P01

Aquasearch: a new software for fast proteomic characterization and classification of wastewater samples analyzed using MALDI-TOF.

Montserrat Carrascal[2], Antoni Ginebreda[1], Ester Sanchez[2], Damia Barcelo[1], Joaquin Abian[1]

Presenting author: Carlos Perez-Lopez

[1] Institute of Environmental Assessment and Water Studies (IDAEA-CSIC), Barcelona, Spain

[2] Institute of Biomedical Research of Barcelona (IIBB-CSIC/IDIBAPS), Barcelona, Spain

The study of wastewater is a valuable source of information about the environment, health and industrial activities of the inhabitants of an area. Although the study of wastewater has traditionally focused on small molecules such as pharmaceuticals or illegal drugs, recent studies have reported the valuable information that can be obtained from large molecules in wastewater, introducing proteomics as an emerging field in environmental monitoring.

Liquid Chromatography coupled with High-Resolution Mass Spectrometry (LC-HRMS) instrument was used to identify the proteins in wastewater in previous studies with a shotgun proteomics approach. Although the entire process reports comprehensive and accurate results, it is expensive and time-consuming. Therefore, Matrix-Assisted Laser Desorption/Ionization coupled with Time of Flight (MALDI-TOF) is proposed as a high-throughput instrumental approach for faster and more cost-effective sample characterization. In this work, we present Aquasearch, a newly developed software in Python for the characterization and classification of samples in a multisampling analysis. Aquasearch primarily performs two tasks: 1) signal filtering from wastewater proteomics samples analyzed by MALDI-TOF and identification of peptides belonging to livestock and human biomarkers using an in-house database and 2) using the identification results to classify the samples based on their proteomic profile in a non-supervised analysis. To facilitate the use of Aquasearch, including the parameter selection and result visualization, the program can be run through a graphic user interface (GUI).

To test the program, 4 wastewater samples collected from 4 WWTPs in Catalonia, Spain (Besòs, Girona, Vic and Figueres), were analyzed by MALDI-TOF. The Aquasearch analysis of the corresponding protein profiles showed the dominance of human biomarkers in Besòs and Girona, while pig and chicken biomarkers were the major components in Vic and Figueres. Finally, these proteomic profiles clustered the samples in the non-supervised multisampling analysis based on their origin.

P02

Glioblastoma TME Analyses by SpiderMass for Surgery Decision Making and Patient Management

Yanis Zirem [1], Léa Ledoux [1], Lucas Roussel [1], Claude Alain Maurage [2], Pierre Tirilly [3], Émilie Le Rhun [1,4], Bertrand Meresse [5], Gargey Yagnik [6], Mark J. Lim [6], Kenneth J. Rothschild [6,7], Marie Duhamel [1], Michel Salzet [1,8], Isabelle Fournier [1,8]

Presenting author: Yanis ZIREM

[1] Univ. Lille, Inserm, CHU Lille, U1192 - Protéomique Réponse Inflammatoire Spectrométrie de Masse - PRISM, F-59000 Lille, France

[2] CHU Lille, Service de biochimie et biologie moléculaire, CHU Lille, F-59000 Lille, France

[3] Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France

[4] Department of Neurology, Clinical Neuroscience Center, University Hospital of Zurich, University of Zurich, Zurich, Switzerland

[5] Univ. Lille, Inserm, CHU Lille, U1286 - INFINITE - Institute for Translational Research in Inflammation, F-59000 Lille, France

[6] AmberGen, Inc., Billerica, MA, United States

[7] Department of Physics and Photonics Center, Boston University, Boston, MA, United States

[8] Institut Universitaire de France (IUF), Paris, France

Glioblastoma is a highly heterogeneous and infiltrative form of brain cancer associated to a poor outcome with a limited efficiency of therapies. The extend of the surgery is known to be related to the patient survival. Reaching an accurate diagnostic and prognostic by the time of the initial surgery is therefore paramount in the management of glioblastoma. To this end, we have studied the performances of the SpiderMass, an ambient ionization mass spectrometry technology that can be used in vivo without invasiveness, coupled to a newly developed AI pipeline. We demonstrated that we can, both stratify IDH-wild type glioblastoma patients into molecular sub-groups and get an accurate diagnostic with over 90% accuracy after cross-validation. Interestingly, the developed method offers the same accuracy for prognosis. Additionally, we tested the potential of an immune-scoring strategy based on SpiderMass fingerprints, showing the association between prognosis and immune cell infiltration, to predict the patient outcome.

P03

Integrated view of baseline protein expression in human & model organisms from reanalysis of public proteomics experiments

Ananth Prakash[1][2], David Garcia Seisdedos[1], Shengbo Wang[1], Deepti Jaiswal Kundu[1], Andrew Collins[3], Nancy George[1], Silvie Fexova[1], Pablo Moreno[1], Irene Papatheodorou[1][2], Andrew R. Jones[3], Juan Antonio Vizcaíno [1][2].

Presenting author: Ananth Prakash

[1] EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridge. CB10 1SD

[2] Open Targets, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD. United Kingdom

[3] Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool L69 7ZB, United Kingdom

The availability of proteomics datasets in the public domain, and in the PRIDE database in particular, has increased dramatically in recent years. This unprecedented large-scale availability of data provides an opportunity for combined analyses of datasets to get organism-wide protein expression data in a consistent manner. We have reanalysed 61public proteomics datasets from healthy human, mouse, rat and pig samples to assess baseline protein abundance in 42 organs. We defined tissue as a distinct functional or structural region within an organ. Overall, the aggregated dataset contains 99 healthy tissues, corresponding to 4891 mass spectrometry runs covering 837 samples.

In all cases, we studied the distribution of canonical proteins between different organs and the distribution of proteins across organs. We also compared the results with data generated in analogous studies. We carried out a comparative analysis of protein expression between mouse, rat, pig and human tissues. We observed a high level of correlation of protein expression among orthologs between all three species in brain, kidney, heart and liver samples, whereas the correlation of protein expression was generally slightly lower between organs within the same species. We also performed gene ontology and pathway enrichment analyses to identify organ-specific enriched biological processes and pathways. We also compared protein abundances with RNA-seq expression. As a key point, we have integrated the protein expression results into the resource Expression Atlas, where it can be accessed and visualised either individually or together with gene expression data coming from transcriptomics datasets. We believe this is a good mechanism to make proteomics data more accessible for scientists, especially those non-experts in proteomics.

P04

A Workflow towards the Reproducible Identification and Quantitation of Protein Carbonylation Sites in Human Plasma

Juan Camilo Rojas Echeverri [1, 2] Sanja Milkovska-Stamenova [1, 2] Ralf Hoffmann [1, 2]

Presenting author: Juan Camilo Rojas Echeverri

[1] Institute for Bioanalytical Chemistry, Faculty of Chemistry and Mineralogy, Universität Leipzig

[2] Center for Biotechnology and Biomedicine, Universität Leipzig

Protein carbonylation, a marker of excessive oxidative stress, has been studied in the context of multiple human diseases related to oxidative stress. The variety of post-translational carbonyl modifications (carbonyl PTMs) and their low concentrations in plasma challenge their reproducible identification and quantitation. However, carbonyl-specific biotinylated derivatization tags (e.g., aldehyde reactive probe, ARP) allow for targeting carbonyl PTMs by enriching proteins and peptides carrying these modifications. In this study, an oxidized human serum albumin protein model (OxHSA) and plasma from a healthy donor were derivatized with ARP, digested with trypsin, and enriched using biotin-avidin affinity chromatography prior to nano reversed-phase chromatography coupled online to electrospray ionization tandem mass spectrometry with travelling wave ion mobility spectrometry (nRPC-ESI-MS/MS-TWIMS). The presented workflow addresses several analytical challenges by using ARP-specific fragment ions to reliably identify ARP peptides. Furthermore, the reproducible recovery and relative quantitation of ARP peptides were validated. Human serum albumin (HSA) in plasma was heavily modified by a variety of direct amino acid oxidation products and adducts from reactive carbonyl species (RCS), with most RCS modifications being detected in six hotspots, i.e., Lys10, Lys190, Lys199, Lys281, Lys432, and Lys525 of mature HSA. These results have been replicated recently in a larger cohort (n = 64) where the ubiquitous has become clear. However, MS evidence also suggests that a large portion of the data remains unidentified necessitating the exploration of open mass search strategies.

P05

Is scanning swath the holy grail for histone analysis?

Sigrid Verhelst [1], Laura Corveleyn [1], Bart Van Puyvelde [1], Dieter Deforce [1] and Maarten Dhaenens [1]

Presenting author: Sigrid Verhelst

[1] ProGenTomics, Laboratory of Pharmaceutical Biotechnology, Ghent University, Ghent, Belgium

Epigenetic control mechanisms allow the differentiation of stable gene-expression patterns in various cell types and thus the very existence of multicellular organisms. One of the predominant mechanisms for epigenetic regulation are the histone post-translational modifications (hPTMs). Because of their sophisticated function, histones are among the most complexly modified molecules in the biotic world, in turn making these key regulators of multicellular organization notoriously hard to analyze. As a result, our knowledge of the histone code remains highly fragmented, which makes it one of the most frustrating 'known unknowns' in biology. One of the primary reasons for the lag in the study of intricately modified histones compared to the broader proteomics field is the absence of a working target-decoy strategy. Consequently, popular data analysis algorithms for data-independent acquisition (DIA), such as DIA-NN, cannot be readily applied to investigate histone extracts as they depend on a target-decoy strategy for empirically determining false discovery rates (FDR) and re-scoring peptide ion feature weights. The quest for a working target-decoy strategy tailored to histones has been ongoing for years but has made limited progress. On the other hand, DIA data analysis can be performed in for example Skyline where manual validation can be used instead of automated target-decoy FDR control. The crucial step in this process will be the generation of an extensive spectral library. In the broader proteomics field, spectral libraries are by now mainly predicted, using tools like MS2PIP or Prosit. Unfortunately, these predictions are not yet applicable to highly modified histones. Therefore, a spectral library for histones should still be build based on experimental data. This poster focusses on leveraging an enhanced iteration of one of the most popular DIA technologies sequential window acquisition of all theoretical fragment ion spectra (SWATH), referred to as Scanning SWATH. The challenges associated with histone analysis are summarized together with a proposal of what currently seems to be the best approach to decipher this histone code. Nevertheless, there is so much information in the histone code, and with the appropriate support in terms of data analysis, this can lead to significant breakthroughs. Therefore, a call: if anyone has ideas, interest, or tools to make this possible, do not hesitate to visit my poster.

P06

Multi-omics investigation of central metabolism shifts in tuberous sclerosis complex

Anna-Sophia Egger [1, ‡], Madlen Hotze [1, ‡], Alienke van Pijkeren [1, 2], Tobias Kipura [1], Alexa Hofer [1], Ulrike Rehbein [1], Florian Hatzmann [1], Anna Jansen [2], Liesbeth De Waele [3], Francois Jouret [4], Djalila Mekahli [5, *], Peter Janssens [6,*] , Marcel Kwiatkowski [1,*], Kathrin Thedieck [1, 7, 8, 9, *]

Presenting author: Anna-Sophia Egger

‡, * authors contributed equally

1 Institute of Biochemistry and Center for Molecular Biosciences Innsbruck, University of Innsbruck, Austria

2 Translational Neurosciences, University of Antwerp, Belgium

3 Department of Paediatric Neurology, University Hospitals Leuven, Belgium

4 Division of Nephrology, Department of Internal Medicine, University of Liège Hospital (ULiège CHU), Belgium

5 Department of Pediatric Nephrology, University Hospitals Leuven, Belgium

6 Department of Nephrology and Arterial Hypertension, Universitair Ziekenhuis Brussel (UZ Brussel), Vrije Universiteit Brussel, Belgium

7 Laboratory of Pediatrics, Section Systems Medicine of Metabolism and Signaling, University of Groningen, University Medical Center Groningen, the Netherlands.

8 Department for Neuroscience, School of Medicine and Health Sciences, Carl von Ossietzky University Oldenburg, Germany.

9 FMF - Freiburg Materials Research Center, University of Freiburg, Germany

Tuberous sclerosis complex (TSC) is a congenital disease that affects approximately 1 in 6000 newborns (1). The disease is caused by mutations in the genes coding for the proteins tuberin (TSC1) or hamartin (TSC2). They form the TSC protein complex (2) which serves as an inhibitor of the serine/threonine kinase mechanistic target of rapamycin complex 1 (mTORC1)(3). TSC patients experience a variety of symptoms including tumor formation, lung disease (4) and neurological manifestations (5). mTORC1 is a master regulator of metabolism that controls glucose and lipid metabolism. It stands to reason that a loss of TSC could dysregulate central metabolism, and in support epilepsy symptoms in TSC patients can be ameliorated by a ketogenic diet.

In our study, we set out to investigate alterations in central metabolism in a TSC model system by simultaneous proteo-metabolomics. We conducted quantitative proteome analyses as well as targeted metabolomics analyses in TSC2 knockout and control cells, investigating a fundamental shift in energy metabolism induced by a loss of TSC2. We validated the findings in a TSC patient cohort.

P07

ProteoBench: a community-curated platform for comparison of proteomics data analysis workflows

Holda Anagho[1], Nadezhda T. Doncheva[1], Viktoria Dorfer[2], Ralf Gabriels[3][4], Vedran Kasalica[5], Caroline Lennartsson[1], Matthias Mattanovich[6], Emmanuelle Mouton-Barbosa[7], Martin Rykær[1], Veit Schwämmle[8], Maximillian Strauss[1], Tim Van Den Bossche[3][4], Bart Van Puyvelde[9], Henry Webel[1], Jakub Vasicek[10][11], Julian Uszkoreit[12], Witold Wolski[13][14], Robbin Bouwmeester[3][4], Marie Locard-Paulet[7][15]

Presenting author: Marie Locard-Paulet

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

[2] University of Applied Sciences Upper Austria, Bioinformatics Research Group, Hagenberg, Austria

[3] VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium

[4] Department of Biomolecular Medicine, UGent, Ghent, Belgium

[5] Netherlands eScience Center, Science Park 402, 1098 XH, Amsterdam, The Netherlands

[6] Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Blegdamsvej 3B, DK-2200 Copenhagen, Denmark

[7] Institut de Pharmacologie et de Biologie Structurale (IPBS), Université de Toulouse, CNRS, Université Toulouse III - Paul Sabatier (UT3), Toulouse, France

[8] Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense, Denmark

[9] ProGenTomics, Laboratory of Pharmaceutical Biotechnology, Ghent University, 9000 Ghent, Belgium

[10] Mohn Center for Diabetes Precision Medicine, Department of Clinical Science, University of Bergen, Bergen, Norway

[11] Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway

[12] Ruhr University Bochum, Medical Faculty, Medical Bioinformatics

[13] Functional Genomics Center Zurich (FGCZ)-University of Zurich/ETH Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

[14] Swiss Institute of Bioinformatics (SIB) Quartier Sorge-Batiment Amphipole, 1015 Lausanne, Switzerland

[15] Infrastructure nationale de protéomique, ProFI, FR 2048, Toulouse, France

Mass spectrometry (MS)-based proteomics is a go-to strategy for analyzing complex biological mixtures used extensively in biology, biomedicine, and clinical research. With the abundance of dedicated data analysis pipelines, continuously updated and developed, the community would benefit from a platform to compare their performance in an objective and unbiased manner.

Here we propose ProteoBench, a comprehensive and open web platform for comparing MS-based proteomics data analysis tools. It allows for easy and controlled comparison of proteomics tools developed or used by the participants to other state-of-the-art pipelines. ProteoBench originated as a community project by the European Bioinformatics Community for Mass Spectrometry (EuBIC-MS) and its design and development will be open to all interested researchers. It will provide a centralized resource, easy to use for non-coders.

(continued on next page)

This platform will allow the:
1. end-users to select a well-performing workflow to meet their needs
2. developers to discern the strengths and limitations of their workflow, thereby directing further development
3. publishers and reviewers to easily position workflows in the context of existing state-of-the-art workflows

ProteoBench will be composed of several benchmarking modules dedicated to specific types of analysis tools (developed for identification, quantification, statistical analysis, for DDA or DIA data, etc...). Each module should provide a set of input file- and module-specific parameters needed for reproducibility. Users shall be able to provide their analysis results in a specified format with the option of making this result publicly available. This input will be parsed by ProteoBench, and module-specific metrics for comparison will be calculated and visualized. Currently, we have five modules in preparation, of which two are in active development. A first benchmarking module is already available as a prototype that we intend to open for beta-testing at the Winter School.

ProteoBench is open to all, for software tool comparison as well as for contributing (improving existing or developing new benchmark modules). The submitted data will continuously grow and should remain up-to-date with the latest developments. It will create a frame of reference for evaluating the performance of new and/or custom tools when publishing results and should increase transparency and reproducibility between data analysis pipelines developed in the field.

P08

Global Analysis of Protein Methylation in The Mitochondrial Compartment of Cancer Cells: A Proteomic Approach

Ayusi Mondal [1,2], Alessandro Vai [1,2], Silvia Pedretti [1,3], Nico Mitro [1,3], Tiziana Bonaldi [1,4]

Presenting author: Ayusi Mondal

[1] Department of Experimental Oncology, European Institute of Oncology (IEO), IRCCS Milano, Italy

[2] European School of Molecular Medicine (SEMM)

[3] Department of Pharmacological and Biomolecular Sciences, University of Milan, Italy

[4] Department of Oncology and Hematology-Oncology, University of Milan, Milan, Italy

Protein methylation, catalyzed by a number of methyltransferases (MTases), has been established to be a crucial post-translational modification (PTM) in the human proteome. In the past decade, a plethora of methylation sites have been annotated in human proteome, occurring mainly on K and R sites. Protein methylation has been shown to affect protein subcellular localization, protein-protein and protein-nucleic acids interactions, ultimately regulating various cellular and processes. Thus, it is not so surprising that MTase dysregulation, mainly due to mutation and/or aberrant expression, has been linked to protein dysfunction and disease, such as neurological disorders and cancer. Recently our group has collected a wealth of published and preliminary evidence suggestive of a particular role of protein-methylation in modulating metabolic enzymes, in particular the mitochondrial ones, such as glutathione, glutamine, leucine metabolism, as well as electron transport chain.

This has raised the question of the presence, extent and dynamicity of protein-methylation in the mitochondrial compartment, an issue which has not yet been investigated systematically, at a single-site resolution.

Mass spectrometry (MS) has emerged as the most powerful tool to analyze PTMs. Our group, in particular, has developed complete workflows to systematically identify and profile different PTMs on both histones and non-histone protein. Hence, we have set up a MS-based proteomic approach to investigate and annotate genuine in vivo methylation sites which are enzymatically incorporated post-translationally on mitochondrial proteins. The setup of a methyl-proteomics workflow tailored for the analysis mitochondrial compartment required various optimization steps to overcome various technical challenges, such as: the capability to discern in vivo protein methylation from isobaric artefacts; the sub-stoichiometric nature of this PTM; the lower representation of less abundant mitochondrial methyl-proteins versus highly abundant modified nuclear and cytosolic ones. The crucial steps of this implementation phase will be described, together with the presentation of preliminary data on the extent and features of mitochondrial methyl-proteomes of cervical and ovarian cancer cell lines.

P09

Cleavable crosslink identification from MS2 and MS3 spectra with MS Annika 2.0

Micha Johannes Birklbauer [1], Manuel Matzinger [2], Fränze Müller [2], Karl Mechtler [2, 3, 4], Viktoria Dorfer [1]

Presenting author: Micha Johannes Birklbauer

[1] University of Applied Sciences Upper Austria, Bioinformatics Research Group, Softwarepark 11, 4232 Hagenberg, Austria

[2] Institute of Molecular Pathology (IMP), Vienna BioCenter (VBC), Campus-Vienna-Biocenter 1, 1030 Vienna, Austria

[3] Institute of Molecular Biotechnology (IMBA), Austrian Academy of Sciences, Vienna BioCenter (VBC), Dr. Bohr-Gasse 3, 1030 Vienna, Austria

[4] Gregor Mendel Institute (GMI), Austrian Academy of Sciences, Vienna BioCenter (VBC), Dr. Bohr-Gasse 3, 1030 Vienna, Austria

Introduction:

Cross-linking mass spectrometry (XLMS) has emerged as a prominent tool for identification of protein-protein interactions and for gaining insights into the structures of proteins. Over the last decade XLMS has seen continuous growth and the development of new cross-linkers, enrichment strategies and data acquisition methods led to the establishment of numerous new software tools specifically for the analysis and interpretation of cross-linking data. We here present MS Annika 2.0, an updated and improved version of our cross-linking search engine MS Annika that additionally to MS2-only supports processing of data from MS2-MS3-based approaches and identification of peptides from MS3 spectra.

Methods & Results:

In the new MS2-MS3 search, MS Annika 2.0 first matches each MS3 spectrum to the corresponding doublet peak of the precursor MS2 spectrum to identify the crosslink modifications and monoisotopic masses of the cross-linked peptides. MS3 spectra are then adjusted accordingly for search with MS Amanda, our in-house developed peptide search engine which is used to identify the cross-linked peptides. Peptides that are identified in the MS2 spectrum and one or more corresponding MS3 spectra are re-scored with a novel scoring function to reflect the increased confidence. Finally, the detected cross-links are validated by estimating the false discovery rate (FDR) using a target-decoy approach. We evaluated the MS3-search-capabilities of MS Annika 2.0 on different datasets covering a variety of experimental approaches and compared it to XlinkX and MaXLinker, two other cross-linking search engines that support MS3 crosslink identification. We show that MS Annika detects up to 4 times more true crosslinks while also providing a more accurate FDR estimation than the other two search engines.

P10

Evaluating the capabilities of the Astral mass analyzer for single-cell proteomics

Pedro Aragon-Fernandez [1], Valdemaras Petrosius [1], Tabiwang N. Arrey [2], Nil Üresin [3], Benjamin Furtwängler [3], Hamish Stewart [2], Eduard Denisov [2], Johannes Petzoldt [2], Amelia C. Peterson [2], Christian Hock [2], Eugen Damoc [2], Alexander Makarov [2], Vlad Zabrouskov [2], Bo T. Porse [3], Erwin M. Schoof [1]

Presenting author: Pedro Aragon Fernandez

[1] Technical University of Denmark, Lyngby, Denmark

[2] Thermo Fisher Scientific, Bremen, Germany

[3] University of Copenhagen, Copenhagen, Denmark

The complexity of human physiology arises from well-orchestrated interactions between trillions of single cells in the body. While single-cell RNA sequencing (scRNA-seq) has enhanced our understanding of cell diversity, gene expression alone does not fully characterize cell phenotypes. Additional molecular dimensions, such as proteins, are needed to define cellular states accurately. Mass spectrometry (MS)-based proteomics has emerged as a powerful tool for comprehensive protein analysis, including single-cell applications. However, challenges remain in terms of throughput and proteomic depth, in order to maximize the biological impact of single-cell proteomics by Mass Spectrometry (scp-MS) workflows. This study leverages a novel high-resolution, accurate mass (HRAM) instrument platform, consisting of both an Orbitrap and an innovative HRAM Asymmetric Track Lossless (Astral) analyzer. The Astral analyzer offers high sensitivity and resolution through lossless ion transfer and a unique flight track design. We evaluate the performance of the Thermo Scientific Orbitrap Astral MS using Data-Independent Acquisition (DIA) and assess proteome depth and quantitative precision for ultra-low input samples. Optimal DIA method parameters for single-cell proteomics are identified, and we demonstrate the ability of the instrument to study cell cycle dynamics in Human Embryonic Kidney (HEK293) cells, and cancer cell heterogeneity in a primary Acute Myeloid Leukemia (AML) culture model.

P11

Elucidating the Potential of Open Modification Search in Peptide Identification

Louise Marie Buur [1], Viktoria Dorfer [1]

Presenting author: Louise Marie Buur

[1] Bioinformatics Research Group, University of Applied Sciences Upper Austria, Hagenberg, Austria

In contrast to conventional closed database searches, open modification searches enable the simultaneous exploration of hundreds of post-translational modifications, potentially increasing the identification rate of mass spectra. However, the inclusion of numerous modifications expands the search space, leading to prolonged run times and an increased possibility of false positive hits. To address this, we present initial developments in our novel open modification search engine.

The first step in our approach involves spectral pre-processing, encompassing the removal of precursor peaks and simple deconvolution. Subsequently, peaks are selected from 100 m/z windows based on the precursor mass. Using the selected peaks, sequence tags consisting of 3 and 4 residues are extracted from each spectrum. Peptides from the protein database are then organized according to the sequence tags they contain making it possible to extract only those peptides that contain a sequence tag found in the spectrum. This peptide candidate pool is then further filtered using a wide precursor mass tolerance and subsequently scored using the MS Amanda scoring function. In the current version, our open search engine can consider fixed modifications and up to two variable modifications per peptide.

To evaluate the performance of our newly developed open modification search engine, we conducted an analysis on two datasets—phosphopeptides and peptides derived from HeLa cells. We compared the identification results with those obtained using MS Amanda. Initial findings indicate a substantial overlap in peptide-spectrum matches between the two search engines across both datasets at 1% estimated false discovery rate. However, there are instances where the two search engines identify different peptides and peptidoforms. We anticipate that the of overlap in identifications will increase as we fully implement the open search functionality. Our ongoing efforts involve continuous refinement and development of our open modification search engine to achieve rapid and accurate identification of modified peptides.

P12

ReCon-CETSA: A new data imputation approach in IMPRINTS-CETSA dataset using deep learning

Marc-Antoine Gerault [1], Pär Nordlund [1,2]

Presenting author: Marc-Antoine Gerault

[1] Department of Oncology and Pathology, Karolinska Institutet, 171 77 Stockholm, Sweden

[2] Institute of Molecular and Cell Biology, A*STAR, 138673, Singapore

Proteins and their interactions with other biomolecules underlie the biochemical operations of almost all cellular events. To measure protein-ligand interactions in intact cells and extract their modulations, we use an assay named IMPRINTS (Integrated Modulation of Protein Interaction States)-CETSA (Cellular Thermal Shift Assay) which is based on ligand-induced thermal stabilization of target proteins. This experiment involves multiple biological replicates of samples that will be compared after heat challenge at an isothermal temperature using a tandem mass tag (TMT) multiplex set, and the arrangement of several TMT sets from numerous isotherms (typically six) along the melting range (typically 37-64°C). As the temperature increases, the protein abundance decreases overall, resulting in a higher proportion of protein missing values in higher temperature. Moreover, integrating multiple TMT batches in a single analysis usually results in an inflation rate of protein missing values. The same effect is observed when comparing different IMPRINTS-CETSA datasets which can be of high interest when comparing different drugs or cell lines.

To overcome this challenge, we developed a deep learning architecture, ReCon-CETSA, to perform data imputation in IMPRINTS-CETSA datasets. Instead of training ReCon-CETSA on the distribution of each variable as we would do in more classical statistics method, we chose to see each protein abundances like a time-dependant sequence, where the first time point would be the lowest temperature. Our architecture uses an LSTM layer coupled with a convolutional layer, based on the idea of attention. It is then trained to predict the next value based on the two previous ones (forward) for each protein for a given treatment and replicate. This same idea is also applied to predict the previous value based on the two next ones (backward). ReCon-CETSA uses preferably the backward model when possible, as we saw experimentally that it performs slightly better than the forward model. We tested ReCon -CETSA to predict missing values in four published IMPRINTS-CETSA datasets and showed that ReCon-CETSA always performs better than classical data imputation algorithm like QRILC, random forest or knn using the RMSE as the performance metric.

This could help identify new protein targets that couldn’t be measured but also help compare a high number of different IMPRINTS-CETSA datasets to potentially extract new biological knowledge.

P13

Single-cell profiling of histone post-translational modifications using label-free LC-MS/MS

Laura Corveleyn [1], Ronald Cutler [2], Claudia Ctortecka [3], Maarten Dhaenens [1], Dieter Deforce [1], Malvina Papanastasiou [3], Simone Sidoli [2]

Presenting author: Laura Corveleyn

[1] Laboratory of Pharmaceutical Biotechnology, Ghent University, Belgium

[2] Department of Biochemistry, Albert Einstein College of Medicine, The Bronx, NY, USA

[3] Broad Institute of MIT and Harvard, Cambridge, MA, USA

Histones can be modified by a diversity of dynamic post-translational modifications (PTMs) that collectively make up the "histone code". Histone PTMs (hPTMs) contribute to the modulation of chromatin structure through their distinct chemical properties and their capacity to attract chromatin-modifying enzymes and binding proteins, thereby influencing gene expression. Bulk analysis of histone PTMs has already significantly enriched our understanding of their pivotal role in health and disease. However, these bulk methods often require substantial starting material (few millions of cells), which limits their use in cases where sample size is small, e.g. in clinical settings. Additionally, traditional bulk analyses mask the variability inherent to cell populations, limiting the ability to uncover subtle differences that may drive disease progression. Single-cell analysis enables the identification of rare cell subpopulations, thereby offering insights into cellular heterogeneity, transitions, and responses to stimuli.

Several methods have already been developed successfully analyzing histone PTMs from single cells, such as scChiPseq, scCUT&TAG and EpiTOF. However, all of the abovementioned approaches rely on antibodies, which involve several challenges, such as i) cross-reactivity, ii) limited number of targets, iii) differences in antibody affinity resulting in quantitative challenges, iv) lack of multiplexing and v) limited detection of novel modifications. Contrarily, liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) allows direct and quantitative measurement of multiple histone modifications simultaneously, providing a comprehensive view of the epigenetic landscape.

Here, we present a method that accurately quantifies hundreds of histone modifications simultaneously from single cells using label-free LC-MS/MS. We combined single cell sorting and automated sample preparation using cellenONE with data acquisition on the TimsTOF Ultra system. To show the technical robustness of the method, we measured increasing amounts of bovine histone standards equivalent to 1, 2, 4, and 8 cells. Subsequently, we successfully generated calibration curves for all histone peptides, achieving an average R2 value of 0.95. Finally, we demonstrate the ability to differentiate between individual cells treated with sodium butyrate, a histone deacetylase (HDAC) inhibitor, and control cells by assessing the levels of acetylated histone peptides.

P14

MS1-base peptide identification and quantification by sparse encoding

Zixuan Xiao [1], Mathias Wilhelm [1]

Presenting author: Zixuan Xiao

[1] Computational Mass Spectrometry, TUM School of Life Sciences, Technical University of Munich, Freising, Germany

Peptide identification and quantification using Accurate Mass and Time (AMT) tags in liquid chromatography (LC) tandem mass spectrometry (MS/MS) data was proposed as an approach to partially or even completely obviate the need for subsequent MS/MS analysis. Recent developments in high-resolution mass spectrometers and deep learning offer higher accuracy of mass and retention time mapping, thus facilitating the further potential of AMT. Here we present a proof of concept experiment for joint peptide identification and quantification by expanding the AMT approach. We propose to de-convolute MS1 signals in a scan-by-scan manner. For each MS1 scan, we construct a set of peptides from either an in silico digestion or prior deep proteome measurements, described by their accurate mass and predicted retention time. The joint identification and quantification is formulated as a spare encoding problem with LASSO loss, or a multi-variate regression problem focusing on inference of positive coefficients and sparsity regularization. We demonstrated that on the same dataset, the correlation between the quantification results from our method (MS1 only) and Maxquant reached a Pearson's correlation of 0.948. In addition, our approach recovered and quantified 82.4% of the peptide sequences detected in the deep proteome measurement, while only 28.5% were identified by MaxQuant in the shallow measurement, highlighting the potential of MS1-based signal deconvolution.

P15

A modular, flexible, and user-friendly toolbox for deep exploration, processing, and analysis of quantitative proteomics data

Thierry Balliau[1], Olivier Langella[1], Anne Aubert-Frambourg[2], Michel Zivy[1], Mélisande Blein-Nicolas [1]

Presenting author: Balliau Thierry

[1] GQE-Le Moulon, Université Paris-Saclay, INRAE, CNRS, AgroParisTech, IDEEV 12, route 128 Gif-sur-Yvette F-91272 France

[2]Université Paris-Saclay, INRAE, ENVA, BREED, 78350, Jouy-en-Josas, France

MCQR is an R package designed to standardize the processing and analysis of large quantitative data sets obtained from bottom-up proteomics. It is applicable to a wide range of proteomics experiments, including label-free shotgun experiments, isotopic labeling experiments, fractionation-based experiments, and PTM enrichment experiments, and includes all the necessary functions to import, check, filter, normalize, impute, describe, and statistically analyze quantitative data based on either spectral counts or extracted ion currents. In addition to being one of the most comprehensive tools available, with unique functions that make the most of the information available in the input data, MCQR stands out from other similar R packages currently available because of its modular and flexible architecture, which allows the users to build the processing and analysis scenario best suited to their data. These strengths are further enhanced by its simplicity and ease of use: MCQR requires only basic knowledge of R and no programming skills, making it accessible to all users, especially proteomics biologists who want to analyze their own data and proteomics platform analysts who want to ensure the quality of their MS acquisitions. MCQR is actively maintained and freely available at http://pappso.inrae.fr/bioinfo/mcqr/.

P16

Wheat proteomics for analysing baking quality

Christine Kaemper [1], Sabrina Geisslitz [1], Katharina A. Scherf [1]

Presenting author: Christine Kaemper

[1] Karlsruhe Institute of Technology, Karlsruhe, Germany

Wheat is one of the most important cereals in the human diet. It is characterised by its outstanding baking quality, which is influenced by various factors, particularly the quality and quantity of the proteins in the grains. These are mainly storage proteins, which are collectively referred to as gluten. Gluten can be divided into gliadins and glutenins. Both the total amount of protein and the ratio of gliadins to glutenins are relevant for baking quality. Furthermore, individual gluten proteins (e.g. high molecular weight glutenins) can also have an impact.

The initial aim of this work is to establish high-throughput bottom-up proteomics methods (targeted and untargeted) to analyse a large number of wheat flour samples. The proteomic data obtained and other results from baking trials can then be used to link individual gluten proteins to baking quality. The sample set to be analysed is the Bavarian Multiple Advanced Generation Intercross wheat population (BMWpop). It consists of 394 inbred lines covering 72 % of the allelic diversity of the German wheat breeding gene pool. With its high variability in terms of baking quality, the BMWpop is well suited for identifying proteins that are associated with high loaf volume, for example.

Firstly, the proteins were isolated from the wheat flours in a one-step extraction. After reduction and alkylation of the cysteine groups, the samples were enzymatically digested with trypsin. The work-up was completed by purification using solid phase extraction in 96-well plates. The subsequent untargeted LC-MS/MS measurement was done in data dependant acquisition mode. Evaluations were performed using the Software MaxQuant, Perseus and the UniProtKB database.

A total of 3,600 peptides corresponding to almost 1,000 different protein groups were identified in the BMWpop parental wheat lines. 129 groups were gluten protein groups, which were divided into 73 gliadin and 56 glutenin protein groups. The remaining protein groups mainly corresponded to enzymes. 250 of the identified protein groups could not be characterised. The parental lines were also distinguishable in terms of their amounts of gliadin and glutenin based on their relative protein composition. It was noticeable that two wheat varieties (cv Ambition and cv BAYP4535) separated themselves particularly clearly from the others. Based on these results, a targeted proteomics method for the absolute quantification of individual gluten protein groups is being developed.

P17

DeepLC enables CCS prediction of peptides carrying modifications not seen during training

Robbe Devreese [1,2], Alireza Nameni [1,2], Arthur Declercq [1,2], Ralf Gabriels [1,2], Sven Degroeve [1,2], Lennart Martens [1,2], Robbin Bouwmeester [1,2]

Presenting author: Robbe Devreese

[1] VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium

[2] Department of Biomolecular Medicine, Ghent University, Ghent, Belgium

Instruments capable of measuring the ion mobility of peptides are now part of the standard proteomics workflows. While ion mobility separation allows for the acquisition of much cleaner and more reproducible data, it is still rarely used for identification purposes. This in part shows the need for a highly performant collisional cross-section (CCS) prediction models that enable improved rescoring of database search results. For example, it is currently common to perform rescoring with predicted retention times and fragment intensities. One notable predictor for retention is DeepLC, which allows for the prediction of retention times for modified peptides without explicit training. Interestingly, DeepLC's architecture and logic can also be repurposed to predict the mobility of modified peptides. While the separation mechanism between LC and ion mobility are nearly orthogonal, the atomic composition used to encode peptides in DeepLC can also be directly related to a peptide's shape and size. This shape and size, in combination with charge, are the main properties that lead to separation in ion mobility. In this study, we show that DeepLC performs on par with other CCS prediction algorithms, and accurately predicts CCS values for modifications and amino acids not seen during training.

P18

JLspec: A Web Application for Streamlined Processing and Analysis of Large-Scale Untargeted Metabolomics Data

Ana Mendes [1], Jesper Foged Havelund [1], Jonas Lemvig [2], Veit Schwämmle [1], Nils J. Færgeman [1]

Presenting author: Ana Mendes

[1] Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense, Denmark

[2] Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark

The post-processing and analysis of large-scale untargeted metabolomics data face significant challenges due to the intricate nature of correction, filtration, imputation, and normalization steps. Manual execution across various applications often leads to inefficiencies, human-induced errors, and inconsistencies within the workflow. Addressing these issues, we introduce JLspec, a novel web application designed to combine established methodologies, offering flexibility and ease of implementation.

JLspec produced promising outcomes, proving its efficacy in consistently producing results across diverse instances. It simplifies the complexity of current metabolomics workflows. JLspec stands out in its ability to perform quality control normalization, providing diverse visualization methods for comprehensive data interpretation and comparison of data processing methods. Moreover, it provides an interface for conducting statistical tests, using the Limma package in R. This is complemented by integration with other applications such as PolySTest for robust statistical analysis and VSClust for clustering.

The successful deployment of JLspec provides a leap forward in mitigating the manual intervention, enhancing reproducibility, and advancing the untargeted metabolomics data processing workflow. This user-friendly and comprehensive tool holds promise in contributing to the field by facilitating data analysis, thus fostering advancements in metabolomics research.

P19

Sensitive differential PTM Analysis in E. coli

Caroline Jachmann [1], Aurélie Hirschler [1], Florence Arsène-Ploetze [2], Christine Carapito [1], Lennart Martens [3, 4]

Presenting author: Caroline Jachmann

[1] Laboratoire de Spectrométrie de Masse BioOrganique (LSMBO), Université de Strasbourg, CNRS, Strasbourg, France

[2] Institut de Biologie Moléculaire des Plantes, CNRS, Université de Strasbourg, 67084 Strasbourg, France

[3] VIB-UGent Center for Medical Biotechnology, VIB, 9000 Ghent, Belgium

[4] Department of Biomolecular Medicine, Faculty of Health Sciences and Medicine, Ghent University, 9000 Ghent, Belgium

We introduce an exploratory approach aimed at detecting changes in post-translational modification abundance at the modification as well as the peptidoform level in E. coli by combining a robust quantitative method (MSqRobPTM) with a sensitive data acquisition method (DIA-PASEF). Traditional PTM analysis methods often struggle with capturing subtle variations in modification levels, especially in lower abundant proteoforms. In response, our initial investigation merges the statistical power of msqrobPTM with the comprehensive and unbiased sampling capabilities of DIA-PASEF mass spectrometry, comparing the results with a standard DDA open search analysis.

This preliminary study offers a glimpse into potential advancements in understanding dynamic PTM regulatory mechanisms, showcased on a comprehensive dataset of E. coli cultures grown and sampled under twelve different conditions.

P20

PAC·MASS: a versatile proteogenomic-based pipeline for the analysis of cancer immunopeptidomes.

Laura Ossorio Carballo [1], Virginia Sánchez Quiles [1], Thibault Chaze [1], Benjamin Sadacca [1]
Presenting author: Laura Ossorio Carballo
[1] Mnemo Therapeutics

The coding portion of the human genome represents barely around 5% of the whole genome, with the non-coding regions of the genome being its vast majority. These non-coding regions are usually referred as the “dark genome”. They contain both unannotated (around 50% of the whole genome) and annotated regions. Traditionally believed as not translated, increasing evidence point to the translation of certain non-coding genomic regions into proteins.

Some of these novel translational events can occur as a result of epigenetic and splicing defects in tumors, generating novel chimeric proteins that can subsequently be presented in form of short, chimeric peptides, linked to Major Histocompatibility Complex type I (MHC-I) at the surface of tumoral cells. Peptides presented by the MHC-I are hence recognized by the immune system, triggering a response against the tumor cells displaying the chimeric peptide.
At Mnemo, we investigate the non-canonical chimeric proteins and their MHC I-presented peptides, leveraging the annotated non-coding genome, to which we refer as the “grey genome”, for targeting cancer.

The validation of antigens presented by MHC I through mass spectrometry is a pivotal step in selecting candidates for therapeutic purposes. This is challenging, given that the chimeric peptide sequences are not described in canonical proteomes, and surface presentation depends on the peptide's ability to bind MHC I.

Therefore, at Mnemo we use a mass spectrometry-based proteogenomic approach for the detection of chimeric MHC-presented peptides. First, we construct, assemble and annotate RNA-seq based databases which are then utilized in mass spectrometry interrogations. We then apply our customized MS data analysis pipeline, PAC·MASS, on both in-house and public immunopeptidomic datasets. PAC·MASS allows the harmonization of multi-search engine outputs annotates the chimeric peptides within the corresponding proteins of origin and interprets the peptides in the context of their MHC-I presentation. Furthermore, PAC·MASS generate comprehensive reports, complete with graphical representations, enabling users to explore data quality and biological attributes. Despite the complexities associated with data formats, we advocate for collaborative efforts between industry and academia to advance tools such as PAC·MASS that prioritize reproducibility and compatibility.

P21

Molecular graph representations allow accurate retention time predictions for peptides carrying previously unseen PTMs

Ceder Dens [1], Kris Laukens [1], Wout Bittremieux [1]

Presenting author: Ceder Dens

[1] Adrem Data Lab, Department of Computer Science, University of Antwerp, Middelheimlaan 1, 2020 Antwerpen, Belgium

Mass spectrometry-based proteomics has sparked a series of breakthroughs, enhancing our understanding of cellular processes, disease mechanisms, and drug target screening. Despite these strides, a significant challenge persists: identifying proteoforms carrying diverse Post Translational Modifications (PTMs). PTMs play critical functional roles, ranging from the regulation of protein activity and interactions to impacting protein stability and localization. The limited understanding of the myriad PTMs, their potential occurrence, and their influence on experimental properties, such as liquid chromatography (LC) behavior presents a significant hurdle. In recent years, peptide property prediction has emerged as a pivotal tool, significantly enhancing peptide identification performance. These advancements, however, are primarily focused on unmodified peptides or those with a limited diversity in modifications. We present a transformer-based Retention Time (RT) prediction model for peptides with any PTM. Our novel input representation converts the modified peptides into molecular graphs, resulting in a more information rich embedding and simultaneously allowing predictions for peptides containing any PTM, even those not seen during training. We show that we outperform a baseline model that ignores unseen PTMs and the current state-of-the-art, DeepLC, that encodes the PTMs by their atom counts.

P22

Functional analysis of MS-based proteomics data: from protein groups to networks

Nadezhda T. Doncheva [1], Marie Locard-Paulet [1, 2], John Scooter Morris [3], Lars Juhl Jensen [1]

Presenting author: Nadezhda Doncheva

[1] Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark

[2] Institut de Pharmacologie et de Biologie Structurale, Université de Toulouse, CNRS, Université Toulouse III—Paul Sabatier, Toulouse, France

[3] Resource on Biocomputing, Visualization, and Informatics, University of California, San Francisco, San Francisco, USA

In bottom-up mass spectrometry (MS), proteins are digested into peptides and the peptide MS signals are then used to infer protein relative quantities across samples. Proteins that cannot be unambiguously distinguished based on the available set of peptides are reported as protein groups containing several protein accessions. However, typical follow-up analysis such as gene set enrichment and protein interaction networks are based on gene-level annotation. Thus, they can only be performed on single proteins or genes, rendering such analysis incompatible with protein group outputs. Currently, there is no best practice on how to handle this and its impact on functional analysis has not been studied yet. Here, we investigate the composition of protein groups identified in 14 published proteomics data sets, including deep proteomes, phosphoproteomics data, single-cell proteomics and pull downs from different species. According to our analysis, some of the factors that strongly impact the number and size of protein groups include sensitivity of the experiment and amount of input material as well as the set of sequences used for searching the MS data. We also show that gene set enrichment and network analysis can be affected to a different extent by the choice of which single protein is selected from each protein group, and that this selection should not be overlooked. To this end, we are developing a new Cytoscape app that will complement the widely used stringApp by creating STRING networks from protein groups input instead of single protein accessions. In the resulting networks, each protein group will be represented as a single node that will inherit all existing edges of the group members. In addition, all relevant node and edge attributes will be aggregated. This app will open new avenues for performing network analysis with protein groups from bottom-up MS studies.

P23

PathwayPilot: A User-Friendly Tool for Visualizing Metabolic Pathways and Navigating the KEGG Database

Tibo Vande Moortele [1], Pieter Verschaffelt [1, 2], Qingyao Huang [3], Nadezhda Tsankova Doncheva [4], Tanja Holstein [2, 5, 6], Caroline Jachmann [7], Lennart Martens [2, 5], Bart Mesuere [1], Tim Van Den Bossche [2, 5]

Presenting author: Tibo Vande Moortele

[1] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

[2] VIB - UGent Center for Medical Biotechnology, VIB, Ghent, Belgium

[3] Swiss Institute of Bioinformatics and University of Zurich

[4] Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Denmark

[5] Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium

[6] Bundesanstalt für Materialforschung -und Prüfung, Berlin, Germany

[7] Institute for Bioinformatics and Medical Informatics, University of Tuebingen, Germany

Understanding metabolic pathways is critical in deciphering biological processes, yet navigating the vast information within the Kyoto Encyclopedia of Genes and Genomes (KEGG) database presents challenges. Indeed, transitioning from identified peptides to these metabolic pathways can be a daunting task. To address this, we here introduce PathwayPilot, a user-friendly web application streamlining the exploration and visualization of metabolic KEGG pathways from metaproteomics data.

PathwayPilot seamlessly aligns identified peptides or proteins with Enzyme Commission numbers and taxon identifiers, providing users the flexibility for peptide-centric or protein-centric analyses. Through intuitive visualizations, this tool highlights identified proteins across pathways, offering clear insights into metaproteomics data. Additionally, researchers can precisely target organisms of interest, facilitating intra- and inter-sample comparisons.

P24

Deciphering APC/C-Cdc20 phosphorylation dynamics by advanced quantitative mass-spectrometry.

Sénécaut Nicolas [1,2], Han Ying [1], Ho Franz [3], Kamenz Julia [1]

Presenting author: Nicolas Sénécaut

[1] Molecular Systems Biology (MSB), Groningen Biomolecular Sciences and Biotechnology Institute (GBB), University of Groningen, Groningen, The Netherlands

[2] (previously) Camadro Lab, ProteoSeine@IJM, Institut Jacques Monod, Paris, France

[3] Proteomics Platform, Groningen Biomolecular Sciences and Biotechnology Institute (GBB), University of Groningen, Groningen, The Netherlands

Cell division is one of the most fundamental processes of life. Errors during cell division can result in the loss or gain of genomic information and promote diseases, such as cancer. The anaphase-promoting complex/cyclosome (APC/C), together with its coactivator Cdc20, coordinates progression through mitosis, which is tightly regulated via temporally controlled polyubiquitination. During mitosis, APC/CCdc20 is phosphorylated at over a hundred residues distributed over at least 10 subunits, most of which are highly conserved, suggesting they are functionally relevant. Despite the large number of phospho-sites, recent studies suggest that only a handful of these sites are necessary and sufficient for conferring APC/CCdc20 activity. We hypothesize that additional phosphorylations play a role in the temporal regulation of APC/CCdc20 activity and are responsible for the observed time delay between mitotic entry and APC/C activation. However, to explore how phosphorylation precisely contributes to the dynamics of APC/C activity, we lack the fundamental knowledge on which phosphorylation site becomes phosphorylated exactly when and with what kinetics. To address this question, we will isolate the APC/CCdc20 at different time points from homogenously cycling extracts and analyze the phosphorylation state of the APC/CCdc20 using advanced quantitative mass spectrometry. Furthermore, we will determine the overall extent of phosphorylation of full-length APC/C subunits using native MS. By linking this data to structural and functional information, we will develop a precise kinetic model of APC/C phosphorylation at single amino acid resolution. Leveraging my extensive expertise gained during my doctoral studies in mass spectrometry-based omics and the successful development of the innovative method for quantitative proteomics in bottom-up and top-down, the "Simple Light Isotope Metabolic Labeling" (SLIM-Labeling). Employing state-of-the-art biochemical and quantitative mass spectrometry techniques, complemented by advanced home-developed bioinformatics tools, this postdoctoral project aims to dissect the intricate phosphorylation dynamics governing APC/CCdc20. The project will not only provide a comprehensive model of the phospho-regulation of a critical cell cycle regulator but more generally provide fundamental insights into how multisite phosphorylation contributes to the temporal regulation of enzyme function.

P25

Aligning DIA Proteomics Data in Space: A Large-Scale Citizen Science Project

Toon Callens [1], Maarten Dhaenens [2], Tine Claeys [1], Sander Willems [3], Lennart Martens [1]

Presenting author: Toon Callens

[1] VIB-UGent Center for Medical Biotechnology, Department of Biomolecular Medicine, Ghent University, VIB, 9000 Ghent, Belgium

[2] ProGenTomics, Laboratory of Pharmaceutical Biotechnology, Ghent University, 9000 Ghent, Belgium

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

Identification and quantification of large amounts of proteins from a sample has major clinical importance and mass spectrometry (MS) provides a means to perform this in a sensitive way. The majority of proteomic analyses conducted by MS is data dependent acquisition (DDA). Recently, however, data independent acquisition (DIA) became increasingly popular to try and overcome certain problems in DDA by not selecting specific ions, but instead acquiring all fragments simultaneously in a continuous manner, at the price of leading to complex, chimeric data. Hidden in this complex data cluster lies a great amount of information on the proteome of the specific sample. One of the components of correct protein identification and noise reduction is the alignment of data of different mass spectrometry runs. We therefore set up a collaboration with MMOS and NetEase to implement DIA data alignment in the highly popular mobile game EVE Echoes. For this project, data is used in four dimensions: precursor retention time, precursor drift time, fragment mass over charge and fragment intensity. Players of EVE Echoes will align data of two runs of the same sample based on the first three dimensions. We will acquire the exact matches, as well the translation vectors the players made to create these matches. Later, a scoring function based on all dimensions will be used as a means to compare the gamers' alignment to already existing and our newly developed alignment algorithm. Eventually, a machine learning algorithm will be trained on this acquired data to achieve better alignment and filter out noise from these complex datasets.

P26

Unraveling The Protein Phosphorylation and Glycosylation Landscape of Diffuse Large B-Cell Lymphoma Using Multilayer Proteomics and Integrative Multiomics

Sebastian P. Perner [1, 2, 3], Yanlong Ji [2], Chih-Hsuan Yeh [2], Ralf Pflanz [2], Sabine Koenig [2], Monika Raabe [2], Julius C. Enssle [3], Carmen Doebele [3], Bjoern Haeupl [3,4], Thomas Oellerich [1, 3, 4], Kuan-Ting Pan [1, 3], Henning Urlaub [2]

Presenting author: Sebastian P. Perner

[1] Frankfurt Cancer Institute, Frankfurt am Main, Germany

[2] Max-Planck-Institute for Multidisciplinary Sciences, Goettingen, Germany

[3] University Hospital Frankfurt, Department of Hematology/Oncology, Frankfurt am Main, Germany

[4] German Cancer Consortium/German Cancer Research Center, Heidelberg, Germany

Introduction: Diffuse large B-cell lymphoma (DLBCL) is the most common type of non-Hodgkin lymphoma encountered clinically. Extensive genomics studies have enabled the stratification of DLBCL subtypes which exhibit significantly different clinical outcomes and responses to the most common treatment regimes. However, the functional consequences of these heterogeneous genetic aberrations on protein dynamics remain largely unknown. We applied multilayer proteomics and integrative multi-omics on DLBCL cell lines to investigate the underlying mechanisms of the disease.

Methods: 20 DLBCL cell lines were quantitatively profiled using our streamlined TMT-based pipeline, allowing for multiplexed proteomics, phospho- and glycoproteomics being analyzed from one consecutive sample preparation workflow. Data was analyzed on single layer level (e.g. differential expression and functional enrichment) and afterwards combined in an integrative approach using other layers of omics data (e.g. RNAseq and drug sensitivity) to identify interlayer-correlations.

Results: Our data clustered the cell lines into 4 subgroups showing distinct proteomic patterns of 772 proteins total. Two clusters showed significantly regulated proteins in B-cell receptor and NF-kappaB signalling pathways, respectively, in agreement with the well-established DLBCL genetic subtypes. Quantification of 18278 phospho-sites revealed distinct molecular regulations in the cell line clusters and allowed for activity-based prediction of upstream kinases and phosphatases, further stratifying the clusters. Quantification of 7800 glycoforms showed a distinct pattern compared to proteomics-based clustering, establishing three glycosylation-specific clusters exhibiting unique characteristics. The integrative multi-omics approach allowed us to identify additional pathways of interest in the respective clusters and provides further validation for our findings.

Conclusion: We established characteristic clusters of DLBCL cell lines with distinct biological and functional features. Integrated multi-omics analysis provides a valuable insight into the underlying mechanisms of different DLBCL subtypes and identifies potential targets for functional analysis guiding towards new treatment approaches, e.g. interfering with critical upstream enzymes of the respective subtypes. Soon, more DLBCL cell lines will be characterised using other mass spectrometry approaches and more in-depth integration is to be established.

P27

Dual phosphoenrichment method for the recovery of pSer/Thr/Tyr-peptides from cisplatin-resistant esophageal carcinoma cells to study the modulation of CDK4/6 inhibition

Marta Ávalos Moreno [1], Virginie Imbault [1], Benjamin Beck [1, 2], Xavier Bisteau [1]

Presenting author: Marta Ávalos Moreno

[1] Institut de Recherche Interdisciplinaire en Biologie Humaine et Moléculaire (IRIBHM), Faculty of Medecine, Université Libre de Bruxelles (ULB), 1070 Brussels, Belgium

[2] Welbio/FNRS Principal investigator at IRIBHM, Faculty of Medecine, Université Libre de Bruxelles (ULB), 1070 Brussels, Belgium

Esophageal squamous cell carcinoma (eSCC) accounts for 90% of esophageal cancer (EC) cases worldwide and has an overall survival below 20%. Although the standard of care – a combination of platinum salts and taxanes – has not changed over the last ten years, patients following this regimen often relapse and develop resistance. Among the heterogeneous landscape of oncogenic alterations, the cell cycle has appeared frequently deregulated, especially at the level of the cyclin D-CDK4/6-pRB axis, making this pathway a promising target for therapy. CDK4/6 inhibitors are already being used in first-line treatment for advanced ER+/HER2- breast cancer, but some clinical trials testing CDK4/6i as monotherapy against advanced EC (including eSCC) reported only a limited clinical activity. It should be noted all enrolled patients with eSCC tumors had been previously treated with genotoxic drugs such as platinum salts before CDK4/6i treatment. We wonder how resistance to genotoxic treatment (e.g. cisplatin) may modulate the tumors' molecular landscape and, ultimately, the response to CDK4/6i. In this context, characterization of the phosphoproteome can provide key information to comprehend the changes that allow eSCC cisplatin-resistant tumors to adapt to treatment and survive. To this end, we are implementing a dual phosphoenrichment method. This method includes a first enrichment step with classical TiO2 and Fe-IMAC beads, for which parameters such as buffer composition, type of column, and peptide-to-beads ratio have been optimized to maximize recovery. Additionally, due to prevalence of alterations identified among tyrosine kinase receptors in eSCC, a second phosphoenrichment step which combines a 4G10 antibody and a SH2-modified superbinder coupled to NHS beads will allow to efficiently recover the pTyr-peptides that the classical beads fail to do. We have successfully established the protocols for protein production, purification and NHS-beads coupling, showing the capacity of the new beads to bind pTyr peptides, and we are currently optimizing the phosphoenrichment protocol. Once optimized, we will acquire the phosphoproteome of four cisplatin-resistant eSCC cell lines to determine the altered molecular pathways via gene ontology enrichment, and to identify changes in kinase activity via kinase set enrichment analysis (KSEA). These analyses will potentially reveal strategies to rescue the sensitivity to CDK4/6i.

P28

Multi-omics investigation of glioblastoma using DESI imaging and LC-MS

Brittannie Willis [1], Elizabeth Want [2], Harry Whitwell [2], Nelofer Syed [2]

Presenting author: Brittannie Willis

[1] Imperial College London

[2] Imperial College London

Despite extensive research, glioblastoma (GBM) remains one of the most challenging primary brain tumours, characterised by its heterogeneity, aggressive nature, high recurrence rate, and the absence of effective treatment strategies, contributing to its high mortality rate. My research focuses on identifying proteomic markers associated with the heterogeneity and metabolic behaviour of the disease, particularly in response to metabolic therapies. We have access to both control mouse brains and brain samples from GBM mouse models. To maximise the protein extraction efficiency from mouse brain tissue, I employed iTRAQ-tagging (Isobaric Tags for Relative and Absolute Quantification) using Acquity M-Class ultra-performance liquid chromatography (Waters) coupled with a ZenoTOF 7600 mass spectrometer (Sciex).

Laser resection was performed on fresh-frozen mouse brain slices (8 µm thick) to create regions of various sizes (1,000,000 µm², 250,000 µm², 10,000 µm², and 100 µm²). Each region was extracted using one of three lysis methods: 1) 6M Urea, 100 mM triethylammonium bicarbonate (TEAB); 2) 1% sodium deoxycholate, 100 mM TEAB, 10% isopropanol, and 50 mM sodium chloride; and 3) PBS, which included freezing at -80°C for 5 min and heating at 90°C for 10 min (Minimal ProteOmic sample Preparation). Subsequently, these regions were subjected to iTRAQ labelling to determine the smallest sample size suitable for proteomic analysis. The peptides were separated on a Kinetex 2.6 µm XB-C18 100 LC Column (Phenomenex), with mobile phase A (0.1% formic acid in water) and mobile phase B (0.1% formic acid in acetonitrile) as solvents.

Liquid Chromatography-Mass Spectrometry (LC-MS) parameters were carefully chosen to improve the chromatographic resolution and mass spectrometric sensitivity for proteomic profiling. The gradients ranged from a 20-min gradient at 30°C with a 0.01 sec MS/MS accumulation time to a 40-min gradient at 50°C with a 0.02 sec MS/MS accumulation time. Further changes included MS/MS accumulation times of 0.2 and 0.03 secs, as well as setting Q1 resolution to HIGH. Mascot and Fragpipe, combined with statistical analysis in R, enabled the exploration of iTRAQ proteomics, revealing the optimal lysis buffer and aiding the identification of the most effective MS/MS method. These findings form the basis for further research during the PhD, facilitating the identification of molecular signalling pathways associated with GBM.

P29

A Multi-Proteomic Approach to Unravel New Players in Metastatic Melanoma Progression

Alessandra Morelli [1], Vittoria Matafora [1], Angela Bachi [1]

Presenting author: Alessandra Morelli

[1] IFOM ETS - The AIRC Institute of Molecular Oncology, Via Adamello 16, 20139 Milano, Italy

Amyloid-like fibrils have been recently discovered as characteristic of metastatic melanoma, both in vitro and in patient samples. Although amyloids are typically harmful, physiological fibrils of melanocyte-specific protein (PMEL) in melanocytes serve as a scaffold for melanin deposition. Some proteases, including the beta-secretase BACE2, generate the amyloidogenic peptides, and in normal conditions this process is finely regulated. Elevated expression and activity of BACE2, correlating with a poorer prognosis, have been reported in numerous tumors, including melanoma. Our group demonstrated that BACE2-dependent PMEL fibrils in metastatic melanoma promote cancer cell growth and invasion through mechanotransduction activation. Preliminary results had also detected increased secretion of ECM components in high BACE2 cells. To explore the structural rearrangements occurring in the secretome of high-BACE2 metastatic melanoma cells, we employed Limited Proteolysis coupled to Mass Spectrometry (LiP-MS). The R package protti was exploited to normalize the peptide abundance level on the protein one, and to identify and visualize conformotypic peptides, assessing their proximity to critical functional sites. Moreover, to reveal BACE2-dependent cleavage targets, and clarify whether their processing might influence the metastatic phenotype, we applied the N-Tails degradomics approach. Our N-Tails experiments unveiled numerous N-Termini regulated by BACE2 activity, many of which are found in proteins involved in cell adhesion. N-Tails data were elaborated with a Perl tool named MaxQuant Advanced N-termini Interpreter (MANTI), that permits to perform limma-based statistical test, further integrating important annotation from external sources, such as TopFinder and Uniprot. Interestingly, LiP-MS analysis demonstrated structural rearrangements in some BACE2-dependent cleaved proteins. Among those, amyloidogenic PMEL peptides resulted conformationally perturbed in metastatic melanoma models. Other known amyloidogenic proteins were also detected by LiP-MS. This suggests a potential role of PMEL as seed, influencing the stability of other proteins and their fibrillation. Further studies are required to elucidate whether processing of BACE2 targets may sustain a pro-metastatic phenotype and to clarify whether fibrils contribute to the creation of a solid fibril network, which could impede drug diffusion or create a hypoxic and stressful microenvironment.

P30

i2MassChroQ : full native timsTOF PASEF-enabled quantitative proteomics

Olivier Langella [1], Thierry Balliau [1], Marlène Davanture [1], Mélisande Blein-Nicolas [1], Filippo Rusconi [1]

Presenting author: Olivier Langella

[1] PAPPSO, Université Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE-Le Moulon

#Introduction

X!TandemPipeline (Langella et al. 2017) is a proteomics free and open source Java software program designed to filter and group peptide/protein identifications from MS/MS mass spectra that is used with the MassChroQ quantitative proteomics software (Valot et al. 2011). After a complete rewrite in C++17, the new software, named i2MassChroQ, features native support for the timsTOF raw data format, peptide/protein quantification by merging the features initially in X!TandemPipeline and MassChroQ, and ultimately statistical analysis using either the MSstats or MCQR GNU R package.

#Methods

i2MassChroQ is written in portable C++17 and makes use of the Qt libraries for the graphical user interface. Binary packages are available for Linux and MS Windows. The timsTOF native raw data reader was developed in-house with the technical specifications provided by Bruker. The sofware was tightly optimized to ensure very fast access to the binary data. The current version provides real time MS/MS peptide annotation and performs extremely fast ion current extractions and XIC chromatogram visualizations.

#Results

The Bruker timsTOF line of instruments improves the identification of peptides and proteins in complex mixtures by implementing a peculiar ion mobility technology. i2MassChroQ has the distinct feature, with respect to the MaxQuant and MSFragger competitors, of natively parsing the timsTOF raw data with original software code that puts us in total control of the data processing, leveraging speed and accuracy.

Our software identifies roughly the same amount of peptides as when using the X!Tandem engine on data provided by the Bruker Data analysis software but performs much faster, reducing processing time from 55 to 12 minutes on a typical HeLa sample quality control LC-MS/MS run. For peptide quantifications, i2MassChroQ uses the ion mobility-enabled version of MassChroQ. Peptide quantification is 20% faster than IonQuant on the same computer.

Using benchmark datasets (PXDO1OO12 and PXD014777), we demonstrate that i2MassChroQ identifies and quantifies significantly more proteins, in particular with a better capability to quantify peptides and proteins of lesser abundance, while maintaining extremely low missing data percentages at the feature level (3% in each condition), compared to 56% in some conditions for MS-Fragger.

This leads to better protein quantification values by drastically reducing the requirement of missing value imputations.

P31

The open-source Alpha universe for all processing steps of proteomics data

Maximilian T. Strauss [1, 2], Sander Willems [1, 3], Isabell Bludau [1], Wen-Feng Zeng [1], Eugenia Voytik [1], Constantin Ammar [1], Elena Krismer [2], Georg Wallmann [1], Xie-Xuan Zhou [1], Julia Schessner [1], Florian Meier [1, 4], Matthias Mann [1, 2]

Presenting author: Julia Schessner

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

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

[3] Research and Development, Bruker Belgium nv., Kontich, Belgium

[4] Functional Proteomics, Jena University Hospital, Jena, Germany

MS-based shotgun proteomics is used in a wide array of biological and clinical studies. A major aspect and challenge is data analysis, from identifying and quantifying peptides and proteins to differential analysis and result visualization. While there are freely available tools for many individual applications, they do not always interoperate well. Additionally, only few of them live up to open-science standards, such as providing open-source code, extensive testing and documentation. This makes it challenging to leverage latest developments in machine and deep learning or implementing novel ideas across the pipeline. Our universe of AlphaX tools aims to alleviate this by providing all processing steps, entirely in the form of well-documented python code available on GitHub. To achieve high processing speeds despite using Python, we use numba just-in-time compilation. To facilitate interoperability, all tools share utilities, like fast accession of raw files from all vendors, in the form of AlphaBase and generally use efficient and widely used file formats like hdf5 and csv. The two search engines AlphaPept and AlphaDIA enable a variety of fast raw data processing workflows, including multiplexed DIA. AlphaPeptDeep can be used for deep learning powered prediction of peptide and fragment properties, enabling among others HLA peptidomics. DirectLFQ and AlphaQuant tackle scalable and accurate peptide and protein quantification, followed by AlphaPeptStats for differential protein analysis and result visualization. The dedicated visualization tools AlphaViz and AlphaMap enable exploration of raw data and identified peptides, for instance in the context of the protein structures predicted by AlphaFold. By making our well-documented and tested code-base open-source with a permissive license, we support open-science concepts and enable community contributions.

P32

Enhancing Proteomics Research through the National Health Data Science Sandbox

Jacob Fredegaard Hansen [1], Ole Nørregaaard Jensen [1], Veit Schwämmle [1]

Presenting author: Jacob Fredegaard Hansen

[1] Department of Biochemistry and Molecular Biology, University of Southern Denmark, DK-5230 Odense M

The National Health Data Science Sandbox, hosted on UCloud and Computerome's high-performance cloud services, offers a national infrastructure for health data science education, featuring non-sensitive health data and advanced tools within a supercomputing environment. While the Sandbox encompasses various areas, our emphasis here is on the clinical proteomics training module, which is part of a broader curriculum that also covers genomics and transcriptomics. Proteomics is the large-scale study of proteins, particularly their structures and functions, within a given organism or biological context.

The clinical proteomics module is equipped with cutting-edge software tools essential for a comprehensive proteomics pipeline. This includes FragPipe, MaxQuant, PDV, SearchGUI, PeptideShaker, MZmine 3, and DIA-NN. Available on UCloud as the 'Proteomics Sandbox', these tools facilitate both independent and guided learning in clinical proteomics analysis. Another key innovation is the integration of ColabFold with AlphaFold2 on the UCloud platform, making advanced protein structure prediction accessible to Danish researchers and students. This incorporation empowers in-depth learning of advanced computational methodologies requiring significant extensive storage and computational resources.

These tools collectively form the core of training in proteomics-based research and innovation in the National Health Data Science Sandbox. The deployment of the 'Proteomics Sandbox' and 'ColabFold' within UCloud's supercomputing environment offers high-performance computing capabilities, combined with a user-friendly graphical user interface. These resources provide open-source software tools for clinical proteomics data analysis and protein structure prediction, leveraging supercomputing power. Additionally, these training modules have been effectively incorporated into applied bioinformatics curricula for both master's and PhD students.

The modules are continuously evolving to incorporate the latest tools and techniques in clinical proteomics with feedback from students and researchers. Hence, we are open to and actively seeking feedback and collaboration proposals to further enrich and expand the scope of clinical proteomics offerings within these modules in the Sandbox.

P33

MOLECULAR CHARACTERIZATION OF COLLAGEN-BASED ANIMAL GLUES BY PROTEOMICS AND SPECTROSCOPIC ANALYSES

Georgia Ntasi [1], Brunella Cipolletta [1], Carmen Aprea [1], Laura Dello Ioio [2], Celia Duce [3], Emanuele Crisci [3], Emilia Bramanti [4], Alessandro Vergara [1,5], Ilaria Bonaduce [3], Leila Birolo [1,5]

Presenting author: Brunella Cipolletta

[1] Department of Chemical Sciences, University of Naples Federico II, 80126 Naples, Italy

[2] Dello Ioio Restauri, Vico Equense, 80069 Naples, Italy

[3] Department of Chemistry and Industrial Chemistry, University of Pisa, 56126 Pisa, Italy

[4] Institute of Chemistry of Organo Metallic Compounds, CNR, 56124, Pisa, Italy

[5] Task Force "Metodologie Analitiche per la Salvaguardia dei Beni Culturali", University of Naples Federico II, 80126 –Naples, Italy

Animal glues prepared from connective tissues are widely used in restoration as adhesives, binders and consolidants. Collagen in its natural state is a triple helix protein characterized by a Gly−X−Y repetitive sequence and a unique high content of Pro and HyPro that make collagen easily recognizable in the protein universe. Upon treatment with acids or alkalis in hot water, the insoluble collagen becomes a soluble gelatin to be used as animal glue. The performance of the glue strongly depends on the original source of collagen but is also influenced by the extraction and preparation procedures. Molecular characterization of animal glues is crucial to help restorers in selecting the most appropriate materials to be used in conservation treatments. A multi-analytical approach based on proteomics and spectroscopic analyses is used in this work to characterize animal glues from different sources provided by restoration workshops. First, a shotgun proteomics approach was used for protein identification, allowing to establish the glue source and distinguish between hide and bone glues. Then, proteomics and analytical pyrolysis coupled to GC-MS were used to analyse chemical modifications in collagen. In particular, backbone cleavage of the polypeptide chain and deamidation of Asn and Gln were evaluated as expected chemical modifications in proteins from animal glues. Backbone cleavage was evaluated as semitryptic peptides generated upon trypsin hydrolysis, with a trypsin cleavage site only at one end. Pyrolysis coupled with GC-MS was also carried out to evaluate the most characteristic pyrolysis products of proteins: DKPs. High yield of DKPs can be ascribed to a high degree of protein hydrolysis. Data obtained reveal that generally bone glues are more fragmented than hide ones. Lastly, deamidation was evaluated from raw LC-MS/MS data by MaxQuant software with an in-house script based on PSMs intensities for semiquantitative evaluation. Data collected show that on average bone glues are less deamidated than hide ones, while mixed glues are overall less deamidated than pure ones. How these molecular details are reflected in the 3D structures and in the rheological properties of animal glues is now under investigation by a combination of TGA, DSC and spectroscopic analyses.

P34

CPred: Charge State Prediction for Modified and Unmodified Peptides in Electrospray Ionization

Frédérique Vilenne [1,2], Simon Appeltans [1], Gökhan Ertaylan [2], Dirk Valkenborg [1]

Presenting author: Frédérique Vilenne

[1] University Hasselt, Hasselt, Belgium

[2] VITO, Mol, Belgium

Mass spectrometry-based proteomics is becoming a more indispensable tool as time passes, capable of identifying and detecting (un)modified peptides. This makes mass spectrometry a key element in shifting the current curative medicine towards personalised medicine. The mass-to-charge ratio is pivotal to identifying these peptides, especially when peptides are ionised by tools like electrospray ionisation, which produces multiply charged ions. During this research, we developed a neural network called CPred, which can accurately predict the charge state distribution from +1 to +7 for modified and unmodified peptides. The model was unrestricted about the protease and fragmentation methods. CPred was trained on the large-scale synthetic ProteomeTools project and further evaluated on independent test datasets. Results were evaluated through the Pearson correlation coefficient and showed high correlations up to 0.9998978 between the predicted and acquired charge state distributions. The effect of specifying modifications in the neural network and feature importance was further investigated, revealing the value of modifications and vital elements in holding on to protons. CPreds' accurate predictions of the charge state distribution can play a pivotal role in boosting confidence in peptide identifications during rescoring as a novel feature.

P35

XL-Prosit: Transfer learning to predict fragment intensities of cleavable cross-linked peptides

Mostafa Kalhor [1], Mathias Wilhelm [1]

Presenting author: Mostafa Kalhor

[1] Computational Mass Spectrometry, TUM School of Life Sciences, Technical University of Munich, Germany

Chemical cross-linking (XL) mass spectrometry (MS) is highly effective for studying protein structures and interactions. Despite its effectiveness, identifying cross-linked peptides remains a complex task. Previous studies suggest that incorporating fragment intensity data into the matching process can improve cross-linked peptide identification. Hence, we aim to extend Prosit's capabilities to predict fragment ion intensities for cleavable cross-linked peptides. For this purpose, training and test data were obtained from PRIDE and processed using Plink2 and XlinkX. In total, 130k MS2 and 37k MS3 spectra for cleavable cross-linked peptides were collected. However, one major challenges is the limited availability of training data. To address this, we systematically assessed methods to fine-tune the pre-trained Prosit model. To determine the optimal deep learning model for predicting MS2 spectra of cleavable cross-linked peptides, we explored ~20 architectures. To increase the prediction accuracy, we used augmentation by swapping the position of two cross-linked peptides in the context of a model that focuses on the prediction of the intensity pattern of one. The refined XL-Prosit model achieves high accuracy on the test data for MS3 (median spectral angle of 0.84 and Pearson’s correlation of ~ 0.95) and MS2 (median spectral angle of ~ 0.82 and Pearson’s correlation > 0.9). Next, we aim to integrate XL-Prosit into Oktoberfest, a data-driven rescoring pipeline, via Koina, a publicly available proteomics inference service, to enhance the sensitivity and specificity of identifying cross-linked peptides.

P36

Fragment ion intensity prediction improves the identification rate of non-tryptic peptides in timsTOF

Charlotte Adams [1, 2], Wassim Gabriel [3], Mario Picciani [3], Kris Laukens [1], Mathias Wilhelm [3], Wout Bittremieux [1], Kurt Boonen [2]

Presenting author: Charlotte Adams

[1] Adrem Data Lab, Department of Computer Science, University of Antwerp, Antwerp, Belgium

[2] Laboratory of Protein Science, Proteomics and Epigenetic Signaling (PPES), Department of Biomedical Sciences, University of Antwerp, Antwerp, Belgium

[3] Computational Mass Spectrometry, Technical University of Munich, Freising, Germany

The adaptive immune system can eradicate cancerous cells by recognising peptides bound to HLA-molecules present on the cell surfaces. In immunopeptidomics, these peptides—commonly termed immunopeptides—are isolated and characterized using mass spectrometry. To minimize false positives and improve spectrum annotation rates, peptide-spectrum match (PSM) rescoring can be used. This involves post-processing results from an unfiltered database search, during which multiple PSM features are used to distinguish between correct and incorrect PSMs. Recently, there has been significant interest in using additional features for PSM rescoring, including spectral features based on the similarity between experimental and predicted fragment ion intensities.
Because low abundant immunopeptides often occur, highly sensitive timsTOF instruments are increasingly gaining popularity. To improve PSM rescoring for immunopeptides measured using timsTOF instruments, we fine-tuned Prosit, a deep neural network that can predict the fragment ion intensities for a given peptide sequence. To fine-tune Prosit, a dataset was generated by analyzing over 300,000 synthesized non-tryptic peptides on a timsTOF-Pro.

After fine-tuning the Prosit model we were able to improve the prediction accuracy of tryptic and non-tryptic peptides measured on a timsTOF. By applying this new 2023 Prosit timsTOF model during PSM rescoring, we achieved an up to 3-fold increase in the identification rate of immunopeptides compared to standard database searching. Furthermore, our approach increased the detection of immunopeptides even from low input samples. Importantly, the immunopeptides identified after PSM rescoring are likely to bind HLA-molecules, as supported by motif analysis and binding affinity assessment, providing an orthogonal validation of their veracity and demonstrating the powerful benefits of PSM rescoring using highly accurately predicted fragment ion intensities.

To conclude, by applying our new fragment ion intensity prediction model for PSM rescoring, we can drastically increase the detection of immunopeptides, which hold the potential to serve as valuable targets for immunotherapy. The 2023 Prosit timsTOF fragment ion prediction model is freely available for community use through Koina.

P37

Improved protein quantification by using bipartite peptide-protein graphs

Karin Schork [1,2], Michael Turewicz [1,3], Julian Uszkoreit [1,4], Jörg Rahnenführer [2], Martin Eisenacher [1]

Presenting author: Karin Schork

[1] Medizinisches Proteom-Center, Medical Faculty, Ruhr-University Bochum, Bochum, Germany and Medical Proteome Analysis, Center for Protein Diagnostics (PRODI), Ruhr-University Bochum, Bochum, Germany

[2] Department of Statistics, TU Dortmund University, Dortmund, Germany

[3] Institute for Clinical Biochemistry and Pathobiochemistry, German Diabetes Center (DDZ), Leibniz Center for Diabetes Research at the Heinrich Heine University Düsseldorf, Düsseldorf, Germany and German Center for Diabetes Research (DZD), Partner Düsseldorf,
München-Neuherberg, Germany

[4] Medical Bioinformatics, Medical Faculty, Ruhr-University Bochum, Bochum, Germany

In bottom-up proteomics, proteins are enzymatically digested to peptides (smaller amino acid chains) before measurement with mass spectrometry (MS), often using the enzyme trypsin. Because of this, peptides are identified and quantified directly from the MS measurements. Quantification of proteins from this peptide-level data remains a challenge, especially due to the occurrence of shared peptides, which could originate from multiple different protein sequences.

The relationship between proteins and their corresponding peptides can be represented by bipartite graphs. In this data structure, there are two types of nodes (peptides and proteins). Each edge connects a peptide node with a protein node, if and only if the peptide could originate from a tryptic digestion of the protein. The aim of this study (Schork et al, 2022, PLOS ONE) is to characterize and structure the different types of graphs that occur and to compare them between different data sets. Furthermore, we want to show how this knowledge can aid relative protein quantification. Our focus is especially on gaining quantitative information about proteins with only shared peptides, as they are neglected by many current algorithms.

We constructed bipartite peptide-protein graphs using quantified peptides from four gold standard data sets and characterized them, especially regarding the protein nodes without unique peptides. Based on these findings, we developed and applied a novel method that calculates protein ratios from peptide ratios by making use of the bipartite graph structures. For each peptide node, an equation is formed based on the bipartite graph structures and the measured peptide ratios. Protein ratios are estimated by using an optimization method to find solutions with a minimal error term. Special focus lies on the proteins with only shared peptides, which often lead to a range of optimal solutions instead of a point estimate.

Depending on the data set between 7.1 % and 72.8 % of protein nodes do not have any unique peptide. With the new protein quantification method, we can obtain quantitative information also for these protein nodes that are in line with the expected ratios in the gold standard data sets. Further improvement may include the handling of missing values and outliers.

P38

Haplotypes and Population Diversity in Proteomics

Jakub Vašíček [1, 2], Ksenia G. Kuznetsova [1, 2], Dafni Skiadopoulou [1, 2], Pål R. Njølstad [1, 3], Stefan Johansson [1, 4], Stefan Bruckner [5], Lukas Käll [6], Marc Vaudel [1, 2, 7]

Presenting author: Jakub Vašíček

[1] Mohn Center for Diabetes Precision Medicine, Department of Clinical Science, University of Bergen, Bergen, Norway

[2] Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway

[3] Children and Youth Clinic, Haukeland University Hospital, Bergen, Norway

[4] Department of Medical Genetics, Haukeland University Hospital, Bergen, Norway

[5] Chair of Visual Analytics, Institute for Visual and Analytic Computing, University of Rostock, Rostock, Germany

[6] Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH - Royal Institute of Technology, Stockholm, Sweden

[7] Department of Genetics and Bioinformatics, Health Data and Digitalization, Norwegian Institute of Public Health, Oslo, Norway

Haplotypes are sets of alleles inherited together from a parent that are found in different populations with different frequencies. Using a single reference genome in analyses thus results in a bias against underrepresented populations (reference bias). Alleles co-occurring in the protein-coding regions of the same gene produce a unique protein sequence - protein haplotype. These haplotypes are present in biological samples, and detectable by mass spectrometry, but are not accounted for in proteomic searches. Consequently, the impact of haplotypic variation on the results of proteomic searches and the degree of reference bias in proteomics remain unknown. To address this gap, we introduce ProHap, a python-based tool that constructs protein sequence databases from phased genotypes of reference panels. ProHap empowers researchers to account for both common and rare genetic variation on protein sequences, enabling them to account for population diversity, and providing the opportunity to make visible the influence of haplotypes on protein abundance and regulation.

P39

Evaluating phospho-peptides quantification across various acquisition methods and bioinformatic tools: novel standards and their utility in algorithm development

Pinar Altiner [1], Carine Froment [1], Anne Gonzalez de Peredo [1], David Bouyssié [1], Odile Schiltz [1]

Presenting author: Pinar Altiner

[1] Institut de Pharmacologie et de Biologie Structurale, Université Paul Sabatier, CNRS, 205 Rte de Narbonne, 31400, Toulouse, France

Phospho-proteomic analysis seeks to identify, localize and quantify phospho-peptides purified from complex protein samples. Despite advances in analytical techniques, determining the exact localization of phosphorylation site(s) for a given phospho-peptide remains challenging due to the ambiguity in experimental fragmentation patterns within MS/MS spectra. While efforts have been made to create accurate localization algorithms, the existing implementations are still generating identification and localization errors. Evaluating these algorithms is crucial to assess their relative sensitivity and specificity. Numerous studies were published to that aim, but they solely focused on localization accuracy at identification level. However, the precise evaluation of the relationship between localization and label-free quantification accuracies remains still unclear to our knowledge.
In this regard, we performed several experiments designed for a quantitative evaluation across various acquisition methods and bioinformatic tools. We performed three different experiments by using either phosphorylated standards, based on a library of 180 synthetic phospho-peptides with known localizations, or on phospho-peptides enriched from mouse T -cells, spiked at various concentrations into an E. coli background. Those samples were analyzed using different MS instruments (Thermo Exploris and Bruker timsTOF) and various acquisition methods (Data Dependent Acquisition (DDA), Data Independent Acquisition (DIA)), with/out ion mobility separation. Several bioinformatic tools (Proline, MaxQuant, Proteome Discoverer (PD), DIA-NN, Spectronaut) were utilized for the post-processing analysis. According to the preliminary results, PD was found to have the highest sensitivity in the DDA datasets, although it also introduces a higher number of quantitative false positives. Moreover, when comparing DDA to DIA, Spectronaut exhibited the best sensitivity/ specificity ratio. Overall, the obtained results provide evidence that label-free quantification errors are more important for phospho-peptides present in the sample as different isomers, both for DDA and DIA experiments. Finally, the created scripts, automating the whole comparison of the spiked-in dataset, will be implemented inside WOMBAT-P. This should ease further evaluations of the provided data by the proteomics community.

P40

PPM facility : 15 years of expertise and best practice in proteomics data management in a FAIR mood

Oana Vigy [1,2], Serge Urbach [1,2], Cherine Bechara [1,2], Mathilde Decourcelle [1,3], Tristan Girbau [1,2], Khadija El Koulali [1,3], Séverine Chaumont-Dubel [1,2], Martial Séveno [1,3]

Presenting author: Vigy Oana

[1] Proteome Pole of Montpellier (PPM), Montpellier, France

[2] Institut de Génomique Fonctionnelle (IGF), Univ. Montpellier, CNRS, INSERM, Montpellier, France

[3] BioCampus Montpellier (BCM), Univ. Montpellier, CNRS, INSERM, Montpellier, France

Proteomics is a highly dynamic field requiring continuous development of new technological approaches and methods for analysis and processing data. Hence, there is a major need for a high-performance infrastructure to manage scientific data.

The Proteome Pole of Montpellier (PPM), founded in 2007, is a four-site proteomics facility gathering technologies and complementary expertise in various fields of proteomics and mass spectrometry. Our aim is to offer our users solutions in proteomics that meet the highest international standards. To achieve this, the platform is strongly involved in methodological developments and the implementation of new technologies and equipment. For almost 15 years, PPM has been implementing a data management strategy which is constantly evolving to enhance the Findability, Accessibility, Interoperability and Reusability (FAIR) of the scientific data. All stages of the proteomics data lifecycle are managed according to this strategy, which takes into account the extent and evolution of the proteomics landscape (currently on the FPP site):

- Collect data from various suppliers of mass spectrometry instruments and deal with each of their proprietary formats and tools (Thermo Fisher Scientific, Waters, etc.).
- Process data using a variety of analysis softwares that fit a particular purpose, which emerge, continue to evolve or decline according to the strengths of the developer and user communities (MaxQuant, Perseus, Cytoscape, Skyline, DIA-NN, DAPAR-ProStar, etc.). Further “in-house” developments are often necessary to improve the analysis, as we did with a particular MaxQuant output file to highlight automatically a representative “leading” protein among groups of protein.
- Preserve and store increasingly voluminous proteomics data (the amount of data generated has doubled over the last 5 years with 11 Tb in 2022) by evolving the data management strategy.
- Publish and share data, guiding our users through the process of submitting their data to international proteomics data repositories (ProteomeXchange, HUPO PSI standards, etc.).

We present how we apply the FAIR principles throughout the data lifecycle and highlight some of the features we have developed to improve the proteomics data analysis workflow. The aim is that our expertise may inspire those wishing to improve the management of their data and generate discussions on different ways to achieve that goal.

P41

lesSDRF is more: maximizing the value of proteomics data through streamlined metadata annotation

Tine Claeys[1], Tim Van Den Bossche[1], Yasset Perez-Riverol[2], Kris Gevaert[1], Juan Antonio Vizcaíno[2] & Lennart Martens [1]

Presenting author: Tine Claeys

[1] VIB-UGent Center for Medical Biotechnology, VIB, 9000, Ghent, Belgium

[2] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge

Open science in life sciences, exemplified by the Protein Data Bank and AlphaFold, has fostered global collaboration and advanced understanding. In proteomics, the ProteomeXchange consortium, established in 2011, has standardized data sharing and enforced FAIR principles, enabling the reuse and integration of diverse experimental data. However, the lack of sufficient metadata annotation remains a major barrier to fully leveraging public proteomics data, especially in a biological context.

Introduced in 2021, the Sample and Data Relationship Format (SDRF) for Proteomics maps data files to sample characteristics using standardized terms. Despite its potential, its complexity and the absence of streamlined annotation methods have led to limited use, with a mere 3.9% of datasets in PRIDE containing SDRF annotation since mid-2022. To address this issue, we developed lesSDRF, a user-friendly, web-accessible application that facilitates the SDRF annotation process. The app is structured into five intuitive steps: selecting species, uploading local metadata files, adding labeling information, and entering required and additional columns using ontology terms, all designed to ensure SDRF compliance and user ease.

lesSDRF's development and ongoing improvement are deeply rooted in community collaboration. This tool exemplifies the shift from data hoarding to open sharing, with continual updates based on feedback from proteomics experts and the broader community.. Moreover, we actively collaborate with leading proteomics tools to incorporate built-in SDRF outputs, setting a standard for data reusability and effective management across the field. This collective approach aims to unlock the full potential of public data, ensuring that accurate metadata becomes a cornerstone of good scientific practice and laying the groundwork for more significant scientific advancements.

P42

iDeepLC: An effective retention time predictor for unseen modified peptides that can differentiate between isomers

Alireza Nameni[1,2], Sven Degroeve[1,2], Lennart Martens[1,2],
and Robbin Bouwmeester[1,2]
Presenting author: Alireza Nameni
[1] VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium
[2] Department of Biomolecular Medicine, Ghent University, Ghent, Belgium

This research delves into proteomics, aiming to refine peptide and protein behavior modeling in high-throughput data. It addresses the challenge of identifying modified peptides, crucial for understanding proteomics, using advanced machine-learning techniques.
The motivation driving this study arises from the need for an improved and reliable method of identifying modified peptides with post-translational modifications (PTMs) via LC-MS/MS data analysis. Its predictive capability significantly aids in ordering peptides by likelihood, thereby facilitating the detection of potential errors in Open Modification Search (OMS) protocols. Essentially, it helps in sorting through a large amount of data and identifying peptides that might have been misidentified, contributing to more accurate and reliable results.

In a preceding work, the DeepLC, we introduced a method predicting retention times for unseen peptide modifications through atomic composition encoding. However, this method struggled to differentiate between structures with similar compositions and had limited extrapolation capabilities. In our iDeepLC model, we achieve a substantial elevation in precision for predicting the retention times of modified peptides.

This improvement is achieved through the utilization of chemical descriptors to encode amino acids and their corresponding modifications, surpassing atom counting by incorporating computed features from the chemical architecture of amino acids. By incorporating insights into the bonding patterns among atoms within amino acids, the model can more properly distinguish divergences among various amino acids and their modifications by better generalization of the model which leads to more sophisticated predictions for unseen modifications.

Moreover, iDeepLC stands out as the first retention time predictor capable of distinguishing between isomeric structures. For instance, it distinguishes between symmetric dimethyl arginine and asymmetric dimethyl arginine, showcasing its unparalleled ability among retention time predictors. This advancement is further validated through evaluations against ProteomeTools PTM datasets.

In essence, iDeepLC improves peptide modification identification, using a novel approach that not only enhances predictive accuracy and increases generalization but also enables distinguishability between structurally similar amino acids and modifications.

P43

MoDPA: Investigating diseases through Modification-Dependent Protein Associations

Enrico Massignani [1,2], Lennart Martens [1,2]

Presenting author: Enrico Massignani

[1] VIB-UGent Center for Medical Biotechnology, VIB, 9052 Ghent, Belgium.

[2] Department of Biomolecular Medicine, Ghent University, 9052 Ghent, Belgium.

Protein-Protein Interactions (PPIs) and Post-Translational Modifications (PTMs) are vital in regulating various biological processes. However, the interplay between these two regulatory mechanisms is still not well understood. To address this, we are developing a computational approach called MoDPA, which aims to identify co-occurring modified proteins by reanalysing a large number of public MS-based proteomics datasets. To simplify the analysis of high-dimensional quantitative PTM data, we employ a Variational Autoencoder. This deep-learning model projects the data to a low-dimensional latent space. To measure the similarity between two PTMs, we calculate their Pearson correlation coefficient in the latent space, following an approach that is conceptually similar to gene co-expression analysis.

The CompOmics, where I am currently doing my post-doctoral work, has access to a unique and comprehensive dataset of proteome-wide post-translational modifications (PTMs). This dataset was generated by reprocessing proteomics data from the PRIDE public repository using ionbot, a machine learning-based peptide identification engine, which can perform sensitive yet specific open modification searches to identify all PTMs present in a sample. So far, our reprocessing pipeline has processed 633 proteomics datasets, resulting in over 240 million peptide identifications. This corresponds to approximately 7 million unique peptidoforms and 4 million unique PTMs.

After applying the VAE-based dimensionality reduction approach to this dataset, I was able to generate a network of PTM associations. Upon a preliminary examination of the network, I identified a cluster of strongly interconnected PTMs on proteins associated with immune response, as well as two distinct clusters of tyrosine kinases and serine/threonine kinases, respectively.

As part of my post-doctoral work, I intend to conduct a thorough examination of this PTM association network using various network analysis methods, such as information flow analysis (IF). By simulating how biological information spreads throughout the network, the IF analysis will enable me to distinguish between "driver" and "passenger" proteins and PTMs.

Finally, I plan to study how PTMs affect the interactions between chaperones and their client proteins, with a focus on spontaneously occurring modifications related to molecular ageing.

P44

Differential usage analysis of the histone code at the consensus residue level

Ruben Almey [1], Nina Demeulemeester [2, 3, 4], Lieven Clement [4], Maarten Dhaenens [1]

Presenting author: Ruben Almey

[1] ProGenTomics, Laboratory of Pharmaceutical Biotechnology, Ghent University, Ghent, Belgium

[2] VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium

[3] Department of Biomolecular Medicine, Ghent University, Ghent, Belgium

[4] StatOmics, Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

The histone code is the abstraction of the myriad post-translational modifications on histones (hPTMs) that together regulate gene expression. Increasingly, this epigenetic phenomenon is thought to be an ancient sense and respond system for the metabolic and energetic state of the eukaryotic cell. Current understanding of the histone code largely derives from targeted, antibody-based technologies; however, attaining an overarching, unbiased image of the combinatorial nightmare it depicts is currently only possible through mass spectrometry (MS). While proven to be an extremely valuable technology, MS-based histone code analysis still suffers from underdeveloped data analysis tools. Here, we present a bottom-up histone data analysis pipeline that leverages both the evolutionary conservation of histones and PTM-tailored robust statistics to deeply mine histone code dynamics at consensus amino acid resolution. Through multiple sequence alignment, peptidoforms from all variants of the five histone families (H1, H2A, H2B, H3, and H4) are redefined against their respective consensus backbones. This deconstruction of variant information in turn enables a more comprehensive and biologically relevant testing of differential hPTM usage (DhPU) using a bespoke msqrob2PTM histone workflow. The resulting map of the histone code, detailing hotspots of change and hPTMs of interest, is finally presented in a browsable and user-friendly format, stimulating a more nuanced interpretation of MS-based histone data. This poster is supposed to elicit a discussion on potential pitfalls of the pipeline and how to address them, as well as ways to improve the end user experience.

P45

Species identification through peptide sequence ambiguation on ancient bone samples

Ian Engels [1], Alexandra Burnett [1], Simon Daled [1], Robbin Bouwmeester [2], Dieter Deforce [1], Isabelle De Groote [3], Maarten Dhaenens [1]

Presenting author: Ian Engels

[1] ProGenTomics, Laboratory of Pharmaceutical Biotechnology, Ghent University, Ghent, Belgium

[2] Compomics, VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium

[3] ArcheOs–Research Laboratory for Biological Anthropology, Department of Archaeology, Ghent University, Ghent, East-Flanders, Belgium

For centuries archaeologists have tried to identify species based on bone morphology. Although in the case of bone fragments this becomes more of a challenge. ZooMS approaches via MALDI-TOF-MS have proven to be able to tackle this problem based on species-specific PMFs. Some attempts have been made to use LC-MS for the classification of bones of unknown origins. However, the use of database search algorithms like Mascot to find peptide-spectrum matches is accompanied with a multitude of well-known issues peculiar to palaeoroteomic samples. These challenges include the selection of appropriate databases, which differ in size and composition; and the absence of sequences from extinct or unsequenced animals. The high similarity in tryptic peptides across the collagen homologues and orthologues can also result in a high number of ambiguous results, making the use of algorithms like Unipept not possible. Altogether the ambiguousness of the data makes the taxonomic classification of samples difficult.

With this poster we propose a novel downstream processing tool ‘ClassiCol’ that takes advantage of the ambiguity in the data to make a final taxonomic classification based on collagen sequences, taking into account the possibility of a sample mixture and/or missingness of the animal of origin in the database. This tool is accompanied by a comprehensive collagen database, covering over 10,000 sequences of more than 250 animals. ClassiCol takes a Mascot result file as input and matches every peptide to the collagen database, allowing for isobaric switches (i.e. positional isomers and isobaric substitutions, including PTMs). All isobaric peptide possibilities are subsequently filtered via retention time prediction and intensity prediction using DeepLC and MS2PIP respectively. All retained protein sequences and animals are clustered according to their homology and taxonomy respectively. This matrix is than separated into blocks, followed by discard of all single hit wonders. Each block is subsequently analysed in a leave-one-out approach in order to find the taxonomic level at which each (group of) species has the ability to distinguish itself from others. Finally the algorithm compares each group of animals in a 1-vs -many approach to find unique peptides for each group. In this manner, the algorithm is able to classify ancient bone samples, opening up the field of Paleoproteomics to the use of LC-MS as a more comprehensive and precise means of species classification.

P46

Development of Intelligent Mass Spectrometry Acquisition Methods Using Real-Time Control of Instruments

Zoltan Udvardy[1], Mathias Wilhelm[2], Odile Burlet-Schiltz[1], David Bouyssié[1]

Presenting author: Zoltan Udvardy

[1] Institut de Pharmacologie et de Biologie Structurale, Université de Toulouse, CNRS

[2] Computational Mass Spectrometry, TUM School of Life Sciences, Technical University of Munich

Introduction:
Complex proteomics samples necessitate sophisticated MS acquisition procedures. Despite the recent advances in optimising current workflows, inherent limitations persist, driving the creation of novel intelligent acquisition strategies. Manipulating mass spectrometers in real-time is a cornerstone in building such methods[1][2]. Although interfacing with MS instruments requires expertise and substantial human and MS instrument time, there are no open-source solutions alleviating the creation process. We are developing MSReact, a practical framework simplifying the process of method building and demonstrating its capabilities with a new intelligent targeted acquisition workflow.

Methods:
MSReact consists of two units, a server implemented in .NET interacting with mass spectrometers using an Instrument Application Programming Interface (IAPI) provided by Thermo and a client written in Python executing acquisition workflows. The two components are connected via websocket and a custom protocol. The intelligent targeted strategy is built on a novel real-time retention time (RT) adjustment procedure. Its throughput and sensitivity were assessed by analysing HeLa lysate and JPT SpikeMix samples on different Thermo instruments.

Results:
The MSReact server interfaces with three Thermo instrument families (Exactive, Exploris and Tribrid), and presents a simulation mode, aiding offline acquisition workflow development and testing. The Python client promotes the use of popular data analysis and machine learning libraries for real-time data analysis and decision making. The implemented workflow using a tailored retention time calibration algorithm enables peptide monitoring with narrow time windows, which is pivotal for high-throughput targeted acquisitions[3]. Throughput test of a 95 minutes acquisition on Exploris instrument using HeLa lysate showed that from 2146 targeted peptides 2069 were identified and 1436 MS1 level quantified.

Conclusion: Novel intelligent MS acquisition workflows may be readily created with MSReact. It's a powerful open-source framework, thanks to its client-server architecture and built-in simulator. The presented intelligent targeted method is capable of targeting considerably more peptides than what is reported in state of the art targeted proteomics studies on Thermo instruments, establishing a major improvement.

References:
1. Wichmann, C. et al., MCP, 2019
2. Erickson, B. K. et al., JPR, 2019

P47

ProteomicsDB: Connecting proteomes across species

Armin Soleymaniniya [1,2], Sarah Brajkovic [2, 3], Miriam Abele [3, 4], Christina Ludwig [2, 3, 4], Bernhard Küster [2, 3], Mathias Wilhelm [1,2]

Presenting author: Armin Soleymaniniya

[1] Computational Mass Spectrometry department; TUM School of Life Sciences, Technical University of Munich, Freising, Germany

[2] Elite Network of Bavaria; TUM School of Life Sciences, Technical University of Munich, Freising, Germany

[3] Chair of Proteomics and Bioanalytics, TUM School of Life Sciences, Technical University of Munich, Freising, Germany

[4] Bavarian Center for Biomolecular Mass Spectrometry (BayBioMS), TUM School of Life Sciences, Technical University of Munich, Freising, Germany

ProteomicsDB, initially established in 2014 as a protein-centric in-memory database for exploring the first draft of the human proteome [1], has evolved into a multi-omics and multi-organism resource for life sciences research encompassing more than 198 projects totaling over 29k LC-MS/MS experiments. Over the years, it has grown substantially, incorporating data from various omics sources and expanding its functionalities, such as integrating diverse omics data, introducing new APIs for systematic data access, enhancing user interfaces for better data visualization, and integrating deep-neural-network Prosit for peptides' fragments prediction, ultimately enabling reevaluation of stored search engine results with state-of-the-art methods to boost data utilization and identification rate in proteomics studies [2]. This continuous development and integration aligns with its commitment to improving data accessibility, interoperability, and usability while broadening its content to encompass various human biology experiments and supporting additional organisms [3]. Most recently, we integrated the proteome of Triticum aestivum (common wheat plant) acquired using a timsTOF Pro machine into ProteomicsDB, expanding the list of file formats and vendors supported by the database.

ProteomicsDB is set to integrate two groundbreaking projects into its framework: a proteomics map of the bacterial kingdom, encompassing data from approximately 400 bacterial strains, and "The Proteomes that Feed the World" project, aiming to provide comprehensive tissue-resolved data for almost 100 most important plants for human nutrition. These projects are characterized by their suitability for cross-species analysis, with the former extending ProteomicsDB's reach to the bacterial kingdom and the latter, shedding light for the first time on many important plant species' proteomes. ProteomicsDB will soon provide statistics and visualizations for cross-species analysis which empowers the community with unprecedented information. These integrations will significantly enhance ProteomicsDB's capabilities, fostering comprehensive insights into diverse species and facilitating cross-species comparisons for in-depth scientific exploration.

[1] Schmidt et al., 2018 (doi.org/10.1093/nar/gkx1029)
[2] Samaras et al., 2020 (doi.org/10.1093/nar/gkz974)
[3] Lautenbacher et al., 2022 (doi.org/10.1093/nar/gkab1026)

P48

Improving peptide identification rates of single cell experiments by utilizing single-cell specific features for rescoring with MS²Rescore.

Sam van Puyenbroeck [1,2], Arthur Declerq [1,2], Tine Claeys [1,2], Lennart Martens [1,2]

Presenting author: Sam van Puyenbroeck

[1] VIB-UGent Center for Medical Biotechnology, VIB, Zwijnaarde, Belgium

[2] Department of Biomolecular Medicine, Ghent University, Ghent, Belgium

Mass spectrometry-based proteomics has proven to be invaluable to study the expression level of proteins, offering insights into disease phenotypes and cellular responses to stress in the human body. However, currently protein expression levels are mainly studied on the tissue level. Recent advances in single cell proteomics on both the technology and sample preparation side has facilitated a deeper understanding of the biological intricacies at the single cell level, and allows to study nuanced differences between cells. Despite these advancements, the low sample amount still provides severe challenges for identification. While for now many efforts have been focused on optimizing the sample preparation and acquisition, novel bio-informatics approaches that tackle the challenges inherent to single cell proteomics are lacking.

Recently, data-driven rescoring has proven to boost peptide identifications in challenging workflows, and thus might also improve identification rates in single cell experiments. Notably, MS²Rescore, can improve the identification rate of peptides by calculating extra features that can help differentiate true from false positives. Unlike traditional PSM-scoring, MS²Rescore not only uses classical search engine scores, but also calculates orthogonal features using machine learning models like MS²PIP and DeepLC. Nonetheless, as these features are not single-cell specific, there is still substantial room for improvement. Therefore, including more features like signal to noise metrics or including other ion types might further boost peptide identifications specifically in single cell experiments. By reanalyzing public single cell projects, we have observed that the addition of these simple features on top of the more complex MS²PIP and DeepLC features can boost identifications with 1%. This is an encouraging sign that single-cell specific features can further improve the identification rate of peptides.

Another area of improvement in peptide identification for single cell LC-MS/MS experiments is related to the need for high-throughput analyses. This high-throughput is generally achieved by substantially reducing the elution time. This will result in an increase of chimeric spectra, which to date is still not well researched.

Therefore, in my research, I aim to boost peptide identifications by focusing on these chimeric spectra and rescoring matches by including single-cell specific features with the MS²Rescore algorithm.

P49

A subcellular specific PTM map of the human proteome

Pathmanaban Ramasamy [1,2,3,4], Lennart Martens [1,2], Wim F. Vranken [3,4]

Presenting author: Pathmanaban Ramasamy

[1] VIB-UGent Center for Medical Biotechnology, Ghent, Flanders, Belgium

[2] Department of Biomolecular Medicine, Faculty of Health Sciences and Medicine, Ghent, Flanders, Belgium

[3] Interuniversity Institute of Bioinformatics in Brussels, ULB-VUB, Brussels, Brussels, Belgium

[4] Structural Biology Brussels, Vrije Universiteit Brussel, Brussels, Brussels, Belgium

Protein function is highly regulated by the co-ordinated expression and interaction in different tissues and the correct localization within different sub-cellular compartments. Proteins must localize to their intended subcellular niche, with the functionality of some proteins only relevant and required in specific cell compartments; mis-localization of proteins can cause disease. Active transportation mechanisms, such as the nuclear import mechanism, utilize sequence targeting signals to transport proteins to specific locations, with N- and C-terminal sequence signals often cleaved off after relocation. Interactions of the cargo proteins carrying such signals with receptor proteins, which realize the relocation, are essential for this process. Proteins can also phase-separate to form often temporary biomolecular condensates with specific characteristics and function. In all cases, the surface residues of such proteins seem to be adapted to the particular chemical environment(s) present at their correct sub-cellular location(s); their sequences and by extension biophysical properties are tuned to their particular function at that location. Finally, PTMs are essential in regulating protein localization, and might play a role in adapting their biophysical characteristics to a particular sub-cellular environment. This highly complex relationship between a protein's sequence, biophysical characteristics, interactions and PTMs on the one hand, and its sub-cellular localization on the other hand, can now be pursued at the proteomics scale thanks to the advent of high-throughput (HTP) methods and computational approaches to analyse and integrate HTP data. In this work, we report the proteome wide identification of PTM landscapes in different subcellular specific and multi-localizing proteins. Moreover, we show how the changes in PTMs in different subcellular compartments of the cells relates to protein structures and their biophysical properties.

P50

DEFINING A MOLECULAR LANDSCAPE OF REPARATIVE CARDIAC CELLS BY MS-IMAGING AND ADVANCED MS-PROTEOMICS APPROACHES

Glenda Oliveira [1], Andrew Smith [1], Lisa Pagani [1], Patrizia Camelliti [2] Kenza Sackho [2], Paola Campgnolo [2], Fulvio Magni [1]

Presenting author: Glenda Oliveira

[1] University of Milano-Bicocca, Italy. [2] University of Surrey, United Kingdom .

Understanding the importance of the epicardium in heart repair has increased the interest in developing strategies to explore its regenerative potential for human therapies [1,2]. In this context, MALDI Mass Spectrometry Imaging is a powerful tool that enables exploring the molecular complexity of tissue at a proteomic level[3]. This work aims to evaluate the proteomic response of novel compounds targeting the proliferation of epicardial (ECs) and epicardial-derived cells (EPDCs) through MS-proteomic approaches. Epicardial slices were obtained from the left ventricle of porcine heart, cultured and treated with pharmacological compounds. Formalin-fixed paraffin-embedded epicardial slices were analysed by MALDI-MSI, targeting the generation of protein profiles from epicardial and myocardial as regions of interest. Epicardial slices provided an outstanding model for studying epicardium physiology and applying pharmacological therapies. The MALDI-MSI analysis provided a preliminary region-specific protein profile, discriminating epicardial from myocardial area. Complementary, nLCMS/MS identifies relevant proteins, such as collagen alpha-1, fibronectin, cadherin-2, vimentin, prostaglandin E synthase-3, Moesin and proliferating cell nuclear antigen (PCNA). Protein-protein interaction revealed relevant pathways involved in those responses: VEGFA-VEGFR2 signalling, cellular response to stress, fibroblast metabolic pathways, and hypoxia-induced factor 1signalling. Epicardial slices proved to be a suitable model for a deeper understanding of the molecular pathways involved in the regenerative response of ischemic myocardium driven by ECs. MALDI-MSI and complementary proteomics approaches highlighted relevant proteins and corresponding pathways involved in this response.

Keywords: Proteomics; MALDI MSI; Epicardial-derived cells; Cardiovascular research.

References

[1] Maselli D, et al. Porcine Organotypic Epicardial Slice Protocol: A Tool for the Study of Epicardium in Cardiovascular Research. Front Cardiovasc Med., 2022;9.

[2] Egido J et al. Animal models of cardiovascular diseases. Journal of Biomedicine and Biotechnology, vol. 2011. 2011.

[3] Stella M, et al. Histology-guided proteomic analysis to investigate the molecular profiles of clear cell Renal Cell Carcinoma. J Proteomics. 2019;191:38-47.

P51

The Peptonizer2000 for taxonomic identification of metaproteomic samples with a new taxonomic quantification method based on MaxLFQ

Tanja Holstein [1,2], Pieter Verschaffelt [1,3], Lennart Martens [1,2], Thilo Muth [4]

Presenting author: Tanja Holstein

[1]VIB-Ugent Center for Medical Biotechnology, 9052, Zwijnaarde, Belgium

[2]Department of Biomolecular Medicine, Ghent University, 9000, Ghent, Belgium

[3]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, 9000 Ghent, Belgium

[4]Domain Data Competence Group, Robert Koch Institute, Berlin, Germany

Taxonomic inference in mass spectrometry-based metaproteomics is complex. The exact composition of metaproteomic samples is unknown, requiring large reference databases for peptide identification. The presence of proteins and corresponding taxa has to be inferred from the identified peptides. This is complicated by protein homology: many proteins share peptides not only within a single taxon but also across different taxa. Current taxonomic profiling approaches rely on heuristics and strategies such as peptide-spectrum-match counting or the use of unique peptides, both for taxonomic annotation and quantification. The Peptonizer2000 is a taxonomic annotation tool for metaproteomics addressing some of the aforementioned challenges. It leverages the peptide-taxon relationships encoded in the Unipept database and combines it with a graphical model-based statistical inference method. This enables the Peptonizer2000 to provide taxonomic profiles of samples with greater taxonomic resolution and statistically computed confidence estimates.

Until now, the Peptonizer2000 has not performed quantification of the identified taxa. We introduce a first estimate of taxonomic biomass contributions based on peptide-taxon matches. However, this heuristic approach lacks accuracy and does not ensure comparability between samples. For protein samples, MaxLFQ is a popular, label-free quantification algorithm that leverages MS1 peptide intensities through delayed normalization and maximal peptide ratio extraction to compare relative protein quantities across proteomic samples. We introduce an algorithm, based on MaxLFQ, for the label-free relative quantification of taxa. Analogous to MaxLFQ, peptide intensities are normalized across samples in a delayed manner. Instead of mapping the peptide intensities to the proteins, we propose to map them directly to the taxa according to the taxonomy previously determined by the Peptonizer.

We evaluate the Peptonizer2000 with various metaproteomic datasets. We show that it provides taxonomic profiles with statistically computed scores, heuristic biomass estimates, and offer initial insights into the new, label-free taxonomic quantification approach.

P52

Genome assembly and annotation using Brownotate, a newly developed automated tool

Adrien BROWN [1, 2], Alexandre BUREL [1, 2], Sarah CIANFERANI [1, 2], Christine CARAPITO [1, 2], Fabrice BERTILE [1, 2]

Presenting author: Adrien BROWN

[1] Laboratoire de Spectrométrie de Masse BioOrganique (LSMBO), IPHC UMR 7178, Université de Strasbourg, CNRS, 25 rue Becquerel, 67087 Strasbourg, France.

[2] Infrastructure Nationale de Protéomique ProFI – FR2048, 67087 Strasbourg, France

Mass spectrometry (MS)-based proteomics most often relies on database searching to identify peptides or proteins. Ideally, protein sequences come from publicly available and reviewed reference databases, such as NCBI RefSeq or UniProtKB-Swissprot. However, 98.1% of the 1,758,200 sequencing datasets (with full genome representation; April 2023) for eukaryotes in NCBI are not assembled, and only 0.4% are fully assembled and annotated (0.1% only in Refseq). To overcome this bottleneck and enable the proteomist to assemble and annotate any genome on their own, we are developing a bioinformatics pipeline named Brownotate (BR).

BR gathers existing tools to download DNA sequencing data, filter out unreliable reads, assemble the retained reads, predict protein-coding genes, perform translation into protein sequences, and name them. An evaluation of the completeness of BR-derived assemblies or annotated genes is performed at runtime using the single-copy orthologs expected for given species (OrthoDB). To asses BR performance, BR-derived assemblies and annotations for 27 species belonging to 11 different taxons were compared to reference (REF) data extracted from NCBI SRA and RefSeq. MS raw data downloaded from PRIDE for these same species were also processed using Maxquant against either the BR-derived or the REF protein database, and datasets of identified proteins (FDR 1%) were compared. The length of BR and REF assemblies was more or less comparable, with the REF assemblies well covered by BR assemblies, but completeness (ortholog groups) of BR assemblies and annotations was slightly lower in the case of large genomes. On average, a higher number of proteins, and often shorter ones, was present in the BR than in REF dataset. The analysis of PRIDE-extracted raw MS/MS data using BR and REF protein databases reveals that, in general, a similar number of protein groups are identified, with, on average, a 80% overlap. 9% of the proteins identified only derived from the BR-derived sequences and this value reached 11% for the REF database.

BR proved capable of generating protein sequence databases of satisfactory quality for MS/MS data interpretation. BR is still undergoing further development, e.g. to adapt it to RNA sequence data and use it in a user-friendly graphical interface. Ultimately, BR should be useful for studies on species for which only DNA or RNA sequence data are available, or in the context of personalized medicine.

P53

The role of distributed databases in research application on the example of MaCPepDB & MaCcoyS

Dirk Winkelhardt [1,2], Karin Schork [2,3,4], Martin Eisenacher [2,3,4], Julian Uszkoreit [1,2]

Presenting author: Dirk Winkelhardt

[1] Ruhr-University Bochum, Medical Faculty, Medical Bioinformatics

[2] Ruhr-University Bochum, Medical Faculty, Core Unit for Bioinformatics in Medicine RUB

[3] Ruhr-Universität Bochum, Medizinische Fakultät, Medizinisched Proteom Center

[4] Ruhr-University Bochum, Medical Faculty, Medizinische Proteomanalyse, Protein, Diagnostic Center

Databases provide a way to store structural data of different sizes. In the research field of proteomics, this may be a mzDB file with a couple of 100MB, storing mass spectra, or MaCPepDB (Mass Centric Peptide Database) with a total of 10 TB which includes a tryptic digest of UniProt (SwissProt + TrEMBL).

The structure of the stored data is heavily depending on the usage and chosen database engine, predefined in a schema. The schema of a database is usually defined to efficiently access to the stored data.

However, regardless of how well these schemas are defined, an ever growing database will, at some, outgrow the capabilities of the underlaying hardware. This normally starts with reduced query performance as the file I/O cannot load the data fast enough or the memory is too small for the full index which normally would increase the data matching. The first version of MaCPepDB (Mass Centric Peptide Database), which contains the tryptic digest of all proteins in UniProt, suffered from exactly this problems. To overcome this issues MaCPepDB was moved to a distributed database engine and multiple servers. As a result it is now able to serve houndreds of queries more than before in a fraction of time. This of course does not matter for the average researcher when looking up the origin of a peptide or a target for a SRM assay but the integration in various analysis workflows. One example is MaCcoyS (Mass Centric Decoy Search), which uses MaCPepDB to build a search space for each MS2 spectrum in a MS-run with all peptides of MaCPepDB matching the mass of the spectrum's precursor. By refining those queries with post translational modifications (PSM), the number of queries is increasing even more with each PSM, as they change the weight of the plain peptides as stored by MaCPepDB. A MS-run with about 27000 MS2 spectra will yield around 9 Mio. queries.

The benefit of these method is the identification of previously unidentified MS2 spectra due to missing peptides in the search spaces. This method is very interesting for samples where the origin or content is not exactly known, like the mixed microbiome of a gut. While the distributed database system is helping with the performance of MaCPepDB and analysis software based on it, it also introduces the option to store metadata for each peptide, collected from the containing proteins of origin, like the taxonomy or the uniqueness state of peptide within a taxonomy.

P54

Triple negative breast cancer: novel therapeutic strategies investigated by proteomics

Sveva Germini [1,2], Serena Camerini [1], Irene Ruspantini [1], Maria E. Pisanu [1], Mattea Chirico [1], Sara Baccarini [1], Rosa Vona [1], Donatella Pietraforte [1], Egidio Iorio [1], Marialuisa Casella [1]

Presenting author: Sveva Germini

[1] National Italian Institute of Health, Rome, Italy

[2] University of Rome Tor Vergata, Department of System and Experimental Medicine, Italy

Triple negative breast cancer (TNBC) is the most aggressive, with the poorest prognosis, of all BC subtypes. TNBC lacks estrogen receptor, progesterone receptor and human epidermal growth factor receptor 2; hormonal therapies have no effect and chemotherapy resistance develops. Metformin (MF), an anti-diabetic drug, is known to lower cancer incidence. This anti-cancer activity is not completely understood. We are investigating a novel drug repurposing strategy coupling MF with D609 (inhibitor of phosphatidylcholine-specific phospholipase C, PC-PLC). Preliminary results on MDA-MB-231 cells reveal that the combination treatment (MF and D609) inhibits cells migration more powerfully compared to the effect due to MF alone and that PC-PLC inhibition by D609 enhances the anti-proliferative effects exerted by MF singularly. We investigated the proteome changes following MF treatment alone and/or in association with D609, using qualitative and quantitative (label-free) LC-MS/MS bottom-up proteomic approaches carried on applying both data-dependent (DDA) and data-independent (DIA) acquisition methods. We identified and quantified more proteins in samples acquired in DIA mode with a reduction in the number of missing values, compared to DDA mode. Principal Component Analysis of the acquired data groups together MF and combination treatment, separating them from control and D609; this result suggests that MF is the key modulator of the proteome in the combined treatment. MF and combination treatment up- and down-regulate the concentration of several proteins with a greater number recognized from DIA data. Enrichment analysis of the regulated proteins derived from DIA shows more significantly enriched pathways than DDA. Oxidative phosphorylation and cell cycle are two of the most significant up- and down-regulated pathways respectively, for both MF and combination treatment. Moreover, these two conditions share several pathways in agreement with the observed predominant MF effect. Additional proteomic and phosphoproteomic experiments are in progress together with metabolomic analyses: the integration of these –omics data will show a multifaceted mechanism of action of the drug repurposing strategy on TNBC cells.

P55

Carboxyl ester lipase (CEL) in pancreatic disease: Effects of cysteine residues in pathogenic CEL variants

Miguel A. Juárez Garzón [1], Khadija El Jellas [1], Janniche Torsvik [1], Mark E. Lowe [2], Xunjun Xiao [2], Karianne Fjeld [1], Anders Molven [1,3]

Presenting author: Miguel A. Juárez Garzón

[1] Gade Laboratory for Pathology, Dept. of Clinical Medicine, University of Bergen, Bergen, Norway

[2] Washington University School of Medicine, St. Louis, MO, USA

[3] Section for Cancer Genomics, Haukeland University Hospital, Bergen, Norway

Introduction

Carboxyl ester lipase (CEL) is a digestive enzyme mainly expressed in pancreatic acinar cells. The protein is structurally divided into two regions: an N-terminal globular domain and a C-terminal tail encoded by a variable number of tandem repeats (VNTR), consisting of 3 to 23 repeats that each encodes an 11-amino acid polypeptide. All verified pathogenic mutations of CEL are located within the highly polymorphic VNTR and involve the generation of cysteine residues in the polypeptide repeats. These mutations either cause the endocrine/exocrine pancreatic disorder MODY8 (Ræder et al., Nat. Genet. 2006) or confer 5-fold increased risk for chronic pancreatitis (CEL-HYB1; Fjeld et al., Nat. Genet. 2015). Our project aims to understand how de novo cysteine residues may determine the pathogenic properties of CEL-MODY and CEL-HYB1.

Methods

Cysteine residues of the variants CEL-MODY and CEL-HYB1 were changed to alanine by in vitro mutagenesis of plasmid constructs. The plasmids were expressed in HEK293 and Cosm KO HEK293 cells, and effects were studied by immunoblotting and immunofluorescence.

Results

The CEL-MODY variant showed elevated intracellular accumulation, reduced secretory levels, co-localization with ER stress markers and impaired O-glycosylation when compared to normal CEL. After mutating the 10 cysteines in CEL-MODY to alanine, properties of this variant became normalized and, in some cases, identical to those of normal CEL. For CEL-HYB1, which is encoded by a short VNTR of 3 repeats and has only 2 cysteines, properties did not change when mutating the cysteine residues to alanine and remained similar to what is described in the literature.

Conclusions

We revealed clear differences in the impact of cysteine residues present in the pathogenic variants CEL-MODY and CEL-HYB1. For CEL-MODY, cysteines appear central in the pathogenesis, whereas cysteine-mediated effects in CEL-HYB1 could not be revealed. We propose that the high number of cysteines in the CEL-MODY tail region enables intra- and intermolecular disulfide bonds that induce protein aggregation, ER stress and impaired O-glycosylation. For CEL-HYB1, we propose that it is primarily the unusually short tail length that determines its pathogenicity.

P56

Information content assessment of peptide fragmentation spectra using deep learning models

Zahra ELHAMRAOUI[1, 2], Eva Borras[1, 2], Mathias Wilhelm[3],Eduard Sabido[1, 2]

Presenting author: Zahra ELHAMRAOUI

1 Centre for Genomic Regulation, Barcelona, Spain,

2 Universitat Pompeu Fabra, Barcelona, Spain,

3 Computational Mass Spectrometry, Technical University of Munich, Freising, Germany

Peptide identification by mass spectrometry relies on the interpretation of fragmentation spectra based on the m/z pattern, relative intensities, and retention time (RT). Given a proteome, we wondered how many peptides generate identical fragmentation spectra with current MS methods. Calculating an information-content index for all peptides in a given proteome would enable us to design data acquisition and data analysis strategies that generate and prioritize the most informative fragment ions to be queried for peptide quantification.

We predicted nearly half million MS2 fragment spectra of human tryptic peptides using deep learning model PROSIT. In order to assess the information content of a peptide, we compared the fragmentation pattern of any two peptides that may be co-isolated in low- and high-resolution MS1 (precursor level) and MS2 (fragment level) data. We binned the predicted spectra for peptides with the same precursor mass and RT within a tolerance of 10 ppm m/z and 10 units indexed retention time (iRT) (high-resolution) and 1 Da m/z, 10 units iRT (low-resolution). On the MS2 level, m/z tolerance of 10 ppm, and 1 Da was used respectively for high and low-resolution setups. We used normalized spectral contrast angle (SA) to assess the similarity.

We selected the human proteome, and calculated the similarity score among its tryptic peptides in order to establish distinguishable peptides within the human proteome given a set of data acquisition parameters.

We used the deep learning prediction model Prosit - including the relative intensity pattern of the fragment ions - and iRT values for 416,813 human unmodified tryptic peptides with no missed cleavage, charge +2 and a normalized collision energy of 28. For each peptide pair within the m/z and iRT tolerance (e.g., 10 ppm and 10 iRT), we calculated the spectral angle of the predicted fragmentation spectrum.

The distribution of the similarity scores we observed shows that more than 98% of peptide pairs are distinguishable from their fragmentation pattern. However, several pairs have a high similarity score, which makes them difficult to distinguish with current data acquisition approaches. Similar analyses were performed accounting for high and low resolution at MS2 fragment level, as well as using only the 6 most intense fragments (Top6), or removing all the fragments with an intensity below 5 %.

Future work includes the exploration of charges, post-translational modifications...