (2022).
208. Chandler, K. B., Marrero Roche, D. E. & Sackstein, R. Multidimensional separation and analysis of alpha-1-acid glycoprotein N-glycopeptides using high-field asymmetric waveform ion mobility spectrometry (FAIMS) and nano-liquid chromatography tandem mass spectrometry. *Anal. Bioanal. Chem.* **415**, 379–390 (2023).
209. Staudt, D. E. *et al.* Phospho-heavy-labeled-spiketide FAIMS stepped-CV DDA (pHASED) provides real-time phosphoproteomics data to aid in cancer drug selection. *Clin. Proteomics* **19**, 48 (2022).
210. Li, J. *et al.* Development of LC-FAIMS-MS and its application to lipidomics study of *Acinetobacter baumannii* infection. *J. Lipid Res.* 100668 (2024) doi:10.1016/j.jlr.2024.100668.
211. Cong, Y. *et al.* Ultrasensitive single-cell proteomics workflow identifies >1000 protein groups per mammalian cell. *Chem. Sci.* **12**, 1001 (2020).
212. Woo, J. *et al.* Three-dimensional feature matching improves coverage for single-cell proteomics based on ion mobility filtering. *Cell Syst.* **13**, 426-434.e4 (2022).
213. Wörner, T. P., Thurman, H. A., Makarov, A. A. & Shvartsburg, A. A. Expanding Differential Ion Mobility Separations into the MegaDalton Range. *Anal. Chem.* **96**, 5392–5398 (2024).
214. Petrosius, V. *et al.* Evaluating the capabilities of the Astral mass analyzer for single-cell proteomics. 2023.06.06.543943 Preprint at https://doi.org/10.1101/2023.06.06.543943 (2023).
215. Bubis, J. A. *et al.* Challenging the Astral ™ mass analyzer - up to 5300 proteins per single-cell at unseen quantitative accuracy to study cellular heterogeneity. Preprint at https://doi.org/10.1101/2024.02.01.578358 (2024).
216. Truong, T. & Kelly, R. T. What's new in single-cell proteomics. *Curr. Opin. Biotechnol.* **86**, 103077 (2024).
217. Ye, Z. *et al.* High-throughput and scalable single cell proteomics identifies over 5000 proteins per cell. Preprint at https://doi.org/10.1101/2023.11.27.568953 (2023).
218. Dodds, J. N. & Baker, E. S. Ion Mobility Spectrometry: Fundamental Concepts, Instrumentation, Applications, and the Road Ahead. *J. Am. Soc. Mass Spectrom.* **30**, 2185–2195 (2019).

219. Delafield, D. G., Lu, G., Kaminsky, C. J. & Li, L. High-end ion mobility mass spectrometry: A current review of analytical capacity in omics applications and structural investigations. *TrAC Trends Anal. Chem.* **157**, 116761 (2022).
220. Giles, K., Williams, J. P. & Campuzano, I. Enhancements in travelling wave ion mobility resolution. *Rapid Commun. Mass Spectrom. RCM* **25**, 1559–1566 (2011).
221. Shvartsburg, A. A. & Smith, R. D. Fundamentals of traveling wave ion mobility spectrometry. *Anal. Chem.* **80**, 9689–9699 (2008).
222. Hernández-Mesa, M. *et al.* Ion Mobility Spectrometry in Food Analysis: Principles, Current Applications and Future Trends. *Molecules* **24**, 2706 (2019).
223. Fernandez-Lima, F. A., Kaplan, D. A. & Park, M. A. Note: Integration of trapped ion mobility spectrometry with mass spectrometry. *Rev. Sci. Instrum.* **82**, 126106 (2011).
224. 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).
225. Ridgeway, M. E., Lubeck, M., Jordens, J., Mann, M. & Park, M. A. Trapped ion mobility spectrometry: A short review. *Int. J. Mass Spectrom.* **425**, 22–35 (2018).
226. Michelmann, K., Silveira, J. A., Ridgeway, M. E. & Park, M. A. Fundamentals of Trapped Ion Mobility Spectrometry. *J. Am. Soc. Mass Spectrom.* **26**, 14–24 (2015).
227. Meier, F., Park, M. A. & Mann, M. Trapped Ion Mobility Spectrometry and Parallel Accumulation–Serial Fragmentation in Proteomics. *Mol. Cell. Proteomics* **20**, 100138 (2021).
228. Liu, F. C., Ridgeway, M. E., Park, M. A. & Bleiholder, C. Tandem trapped ion mobility spectrometry. *The Analyst* **143**, 2249–2258 (2018).
229. Naylor, C. N., Reinecke, T., Ridgeway, M. E., Park, M. A. & Clowers, B. H. Validation of Calibration Parameters for Trapped Ion Mobility Spectrometry. *J. Am. Soc. Mass Spectrom.* **30**, 2152–2162 (2019).
230. Thoben, C. *et al.* Ultra-Fast Ion Mobility Spectrometer for High-Throughput Chromatography. *Anal. Chem.* **95**, 17073–17081 (2023).
231. Wang, K., Qiu, R., Zhang, X., Gillig, K. J. & Sun, W. U-Shaped Mobility Analyzer: A Compact and High-Resolution Counter-Flow Ion Mobility Spectrometer. *Anal. Chem.* **92**, 8356–8363 (2020).
232. Bansal, P. *et al.* Using SLIM-Based IMS-IMS Together with Cryogenic Infrared Spectroscopy for Glycan Analysis. *Anal. Chem.* **92**, 9079–9085 (2020).
233. Li, A. *et al.* Ion Mobility Spectrometry with High Ion Utilization Efficiency Using Traveling Wave-Based Structures for Lossless Ion Manipulations. *Anal. Chem.* **92**, 14930 (2020).

234. Giles, K. *et al.* A Cyclic Ion Mobility-Mass Spectrometry System. *Anal. Chem.* **91**, 8564–8573 (2019).
235. Eldrid, C. & Thalassinos, K. Developments in tandem ion mobility mass spectrometry. *Biochem. Soc. Trans.* **48**, 2457 (2020).
236. Distler, U. *et al.* Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. *Nat. Methods* **11**, 167–170 (2014).
237. Thermo Fisher extends mass spectrometry portfolio - 2019 - Wiley Analytical Science. *Analytical Science News* https://analyticalscience.wiley.com/content/news-do/thermo-fisher-extends-mass-spectrometry-portfolio.
238. Denisov, E., Damoc, E. & Makarov, A. Exploring frontiers of orbitrap performance for long transients. *Int. J. Mass Spectrom.* **466**, 116607 (2021).
239. Hauschild, J.-P. *et al.* A Novel Family of Quadrupole-Orbitrap Mass Spectrometers for a Broad Range of Analytical Applications. Preprint at https://doi.org/10.20944/preprints202006.0111.v1 (2020).
240. Senko, M. W. *et al.* Novel Parallelized Quadrupole/Linear Ion Trap/Orbitrap Tribrid Mass Spectrometer Improving Proteome Coverage and Peptide Identification Rates. *Anal. Chem.* **85**, 11710–11714 (2013).
241. Kelstrup, C. D. *et al.* Limits for Resolving Isobaric Tandem Mass Tag Reporter Ions Using Phase-Constrained Spectrum Deconvolution. *J. Proteome Res.* **17**, 4008–4016 (2018).
242. Grinfeld, D., Aizikov, K., Kreutzmann, A., Damoc, E. & Makarov, A. Phase-Constrained Spectrum Deconvolution for Fourier Transform Mass Spectrometry. *Anal. Chem.* **89**, 1202–1211 (2017).
243. Arrey, T. *et al. Evaluation of the Phase-Constrained Spectrum Deconvolution Method (ΦSDM) for Multiplex TMT Applications*. (2018).
244. Stejskal, K., Op de Beeck, J., Dürnberger, G., Jacobs, P. & Mechtler, K. Ultrasensitive NanoLC-MS of Subnanogram Protein Samples Using Second Generation Micropillar Array LC Technology with Orbitrap Exploris 480 and FAIMS PRO. *Anal. Chem.* **93**, 8704–8710 (2021).
245. Martinez-Val, A. *et al.* Spatial-proteomics reveals phospho-signaling dynamics at subcellular resolution. *Nat. Commun.* **12**, 7113 (2021).
246. Theurillat, I. *et al.* Extensive SUMO Modification of Repressive Chromatin Factors Distinguishes Pluripotent from Somatic Cells. *Cell Rep.* **32**, 108146 (2020).
247. Ctortecka, C., Stejskal, K., Krššáková, G., Mendjan, S. & Mechtler, K. Quantitative Accuracy and Precision in Multiplexed Single-Cell Proteomics. *Anal. Chem.* **94**, 2434–2443 (2022).

248. Ctortecka, C. *et al.* An Automated Nanowell-Array Workflow for Quantitative Multiplexed Single-Cell Proteomics Sample Preparation at High Sensitivity. *Mol. Cell. Proteomics* **22**, 100665 (2023).
249. Sanchez-Avila, X. *et al.* Easy and Accessible Workflow for Label-Free Single-Cell Proteomics. *J. Am. Soc. Mass Spectrom.* **34**, 2374–2380 (2023).
250. Murgia, M. *et al.* Plasma proteome profiling of healthy subjects undergoing bed rest reveals unloading-dependent changes linked to muscle atrophy. *J. Cachexia Sarcopenia Muscle* **14**, 439–451 (2023).
251. Sølberg, J. B. K. *et al.* The Proteome of Hand Eczema Assessed by Tape Stripping. *J. Invest. Dermatol.* **143**, 1559-1568.e5 (2023).
252. Keating, M. F. *et al.* Data Acquisition and Intraoperative Tissue Analysis on a Mobile, Battery-Operated, Orbitrap Mass Spectrometer. *Anal. Chem.* **96**, 8234–8242 (2024).
253. Zhang, J. *et al.* Clinical Translation and Evaluation of a Handheld and Biocompatible Mass Spectrometry Probe for Surgical Use. *Clin. Chem.* **67**, 1271–1280 (2021).
254. Zhang, J. *et al.* Nondestructive tissue analysis for ex vivo and in vivo cancer diagnosis using a handheld mass spectrometry system. *Sci. Transl. Med.* **9**, eaan3968 (2017).
255. Salek, M. *et al.* optiPRM: A Targeted Immunopeptidomics LC-MS Workflow With Ultra-High Sensitivity for the Detection of Mutation-Derived Tumor Neoepitopes From Limited Input Material. *Mol. Cell. Proteomics* **23**, 100825 (2024).
256. Castañeda-Monsalve, V. *et al.* High-throughput screening of the effects of 90 xenobiotics on the simplified human gut microbiota model (SIHUMIx): a metaproteomic and metabolomic study. *Front. Microbiol.* **15**, (2024).
257. Wang, R. *et al.* Temporal Proteomic and Lipidomic Profiles of Cerulein-Induced Acute Pancreatitis Reveal Novel Insights for Metabolic Alterations in the Disease Pathogenesis. *ACS Omega* **8**, 12310–12326 (2023).
258. Lu, W. *et al.* Selected Ion Monitoring for Orbitrap-Based Metabolomics. *Metabolites* **14**, 184 (2024).
259. Assress, H. A., Ferruzzi, M. G. & Lan, R. S. Optimization of Mass Spectrometric Parameters in Data Dependent Acquisition for Untargeted Metabolomics on the Basis of Putative Assignments. *J. Am. Soc. Mass Spectrom.* **34**, 1621–1631 (2023).
260. Thermo Fisher launches new mass detector for biopharmaceuticals - 2021 - Wiley Analytical Science. *Analytical Science News*

https://analyticalscience.wiley.com/content/news-do/thermo-fisher-launches-new-mass-detector-biopharmaceuticals.

261. James, V. K. *et al.* Advancing Orbitrap Measurements of Collision Cross Sections to Multiple Species for Broad Applications. *Anal. Chem.* **94**, 15613–15620 (2022).
262. Meier, F. *et al.* Deep learning the collisional cross sections of the peptide universe from a million experimental values. *Nat. Commun.* **12**, 1185 (2021).
263. Fisher, N. P. *et al.* Determining Collisional Cross Sections from Ion Decay with Individual Ion Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **34**, 2625–2629 (2023).
264. Guzman, U. H. *et al.* Exploring linear sequence determinants of inferred Collisional Cross-Sections of unmodified and phosphorylated peptides in an Orbitrap Mass Analyzer.
265. Guzman, U. H. Inference of Collisional Cross Sections.
266. Wichmann, C. *et al.* MaxQuant.Live Enables Global Targeting of More Than 25,000 Peptides. *Mol. Cell. Proteomics* **18**, 982a–9994 (2019).
267. Bentum, M. van & Selbach, M. An Introduction to Advanced Targeted Acquisition Methods. *Mol. Cell. Proteomics MCP* **20**, 100165 (2021).
268. Kalogeropoulos, K. *et al.* High-throughput and high-sensitivity biomarker monitoring in body fluid by FAIMS-enhanced fast LC SureQuant™ IS targeted quantitation. *Mol. Cell. Proteomics* 100251 (2022) doi:10.1016/j.mcpro.2022.100251.
269. Antelo-Varela, M., Bumann, D. & Schmidt, A. Optimizing SureQuant for Targeted Peptide Quantification: a Technical Comparison with PRM and SWATH-MS Methods. *Anal. Chem.* (2024) doi:10.1021/acs.analchem.4c03622.
270. Martínez-Val, A. *et al.* Hybrid-DIA: intelligent data acquisition integrates targeted and discovery proteomics to analyze phospho-signaling in single spheroids. *Nat. Commun.* **14**, 3599 (2023).
271. Arrey, T., Stewart, H. & Harder, A. *Ion Pre-Accumulation for High Speed Orbitrap Exploris Operation*. (2022).
272. Olsen, J. V. *et al.* A Dual Pressure Linear Ion Trap Orbitrap Instrument with Very High Sequencing Speed*. *Mol. Cell. Proteomics* **8**, 2759–2769 (2009).
273. Levy, M. J., Washburn, M. P. & Florens, L. Probing the Sensitivity of the Orbitrap Lumos Mass Spectrometer using a Standard Reference Protein in a Complex Background. *J. Proteome Res.* **17**, 3586 (2018).
274. Pekar Second, T. *et al.* Dual-Pressure Linear Ion Trap Mass Spectrometer Improving the Analysis of Complex Protein Mixtures. *Anal. Chem.* **81**, 7757–7765 (2009).

275. Huguet, R. *et al.* Proton Transfer Charge Reduction Enables High-Throughput Top-Down Analysis of Large Proteoforms. *Anal. Chem.* **91**, 15732–15739 (2019).
276. Herron, W. J., Goeringer, D. E. & McLuckey, S. A. Ion-ion reactions in the gas phase: Proton transfer reactions of protonated pyridine with multiply charged oligonucleotide anions. *J. Am. Soc. Mass Spectrom.* **6**, 529–532 (1995).
277. Oates, R. N. *et al.* Towards a universal method for middle-down analysis of antibodies via proton transfer charge reduction—Orbitrap mass spectrometry. *Anal. Bioanal. Chem.* (2024) doi:10.1007/s00216-024-05534-z.
278. Dunham, S. D. & Brodbelt, J. S. Enhancing Top-Down Analysis of Proteins by Combining Ultraviolet Photodissociation (UVPD), Proton-Transfer Charge Reduction (PTCR), and Gas-Phase Fractionation to Alleviate the Impact of Nondissociated Precursor Ions. *J. Am. Soc. Mass Spectrom.* **35**, 255–265 (2024).
279. Kline, J. T. *et al.* Sequential Ion-Ion Reactions for Enhanced Gas-Phase Sequencing of Large Intact Proteins in a Tribrid Orbitrap Mass Spectrometer. *J. Am. Soc. Mass Spectrom.* **32**, 2334–2345 (2021).
280. Beaumal, C. *et al.* Improved characterization of trastuzumab deruxtecan with PTCR and internal fragments implemented in middle-down MS workflows. *Anal. Bioanal. Chem.* **416**, 519–532 (2024).
281. Yugandhar, K. *et al.* MaXLinker: Proteome-wide Cross-link Identifications with High Specificity and Sensitivity*. *Mol. Cell. Proteomics* **19**, 554–568 (2020).
282. Furtwängler, B. *et al.* Real-Time Search-Assisted Acquisition on a Tribrid Mass Spectrometer Improves Coverage in Multiplexed Single-Cell Proteomics. *Mol. Cell. Proteomics* **21**, 100219 (2022).
283. Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat. Methods* **8**, 937–940 (2011).
284. McAlister, G. C. *et al.* MultiNotch MS3 Enables Accurate, Sensitive, and Multiplexed Detection of Differential Expression across Cancer Cell Line Proteomes. *Anal. Chem.* **86**, 7150–7158 (2014).
285. Yu, Q. *et al.* Benchmarking the Orbitrap Tribrid Eclipse for Next Generation Multiplexed Proteomics. *Anal. Chem.* **92**, 6478–6485 (2020).
286. Erickson, B. K. *et al.* Active Instrument Engagement Combined with a Real-Time Database Search for Improved Performance of Sample Multiplexing Workflows. *J. Proteome Res.* **18**, 1299–1306 (2019).
287. Schweppe, D. K. *et al.* Full-Featured, Real-Time Database Searching Platform Enables Fast and Accurate Multiplexed Quantitative Proteomics. *J. Proteome Res.* **19**, 2026–2034 (2020).

288. Liu, X., Gygi, S. P. & Paulo, J. A. Isobaric Tag-Based Protein Profiling across Eight Human Cell Lines Using High-Field Asymmetric Ion Mobility Spectrometry and Real-Time Database Searching. *Proteomics* **21**, e2000218 (2021).
289. Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: An open-source MS/MS sequence database search tool. *PROTEOMICS* **13**, 22–24 (2013).
290. He, Y. *et al.* Evaluation of the Orbitrap Ascend Tribrid Mass Spectrometer for Shotgun Proteomics. *Anal. Chem.* **95**, 10655–10663 (2023).
291. Shuken, S. R. *et al.* Deep Proteomic Compound Profiling with the Orbitrap Ascend Tribrid Mass Spectrometer Using Tandem Mass Tags and Real-Time Search. *Anal. Chem.* **95**, 15180–15188 (2023).
292. Peters-Clarke, T. M., Coon, J. J. & Riley, N. M. Instrumentation at the Leading Edge of Proteomics. *Anal. Chem.* **96**, 7976–8010 (2024).
293. Deslignière, E. *et al.* Ultralong transients enhance sensitivity and resolution in Orbitrap-based single-ion mass spectrometry. *Nat. Methods* **21**, 619–622 (2024).
294. Kharchenko, A., Vladimirov, G., Heeren, R. M. A. & Nikolaev, E. N. Performance of Orbitrap Mass Analyzer at Various Space Charge and Non-Ideal Field Conditions: Simulation Approach. *J. Am. Soc. Mass Spectrom.* **23**, 977–987 (2012).
295. Gorshkov, M. V., Good, D. M., Lyutvinskiy, Y., Yang, H. & Zubarev, R. A. Calibration Function for the Orbitrap FTMS Accounting for the Space Charge Effect. *J. Am. Soc. Mass Spectrom.* **21**, 1846–1851 (2010).
296. Beck, S. *et al.* The Impact II, a Very High-Resolution Quadrupole Time-of-Flight Instrument (QTOF) for Deep Shotgun Proteomics. *Mol. Cell. Proteomics MCP* **14**, 2014–2029 (2015).
297. Brunner, A. *et al.* Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. *Mol. Syst. Biol.* **18**, e10798 (2022).
298. Wang, Z. *et al.* High-throughput proteomics of nanogram-scale samples with Zeno SWATH MS. *eLife* **11**, e83947 (2022).
299. Loboda, A. V. & Chernushevich, I. V. A novel ion trap that enables high duty cycle and wide m/z range on an orthogonal injection TOF mass spectrometer. *J. Am. Soc. Mass Spectrom.* **20**, 1342–1348 (2009).
300. Skowronek, P. *et al.* Synchro-PASEF Allows Precursor-Specific Fragment Ion Extraction and Interference Removal in Data-Independent Acquisition. *Mol. Cell. Proteomics* **22**, 100489 (2023).
301. Grinfeld, D. *et al.* Multi-reflection Astral mass spectrometer with isochronous drift in elongated ion mirrors. *Nucl. Instrum. Methods Phys. Res. Sect. Accel. Spectrometers Detect. Assoc. Equip.* **1060**, 169017 (2024).

302. Stewart, H. I. *et al.* Parallelized Acquisition of Orbitrap and Astral Analyzers Enables High-Throughput Quantitative Analysis. *Anal. Chem.* **95**, 15656–15664 (2023).
303. Alikhanov, S. G. A NEW IMPULSE TECHNIQUE FOR ION MASS MEASUREMENTS. *Sov. Phys JETP* **Vol: 4**, (1957).
304. Mamyrin, B. A., Karataev, V. I., Shmikk, D. V. & Zagulin, V. A. The mass-reflectron, a new nonmagnetic time-of-flight mass spectrometer with high resolution.
305. DEPATISnet | Document DE000010116536A1. https://depatisnet.dpma.de/DepatisNet/depatisnet?action=pdf&docid=DE000010116536A1.
306. Verentchikov, A. N., Yavor, M. I., Hasin, Yu. I. & Gavrik, M. A. Multireflection planar time-of-flight mass analyzer. I: An analyzer for a parallel tandem spectrometer. *Tech. Phys.* **50**, 73–81 (2005).
307. Wolf, R. N. *et al.* ISOLTRAP's multi-reflection time-of-flight mass separator/spectrometer. *Int. J. Mass Spectrom.* **349–350**, 123–133 (2013).
308. Verenchikov, A. N. & Yavor, M. I. Imaging properties of a multi-reflection time-of-flight mass analyzer. *Int. J. Mass Spectrom.* **463**, 116547 (2021).
309. Yavor, M. I., Plaß, W. R., Dickel, T., Geissel, H. & Scheidenberger, C. Ion-optical design of a high-performance multiple-reflection time-of-flight mass spectrometer and isobar separator. *Int. J. Mass Spectrom.* **381–382**, 1–9 (2015).
310. Ishida, Y. *et al.* A time-of-flight mass spectrometer to resolve isobars. *Nucl. Instrum. Methods Phys. Res. Sect. B Beam Interact. Mater. At.* **219–220**, 468–472 (2004).
311. Dickel, T. *et al.* A high-performance multiple-reflection time-of-flight mass spectrometer and isobar separator for the research with exotic nuclei. *Nucl. Instrum. Methods Phys. Res. Sect. Accel. Spectrometers Detect. Assoc. Equip.* **777**, 172–188 (2015).
312. Wollnik, H., Casares, A., Radford, D. & Yavor, M. Multi-pass time-of-flight mass spectrometers of high resolving power. *Nucl. Instrum. Methods Phys. Res. Sect. Accel. Spectrometers Detect. Assoc. Equip.* **519**, 373–379 (2004).
313. Grix, R. *et al.* A time-of-flight mass analyzer with high resolving power. *Rapid Commun. Mass Spectrom.* **2**, 83–85 (1988).
314. Cooper-Shepherd, D. A. *et al.* Novel Hybrid Quadrupole-Multireflecting Time-of-Flight Mass Spectrometry System. *J. Am. Soc. Mass Spectrom.* **34**, 264–272 (2023).
315. GRINFELD, D. & Makarov, A. Multi-reflection mass spectrometer. (2016).

316. Hamish, S., Dmitry, G., Bernd, H. & Robert, O. High resolution multi-reflection time-of-flight mass analyser. (2024).
317. Stewart, H., Grinfeld, D. E. & Makarov, A. A. Time of flight mass spectrometer and method of mass spectrometry. (2022).
318. Makarov, A. A., Grinfield, D. E. & Monastyrskiy, M. A. Multireflection Time-Of-Flight Mass Spectrometer. (2011).
319. Hock, C. *et al.* Time-of-flight mass analysers. (2022).
320. Giannakopulos, A. *et al.* The OrbiTOF mass analyzer: Time-of-flight analysis via an orbitrap quadro-logarithmic field with periodic drift focusing. *Int. J. Mass Spectrom.* **505**, 117315 (2024).
321. Stewart, H. *et al.* A High Dynamic Range Ion Detector for the Astral™ Analyzer. Preprint at https://doi.org/10.26434/chemrxiv-2024-49mzs (2024).
322. Stewart, H. *et al.* A Conjoined Rectilinear Collision Cell and Pulsed Extraction Ion Trap with Auxiliary DC Electrodes. *J. Am. Soc. Mass Spectrom.* **35**, 74–81 (2024).
323. Stewart, H. *et al.* Crowd control of ions in the Astral analyzer. *J. Mass Spectrom.* **59**, e5006 (2024).
324. Stewart, H. *et al.* A multi-reflection time-of-flight analyzer with a long focus lens. *Int. J. Mass Spectrom.* **505**, 117329 (2024).
325. Yavor, M. I., Pomozov, T. V., Kirillov, S. N., Khasin, Y. I. & Verenchikov, A. N. High performance gridless ion mirrors for multi-reflection time-of-flight and electrostatic trap mass analyzers. *Int. J. Mass Spectrom.* **426**, 1–11 (2018).
326. Guzman, U. H. *et al.* Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition. *Nat. Biotechnol.* (2024) doi:10.1038/s41587-023-02099-7.
327. Heil, L. R. *et al.* Evaluating the Performance of the Astral Mass Analyzer for Quantitative Proteomics Using Data-Independent Acquisition. *J. Proteome Res.* **22**, 3290–3300 (2023).
328. Lancaster, N. M. *et al.* Fast and deep phosphoproteome analysis with the Orbitrap Astral mass spectrometer. *Nat. Commun.* **15**, 7016 (2024).
329. Orbitrap Astral Literature List. https://view.highspot.com/viewer/cc5c47206cea799aa3d238863cefa052.
330. 6546 LC QTOF high resolution LCMS w wide dynamic range | Agilent. https://www.agilent.com/en/product/liquid-chromatography-mass-spectrometry-lc-ms/lc-ms-instruments/quadrupole-time-of-flight-lc-ms/6546-lc-q-tof#features.
331. 6560 Ion Mobility QTOF LC/MS, Collision Cross Section | Agilent. https://www.agilent.com/en/product/liquid-chromatography-mass-spectrometry-lc-ms/lc-ms-instruments/quadrupole-time-of-flight-lc-ms/6560-ion-mobility-lc-q-tof.

332. 7600 plus system. https://sciex.com/products/mass-spectrometers/qtof-systems/7600-plus-system.
333. timsTOF. https://www.bruker.com/en/products-and-solutions/mass-spectrometry/timstof.html.
334. Rethink what is possible with the Orbitrap Astral mass spectrometer.
335. SYNAPT XS | Ion Mobility Time-of-Flight Mass Spectrometer | Waters. https://www.waters.com/nextgen/us/en/products/mass-spectrometry/mass-spectrometry-systems/synapt-xs.html.
336. Orsburn, B. C., Yuan, Y. & Bumpus, N. N. Insights into protein post-translational modification landscapes of individual human cells by trapped ion mobility time-of-flight mass spectrometry. *Nat. Commun.* **13**, 7246 (2022).
337. Selby, D. S., Mlynski, V. & Guilhaus, M. Reducing grid dispersion of ions in orthogonal acceleration time-of-flight mass spectrometry: advantage of grids with rectangular repeat cells. *Int. J. Mass Spectrom.* **206**, 201–210 (2001).
338. Stewart, H. Tandem mass spectrometer and method of tandem mass spectrometry. (2024).
339. Stewart, H. *et al.* Proof of principle for enhanced resolution multi-pass methods for the Astral analyzer. *Int. J. Mass Spectrom.* **498**, 117203 (2024).
340. Stewart, H., Wagner, A. & Makarov, A. A. Ion guide. (2022).
341. Stewart, H. & GRINFELD, D. Collision cross section measurement in time-of-flight mass analyser. (2024).
342. Wahle, M. *et al.* A novel hybrid high speed mass spectrometer allows rapid translation from biomarker candidates to targeted clinical tests using 15N labeled proteins. Preprint at https://doi.org/10.1101/2024.06.02.597029 (2024).
343. Remes, P. M. *et al.* Hybrid Quadrupole Mass Filter – Radial Ejection Linear Ion Trap and Intelligent Data Acquisition Enable Highly Multiplex Targeted Proteomics. Preprint at https://doi.org/10.1101/2024.05.31.596848 (2024).
344. Geyer, P. E. *et al.* Plasma Proteome Profiling to detect and avoid sample-related biases in biomarker studies. *EMBO Mol. Med.* **11**, e10427 (2019).
345. Zhang, Z. & Chan, D. W. The road from discovery to clinical diagnostics: lessons learned from the first FDA-cleared in vitro diagnostic multivariate index assay of proteomic biomarkers. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.* **19**, 2995–2999 (2010).
346. Messner, C. B. *et al.* Ultra-fast proteomics with Scanning SWATH. *Nat. Biotechnol.* **39**, 846–854 (2021).

347. Rai, A. J. *et al.* Proteomic Approaches to Tumor Marker Discovery: Identification of Biomarkers for Ovarian Cancer. *Arch. Pathol. Lab. Med.* **126**, 1518–1526 (2002).
348. Nakayasu, E. S. *et al.* Tutorial: best practices and considerations for mass-spectrometry-based protein biomarker discovery and validation. *Nat. Protoc.* **16**, 3737–3760 (2021).
349. Wenk, D., Zuo, C., Kislinger, T. & Sepiashvili, L. Recent developments in mass-spectrometry-based targeted proteomics of clinical cancer biomarkers. *Clin. Proteomics* **21**, 6 (2024).
350. Meng, Z. & Veenstra, T. D. Targeted mass spectrometry approaches for protein biomarker verification. *J. Proteomics* **74**, 2650–2659 (2011).
351. Harlan, R. & Zhang, H. Targeted proteomics: a bridge between discovery and validation. *Expert Rev. Proteomics* **11**, 657 (2014).
352. Kim, J., Koo, B.-K. & Knoblich, J. A. Human organoids: model systems for human biology and medicine. *Nat. Rev. Mol. Cell Biol.* **21**, 571–584 (2020).
353. Rae, C., Amato, F. & Braconi, C. Patient-Derived Organoids as a Model for Cancer Drug Discovery. *Int. J. Mol. Sci.* **22**, 3483 (2021).
354. Weng, T., Jenkins, B. J. & Saad, M. I. Patient-Derived Xenografts: A Valuable Preclinical Model for Drug Development and Biomarker Discovery. *Methods Mol. Biol. Clifton NJ* **2806**, 19–30 (2024).
355. Yoshida, G. J. Applications of patient-derived tumor xenograft models and tumor organoids. *J. Hematol. Oncol.J Hematol Oncol* **13**, 4 (2020).
356. Loewa, A., Feng, J. J. & Hedtrich, S. Human disease models in drug development. *Nat. Rev. Bioeng.* **1**, 545–559 (2023).
357. Norman, G. A. V. Limitations of Animal Studies for Predicting Toxicity in Clinical Trials: Is it Time to Rethink Our Current Approach? *JACC Basic Transl. Sci.* **4**, 845 (2019).
358. McGonigle, P. & Ruggeri, B. Animal models of human disease: Challenges in enabling translation. *Biochem. Pharmacol.* **87**, 162–171 (2014).
359. Moorman, A. R. *et al.* Progressive plasticity during colorectal cancer metastasis. *Nature* 1–8 (2024) doi:10.1038/s41586-024-08150-0.
360. Moral, T. T. *et al.* Methods for Stratification and Validation Cohorts: A Scoping Review. *J. Pers. Med.* **12**, 688 (2022).
361. Bader, J. M., Albrecht, V. & Mann, M. MS-Based Proteomics of Body Fluids: The End of the Beginning. *Mol. Cell. Proteomics* **22**, 100577 (2023).
362. Geyer, P. E., Holdt, L. M., Teupser, D. & Mann, M. Revisiting biomarker discovery by plasma proteomics. *Mol. Syst. Biol.* **13**, 942 (2017).

363. Mischak, H. *et al.* Clinical proteomics: A need to define the field and to begin to set adequate standards. *Proteomics Clin. Appl.* **1**, 148–156 (2007).
364. Mi, Y. *et al.* High-throughput mass spectrometry maps the sepsis plasma proteome and differences in patient response. *Sci. Transl. Med.* **16**, eadh0185 (2024).
365. Bruderer, R. *et al.* Analysis of 1508 Plasma Samples by Capillary-Flow Data-Independent Acquisition Profiles Proteomics of Weight Loss and Maintenance. *Mol. Cell. Proteomics MCP* **18**, 1242–1254 (2019).
366. Viode, A. *et al.* A simple, time- and cost-effective, high-throughput depletion strategy for deep plasma proteomics. *Sci. Adv.* **9**, eadf9717 (2023).
367. Anderson, N. L., Ptolemy, A. S. & Rifai, N. The Riddle of Protein Diagnostics: Future Bleak or Bright? *Clin. Chem.* **59**, 194–197 (2013).
368. Malmström, E. *et al.* Large-scale inference of protein tissue origin in gram-positive sepsis plasma using quantitative targeted proteomics. *Nat. Commun.* **7**, 10261 (2016).
369. Shen, Y. *et al.* Characterization of the human blood plasma proteome. *PROTEOMICS* **5**, 4034–4045 (2005).
370. Anderson, N. L. & Anderson, N. G. The Human Plasma Proteome: History, Character, and Diagnostic Prospects*. *Mol. Cell. Proteomics* **1**, 845–867 (2002).
371. Oh, H. S.-H. *et al.* Organ aging signatures in the plasma proteome track health and disease. *Nature* **624**, 164–172 (2023).
372. Thygesen, K. *et al.* Third universal definition of myocardial infarction. *J. Am. Coll. Cardiol.* **60**, 1581–1598 (2012).
373. Geyer, P. E. *et al.* The Circulating Proteome—Technological Developments, Current Challenges, and Future Trends. *J. Proteome Res.* (2024) doi:10.1021/acs.jproteome.4c00586.
374. Loo, J. A., Yan, W., Ramachandran, P. & Wong, D. T. Comparative human salivary and plasma proteomes. *J. Dent. Res.* **89**, 1016–1023 (2010).
375. Ferdosi, S. *et al.* Enhanced Competition at the Nano–Bio Interface Enables Comprehensive Characterization of Protein Corona Dynamics and Deep Coverage of Proteomes. *Adv. Mater.* **34**, 2206008 (2022).
376. Blume, J. E. *et al.* Rapid, deep and precise profiling of the plasma proteome with multi-nanoparticle protein corona. *Nat. Commun.* **11**, 3662 (2020).
377. Wu, C. C. *et al.* Mag-Net: Rapid enrichment of membrane-bound particles enables high coverage quantitative analysis of the plasma proteome. *BioRxiv Prepr. Serv. Biol.* 2023.06.10.544439 (2024) doi:10.1101/2023.06.10.544439.

378. Tognetti, M. *et al.* Biomarker Candidates for Tumors Identified from Deep-Profiled Plasma Stem Predominantly from the Low Abundant Area. *J. Proteome Res.* **21**, 1718–1735 (2022).
379. Ferdosi, S. *et al.* Engineered nanoparticles enable deep proteomics studies at scale by leveraging tunable nano–bio interactions. *Proc. Natl. Acad. Sci.* **119**, e2106053119 (2022).
380. 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).
381. Wang, D. *et al.* A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.* **15**, e8503 (2019).
382. Dyring-Andersen, B. *et al.* Spatially and cell-type resolved quantitative proteomic atlas of healthy human skin. *Nat. Commun.* **11**, 5587 (2020).
383. Doll, S. *et al.* Region and cell-type resolved quantitative proteomic map of the human heart. *Nat. Commun.* **8**, 1469 (2017).
384. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
385. Kavanagh, T. & Drummond, E. Insights from a proteomic atlas of human Alzheimer's disease brain tissue. *Neural Regen. Res.* **20**, 799–800 (2025).
386. Phipps, W. S. *et al.* Clinical Proteomics for Solid Organ Tissues. *Mol. Cell. Proteomics* **22**, 100648 (2023).
387. Jiang, L. *et al.* A Quantitative Proteome Map of the Human Body. *Cell* **183**, 269-283.e19 (2020).
388. Schweizer, L. *et al.* Quantitative multiorgan proteomics of fatal COVID-19 uncovers tissue-specific effects beyond inflammation. *EMBO Mol. Med.* **15**, e17459 (2023).
389. Busso-Lopes, A. F. *et al.* Connecting multiple microenvironment proteomes uncovers the biology in head and neck cancer. *Nat. Commun.* **13**, 6725 (2022).
390. Fomitcheva-Khartchenko, A., Rapsomaniki, M. A., Sobottka, B., Schraml, P. & Kaigala, G. V. Spatial protein heterogeneity analysis in frozen tissues to evaluate tumor heterogeneity. *PLoS ONE* **16**, e0259332 (2021).
391. Zhou, S. *et al.* Proteomics analysis of tumor microenvironment: Implications of metabolic and oxidative stresses in tumorigenesis. *Mass Spectrom. Rev.* **32**, 267–311 (2013).
392. Schäfer, M. *et al.* Spatial tissue proteomics reveals distinct landscapes of heterogeneity in cutaneous papillomavirus-induced keratinocyte carcinomas. *J. Med. Virol.* **95**, e28850 (2023).

393. Hoyer, K. J. R., Dittrich, S., Bartram, M. P. & Rinschen, M. M. Quantification of molecular heterogeneity in kidney tissue by targeted proteomics. *J. Proteomics* **193**, 85–92 (2019).
394. Schweizer, L. *et al.* Spatial proteo-transcriptomic profiling reveals the molecular landscape of borderline ovarian tumors and their invasive progression. *MedRxiv Prepr. Serv. Health Sci.* 2023.11.13.23298409 (2023) doi:10.1101/2023.11.13.23298409.
395. Mund, A. *et al.* Deep Visual Proteomics defines single-cell identity and heterogeneity. *Nat. Biotechnol.* **40**, 1231–1240 (2022).
396. Djambazova, K. V., Van Ardenne, J. M. & Spraggins, J. M. Advances in imaging mass spectrometry for biomedical and clinical research. *TrAC Trends Anal. Chem.* **169**, 117344 (2023).
397. Buchberger, A. R., DeLaney, K., Johnson, J. & Li, L. Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights. *Anal. Chem.* **90**, 240 (2017).
398. Zhang, H., Delafield, D. G. & Li, L. Mass spectrometry imaging: the rise of spatially resolved single-cell omics. *Nat. Methods* **20**, 327–330 (2023).
399. Angelo, M. *et al.* Multiplexed ion beam imaging of human breast tumors. *Nat. Med.* **20**, 436–442 (2014).
400. Kuett, L. *et al.* Three-dimensional imaging mass cytometry for highly multiplexed molecular and cellular mapping of tissues and the tumor microenvironment. *Nat. Cancer* **3**, 122–133 (2022).
401. Thul, P. J. & Lindskog, C. The human protein atlas: A spatial map of the human proteome. *Protein Sci. Publ. Protein Soc.* **27**, 233–244 (2018).
402. The Human Protein Atlas. https://www.proteinatlas.org/.
403. Mund, A., Brunner, A.-D. & Mann, M. Unbiased spatial proteomics with single-cell resolution in tissues. *Mol. Cell* **82**, 2335–2349 (2022).
404. Nordmann, T. M. *et al.* A Standardized and Reproducible Workflow for Membrane Glass Slides in Routine Histology and Spatial Proteomics. *Mol. Cell. Proteomics MCP* **22**, 100643 (2023).
405. Zheng, X., Mund, A. & Mann, M. Deciphering functional tumor-immune crosstalk through highly multiplexed imaging and deep visual proteomics. 2024.05.22.595266 Preprint at https://doi.org/10.1101/2024.05.22.595266 (2024).
406. Pachitariu, M. & Stringer, C. Cellpose 2.0: how to train your own model. *Nat. Methods* **19**, 1634 (2022).
407. MannLabs/scPortrait. Mann Labs (2024).

408. Post, F. *et al.* Deep Visual Proteomics advances human colon organoid models by revealing a switch to an in vivo-like phenotype upon xenotransplantation. 2024.05.13.593888 Preprint at https://doi.org/10.1101/2024.05.13.593888 (2024).
409. Makhmut, A. *et al.* A framework for ultra-low-input spatial tissue proteomics. *Cell Syst.* **14**, 1002-1014.e5 (2023).
410. Nordmann, T. M. *et al.* Spatial proteomics identifies JAKi as treatment for a lethal skin disease. *Nature* 1–9 (2024) doi:10.1038/s41586-024-08061-0.
411. Rosenberger, F. A. *et al.* Spatial single-cell mass spectrometry defines zonation of the hepatocyte proteome. *Nat. Methods* **20**, 1530–1536 (2023).
412. Petelski, A. A. *et al.* Multiplexed single-cell proteomics using SCoPE2. *Nat. Protoc.* **16**, 5398–5425 (2021).
413. Matzinger, M., Müller, E., Dürnberger, G., Pichler, P. & Mechtler, K. Robust and Easy-to-Use One-Pot Workflow for Label-Free Single-Cell Proteomics. *Anal. Chem.* **95**, 4435–4445 (2023).
414. Kabatnik, S. *et al.* Spatial characterization and stratification of colorectal adenomas by deep visual proteomics. *iScience* **27**, 110620 (2024).
415. Zheng, X. *et al.* Deep Visual Proteomics Unveils Precision Medicine Insights in Composite Small Lymphocytic and Classical Hodgkin Lymphoma. 2024.06.12.598635 Preprint at https://doi.org/10.1101/2024.06.12.598635 (2024).
416. Martini, B. R., Aizikov, K. & Mandelshtam, V. A. The filter diagonalization method and its assessment for Fourier transform mass spectrometry. *Int. J. Mass Spectrom.* **373**, 1–14 (2014).
417. Aushev, T., Kozhinov, A. N. & Tsybin, Y. O. Least-Squares Fitting of Time-Domain Signals for Fourier Transform Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **25**, 1263–1273 (2014).
418. Yin, L. *et al.* Skeletal muscle atrophy: From mechanisms to treatments. *Pharmacol. Res.* **172**, 105807 (2021).
419. Cruz-Jentoft, A. J. & Sayer, A. A. Sarcopenia. *The Lancet* **393**, 2636–2646 (2019).
420. Fearon, K. *et al.* Definition and classification of cancer cachexia: an international consensus. *Lancet Oncol.* **12**, 489–495 (2011).
421. Kilroe, S. P. *et al.* Dietary protein intake does not modulate daily myofibrillar protein synthesis rates or loss of muscle mass and function during short-term immobilization in young men: a randomized controlled trial. *Am. J. Clin. Nutr.* **113**, 548–561 (2021).
422. Böcker, J., Schmitz, M.-T., Mittag, U., Jordan, J. & Rittweger, J. Between-Subject and Within-Subject Variaton of Muscle Atrophy and Bone Loss in Response to Experimental Bed Rest. *Front. Physiol.* **12**, 743876 (2022).

423. Pino, L. K., Just, S. C., MacCoss, M. J. & Searle, B. C. Acquiring and Analyzing Data Independent Acquisition Proteomics Experiments without Spectrum Libraries. *Mol. Cell. Proteomics MCP* **19**, 1088–1103 (2020).
424. Moseley, M. A. *et al.* Scanning Quadrupole Data-Independent Acquisition, Part A: Qualitative and Quantitative Characterization. *J. Proteome Res.* **17**, 770–779 (2018).
425. Lou, R. *et al.* Benchmarking commonly used software suites and analysis workflows for DIA proteomics and phosphoproteomics. *Nat. Commun.* **14**, 94 (2023).
426. Ye, Z. *et al.* One-Tip enables comprehensive proteome coverage in minimal cells and single zygotes. *Nat. Commun.* **15**, 2474 (2024).
427. Petrosius, V. *et al.* Quantitative Label-Free Single-Cell Proteomics on the Orbitrap Astral MS. Preprint at https://doi.org/10.1101/2024.07.31.605978 (2024).
428. Benesch, M. G. K. & Mathieson, A. Epidemiology of Signet Ring Cell Adenocarcinomas. *Cancers* **12**, 1544 (2020).
429. El Hussein, S. & Khader, S. N. Primary signet ring cell carcinoma of the pancreas: Cytopathology review of a rare entity. *Diagn. Cytopathol.* **47**, 1314–1320 (2019).
430. Nagtegaal, I. D. *et al.* The 2019 WHO classification of tumours of the digestive system. *Histopathology* **76**, 182–188 (2020).
431. Van Cutsem, E., Sagaert, X., Topal, B., Haustermans, K. & Prenen, H. Gastric cancer. *The Lancet* **388**, 2654–2664 (2016).
432. Odenwald, M. A. & Turner, J. R. The intestinal epithelial barrier: a therapeutic target? *Nat. Rev. Gastroenterol. Hepatol.* **14**, 9–21 (2017).
433. Barker, N. Adult intestinal stem cells: critical drivers of epithelial homeostasis and regeneration. *Nat. Rev. Mol. Cell Biol.* **15**, 19–33 (2014).
434. Martini, E., Krug, S. M., Siegmund, B., Neurath, M. F. & Becker, C. Mend Your Fences. *Cell. Mol. Gastroenterol. Hepatol.* **4**, 33–46 (2017).
435. Jansen, S. A. *et al.* Chemotherapy-induced intestinal epithelial damage directly promotes galectin-9-driven modulation of T cell behavior. *iScience* **27**, 110072 (2024).
436. Sturm, A. & Dignass, A. U. Epithelial restitution and wound healing in inflammatory bowel disease. *World J. Gastroenterol.* **14**, 348 (2008).
437. Greene, C. M. *et al.* α1-Antitrypsin deficiency. *Nat. Rev. Dis. Primer* **2**, 16051 (2016).
438. Lomas, D. A., LI-Evans, D., Finch, J. T. & Carrell, R. W. The mechanism of Z α1-antitrypsin accumulation in the liver. *Nature* **357**, 605–607 (1992).

439. Bortel, P. *et al.* Systematic Optimization of Automated Phosphopeptide Enrichment for High-Sensitivity Phosphoproteomics. *Mol. Cell. Proteomics* **23**, 100754 (2024).
440. Dumas, T. *et al.* The astounding exhaustiveness and speed of the Astral mass analyzer for highly complex samples is a quantum leap in the functional analysis of microbiomes. *Microbiome* **12**, 46 (2024).
441. Ross, C. A. & Poirier, M. A. Protein aggregation and neurodegenerative disease. *Nat. Med.* **10**, S10–S17 (2004).
442. Meier, F., Geyer, P. E., Virreira Winter, S., Cox, J. & Mann, M. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat. Methods* **15**, 440–448 (2018).
443. De Godoy, L. M. F. *et al.* Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**, 1251–1254 (2008).
444. Virant-Klun, I., Leicht, S., Hughes, C. & Krijgsveld, J. Identification of Maturation-Specific Proteins by Single-Cell Proteomics of Human Oocytes. *Mol. Cell. Proteomics* **15**, 2616–2627 (2016).
445. Petricoin, E. F. *et al.* Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet* **359**, 572–577 (2002).
446. Tabb, D. L. Quality assessment for clinical proteomics. *Clin. Biochem.* **46**, 411–420 (2013).
447. Ignjatovic, V. *et al.* Mass Spectrometry-Based Plasma Proteomics: Considerations from Sample Collection to Achieving Translational Data. *J. Proteome Res.* **18**, 4085–4097 (2019).
448. Banerjee, S. Empowering Clinical Diagnostics with Mass Spectrometry. *ACS Omega* **5**, 2041–2048 (2020).
449. Rischke, S. *et al.* Small molecule biomarker discovery: Proposed workflow for LC-MS-based clinical research projects. *J. Mass Spectrom. Adv. Clin. Lab* **28**, 47–55 (2023).

6. Acknowledgements

A PhD, thankfully, does not happen in a vacuum and I am immensely grateful to everyone who contributed to my journey!

First and foremost, I'd like to thank Matthias for the chance to pursue my PhD in his lab. For the great research environment, supervision, and trust over the years and for providing opportunities that are already shaping my future.

For their guidance and input over the course of my PhD a special thank you also goes to my TAC members, Lars Mägdefessel and Florian Meier.

I would like to thank Jesper Olsen and his research group for hosting my master thesis, getting me excited about MS-based proteomics, and setting me on this path. Particularly, Ana, Dorte, Ulises and Patrick for not just being great colleagues but also good friends! Thanks for all the support and encouraging me to pursue my PhD.

Additionally, I want to thank our collaboration partners at Thermo Fisher Scientific for all the insightful discussions and continued support: Kyle Fort, Eugen Damoc, Tabiwang Array, Anna Pashkova, Johannes Petzoldt, Christian Hock, Stevan Horning, and Alexander Makarov.

My office, over the years, including Ankit, Elena, Enes, Marta and Vincent for creating such a fun, kind, and productive environment! And special thanks to Ankit for being so supportive, helping with my analysis questions and always being ready to answer my coding questions until LLMs came to the rescue 😉 As well as Marta, sitting behind me and inadvertently making sure I keep my desk (and her zoom background) tidy, for always being available to bounce ideas and ask stupid questions, whether about figures, presentation titles or non-work-related matters.

My "Little women" group at the MPI: Medini, Lisa, Maria and Patricia.

First of thank you all so much for your support and invaluable input and feedback on this thesis!!

Patricia: Packing columns with someone for 10h twice a month either makes you never want to see them again or very close friends, and I am glad for us it is the latter. From packing columns, to organizing cooking evenings, sewing and embroidering, discussing science and life, to surviving Covid together in a hotel room in Chicago our friendship only grew over the years and I am so grateful for it!

Medini: Thank you for supporting me all throughout my PhD! It was a pleasure to share an office with you in the beginning and support each other through the steep learning curve of writing proposals, grant reporting, and third-party funding for Heart research-related grants! I'll miss our discussions and even the walks to vent if things get too much.

Maria: Thank you for always being supportive, but brutally honest with your feedback. Whether it is about science or being each other's feedback loop for "Can I send the email like this" or "Am I overreacting". I had a great time working on the OA-HLA project with you, bouncing ideas back and forth, and almost losing our minds over acquisition methods!

Lisa: Thank you for enduring all my questions, always having my back, and supporting my PhD journey from before I applied to IMPRS to now. It's invaluable to have you as a

colleague, but even more so to have you as a friend! I am really looking forward being closer to CPH, so we can have art/crafting sessions together again!

A shoutout also to Max and Marvin, for all the scientific discussion, support, and fun over the last 11 (!!!) years. Looking forward to many more.

For their scientific or technological support, a special thanks also goes to Dirk, Igor, Tim and Mario. Moreover, to Ute for all the administrative support from booking flights/hotels and coordinating all the paperwork to planning the retreats and much more.

It's hard to mention everyone, but I would like to also thanks to Sophia, Caro, Feng, Florian, Andreas, Marc, Sonja, Ericka, for all the support, discussions, chats and fun!

Moreover, I'd like to thank the entire Mann Group across Munich and Copenhagen for the amazing time and atmosphere. I feel very lucky to have been part of such a creative, enthusiastic, supportive community over the last years!

My family and especially my parents for always supporting and believing in me. Allowing me to pursue my studies and teaching me to trust myself. That it is ok to fail as long as I get back up and give my best, but also to not work "too much", as I was often reminded. Thanks for encouraging me to be curious, form my own opinions and for enable me to be where I am now!

Florent: We know from all my endeavors writing cards, that I am not great with words and it's hard to express all my gratitude, but you still manage to surprise me all the time with how kind, supportive, and loving you are! From talking about science, fonts, and teaching me GTD, making me dinner and a cup of tea, to staying on the couch with me while I write this thesis till late into the night, I am so grateful to have you in my life and am looking forward to returning the favor ♡

For me personally, still arguably the best outcome of DigiMed.